Deep Networks Through the Lens of Low-Dimensional Structure: Towards Mathematical and Computational Principles for Nonlinear Data

Sam Buchanan

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022
Abstract

Deep Networks Through the Lens of Low-Dimensional Structure: Towards Mathematical and Computational Principles for Nonlinear Data

Sam Buchanan

Across scientific and engineering disciplines, the algorithmic pipeline for processing and understanding data increasingly revolves around deep learning, a data-driven approach to learning features for tasks that uses high-capacity compositionally-structured models, large datasets, and scalable gradient-based optimization. At the same time, modern deep learning models are resource-inefficient, require up to trillions of trainable parameters to succeed on tasks, and their predictions are notoriously susceptible to perceptually-indistinguishable changes to the input, limiting their use in applications where reliability and safety are critical. Fortunately, data in scientific and engineering applications are not generic, but structured—they possess low-dimensional nonlinear structure that enables statistical learning in spite of their inherent high-dimensionality—and studying the interactions between deep learning models, training algorithms, and structured data represents a promising approach to understand practical issues such as resource efficiency, robustness and invariance in deep learning. To begin to realize this program, it is necessary to have mathematical model problems that capture the nonlinear structures of data in deep learning applications and features of practical deep learning pipelines, and there is a question of how to translate mathematical insights into practical progress on the aforementioned issues, as well.

We address these considerations in this thesis. First, we pose and study the multiple
manifold problem, a binary classification task modeled on applications in computer vision, in which a deep fully-connected neural network is trained to separate two low-dimensional submanifolds of the unit sphere. We provide an analysis of the one-dimensional case, proving for a rather general family of configurations that when the network depth is large relative to certain geometric and statistical properties of the data, the network width grows as a sufficiently large polynomial in the depth, and the number of samples from the manifolds is polynomial in the depth, randomly-initialized gradient descent rapidly learns to classify the two manifolds perfectly with high probability. Our analysis demonstrates concrete benefits of depth and width in the context of a practically-motivated model problem: the depth acts as a fitting resource, with larger depths corresponding to smoother networks that can more readily separate the class manifolds, and the width acts as a statistical resource, enabling concentration of the randomly-initialized network and its gradients.

Next, we turn our attention to the design of specific network architectures for achieving invariance to nuisance transformations in vision systems. Existing approaches to invariance scale exponentially with the dimension of the family of transformations, making them unable to cope with natural variabilities in visual data such as changes in pose and perspective. We identify a common limitation of these approaches—they rely on sampling to traverse the high-dimensional space of transformations—and propose a new computational primitive for building invariant networks based instead on optimization, which in many scenarios provides a provably more efficient method for high-dimensional exploration than sampling. We provide empirical and theoretical corroboration of the efficiency gains and soundness of our proposed method, and demonstrate its utility in constructing an efficient invariant network for a simple hierarchical object detection task when combined with unrolled optimization. Together, the results in this thesis establish the first end-to-end theoretical guarantees for training deep neural networks with data with nonlinear low-dimensional structure, and provide a methodology to translate these insights into the design of practical neural network architectures with efficiency and invariance benefits.
# Table of Contents

Acknowledgments .................................................. xi

Dedication ............................................................... xiii

Chapter 1: Introduction .............................................. 1

1.1 The “Classical” Setting: (Piecewise) Linear Data ............... 1

1.2 The “Modern” Setting: (Highly) Nonlinear Data and Models .... 4

1.3 Objectives and Structure of the Thesis ......................... 7

Chapter 2: The Multiple Manifold Problem ........................ 10

2.1 Introduction ...................................................... 10

2.1.1 Related Work .............................................. 14

2.2 Problem Formulation and Main Results ....................... 15

2.2.1 Data Model ................................................. 15

2.2.2 Problem Formulation ....................................... 16

2.2.3 Error Dynamics and Certificates ......................... 17

2.2.4 Main Results and Proof Outline ......................... 18

2.3 Key Proof Elements ............................................ 24

2.3.1 Concentration at Initialization: Martingales and Angle Contraction .... 24

2.3.2 Uniform Concentration and its Consequences ............... 26
Chapter 3: Construction of Certificates: The Roles of Network Depth and Data Geometry  

3.1 Related Work ................................................................. 29  
3.2 Problem Formulation ...................................................... 29  
3.2.1 The Two Curve Problem .............................................. 29  
3.2.2 Key Geometric Properties .......................................... 32  
3.3 Main Results ................................................................. 34  
3.4 Proof Sketch ................................................................. 38  
3.4.1 The Importance of Depth: Localization of the Neural Tangent Kernel ..... 39  
3.5 Discussion ................................................................. 40  

Chapter 4: Designing Network Architectures from Low-Dimensional Structure: Resource-Efficient Invariant Networks ................................................................. 43  

4.1 Introduction ................................................................. 43  
4.2 Related Work ................................................................. 47  
4.3 Invariant Architecture Primitives: Optimization and Unrolling ................. 49  
4.3.1 Conceptual Framework .............................................. 50  
4.3.2 Computational Primitive: Optimization for Domain Transformations ..... 51  
4.3.3 Invariant Networks from Unrolled Optimization .......................... 53  
4.4 Invariant Networks for Hierarchical Object Detection ........................... 54  
4.4.1 Data Model and Problem Formulation ............................... 55  
4.4.2 Aside: Optimization Formulations for Registration of “Spiky” Motifs ... 57  
4.4.3 Invariant Network Architecture ...................................... 58  

2.4 Discussion ................................................................... 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.4.1</td>
<td>Notation and Framework</td>
<td>202</td>
</tr>
<tr>
<td>A.4.2</td>
<td>Pointwise Concentration</td>
<td>203</td>
</tr>
<tr>
<td>A.4.3</td>
<td>Uniformization Estimates</td>
<td>222</td>
</tr>
<tr>
<td>A.4.4</td>
<td>Auxiliary Results</td>
<td>246</td>
</tr>
<tr>
<td>A.5</td>
<td>Sharp Bounds on the One-Step Angle Process</td>
<td>257</td>
</tr>
<tr>
<td>A.5.1</td>
<td>Definitions and Preliminaries</td>
<td>257</td>
</tr>
<tr>
<td>A.5.2</td>
<td>Main Results</td>
<td>259</td>
</tr>
<tr>
<td>A.5.3</td>
<td>Supporting Results</td>
<td>261</td>
</tr>
<tr>
<td>A.5.4</td>
<td>Deferred Proofs</td>
<td>393</td>
</tr>
<tr>
<td>A.6</td>
<td>Auxiliary Results</td>
<td>399</td>
</tr>
<tr>
<td>B.1</td>
<td>Details of Figures</td>
<td>404</td>
</tr>
<tr>
<td>B.1.1</td>
<td>Figure 3.1</td>
<td>404</td>
</tr>
<tr>
<td>B.1.2</td>
<td>Figure 3.2</td>
<td>406</td>
</tr>
<tr>
<td>B.2</td>
<td>Key Definitions</td>
<td>408</td>
</tr>
<tr>
<td>B.2.1</td>
<td>Problem Formulation</td>
<td>408</td>
</tr>
<tr>
<td>B.2.2</td>
<td>Geometric Properties</td>
<td>410</td>
</tr>
<tr>
<td>B.2.3</td>
<td>Subspace of Smooth Functions and Kernel Derivatives</td>
<td>413</td>
</tr>
<tr>
<td>B.3</td>
<td>Main Results</td>
<td>414</td>
</tr>
<tr>
<td>B.4</td>
<td>Proof for the Certificate Problem</td>
<td>419</td>
</tr>
<tr>
<td>B.4.1</td>
<td>Invertibility Over a Subspace of Smooth Functions</td>
<td>420</td>
</tr>
<tr>
<td>B.5</td>
<td>Bounds for the Skeleton Function $\psi$</td>
<td>449</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Left: Two examples of classes (hedgehogs and hairbrushes) in an image classification problem, where the task is to learn a neural network from labeled examples that correctly classifies new images as corresponding to one of the two classes. Hedgehog examples are taken from ImageNet [23], and hairbrush examples are taken from ObjectNet [39]. Right: From the myriad complex variabilities present in the actual image classification task (appearance and shape variability; changes in illumination; classifying in the presence of structured backgrounds; etc.) we can abstract one particularly tractable type of nonlinear geometric structure: the geometric variability obtained when applying transformations of domain (translations, rotations, etc.) to a base template. The low-dimensional geometric structure of such transformations gives rise to a classification problem between low-dimensional manifolds of the ambient space (the set of all digital images).

2.1 Low-dimensional structure in image data and the two curves problem. Left: Manifold structure in natural images arises due to invariance of the label to continuous domain transformations such as translations and rotations. Right: The two curve problem. We train a neural network to classify points sampled from a density $\rho$ on the submanifolds $M_+, M_-$ of the unit sphere. We illustrate the angle injectivity radius $\Delta$ and curvature $1/\kappa$. These parameters help to control the difficulty of the problem: problems with smaller separation and larger curvature are more readily separated with deeper networks.

2.2 (a) Depth acts as a fitting resource. As $L$ increases, the rotationally-invariant kernel $\hat{\Theta}$ (a slight modification of the deterministic kernel in Theorem 2) decays more rapidly as a function of angle between the inputs $\angle(x, x')$ ($n$ is held constant). Below the curves we show an isometric chart around a point $x \in M_+$. Once the decay scale of $\hat{\Theta}$ is small compared to the inter-manifold distance $\Delta$ and the curvature of $M_-$, the network output can be changed at $x$ while only weakly affecting its value on $M_-$. This is one mechanism that relates the depth required to solve the classification problem to the data geometry. (b) Width acts as a statistical resource. The dynamics at initialization are governed by $\Theta$, a random process over the network parameters. As $n$ is increased, the normalized fluctuations of $\Theta$ around $\hat{\Theta}$ decrease (here $L = 10$). These two phenomena are related, since the fluctuations also grow with depth, as evinced by the scaling in theorem 2.
3.1 The $\mathcal{G}$-number—theory and practice. *Left:* We generate a parametric family of space curves with fixed maximum curvature and length, but decreasing $\mathcal{G}$-number, by reflecting ‘petals’ of a clover about a circumscribing square. We set $\mathcal{M}_+$ to be a fixed circle with large radius that crosses the center of the configurations, then rescale and project the entire geometry onto the sphere to create a two curve problem instance. In the insets, we show a two-dimensional projection of each of the blue $\mathcal{M}_-$ curves as well as a base point $x \in \mathcal{M}_+$ at the center (also highlighted in the three-dimensional plots). The intersection of $\mathcal{M}_-$ with the neighborhood of $x$ denoted in orange represents the set whose covering number gives the $\mathcal{G}$-number of the configuration (see (3.2)). *Top right:* We numerically generate a certificate for each of the four geometries at left and plot its norm as a function of $\mathcal{G}$-number. The trend demonstrates that increasing $\mathcal{G}$-number correlates with increasing classification difficulty, measured through the certificate problem: this is in line with the intuition we have discussed. *Bottom right:* t-SNE projection of MNIST images (top: a “four” digit; bottom: a “one” digit) subject to rotations. Due to the approximate symmetry of the one digit under rotation by an angle $\pi$, the projection appears to nearly intersect itself. This may lead to a higher $\mathcal{G}$-number compared to the embedding of the less-symmetric four digit. For experimental details for all panels, see Section B.1.

3.2 The effect of geometry and depth on the certificate. *Left:* The certificate $g$ computed numerically from the kernel $\Theta$ for depth $L = 50$ (defined in (3.3)) and the geometry from Figure 2.1 with a uniform density, graphed over the manifolds. Control of the norm of the certificate implies rapid progress of gradient descent, as reflected in Theorem 3.3.2. Comparing to Section 2.1, we note that the certificate has large magnitude near the point of minimum distance between the two curves—this is suggestive of the way the geometry sets the difficulty of the fitting problem. *Right:* To visualize the certificate norm more precisely, we graph the log-magnitude of the certificate for kernels $\Theta$ of varying depth $L$, viewing them through the arc-length parameterizations $x_\sigma$ for the curves (left: $\mathcal{M}_+$; right: $\mathcal{M}_-$). At a coarse scale, the maximum magnitude decreases as the depth increases; at a finer scale, curvature-associated defects are ‘smoothed out’. This indicates the role of depth as a fitting resource. See Section B.1 for further experimental details.
4.1 Comparing the complexity of covering-based and optimization-based methods for invariant recognition of a template embedded in visual clutter. (a-d): We consider four different classes of deformations that generate the observation of the template, ranging across shifts, rotations, scale, and skew. The dimension $d$ of the family of transformations increases from left to right. (e): A geometric illustration of the covering and optimization approaches to global invariance: in certifying that a query (labeled with a star) is a transformed instance of the template (at the base point of the solid red/blue lines), optimization can be vastly more efficient than covering, because it effectively covers the space at the scale of the basin of attraction of the optimization problem, which is always larger than the template's associated $\varepsilon_{\text{COVER}}$. (f): Plotting the average number of convolution-like operations necessary to reach a zero-normalized cross-correlation (ZNCC) of 0.9 between the template and a randomly-transformed query across the different deformation classes. Optimization leads to an efficiency gain of several orders of magnitude as the dimensionality of the family of transformations grows. Precise experimental details are recorded in Section C.1.3.

4.2 Motif registration with the formulation (4.1), and an unrolled solver. (a-d): Visualization of components of a registration problem, such as (4.2). We model observations $y$ as comprising an object involving the motif of interest (here, the body of the crab template we experiment with in Section 4.4) on a black background, as in (a), embedded in visual clutter (here, the beach background) and subject to a deformation, which leads to a novel pose. A mask $\Omega$ for the nonzero pixels of the motif, as in (b), is used to avoid having pixels corresponding to clutter enter the registration cost. After solving this optimization problem, we obtain a transformation $\tau$ that registers the observation to the motif, as in (d). In (c-d), we set the red and blue pixels corresponding to the mask $\Omega$ to 1 in order to visualize the relative location of the motif. (e): Optimization formulations imply network architectures, via unrolled optimization. Here we show two iterations of an unrolled solver for (4.2), as we detail in Section 4.3.3; parameters that could be learned from data, à la unrolled optimization, are highlighted with red text. The operations comprising this unrolled network consist of linear maps, convolutions, and interpolations, leading to efficient implementation on standard hardware accelerators.
4.3 An example of a hierarchically-structured template, and the results of an implementation of our detection network. (a): Structure of the crab template, described in Section 4.4.1, and its interaction with our network architecture for detection, described in Section 4.4.3. Left: top-down decomposition of the template into motifs. A template of interest \( y_o \) (here, the crab at top left) is decomposed into a hierarchy of abstractions. The hierarchical structure is captured by a tree \( G = (V, E) \): nodes represent parts or aggregations of parts, and edges represent their relationships. Right: bottom-up detection of the template in a novel scene. To detect the template in a novel scene and pose, the network described in Section 4.4.3 first localizes each of the lowest-level visual motifs at left and their transformation parameters in the input scene \( y \) (bottom right). Motifs and the derived occurrence maps are labeled in agreement with the notation we introduce in Section 4.4.3. The output of each round of optimization is an occurrence map \( \omega_v \) for nodes \( v \in V \); these occurrence maps then become the inputs for detection of the next level of concepts, following the connectivity structure of \( G \), until the top-level template is reached (top right). (b-e): Evaluation of the hierarchical invariant object detection network implemented in Section 4.4.4: the learned transformation at the minimum-error stride for each motif is used to draw the motifs’ transformed bounding boxes. Relative to the canonized template, each observation has the template’s eye motifs independently rotated by 5 degrees, and each claw motif by 15 degrees. Insets at the bottom right corner of each result panel visualize the quality of the final detection trace \( \omega_0 \) for the template, with a value of 1 marked in dashed red (an ideal detection has unit \( \ell^1 \) norm). In this example, detection succeeds until a multiple instance issue between the two claw motifs occurs in panel (e), causing detection of the left claw motif to fail.

4.4 Numerical verification of Theorem 4.5.1. (a): A multichannel spike motif containing 5 spikes. (b): A scene generated by applying a random affine transformation to the motif. (c): The solution to (4.10) with these data. The skewing apparent here is undone by the compensated external gaussian filter, which enables accurate localization in spite of these artifacts. (d): Change in objective value of (4.10) across iterations of proximal gradient descent. Convergence occurs in tens of iterations. (e): Change in normalized cross correlation across iterations (see Section C.1.3). We observe that the method successfully registers the multichannel spike scene. (f): Comparison between the left and right-hand side of equation (4.18) with gradient descent iterates from (4.10) (labeled as \( \varphi \) here). After an initial faster-than-predicted linear rate, the discretized solver saturates at a sub-optimal level. This is because because accurate estimation of the transformation parameters \( (A, \tilde{b}) \) requires subpixel-level preciseness, which is affected by discretization and interpolation artifacts. It does not hinder correct localization of the scene, as (e) shows.
4.5 Plotting a basin of attraction for the textured motif registration formulation (4.4).

(a): Heatmap of the ZNCC at convergence (see Section C.1.3), for translation versus rotation. Optimization conducted with SE(2) motion model. (b): Heatmap of the ZNCC at convergence, for translation versus scale. Optimization conducted in ‘similarity mode’, a SE(2) motion model with an extra global scale parameter. In both experiments, each reported data point is averaged over 10 independent runs.

(c-d): Notably, when the registration target \( y \) is zoomed out relative to the motif \( x_o \), resolution is lost in the detection target, so recovering it will cause interpolation artifacts and blur the image. This prevents the ZNCC value from converging to 1 despite correct alignment with the motif, and accounts for the results shown in (b) at small scales.

B.1 The two curve geometry described in Section B.1.1. The different choices of \( M_- \) that lead to different \( \mathcal{V} \)-number are overlapping. The legend indicates the \( \mathcal{V} \)-number of the two curves problem obtained by considering the same \( M_+ \) but a different \( M_- \) as indicated by the color.

\[ x \]
Acknowledgements

I have been fortunate to have been able to interact with, learn from, and be inspired by a great many wonderful people throughout my six years at Columbia, and certain earlier influences seem also to have played a significant role in my completion of this thesis. I would therefore like to try to be somewhat exhaustive here in my thanks.

First, to my family, and especially to Kelly, for constant support, patience, and guidance. It is safe to say that none of this would have been possible without it. I am also grateful to Charles Schutte, who encouraged me in composition some twenty years ago; to Marla Lindsay, who showed me the joy of critical analysis, and to do it by a combination of rigor and creativity; to Bozenna Pasik-Duncan, who showed me that the same was true of mathematics; and to John Paden, who encouraged me to dig deeper into the mathematics of signal processing.

Thanks are due to the members of my thesis committee, especially to Yi Ma for his passionate intellectual leadership and support and to Boris Hanin for his generosity with his time, encouragement, and experience. To my advisor, John Wright, as well, whose careful guidance over several years helped the work in this thesis come to fruition—it has been a great privilege to develop as a researcher working with someone with such strong technical chops, physical intuition, and ability to deeply motivate problems. I thank Dar Gilboa, for an immensely fruitful time spent working together—my research experience at Columbia was broadly and uniformly influenced by his leadership and example. Thanks as well to Henry Kuo, who took me under his wing not long after I joined John’s research group and was uncompromisingly generous in sharing his knowledge of compressed sensing, applied mathematics, and the practice of
theoretical research; to Tanbir Haque, whose professionalism and natural leadership set a standard throughout my PhD; to Tim Wang, who showed me how to think and communicate with order and logic; and to Yuqian Zhang and Qing Qu, for their support within and beyond the research group.

Finally, I must thank John once more—for quite unexpectedly replying to an impromptu email sent the night before a visit; for immediately accelerating me along the mathematical research path, apparently without any reservations; and all this for a master’s admit without a math background. Perhaps from his perspective there was little downside, but I am prone to sentimentality and instead like to think that there was some kind of subtle recognition of something more. It certainly was not universal at that time; and for the subsequent opportunities and his constant efforts on my behalf since then, I have nothing but gratitude.
Dedication

To Georgia, for curiosity; to Mark, for ambition; to Carol, for patience
Chapter 1: Introduction

This thesis is about data with low-dimensional structure, the modern algorithmic frameworks we use to process them, and the mathematical and computational interplay between the two. Before getting to the “modern” aspect of things, we will briefly discuss low-dimensional models for data at a high level in order to fix concepts and motivate our goals.

1.1 The “Classical” Setting: (Piecewise) Linear Data

The prototypical example of data with low-dimensional structure is a $d$-dimensional real vector $x_o \in \mathbb{R}^d$ lying in a $k$-dimensional linear subspace—if $x_o$ corresponds to some measured signal (e.g., a member of the set of real-valued bandlimited functions), $k$ represents the number of degrees of freedom or “information content” of $x_o$, and $x_o$ can be ‘measured’ with computational cost proportional to $k$ rather than the ambient dimension $d$, which can be significantly larger. However, the first point at which this notion begins to become both broadly applicable and computationally interesting is when we consider signals lying in a union of subspaces. This gives rise to the sparse signals, and we will use our understanding of this class of structured data as a ‘barometer’ for our subsequent discussion.

**Example Data I: Sparsity.** A vector $x_o \in \mathbb{R}^d$ is called $k$-sparse if it has at most $k$ nonzero entries. A ubiquitous practical task is to reconstruct a sparse vector from compressive measurements: in the simplest case, these measurements are linear and given by a matrix $A \in \mathbb{R}^{m \times d}$, and this corresponds to solving an underdetermined system of equations

$$y = Ax,$$  \hspace{1cm} (1.1)
with \( y = Ax_0 \). For example, this setup arises in applications as diverse as wireless communications [1], medical imaging [2, 3], neuroscience [4, 5], microscopy [6], and collaborative filtering [7, 8] (under a slightly generalized notion of sparsity). Because \( d \gg m \), solving (1.1) is hopeless in general; on the other hand, if we knew a priori the set of entries on which \( x_0 \) took nonzero values, the problem would become as trivial as the single subspace version discussed above. The great mathematical insight of a body of work that has come to be known as compressed sensing [9, 10, 11, 12] is that for a large class of sufficiently generic measurement matrices \( A \) (say, random), \textit{the vector} \( x_0 \) \textit{can be exactly recovered by minimizing a convex proxy for sparsity over all solutions to the linear system}. Concretely, a typical result states that when \( A \) has i.i.d. gaussian entries of variance \( 1/m \) and the number of measurements satisfies \( m \gtrsim k \log(d/k) \), the convex optimization problem

\[
\text{minimize} \quad \|x\|_1 \\
\text{subject to} \quad y = Ax
\]

has \( x_0 \) as its unique optimal solution for all \( k \)-sparse signals \( x_0 \) with high probability. In particular, the number of measurement resources need only grow with the intrinsic information content of the signal \( x_0 \) rather than the ambient dimension, and reconstruction is reduced to a question of optimization.

The structure of problems (1.1) and (1.2) is deceptively simple: the study of these objects encompassed a period of intense intellectual activity over the last fifteen years that culminated in impactful insights into high dimensional probability [13, 14], phase transitions [15, 16], and scalable convex optimization algorithms [17, 18], as well as numerous applications [19]. Without digressing further into technical details, we highlight three important aspects of this body of work:

1. **Insights into data geometry and efficient algorithms**: Analysis of the optimization problem (1.2) and its variants depends essentially on the geometry of the set of \( k \)-sparse vectors (via its convex surrogate norm, the \( \ell^1 \) norm) and its interaction with the (random) nullspace.
of the matrix $A$. This leads to a favorable tradeoff between resource efficiency (the number of measurements needed for successful reconstruction) and computational efficiency (there is a natural inefficient reconstruction procedure that succeeds with only $2k$ measurements given $O(n^{2k})$ computation time)—the geometry of the data and measurements is the link.

2. **A union of theory and practice**: Compressed sensing is unique in that it is an example of a situation where theoretical insight inspired and drove practical progress: the initial theoretical promise of sensing schemes vastly more efficient than existing ones led to the development of new algorithms for a variety of applications. One of the most significant successes occurred in magnetic resonance imaging, where compressed sensing technologies find clinical application [20].

3. **A flexible modeling language, robust to the real world**: The objective of (1.2) can be modified to accommodate other families of low-dimensional structure (so-called atomic norms [21]), and the problem can be augmented with additional convex constraints. This leads to a flexible and computationally-efficient modeling language for sensing structured signals in a variety of application domains. Moreover, the types of guarantees we quoted above for (1.2) are stable to various types of measurement noise, leading to a rare example of elegant theory that also works in practice.

From an applied mathematics perspective, these three achievements of the compressed sensing program can be argued to be an ideal research outcome. But in spite of these successes, the dawn of the era of deep learning has largely displaced these methods from the practitioner’s toolkit: obtaining state-of-the-art performance on modern tasks seem to demand a combination of flexible learning from large datasets using high-capacity deep neural networks with a modicum of prior knowledge about the task and data incorporated into the network’s architecture (so-called “inductive biases”). Nevertheless, as we will see below low-dimensional structure continues to play an important role in the practice of deep learning, with even greater potential—it is just a question of bringing it to the surface.
1.2 The “Modern” Setting: (Highly) Nonlinear Data and Models

The rise of deep learning methodologies in the past ten years [22] has followed a markedly different path from the theory-inspired growth of compressed sensing: progress on model development and performance has been driven by empirical research and benchmark datasets [23, 24, 25, 26, 27], and the practical success of these methods initially seemed to contract traditional wisdom in statistical learning theory [28, 29]. This immense progress has been driven by a scaling-oriented approach to model development: the core activity is to train high-capacity deep neural networks with vast computational resources and ever more labeled training data. In this paradigm, the training data are treated as a generic statistical resource, like fuel, with the assumption that arbitrarily strong levels of performance will be achievable given enough ‘fuel’ [30]. And indeed, this approach continues to produce impressive empirical results, especially on consumer internet-oriented tasks like image classification and machine translation [27, 31, 32, 33].

At the same time, there are fundamental computational barriers to consider: standard “curses of dimensionality” dictate that the learning of general functions in high-dimensional spaces is computationally intractable [34, 35, 36, 37, 38]. As ever, this suggests that low-dimensional structures in data are at play: underlying the empirical successes of deep learning seems to be a unique ability to flexibly and adaptively learn the nonlinear low-dimensional structures in large datasets from efficient gradient descent training. To make clear the ubiquity of this phenomenon, we discuss two representative examples of data whose nonlinear structure seems to demand the use of the powerful nonlinear models of deep learning, and where there seems to be significant opportunity for improvement by better exploiting structure in the data.

Example Data II: Nonlinear Variability in Natural Images. Image classification (Figure 1.1) has been a major driver of innovation in deep learning over the past ten years. Learning a deep neural network that correctly classifies unseen images from labeled examples requires the network to learn to cope with a staggering level of nonlinearity in the input data: variations in shape and appearance [40, 41, 42], photometric variability due to illumination changes and cast shadows [43,
Figure 1.1: Left: Two examples of classes (hedgehogs and hairbrushes) in an image classification problem, where the task is to learn a neural network from labeled examples that correctly classifies new images as corresponding to one of the two classes. Hedgehog examples are taken from ImageNet [23], and hairbrush examples are taken from ObjectNet [39]. Right: From the myriad complex variabilities present in the actual image classification task (appearance and shape variability; changes in illumination; classifying in the presence of structured backgrounds; etc.) we can abstract one particularly tractable type of nonlinear geometric structure: the geometric variability obtained when applying transformations of domain (translations, rotations, etc.) to a base template. The low-dimensional geometric structure of such transformations gives rise to a classification problem between low-dimensional manifolds of the ambient space (the set of all digital images).

44], geometric variability due to changes in perspective and pose [45, 46, 47], and more. Although a subset of these types of variability are amenable to mathematical modeling, the predominant approach to dealing with these in practice is data-driven: one relies on a high-capacity generic network (e.g., a ResNet or a Vision Transformer [48]) to learn the necessary invariances from large datasets and data augmentation [24, 49, 50]. While one cannot argue with the practical successes of these methods on popular benchmarks, it is worth noting that this data-driven approach, in largely ignoring known structures present in the data, is wasteful in terms of resources (both in terms of training data and network parameters), and can even lead to a failure to learn simple invariances of the data (e.g., translational invariance) [51, 52, 53, 54, 55, 56, 57].

Example Data III: Nonlinear Physical Structure in Scientific Discovery. Any time a physical system with a small number of degrees of freedom is responsible for generating data in an application, one encounters data with low-dimensional structure. In many cases, the structure is also nonlinear: for example, in gravitational wave astronomy [58, 59], gravitational wave signals emitted by a merger of two binary black holes with distinct masses and spins (parameters $\Gamma$) gives rise to a nonlinear manifold of possible observable signals $\{x_\gamma | \gamma \in \Gamma\}$, by the laws of general
relativity. An important scientific task consists of detecting these signals in noise: a simple model for this problem consists of observations $y = x_\gamma + n$ for some $\gamma \in \Gamma$ or $y = n$ with noise following some distribution, with the goal to train a detector $f$ that determines whether a gravitational wave signal is present in the observation.

An optimal test for detecting a single template $x_\gamma$ in gaussian noise is to use matched filtering: one simply calculates $\langle x_\gamma, y \rangle$, and thresholds at a level commensurate with the noise level. One could use a similar approach for the parametric detection problem for gravitational wave astronomy over $\Gamma$ with many templates $\gamma_i$: thresholding the maximum correlation $\max_i \langle x_{\gamma_i}, y \rangle$ can be thought of as a certain shallow neural network architecture for this problem, and is in fact the standard approach in practice. However, it turns out for the parametric detection problem that this is a suboptimal test in the case of gaussian noise [60]. Moreover, in practice the noise $n$ will not be gaussian, but in fact correlated with different properties of the measurement system; and the use of a ‘shallow’ architecture for detection here may also seem questionable, with the possibility to obtain a more efficient detector using some kind of hierarchical (deeper!) model that captures the structure of the set of observable signals. In this way, the combination of nonlinear, manifold-structured data and the need to use learning to cope with the non-gaussian noise distribution and unknown aspects of the physical system has many features that seem ideal for the use of the flexible nonlinear learning machines of deep learning, but these questions are out of reach by a purely data-driven, scaling-based approach—it seems necessary to understand how the network architecture should relate to the data and noise in order to make progress.

The deficiencies of the “standard pipeline” for deep learning for understanding the important practical issues of resource efficiency, invariance, and robustness that are highlighted in the above two examples persist across similar tasks in science and engineering. At this point, we recall the three achievements of the compressed sensing paradigm for sensing signals with piecewise linear structures that we discussed above: issues of efficiency and robustness are naturally understood in this framework, precisely in terms of the intrinsic structure of the data! Therefore it seems that one of the most promising approaches to understanding these practical issues for deep learning is to
‘put the structure back’, and study the mathematical interactions between deep networks and data with low-dimensional nonlinear structure. This is the task we will take up in the remainder of this thesis.

1.3 Objectives and Structure of the Thesis

Motivated by the above considerations and examples, we will take up two major questions in this thesis:

1. What are the resource requirements for computing with nonlinear, structured data and deep neural networks? What structural properties of nonlinear data play a role in these requirements, in analogue to sparsity in the familiar piecewise linear case?

2. How can we exploit specific known low-dimensional structures in data to design deep neural network architectures for efficient and invariant computation?

In Chapter 2, we will begin to address the first of these two questions. We formulate the multiple manifold problem—a binary classification task that uses a deep fully-connected neural network to classify data drawn from two disjoint smooth curves on the unit sphere. We will write the union of the two curves as $M$. Curves are one-dimensional manifolds, and classification of data on curves under the mild regularity conditions we assume already allows the data to assume highly nontrivial configurations (certainly not linearly separable) that lie beyond the implications of contemporary theoretical analyses of deep learning. We prove that when (i) the network depth $L$ is large relative to certain geometric properties that set the difficulty of the problem, (ii) the network width $n$ and number of samples $N$ are polynomial in the depth, and (iii) a certain technical condition that we call the ‘certificate condition’ is satisfied, randomly-initialized gradient descent quickly learns to correctly classify all points on the two curves with high probability. The argument centers around the “neural tangent kernel” $\Theta(x, x')$ (for $x, x' \in M$) of Jacot et al. and its role in the nonasymptotic analysis of training overparameterized neural networks; to this literature, we contribute essentially optimal rates of concentration for the neural tangent kernel of deep fully-
connected ReLU networks, requiring width \( n \geq L \text{poly}(d_0) \) to achieve uniform concentration of the initial kernel over a \( d_0 \)-dimensional submanifold of the unit sphere \( \mathbb{S}^{n_0-1} \), and a nonasymptotic framework for establishing generalization of networks trained in the “NTK regime” with structured data.

Our results in Chapter 2 can be seen as reducing the analysis of the gradient descent dynamics for a sufficiently overparameterized network to the ‘certificate problem’, which asks us to show that there exists a function \( g : \mathcal{M} \to \mathbb{R} \) of sufficiently small norm such that the integral equation associated to the NTK 
\[
\int_{\mathcal{M}} \Theta(\cdot, x) g(x) \, d\mu(x) = \zeta
\]
is approximately satisfied. Here \( \zeta : \mathcal{M} \to \mathbb{R} \) denotes the neural network predictor’s error from the target labels at the initial random parameters.

In Chapter 3, we will prove that the certificate problem has a suitable solution under sufficient assumptions on the network depth \( L \), giving us an end-to-end generalization guarantee in the one-dimensional case of the multiple manifold problem—this is the first generalization guarantee for deep networks with nonlinear data that depends only on intrinsic data properties. In particular, via fine-grained control of the decay properties of the NTK, we demonstrate that when the network is sufficiently deep, the NTK can be locally approximated by a translationally invariant operator on the manifolds and stably inverted over smooth functions, which guarantees convergence and generalization.

Our studies of the multiple manifold problem in Chapters 2 and 3 provide some of the first algorithmic insights into the resource requirements to classify data with general nonlinear low-dimensional structure using a deep neural network. However, there is a question of how to translate these mathematical insights into practical gains on important issues such as resource efficiency and robustness. In Chapter 4, we will approach the study of low-dimensional structure and deep learning from a more practical perspective: we will ask, for specific families of practical low-dimensional structures, how to design neural network architectures ideally suited to computing with these structures. Specifically, we will consider the issue of resource-efficient invariance to low-dimensional geometric transformations of domain with image data (e.g., translations, rigid body motions, changes in perspective, etc.), and leverage classical insights from computer vision.
and unrolled optimization to design neural network architectures that are simultaneously invariant-by-construction and vastly more parameter efficient than the naive ‘template matching’-based approach to invariance. Our results only begin to scratch the surface of the potential for designing invariant architectures for natural image data—they will open many promising directions for future empirical and theoretical work and demonstrate the strong potential of our approach.
Chapter 2: The Multiple Manifold Problem

2.1 Introduction

In applied machine learning, engineering, and the sciences, we are frequently confronted with the problem of identifying low-dimensional structure in high-dimensional data. In certain well-structured data sets, identifying a good low-dimensional model is the principal task: examples include convolutional sparse models in microscopy [6] and neuroscience [4, 5], and low-rank models in collaborative filtering [7, 8]. Even more complicated datasets from problems such as image classification exhibit some form of low-dimensionality: recent experiments estimate the effective dimension of CIFAR-10 as 26 and the effective dimension of ImageNet as 43 [41]. The variability in these datasets can be thought of as comprising two parts: a “probabilistic” variability induced by the distribution of geometries associated with a given class, and a “geometric” variability associated with physical nuisances such as pose and illumination. The former is challenging to model analytically; virtually all progress on this issue has come through the introduction of large datasets and high-capacity learning machines. The latter induces a much cleaner analytical structure: transformations of a given image lie near a low-dimensional submanifold of the image space (Figure 2.1). The celebrated successes of convolutional neural networks in image classification seem to derive from their ability to simultaneously handle both types of variability. Studying how neural networks compute with data lying near a low-dimensional manifold is an essential step towards understanding how neural networks achieve invariance to continuous transformations of the image domain, and towards the longer term goal of developing a more comprehensive mathematical understanding of how neural networks compute with real data. At the same time, in some scientific and engineering problems, classifying manifold-structured data is the goal—one example is in gravitational wave astronomy [58, 59], where the goal is to distinguish true events from
noise, and the events are generated by relatively simple physical systems with only a few degrees of freedom.

Motivated by these long term goals, in this paper we study the **multiple manifold problem** (Figure 2.1), a mathematical model problem in which we are presented with a finite set of labeled samples lying on disjoint low-dimensional submanifolds of a high-dimensional space, and the goal is to correctly classify every point on each of the submanifolds—a strong form of generalization. The central mathematical question is how the structure of the data (properties of the manifolds such as dimension, curvature, and separation) influences the resources (data samples, and network depth and width) required to guarantee generalization. Our main contribution is the first end-to-end analysis of this problem for a nontrivial class of manifolds: one-dimensional smooth curves that are non-intersecting, cusp-free, and without antipodal pairs of points. Subject to these constraints, the curves can be oriented essentially arbitrarily (say, non-linearly-separably, as in Figure 2.1), and the hypotheses of our results depend only on architectural resources and intrinsic geometric properties of the data. To our knowledge, this is the first generalization result for training a deep nonlinear network to classify structured data that makes no a priori assumptions about the representation

![Figure 2.1: Low-dimensional structure in image data and the two curves problem. Left: Manifold structure in natural images arises due to invariance of the label to continuous domain transformations such as translations and rotations. Right: The two curve problem. We train a neural network to classify points sampled from a density $\rho$ on the submanifolds $M_+, M_-$ of the unit sphere. We illustrate the angle injectivity radius $\Delta$ and curvature $1/\kappa$. These parameters help to control the difficulty of the problem: problems with smaller separation and larger curvature are more readily separated with deeper networks.](image-url)
capacity of the network or about properties of the network after training.

Our main result reduces the analysis of the gradient descent dynamics to the construction of a certificate—showing that a certain deterministic integral equation involving the network architecture and the structure of the data admits a solution of small norm. The construction of certificates turns out to be nontrivial, with little prior art available: we take up this task in Chapter 3.

**Theorem 1** (informal). Let $d_0 = 1$. Suppose a certificate for $M$ exists. Then if the network depth satisfies $L \geq \text{poly}(\kappa, C_p, \log(n_0))$, the width satisfies $n \geq \text{poly}(L, \log(Ln_0))$, and the number of training samples satisfies $N \geq \text{poly}(L)$, randomly-initialized gradient descent on $N$ i.i.d. samples rapidly learns a network that separates the two manifolds with overwhelming probability. The constants $C_p, \kappa$ depend only on the data density and the regularity of the manifolds.

Theorem 1 gives a provable generalization guarantee for a model classification problem with deep networks on structured data that depends only on the architectural hyperparameters and properties of the data. In addition, it provides an interpretable tradeoff between the architectural settings necessary to separate the two manifolds: the network depth needs to be set according to the intrinsic difficulty of the problem, and the network width needs to grow with the depth. Our analysis gives further insight into the independent roles played by each of these parameters in solving the problem, with the depth acting as a ‘fitting resource’, making the network’s output more regular and easier to change, and the width acting as a ‘statistical resource’, granting concentration of the network over the random initialization around a well-behaved object that we can analyze. Moreover, the sample complexity of Theorem 1 is dictated by the intrinsic difficulty of the problem instance which is set by the geometry of the data.

Theorem 1 is modular, in the sense that a generalization guarantee is ensured for any geometry for which one can construct a certificate. The key to our approach will be to approximate the gradient descent dynamics with a linear discrete dynamical system defined in terms of the so-called neural tangent kernel $\Theta(x, x')$ defined on the manifolds. Due to the structure in the data, diagonalizing the operator corresponding to this kernel is intractable in general, but we show that constructing a certificate—arguably an easier task, because it requires producing a bound on the
norm of a solution to an equation rather than producing the solution itself—suffices to guarantee that the error decreases rapidly during training given a suitably structured network.

We summarize the primary contributions of this work below.

• **Generalization in deep networks:** There are few generalization results for deep networks trained efficiently with gradient descent available in the literature.\(^1\) Theorem 1 provides such a guarantee that does not depend on any property of the trained network (e.g., norms of final weights) that is not readily available before training. In this context, the certificate condition is equivalent to the initial network function having a controlled norm in a certain RKHS; this condition is natural in the training regime we consider, and appears ubiquitously in works on generalization in shallower networks [62, 63, 64].

• **Uniform concentration of the neural tangent kernel for deep ReLU networks:** As an intermediate step in the proof of Theorem 1, we establish essentially optimal rates of uniform concentration for the neural tangent kernel of an arbitrarily deep network (Theorem 2) using martingale concentration, where we require the width to grow only linearly with the depth. We expect this martingale approach to be applicable to essentially any other compositionally-structured network architecture. Our uniform result generalizes prior results on pointwise concentration [65, 66], analogous to our Theorem A.2.3, and proves useful in establishing generalization.

• **Strong regularity estimates for random ReLU networks:** As a further consequence of the uniform concentration framework we have developed, we obtain depth-logarithmic Lipschitz estimates for random ReLU networks of arbitrary depth and linear width, as well as (for still wider networks) a uniform approximation for the network output by a constant which improves with depth, both with overwhelming probability (Section 2.3.2). We also control the evolution of the Lipschitz constant during NTK regime training (Lemma A.2.6), showing

---

\(^1\) The closest result we are aware of is [61, Theorem 3.4]; this result involves a-priori assumptions on the trained network weights, which are only resolved for two-layer networks, and entails an unnatural relationship between \(n\) and \(N\) and a possible exponential dependence of \(N\) on \(L\), which Theorem 1 avoids.
that it scales polynomially in the depth. These results may be of interest in applications where guaranteeing a Lipschitz property for networks is important, such as GAN training [67] or denoising [68, 69].

2.1.1 Related Work

Deep networks and low-dimensional structure. The idea of modeling data as low-dimensional submanifolds has been widely studied in the context of clustering [70] and manifold learning [71, 72]. Goldt et al. [73] independently proposed the “hidden manifold model”, a model problem for learning shallow neural networks for binary classification of structured data with motivations very similar to ours and which admits a mean-field analysis [74]. The data model consists of gaussian samples from a low-dimensional subspace passed through a nonlinear function acting coordinate-wise in the standard basis; although this models statistical variations around a base domain, a feature of real data that the model we study here lacks, we believe that the study of an arbitrary density supported on two Riemannian manifolds lends our data model increased structural generality. In the context of kernel regression with the kernel given by the NTK of a two-layer neural network, Ghorbani et al. [62] study a data generating model that consists of uniform samples from a low-dimensional subsphere corrupted additively by independent uniform samples from a subsphere in the orthogonal complement, and a target mapping that depends only on the low-dimensional part. The authors obtain asymptotic generalization guarantees for this data model that reveal conditions under which the corruption degrades the performance of neural tangent methods.

Analyses of neural network training. To reason analytically about the complicated training process, we adopt the neural tangent kernel approach [75]. The first works to instantiate these ideas in a nonasymptotic setting obtained convergence guarantees for training deep neural networks on finite datasets [66, 76]. By exploiting more structure in the data, generalization results have been obtained [63, 77, 78, 79, 80, 81, 82, 83] that apply to shallow networks, teacher-student learning scenarios, and/or hold conditional on the existence of certain small-norm interpolators. Other
works have obtained generalization guarantees using generalization bounds for kernel methods [62, 84, 85, 86] using the fact that the linearized predictor in the NTK regime can be linked to a kernel method [65]. A parallel line of works [87, 88, 89, 90, 91] approach the problem by studying an infinite-width limit of neural network training that yields a different training dynamics. Approaches of this type are of interest because there is no restriction to short-time dynamics, and the limit of the dynamics can often be characterized in terms of a well-structured object, such as a max-margin classifier [90]. On the other hand, it is often difficult to prove finite-time convergence to the limit.

2.2 Problem Formulation and Main Results

2.2.1 Data Model

We consider data supported on the union of two class manifolds \( M = M_+ \cup M_- \), where \( M_+ \) and \( M_- \) are two disjoint, smooth, regular, simple curves taking values in \( S^{n_0 - 1} \), with \( n_0 \geq 3 \). We denote the data measure supported on \( M \) that generates our samples as \( \mu^\infty \), and require that it admits a density \( \rho \) with respect to the Riemannian measure on \( M \). We will need to worst-case aspects of the density \( \rho \) in our argument; we write

\[
\rho_{\text{min}} = \inf_{x \in M} \rho(x); \quad \rho_{\text{max}} = \sup_{x \in M} \rho(x).
\]

For our bounds in certain places to be nonvacuous, we will need \( \rho_{\text{min}} > 0 \). We denote by \( \kappa \) a uniform bound on the curvature of the two curves, and because we consider submanifolds of the unit sphere, \( \kappa \geq 1 \). The separation between class manifolds is written as

\[
\Delta = \min_{x \in M_+, x' \in M_-} \angle(x, x'),
\]

and we require \( \Delta > 0 \). We denote by \( \text{dist}_M(\cdot, \cdot) \) the Riemannian distance between two points on the same connected component of \( M \). Additional details about our assumptions are provided in
Given the data measure $\mu^\infty$ supported on $\mathcal{M}$ described in Section 2.2.1, we can formulate our target function as $f_\ast : \mathcal{M} \to \{\pm 1\}$, with

$$f_\ast(x) = \begin{cases} +1 & x \in \mathcal{M}_+ \\ -1 & x \in \mathcal{M}_- \end{cases},$$

which we learn using a fully-connected neural network with ReLU activations and access to i.i.d. samples from $\mu^\infty$ and their corresponding labels. We parameterize our neural network with weights $W^1 \in \mathbb{R}^{n_0 \times n_0}$, $W^\ell \in \mathbb{R}^{n \times n}$ if $\ell \in \{2, \ldots, L\}$, and $W^{L+1} \in \mathbb{R}^{1 \times n}$, which we collect as $\theta = (W^1, \ldots, W^{L+1})$, and write the iterates of the forward pass as

$$\alpha^0_\theta(x) = x; \quad \alpha^\ell_\theta(x) = \left[W^\ell \alpha^{\ell-1}_\theta(x)\right]_+, \quad \ell = 1, 2, \ldots, L,$$

which we also refer to as features or activations, with the network output written as $f_\theta(x) = W^{L+1}\alpha^L_\theta(x)$, and the prediction error as $\zeta_\theta(x) = f_\theta(x) - f_\ast(x)$. For an i.i.d. sample $(x_1, \ldots, x_N)$ from $\mu^\infty$, we write $\mu^N = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ for the empirical measure associated to the sample, and we consider the training objective

$$\mathcal{L}_{\mu^N}(\theta) = \frac{1}{2} \int_{\mathcal{M}} (\zeta_\theta(x))^2 \, d\mu^N(x) = \frac{1}{2N} \sum_{i=1}^N (f_\theta(x_i) - f_\ast(x_i))^2,$$  \hspace{1cm} (2.1)

i.e. the empirical risk evaluated with the square loss. Our algorithm for optimizing (2.1) is vanilla gradient descent with constant step size $\tau > 0$: after randomly initializing the parameters $\theta^N_0$ as $W^\ell \sim_{\text{i.i.d.}} \mathcal{N}(0, 2/n)$ if $\ell \in [L]$ and $W^{L+1} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$ independently of the sample from $\mu^\infty$, we
consider the sequence of iterates

\[
\theta_{k+1}^N = \theta_k^N - \tau \nabla \mathcal{L}_{\mu}^N(\theta_k^N),
\]  

(2.2)

where \( \nabla \mathcal{L}_{\mu}^N \) represents a ‘formal gradient’ of the empirical loss, which we define in detail in Section A.1.1.\(^2\) This choice of initialization guarantees stable forward propagation prior to training: in expectation, the initial feature norms at each layer are unity, and the network output matches the scale of \( f^* \).

Now we can articulate the quantitative version of the task in theorem 1: we say the parameters obtained at iteration \( k \) of gradient descent (2.2) separate the manifolds \( \mathcal{M} \) if the classifier implemented by the neural network with the parameters \( \theta_k^N \) labels the two manifolds correctly, i.e. if

\[
\forall \mathbf{x} \in \mathcal{M}_+, \ \text{sign} \left( f_{\theta_k^N}(\mathbf{x}) \right) = +1 \quad \text{and} \quad \forall \mathbf{x} \in \mathcal{M}_-, \ \text{sign} \left( f_{\theta_k^N}(\mathbf{x}) \right) = -1.
\]  

(2.3)

In the sequel, we will denote evaluation of quantities such as the features and prediction error at parameters along the gradient descent trajectory using a subscript \( k \), with an omitted subscript denoting evaluation at the initial \( k = 0 \) parameters, and we will add a superscript \( N \) to parameters such as the prediction error to emphasize that they are evaluated at the parameters generated by (2.2). For example, in this notation we express \( \zeta_{\theta_k^N} \) as \( \zeta_{\theta_0^N} \). In addition, we will use \( \theta_0 = \theta_0^N \) to denote the random initial parameters. We will emphasize the dependence of certain quantities on these random initial parameters notationally, including the initial network function \( f_{\theta_0} \).

2.2.3 Error Dynamics and Certificates

Because it is difficult to endow the network parameters generated by the gradient iteration with a particular interpretation, we prefer to reason about how the network error \( \zeta_k^N \) evolves under

\(^2\)We introduce these definitions to cope with nonsmoothness of the ReLU \( [ \cdot ]_+ \). Our formal gradient definitions coincide with the expressions one obtains by applying the chain rule to differentiate \( \mathcal{L}_{\mu}^N \) at points where the ReLU is differentiable, and we make use of this fact to proceed with these formal gradients in a manner almost identical to the differentiable setting.
gradient descent. We calculate (in Lemma A.2.7)

\[
\zeta_{k+1}^N(x) = \zeta_k^N(x) - \tau \int_M \Theta_k^N(x, x') \zeta_k^N(x') \, d\mu^N(x'),
\]

(2.4)

where we have defined the integral kernel \( \Theta_k^N(x, x') = \int_0^1 \langle \nabla \tilde{f}_{\theta_0}^N(x'), \nabla \tilde{f}_{\theta_0}^N(x) \rangle \, dt \), where \( \tilde{f}_{\theta_0} \) denotes a formal gradient of the initial network function with respect to the parameters, which is defined in detail in Section A.1.1. We then define a nominal error evolution by

\[
\zeta_{k+1}^\infty(x) = \zeta_k^\infty(x) - \tau \int_M \Theta(x, x') \zeta_k^\infty(x') \, d\mu^\infty(x'),
\]

(2.5)

with identical initial conditions \( \zeta_0^\infty = \zeta \) and where \( \Theta(x, x') = \langle \nabla \tilde{f}_{\theta_0}(x), \nabla \tilde{f}_{\theta_0}(x') \rangle \) is the so-called neural tangent kernel with associated integral operator \( \Theta \). We prove that the error evolution (2.4) is well-approximated by the nominal error evolution under suitable conditions on the network width, step size, and number of samples, which together ensure that training proceeds in the “NTK regime” where \( \Theta_k^N \) stays close to \( \Theta \). As for the nominal error evolution (2.5), we note that this system is linear, time-invariant, and stable when \( \tau \) is set appropriately small, so the norm of the nominal error is guaranteed to decrease rapidly if the initial error \( \zeta \) aligns well with eigenfunctions of \( \Theta \) corresponding to large eigenvalues. However, computation of these eigenfunctions is intractable for general data geometries and distributions because the operator \( \Theta \) is not generally translationally invariant on \( M \). To overcome this issue, we prove this alignment implicitly by constructing an approximate solution to the linear integral equation \( \Theta[g] = \zeta \) such that \( \|g\|_{L^2_{\mu^\infty}} \) is sufficiently small. To be precise, \( g \in L^2_{\mu^\infty} \) will be called a \( \delta_1, \delta_2 \)-certificate for the dynamics (2.5) if

\[
\|\Theta[g] - \zeta\|_{L^2_{\mu^\infty}} \leq \delta_1; \quad \|g\|_{L^2_{\mu^\infty}} \leq \delta_2.
\]

(2.6)

2.2.4 Main Results and Proof Outline

Our main result is that conditional on the existence of a certificate of suitably small norm for \( M \), gradient descent provably separates the two manifolds in time polynomial in the network
Theorem 1. Let $\mathcal{M}$ be a one-dimensional Riemannian manifold satisfying our regularity assumptions. For any $0 < \delta \leq 1/e$, choose

\[
L \geq C_1 \max \{C_{\mu^\infty} \log^9(1/\delta) \log^2(1/\delta) + K^2 \lambda / c^2 \},
\]
\[
n = C_2 L^{99} \log^9(1/\delta) \log(1/\delta),
\]
\[
N \geq L^{10},
\]

and fix $\tau$ such that $C_4 \frac{9}{nL^2} \leq \tau \leq C_4 \frac{1}{nL}$.

Then if there exists a certificate in the sense of (2.6) with $\delta_1 = C_5 C_{\rho}^{1/2} \sqrt{\log(1/\delta) \log(n \rho_0)/L}$ and $\delta_2 = C_6 \sqrt{\log(1/\delta) \log(n \rho_0)/(n \rho_{\min}^{1/2})}$, with probability at least $1 - \delta$ over the random initialization of the network and the i.i.d. sample from $\mu^\infty$, the parameters obtained at iteration $[L^{39/44}/(n \tau)]$ of gradient descent on the finite sample loss $\mathcal{L}_{\mu^N}$ yield a classifier that separates the two manifolds.

The constants $C_1, \ldots, C_6$ are suitably chosen absolute constants, the constants $\kappa, K, c, c_{\lambda}$ are respectively the extrinsic curvature constant and the global regularity constant defined in Section A.1, the constant $C_{\rho}$ is defined as $\max\{\rho_{\min}, \rho_{\max}^{-1}\}$, and the constant $C_{\mu^\infty}$ is defined as $C_{\rho}^{15} (1 + \rho_{\max})^6 \min\{\mu^\infty(\mathcal{M}_+), \mu^\infty(\mathcal{M}_-))^{-11/2}.$

For one-dimensional instances of the two manifold problem with sufficiently deep and over-parameterized networks trained in the small-step-size regime, Theorem 1 completely reduces the analysis of the gradient iteration to the certificate problem. From a qualitative perspective, the network resource constraints imposed by Theorem 1 are natural:

(i) The network depth $L$ is set by geometric and statistical properties of the data with only a mild polylogarithmic dependence on the ambient dimension $n_0$, which reflects the role of depth in controlling the capability of the network to fit functions.

(ii) The network width $n$ is set by the depth $L$: the inductive structure of the network causes
quantities that depend on the initial random weights \( \theta_0 \) to concentrate worse as the depth is increased, which can be counteracted by setting the width appropriately large.

(iii) The sample complexity of \( N \geq L^{10} \) reflects the capacity of the network via the depth, and is in particular independent of the width \( n \), which can thus be interpreted as purely a statistical resource.

In addition, the conclusion of Theorem 1 implies not just that the expected generalization error with respect to \( \mu^\infty \) of a binary classifier is zero, but the stronger separation property, i.e. that the generalization error will be zero for any choice of test distribution supported on \( \mathcal{M} \) simultaneously.

![Figure 2.2](image_url)

**(a)** Depth acts as a fitting resource. As \( L \) increases, the rotationally-invariant kernel \( \hat{\Theta} \) (a slight modification of the deterministic kernel in Theorem 2) decays more rapidly as a function of angle between the inputs \( \angle (x, x') \) (\( n \) is held constant). Below the curves we show an isometric chart around a point \( x \in \mathcal{M}_e \). Once the decay scale of \( \hat{\Theta} \) is small compared to the inter-manifold distance \( \Delta \) and the curvature of \( \mathcal{M}_e \), the network output can be changed at \( x \) while only weakly affecting its value on \( \mathcal{M}_e \). This is one mechanism that relates the depth required to solve the classification problem to the data geometry. **(b)** Width acts as a statistical resource. The dynamics at initialization are governed by \( \Theta \), a random process over the network parameters. As \( n \) is increased, the normalized fluctuations of \( \Theta \) around \( \hat{\Theta} \) decrease (here \( L = 10 \)). These two phenomena are related, since the fluctuations also grow with depth, as evinced by the scaling in theorem 2.

To prove that the nominal error evolution (2.5) decreases rapidly and approximates the actual error evolution (2.4) throughout training, it is essential to have a precise characterization of the ‘initial’ neural tangent kernel \( \Theta \). One of our main technical contributions is to show concentration of \( \Theta \) in the regime where the width \( n \) scales linearly with the depth \( L \).
Theorem 2. For any \( d_0 \in \mathbb{N} \), let \( \mathcal{M} \) be a \( d_0 \)-dimensional complete Riemannian submanifold of \( \mathbb{S}^{n_0-1} \). Then if \( n \geq C_1 L d_0^4 \log^4 (C_M n_0 L) \), one has with probability at least \( 1 - n^{-10} \) that for every \((x, x') \in \mathcal{M} \times \mathcal{M}\)

\[
\left| \Theta(x, x') - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(v) \right) \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(v)}{\pi} \right) \right| \leq C_2 \sqrt{n L^3 d_0^4 \log^4 (C_M n_0 L)},
\]

where we write \( v = \angle(x, x') \) with an abuse of notation, \( \varphi^{(\ell)} \) denotes the \( \ell \)-fold composition of \( \varphi(v) = \cos^{-1} \left( (1 - \frac{v}{\pi}) \cos v + \frac{\sin v}{\pi} \right) \), the constants \( C_1, C_2 > 0 \) are absolute, and the constant \( C_M > 0 \) depends only on the diameters and curvatures of the class manifolds (Lemma A.3.1).

For networks of uniform width that are wider than they are deep by a certain constant factor, we believe that the scalings in Theorem 2 are essentially optimal: the variance calculations of Hanin and Nica [92] give some heuristic evidence here, and we believe the idea of using diagonal concentration to prove deviation lower bounds could be generalized to rigorously establish optimality. Figure 2.2b illustrates the phenomenon underlying Theorem 2. We discuss the proof of Theorem 2 in more detail in Sections 2.3.1 and 2.3.2.

Proof Sketch

In Section A.2, we prove a slightly more general version of Theorem 1 in Theorem A.2.1. Here, we give a brief outline of the proof of this result. Proving the separation property essentially requires us to obtain control of \( \| \xi_k^N \|_{L^\infty(M)} \), and by an interpolation inequality (Lemma A.2.13) it suffices to control the generalization error \( \| \xi_k^N \|_{L^2_{\mu^\infty}} \) and the smoothness (measured through the Lipschitz constant) of \( \xi_k^N \). We start with the generalization error, picking up from where we left off at the end of Section 2.2.3: the triangle inequality gives

\[
\| \xi_k^N \|_{L^2_{\mu^\infty}} \leq \| \xi_k^\infty \|_{L^2_{\mu^\infty}} + \| \xi_k^\infty - \xi_k^N \|_{L^2_{\mu^\infty}},
\]

---

3Since we do not use the “NTK parameterization”, the norm of our NTK scales like \( nL \) rather than \( L \). Due to our scaling of the weights (Section A.1.3) the contribution of the final layer to the NTK is negligible and can be dropped. This leads to discrepancies between the expression above and similar expressions found in the literature—we show essential equivalence between our NTK and others in Section A.1.3.
which allows us to divide the analysis into two subproblems: characterizing the nominal dynamics (Lemmas A.2.5 and A.2.11), and the nominal-to-finite transition (Lemma A.2.6). Beginning with the nominal dynamics, we use (2.5) to write

$$\zeta_k^\infty = (\text{Id} - \tau \Theta)^k [\zeta],$$

where $\Theta$ denotes the operator on $L^2_{\mu^\infty}$ corresponding to integration against the kernel $\Theta$ and $\text{Id}$ denotes the identity operator. The definition of $\Theta$ and compactness of $\mathcal{M}$ imply that $\Theta$ is a positive, compact operator (Lemma A.2.8), so these dynamics are stable when $\tau$ is chosen larger than the operator norm of $\Theta$. However, the rate of decrease of $\|\zeta_k^\infty\|_{L^2_{\mu^\infty}}$ with $k$ could still be extremely slow if the initial error $\zeta$ has significant components in the direction of eigenfunctions of $\Theta$ corresponding to small eigenvalues, and because $\Theta$ acts roughly like a convolution operator, we expect there to exist eigenvalues arbitrarily close to zero. By solving the certificate problem (2.6), we can assert that misalignment does not occur. To solve the certificate problem, we work with analytically-convenient approximations for $\Theta$ and $\zeta$: the exact definitions of these approximations $\hat{\Theta}$ and $\hat{\zeta}$ are given in Section A.1.4, and we prove their suitability as approximations in Theorem A.2.2 (a slightly more general version of Theorem 2) and Lemma A.4.11, respectively. As we have discussed in Section 2.3.1, our rates of concentration for $\Theta$ about $\hat{\Theta}$ are essentially optimal—the poor rates that end up appearing in Theorem A.2.1 are set by later parts of the argument.

With our approximation to $\Theta$ justified, we show that for any sufficiently small step size $\tau$ and number of iterations $k$, solving the certificate problem (2.6) guarantees appropriate decrease of the nominal generalization error.

To complete the proof, we will justify the nominal-to-finite transition in (2.7). Starting from the update equations (2.4) and (2.5), subtracting and rearranging gives an update equation for the difference:

$$\hat{\zeta}_k^N - \zeta_k^\infty = (\text{Id} - \tau \Theta) \left[ \zeta_k^N - \zeta_{k-1}^\infty \right] - \tau \Theta_k^N \left[ \zeta_k^N \right] + \tau \Theta \left[ \zeta_{k-1}^N \right].$$

In particular, if $\tau$ is chosen less than the operator norm of $\Theta$, we can take norms on both sides of
the previous equation, apply the triangle inequality, then exploit a telescoping series cancellation to obtain the difference bound

\[ \| \phi_k^\infty - \phi_k^N \|_{L^2_{\mu^\infty}} \leq \tau \sum_{s=0}^{k-1} \left\| \int_{\mathcal{M}} \Theta_s^N (\cdot, x') \phi_s^N (x') \, d\mu^N (x') - \int_{\mathcal{M}} \Theta (\cdot, x') \phi_s^N (x') \, d\mu^\infty (x') \right\|_{L^2_{\mu^\infty}} \]  \hspace{1cm} (2.8)

There are two obstacles to controlling the norm terms on the RHS of (2.8): the kernels \( \Theta_s^N \) are distinct from the kernel \( \Theta \) due to changes in the weights that occur during training, and the empirical measure \( \mu^N \) incurs a sampling error relative to the population measure \( \mu^\infty \). To address the first challenge, we measure the changes to the NTK during training in a worst-case fashion as

\[ \Delta_k^N = \max_{i \in \{0, 1, \ldots, k\}} \| \Theta_i^N - \Theta \|_{L^\infty (\mathcal{M} \times \mathcal{M})}, \]

and train in the NTK regime, where the network width \( n \) is larger than a large polynomial in the depth \( L \) and the total training time \( k\tau \) is no larger than \( L/n \). These conditions imply that with high probability \( \Delta_k^N \) is no larger than a constant multiple of \( n^{1-\delta} \text{poly}(L, d_0) \) for a small constant \( \delta > 0 \), so that the amortized changes during training \( k\tau \Delta_k^N \) can be made small by sufficient overparameterization. By the preceding argument, we can use the triangle inequality and Jensen’s inequality to pass from the norm term in (2.8) to a difference-of-measures term which integrates against \( \Theta \), and by Theorem A.2.2, we can replace the integration against \( \Theta \) by an integration against a smooth, deterministic kernel, which leads to a bound

\[ \| \phi_k^\infty - \phi_k^N \|_{L^2_{\mu^\infty}} \leq R_k (n, L, d_0) + \tau \sum_{s=0}^{k-1} \left\| \int_{\mathcal{M}} \psi_1 (\angle (\cdot, x')) \phi_s^N (x') \left( d\mu^N (x') - d\mu^\infty (x') \right) \right\|_{L^2_{\mu^\infty}}, \]

where \( R_k \) is a residual term that we argue is small in the NTK regime with high probability, and for concision we write \( \psi_1 \) to denote the function of \( \angle (x, x') \) that appears in Theorem A.2.2. To control the remaining term, we make use of a basic result from optimal transport theory, which states that for any probability measure \( \mu \) on the Borel sets of a metric space \( X \) and corresponding empirical
measure \( \mu^N \), one has for every Lipschitz function \( f \)

\[
\int_X f(x) \left( \text{d}\mu(x) - \text{d}\mu^N(x) \right) \leq \|f\|_{\text{Lip}} W(\mu, \mu^N),
\]

where \( W(\cdot, \cdot) \) denotes the 1-Wasserstein metric, and concentration inequalities for empirical measures in the 1-Wasserstein metric [93]. To apply this result to our setting, it is necessary to control the change throughout training of the Lipschitz constant of \( \zeta_k^N \), and one must also account for the fact that the metric space in our setting is \( M \), which has two distinct connected components. We treat the first issue using an inductive argument, and our treatment of the second issue (Lemma A.2.12) leads to the dependence on the degree of class imbalance demonstrated in the constant \( C_{\mu^\infty} \) in Theorem A.2.1.

2.3 Key Proof Elements

2.3.1 Concentration at Initialization: Martingales and Angle Contraction

The initial kernel \( \Theta \) is a complicated random process defined over the weights \( (W^1, \ldots, W^{L+1}) \). To control it, we first show for fixed \( (x, x') \) that \( \Theta(x, x') \) concentrates with high probability, and then leverage approximate continuity properties to pass to uniform control of \( \Theta \). Here we describe our approach to pointwise control; uniformization is discussed in Section 2.3.2. The kernel can be written in the form

\[
\Theta(x, x') = \langle \alpha^L(x), \alpha^L(x') \rangle + \sum_{\ell=0}^{L-1} \langle \alpha^\ell(x), \alpha^\ell(x') \rangle \langle \beta^\ell(x), \beta^\ell(x') \rangle,
\]

where \( \beta^\ell(x) = (W^{L+1}P_{f_L(x)} \cdots W^{\ell+2}P_{f_{\ell+1}(x)})^* \) will be referred to as backward features, and \( P_{f_{\ell+1}(x)} \) is a projection onto \( \{\alpha^\ell(x) > 0\} \). We consider \( \langle \beta^0(x), \beta^0(x') \rangle \) as a representative example: up to a small residual term, this random variable can be expressed as a sum of martingale differences. Formally, for \( \ell \in [L] \), let \( \mathcal{F}_\ell^\ell \) denote the \( \sigma \)-algebra generated by all weight matrices up to layer \( \ell \),
with $\mathcal{F}^0$ denoting the trivial $\sigma$-algebra. We can then write

$$\left| \langle \beta^0(\mathbf{x}), \beta^0(\mathbf{x}') \rangle - g_0(\nu^0) \right| \leq \left[ \sum_{\ell=1}^{L+1} g_\ell(W^\ell, \ldots, W^1, \nu^0) - \mathbb{E}[g_\ell(W^\ell, \ldots, W^1, \nu^0) \mid \mathcal{F}^{\ell-1}] \right] + R \quad (2.9)$$

for some functions $g_\ell$ and controllable residual $R$, where $\nu^0 = \angle(\mathbf{x}, \mathbf{x}')$. If we fix all the variables in $\mathcal{F}^{\ell-1}$, the fluctuations in the $\ell$-th summand will be due to $W^\ell$ alone. Intuitively, since each weight matrix appears at most once in $\beta^0(\mathbf{x})$, it will appear at most twice in $g_\ell$, and therefore $g_\ell$ will have a subexponential distribution conditioned on $\mathcal{F}^{\ell-1}$ and concentrate well around its conditional expectation. This property stems from the compositional structure of the network, with independent sources of randomness introduced at every layer, and is essentially agnostic to other details of the architecture. The concentration of the summands in (2.9) implies concentration of the sum: even though the summands are not independent, they can be controlled using concentration inequalities analogous to those for sums of independent variables [94, 95].

Showing that terms of the form $\langle \alpha^\ell(\mathbf{x}), \alpha^\ell(\mathbf{x}') \rangle$ concentrate in the linear regime gives rise to additional challenges. Here we exploit an essential difference between the concentration properties of the angles between features $\nu^\ell = \angle(\alpha^\ell(\mathbf{x}), \alpha^\ell(\mathbf{x}'))$ relative to those of the correlation process $\langle \alpha^\ell(\mathbf{x}), \alpha^\ell(\mathbf{x}') \rangle$ studied in prior works on concentration of $\Theta$: when $\nu^{\ell-1} = 0$, we have that $\nu^\ell = 0$ deterministically, whereas the correlation process behaves like a subexponential random variable with small but nonzero deviations. Together with smoothness, this clamping phenomenon allows us to show concentration of the angle at layer $\ell$ around the function $\varphi^{(\ell)}(\nu^0)$, which is no larger than a constant multiple of $\ell^{-1}$. This contraction of the angles with depth is the key to establishing Theorem 2; in addition, it gives the invariant kernel $\hat{\Theta}$ (see Figure 2.2b) its sharpness at zero and localization properties. We provide full details of our approach in Sections A.4 and A.5.

---

4 Technically, the features $\alpha^\ell(\mathbf{x})$ depend on all the weights up to layer $\ell$ and hence so does the projection matrix $P_{\ell(\mathbf{x})}$, but our analysis shows that this dependence has only a minor effect on the statistical fluctuations.
2.3.2 Uniform Concentration and its Consequences

To uniformize the pointwise estimates of Section 2.3.1, we must overcome the issue that the backward features $\beta'(x)$ are not continuous functions of the input, due to the matrices $P_{l_r}(x)$. Our approach is to discretize the input space, control the number of features that can change sign near each point in the discretization, then extend the pointwise estimates of Section 2.3.1 to the setting where a small number of features have changed sign—again, we find martingale concentration a necessity to achieve linear width-depth scaling. We give full details in Section A.4.3.

Although Theorem 2 is the main application of these estimates—with uniform control of $\Theta$, we can prove operator norm bounds on its corresponding integral operator $\Theta$, which is of great help in proving generalization results—they also imply useful regularity estimates for the initial random network $f_{\theta_0}$. For example, we prove that networks of uniform width $n \asymp n_0^d L^2$ are with high probability $\sqrt{n_0(\log n_0)(\log L)}$-Lipschitz as functions on $\mathbb{R}^{n_0}$ (Theorem A.2.4)—in particular, the Lipschitz constant depends only logarithmically on depth, in contrast to existing results in the literature [96]. For networks of larger width $n \gtrsim d_0^3 L^5$, we prove that with high probability the network $f_{\theta_0}$ is approximately constant on the domain $M \subset S^{n_0-1}$ (Lemma A.4.11):

$$\sup_{x \in M} \left| f_{\theta_0}(x) - \int_M f_{\theta_0}(x') \, d\mu^\infty(x') \right| \lesssim \frac{1}{L}.$$

2.4 Discussion

The importance of being low-dimensional. Ghorbani et al. [84] show that kernel ridge regression with any rotationally invariant kernel on $S^d$ (including that of a deep network) is equivalent to polynomial regression with a degree $p$ polynomial if the number of samples is bounded by $d^{p+1}$ and $d \to \infty$. For data lying on a low-dimensional manifold, as we consider here, one would expect less pessimistic rates; indeed, in a subsequent work [62] the authors establish similar guarantees to Ghorbani et al. [84] for a linear data model with low-dimensional structure in terms of a smaller “effective dimension”. In comparison, the general multiple manifold problem formulation allows
one to model nonlinear structure in the data, and measures fitting difficulty through intrinsic parameters like the curvature and separation. The guarantees in Ghorbani et al. [62, 84] depend on the degree of approximability of the target function by low-degree polynomials, and although this achieves additional generality over our model, it seems more challenging to relate this to geometric or other types of nonlinear low-dimensional structure.

The NTK regime and beyond. In recent years there has been much work devoted to the analysis of networks trained in the regime where the changes in $\Theta^N_k$ remain small and the dynamics in (2.4) are close to linear [65, 75, 97, 98] (referred to as the NTK/“overparametrized”/kernel regime). Concurrently, there have also been results highlighting the limitations of this regime. In [99] the authors coin the term “lazy training” in referring to dynamics where the relative change in the differential of the network function is small compared to the change in the objective during gradient descent. While the dynamics we study indeed fall into this category, the analysis makes it evident that not all lazy training regimes are created equal. Our performance guarantees depend on the structure of the kernel $\hat{\Theta}$, and on controlling the fluctuations of $\Theta^N_k$ around it. We are able to control these only if the width of the network is sufficiently large compared to the depth. In contrast, lazy training can also be achieved in homogeneous models by simply scaling the output of the model [99], in which case one cannot argue that the kernel has the decay properties that enable it to fit data.

Our analysis hinges on staying in the NTK regime during training. We obtain suboptimal scaling of $n$ with $L$ in Theorem 1 because we treat all changes that occur in $\Theta^N_k$ during training as being adversarial to the algorithm’s ability to generalize. It is likely that if an improved understanding of feature learning can be incorporated into an analysis of the dynamics, the resulting scaling requirements would be more realistic.
Chapter 3: Construction of Certificates: The Roles of Network Depth and Data Geometry

Our study of the multiple manifold problem in Chapter 2 reduced the analysis of the complex dynamics of gradient descent training on the empirical square loss to the auxiliary certificate problem (2.6), which asks us to construct a stable solution to a certain linear integral equation involving the initial random network, its gradients, and the data manifold $M$. In our analysis, although some conditions on the network depth were required relative to the data structure, the interaction with the geometry of the data did not fully come through—indeed, the certificate problem captures the interesting aspects of this interaction between network capacity and data geometry.

In this chapter of the thesis, we will take up the solution of the certificate problem for the case of one-dimensional curves satisfying minimal regularity assumptions that we considered in Chapter 2, and understand the role depth plays in solving the multiple manifold problem clearly. Solving the certificate problem amounts to proving that the target label function lies near the stable range of the NTK $\Theta$. The existence of certificates (and more generally, the conditions under which practically-trained neural networks can fit structured data) is open, except for a few very simple geometries which we will review below. Our technical contribution is to show that setting the network depth sufficiently large relative to intrinsic properties of the data guarantees the existence of a certificate (Theorem 3.3.1), resolving the one-dimensional case of the multiple manifold problem for a broad class of curves (Theorem 3.3.2). This leads in turn to a novel perspective on the role of the network depth as a fitting resource in the classification problem, which is inaccessible to shallow networks.
3.1 Related Work

Deep networks and low dimensional structure. Modern applications of deep neural networks include numerous examples of low-dimensional manifold structure, including pose and illumination variations in image classification [42, 43], as well as detection of structured signals such as electrocardiograms [100, 101], gravitational waves [58, 59], audio signals [102], and solutions to the diffusion equation [103]. Conventionally, to compute with such data one might begin by extracting a low-dimensional representation using nonlinear dimensionality reduction (“manifold learning”) algorithms [70, 104, 105, 106, 107, 108, 109]. For supervised tasks, there is also theoretical work on kernel regression over manifolds [110, 111, 112, 113]. These results rely on very general Sobolev embedding theorems, which are not precise enough to specify the interplay between regularity of the kernel and properties of the data need to obtain concrete resource tradeoffs in the two curve problem. There is also a literature which studies the resource requirements associated with approximating functions over low-dimensional manifolds [34, 35, 37, 38]: a typical result is that for a sufficiently smooth function there exists an approximating network whose complexity is controlled by intrinsic properties such as the dimension. In contrast, we seek algorithmic guarantees that prove that we can efficiently train deep neural networks for tasks with low-dimensional structure. This requires us to grapple with how the geometry of the data influences the dynamics of optimization methods.

3.2 Problem Formulation

3.2.1 The Two Curve Problem

The content of this section follows the notation, exposition, and assumptions we have discussed in Chapter 2, in large part, but due to some minor notational differences and added/more precise assumptions, we will reproduce it here. This will also allow us to re-motivate the essential nature of the certificate problem to the understanding of gradient descent’s dynamics. We omit some nonessential definitions and derivations for concision; see Section B.2.1 for these details.
A natural model problem for the tasks discussed in Section 2.1 is the classification of low-dimensional submanifolds using a neural network. In this work, we study the one-dimensional, two-class case of this problem, which we refer to as the two curve problem. To fix ideas, let $n_0 \geq 3$ denote the ambient dimension, and let $\mathcal{M}_+$ and $\mathcal{M}_-$ be two disjoint smooth regular simple closed curves taking values in $\mathbb{S}^{n_0-1}$, which represent the two classes (Figure 2.1). In addition, we require that the curves lie in a spherical cap of radius $\pi/2$: for example, the intersection of the sphere and the nonnegative orthant $\{x \in \mathbb{R}^{n_0} \mid x \geq 0\}$.

Given $N$ i.i.d. samples $\{x_i\}_{i=1}^N$ from a density $\rho$ supported on $\mathcal{M} = \mathcal{M}_+ \cup \mathcal{M}_-$, which is bounded above and below by positive constants $\rho_{\text{max}}$ and $\rho_{\text{min}}$ and has associated measure $\mu$, as well as their corresponding $\pm 1$ labels, we train a feedforward neural network $f_\theta : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$ with ReLU nonlinearities, uniform width $n$, and depth $L$ (and parameters $\theta$) by minimizing the empirical mean squared error using randomly-initialized gradient descent. Our goal is to prove that this procedure yields a separator for the geometry given sufficient resources $n$, $L$, and $N$—i.e., that $\text{sign}(f_{\theta_k}) = 1$ on $\mathcal{M}_+$ and $-1$ on $\mathcal{M}_-$ at some iteration $k$ of gradient descent.

To achieve this, we need an understanding of the progress of gradient descent. Let $f_* : \mathcal{M} \rightarrow \{\pm 1\}$ denote the classification function for $\mathcal{M}_+$ and $\mathcal{M}_-$ that generates our labels, write $\zeta_\theta(x) = f_\theta(x) - f_*(x)$ for the network’s prediction error, and let $\theta_{k+1} = \theta_k - (\tau/N) \sum_{i=1}^N \zeta_{\theta_k}(x_i) \nabla_{\theta} f_{\theta_k}(x_i)$ denote the gradient descent parameter sequence, where $\tau > 0$ is the step size and $\theta_0$ represents our Gaussian initialization. Elementary calculus then implies the error dynamics equation $\zeta_{\theta_{k+1}} = \zeta_{\theta_k} - (\tau/N) \sum_{i=1}^N \Theta^N_k(\cdot, x_i) \zeta_{\theta_k}(x_i)$ for $k = 0, 1, \ldots$, where $\Theta^N_k : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$ is the empirical, time-varying NTK we have seen in Chapter 3. The precise expression for this kernel is not important for our purposes: what matters is that (i) making the width $n$ large relative to the depth $L$ guarantees that $\Theta^N_k$ remains close throughout training to its ‘initial value’ $\Theta^\text{NTK}(x, x') = \langle \nabla_{\theta} f_{\theta_0}(x), \nabla_{\theta} f_{\theta_0}(x') \rangle$, the neural tangent kernel; and (ii) taking the sample size $N$ to be sufficiently large relative to the depth $L$ implies that a nominal error evolution defined

---

1The specific value $\pi/2$ is immaterial to our arguments: this constraint is only to avoid technical issues that arise when antipodal points are present in $\mathcal{M}$, so any constant less than $\pi$ would work just as well. This choice allows for some extra technical expediency, and connects with natural modeling assumptions (e.g. data corresponding to image manifolds, with nonnegative pixel intensities).
as \( \zeta_{k+1} = \zeta_k - \tau \Theta^{\text{NTK}}_{\mu}[\zeta_k] \) with \( \zeta_0 = \zeta_{\theta_0} \) uniformly approximates the actual error \( \zeta_{\theta_k} \) throughout training. In other words: to prove that gradient descent yields a neural network classifier that separates the two manifolds, it suffices to overparameterize, sample densely, and show that the norm of \( \zeta_k \) decays sufficiently rapidly with \( k \). This constitutes the “NTK regime” approach to gradient descent dynamics for neural network training [75].

The evolution of \( \zeta_k \) is relatively straightforward: we have \( \zeta_{k+1} = (\text{Id} - \tau \Theta^{\text{NTK}}_{\mu})^k[\zeta_0] \), and \( \Theta^{\text{NTK}}_{\mu} \) is a positive, compact operator, so there exist an orthonormal basis of \( L^2_\mu \) functions \( v_i \) and eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \) such that \( \zeta_{k+1} = \sum_{i=1}^{\infty} (1 - \tau \lambda_i)^k \langle \zeta_0, v_i \rangle_{L^2_\mu} v_i \). In particular, with bounded step size \( \tau < \lambda_1^{-1} \), gradient descent leads to rapid decrease of the error if and only if the initial error \( \zeta_0 \) is well-aligned with the eigenvectors of \( \Theta^{\text{NTK}}_{\mu} \) corresponding to large eigenvalues.

Arguing about this alignment explicitly is a challenging problem in geometry: although closed-form expressions for the functions \( v_i \) exist in cases where \( M \) and \( \mu \) are particularly well-structured, \textit{no such expression is available for general nonlinear geometries}, even in the one-dimensional case we study here. However, this alignment can be guaranteed implicitly if one can show there exists a function \( g : M \to \mathbb{R} \) of small \( L^2_\mu \) norm such that \( \Theta^{\text{NTK}}_{\mu}[g] \approx \zeta_0 \)—in this situation, most of the energy of \( \zeta_0 \) must be concentrated on directions corresponding to large eigenvalues. We call the construction of such a function the certificate problem (2.6):

**Certificate Problem.** Given a two curves problem instance \((M, \rho)\), find conditions on the architectural hyperparameters \((n, L)\) so that there exists \( g : M \to \mathbb{R} \) satisfying \( \| \Theta^{\text{NTK}}_{\mu}[g] - \zeta_0 \|_{L^2_\mu} \lesssim 1/L \) and \( \| g \|_{L^2_\mu} \lesssim 1/n \), with constants depending on the density \( \rho \) and logarithmic factors suppressed.

The construction of certificates demands a fine-grained understanding of the integral operator \( \Theta^{\text{NTK}}_{\mu} \) and its interactions with the geometry \( M \). We therefore proceed by identifying those intrinsic properties of \( M \) that will play a role in our analysis and results.
3.2.2 Key Geometric Properties

In the NTK regime described in Section 3.2.1, gradient descent makes rapid progress if there exists a small certificate $g$ satisfying $\Theta^\text{NTK}_\mu [g] \approx \zeta_0$. The NTK is a function of the network width $n$ and depth $L$—in particular, we will see that the depth $L$ serves as a fitting resource, enabling the network to accommodate more complicated geometries. Our main analytical task is to establish relationships between these architectural resources and the intrinsic geometric properties of the manifolds that guarantee existence of a certificate.

Intuitively, one would expect it to be harder to separate curves that are close together or oscillate wildly. In this section, we formalize these intuitions in terms of the curves’ curvature, and quantities which we term the angle injectivity radius and $\mathcal{G}$-number, which control the separation between the curves and their tendency to self-intersect. Given that the curves are regular, we may parameterize the two curves at unit speed with respect to arc length: for $\sigma \in \{\pm\}$, we write $\text{len}(\mathcal{M}_\sigma)$ to denote the length of each curve, and use $x_\sigma(s) : [0, \text{len}(\mathcal{M}_\sigma)] \to \mathbb{S}^{m_0-1}$ to represent these parameterizations. We let $x_\sigma^{(i)}(s)$ denote the $i$-th derivative of $x_\sigma$ with respect to arc length. Because our parameterization is unit speed, $\|x_\sigma^{(1)}(s)\|_2 = 1$ for all $x_\sigma(s) \in \mathcal{M}$. We provide full details regarding this parameterization in Section B.2.2.

Curvature and Manifold Derivatives. Our curves $\mathcal{M}_\sigma$ are submanifolds of the sphere $\mathbb{S}^{m_0-1}$. The curvature of $\mathcal{M}_\sigma$ at a point $x_\sigma(s)$ is the norm of the component $P_{x_\sigma(s)} x_\sigma^{(2)}(s)$ of the second derivative of $x_\sigma(s)$ that lies tangent to the sphere $\mathbb{S}^{m_0-1}$ at $x_\sigma(s)$. Geometrically, this measures the extent to which the curve $x_\sigma(s)$ deviates from a geodesic (great circle) on the sphere. Our technical results are phrased in terms of the maximum curvature $\kappa = \sup_{\sigma,s} \{\|P_{x_\sigma(s)} x_\sigma^{(2)}(s)\|_2\}$. In stating results, we also use $\hat{\kappa} = \max\{\kappa, \frac{2}{\pi}\}$ to simplify various dependencies on $\kappa$. When $\kappa$ is large, $\mathcal{M}_\sigma$ is highly curved, and we will require a larger network depth $L$. In addition to the maximum curvature $\kappa$, our technical arguments require $x_\sigma(s)$ to be five times continuously differentiable, and use bounds $M_l = \sup_{\sigma,s} \{\|x_\sigma^{(l)}(s)\|_2\}$ on their higher order derivatives.
Angle Injectivity Radius. Another key geometric quantity that determines the hardness of the problem is the separation between manifolds: the problem is more difficult when \( M_+ \) and \( M_- \) are close together. We measure closeness through the extrinsic distance \( \angle(x, x') = \cos^{-1}\langle x, x' \rangle \) between \( x \) and \( x' \) over the sphere, which we call the “angle”. In contrast, we use \( d_M(x, x') \) to denote the intrinsic distance between \( x \) and \( x' \) on \( M \), setting \( d_M(x, x') = \infty \) if \( x \) and \( x' \) reside on different components \( M_+ \) and \( M_- \). We set

\[
\Delta = \inf_{x, x' \in M} \{ \angle(x, x') \mid d_M(x, x') \geq \tau_1 \},
\] (3.1)

where \( \tau_1 = \frac{1}{\sqrt{20k}} \), and call this quantity the angle injectivity radius. In words, the angle injectivity radius is the minimum angle between two points whose intrinsic distance exceeds \( \tau_1 \). The angle injectivity radius \( \Delta \) (i) lower bounds the distance between different components \( M_+ \) and \( M_- \), and (ii) accounts for the possibility that a component will “loop back,” exhibiting points with large intrinsic distance but small angle. This phenomenon is important to account for: the certificate problem is harder when one or both components of \( M \) nearly self-intersect. At an intuitive level, this increases the difficulty of the certificate problem because it introduces nonlocal correlations across the operator \( \Theta^{\text{NTK}}_\mu \), hurting its conditioning. As we will see in Section 3.4, increasing depth \( L \) makes \( \Theta^{\text{NTK}}_\mu \) better localized; setting \( L \) sufficiently large relative to \( \Delta^{-1} \) compensates for these correlations.

\( \heartsuit \)-number The conditioning of \( \Theta^{\text{NTK}}_\mu \) depends not only on how near \( M \) comes to intersecting itself, which is captured by \( \Delta \), but also on the number of times that \( M \) can “loop back” to a particular point. If \( M \) “loops back” many times, \( \Theta^{\text{NTK}}_\mu \) can be highly correlated, leading to a hard certificate problem. The \( \heartsuit \)-number (verbally, “clover number”) reflects the number of near self-intersections:

\[
\heartsuit(M) = \sup_{x \in M} \left\{ N_M \left( \{ x' \mid d_M(x, x') \geq \tau_1, \angle(x, x') \leq \tau_2 \}, \frac{1}{\sqrt{1 + k^2}} \right) \right\}
\] (3.2)
with $\tau_2 = \frac{19}{2020 \sqrt{20}}$. The set $\{x' \mid d_M(x, x') \geq \tau_1, \angle(x, x') \leq \tau_2\}$ is the union of looping pieces, namely points that are close to $x$ in extrinsic distance but far in intrinsic distance. $N_M(T, \delta)$ is the cardinality of a minimal $\delta$ covering of $T \subset M$ in the intrinsic distance on the manifold, serving as a way to count the number of disjoint looping pieces. The $\mathcal{G}$-number accounts for the maximal volume of the curve where the angle injectivity radius $\Delta$ is active. It will generally be large if the manifolds nearly intersect multiple times, as illustrated in Figure 3.1. The $\mathcal{G}$-number is typically small, but can be large when the data are generated in a way that induces certain near symmetries, as in the right panel of Figure 3.1.

### 3.3 Main Results

Our main theorem establishes a set of sufficient resource requirements for the certificate problem under the class of geometries we consider here—by the reductions detailed in Section 3.2.1, this implies that gradient descent rapidly separates the two classes given a neural network of sufficient depth and width. First, we note a convenient aspect of the certificate problem, which is its amenability to approximate solutions: that is, if we have a kernel $\Theta$ that approximates $\Theta_{NTK}$ in the sense that $\|\Theta_\mu - \Theta_{NTK, \mu}\|_{L^2_\mu \to L^2_\mu} \leq n/L$, and a function $\zeta$ such that $\|\zeta - \zeta_0\|_{L^2_\mu} \leq 1/L$, then by the triangle inequality and the Schwarz inequality, it suffices to solve the equation $\Theta_\mu[g] = \zeta$ instead.

In our arguments, we will exploit the fact that the random kernel $\Theta_{NTK}$ concentrates well for wide networks with $n \gtrsim L^5$, depth ‘smooths out’ the initial error $\zeta_0$, choosing $\zeta$ as the piecewise-constant function $\zeta(x) = -f_\star(x) + \int_M f_\theta_0(x') \, d\mu(x')$. We have seen the rigorous justifications for these approximations in Chapter 3, in particular Theorem A.2.2, Lemma A.3.8, and Section A.4.3.

\[
\Theta(x, x') = \frac{n}{2} \sum_{\ell=0}^{L-1} \prod_{\ell' = \ell}^{L-1} \left(1 - \frac{1}{\pi} \varphi(\ell') \angle(x, x')\right),
\]  

(3.3)

where $\varphi(t) = \cos^{-1}((1-t/\pi) \cos t + (1/\pi) \sin t)$ and $\varphi^{(\ell')}$ denotes $\ell'$-fold composition of $\varphi$; as well as the fact that for wide networks with $n \gtrsim L^5$, depth ‘smooths out’ the initial error $\zeta_0$, choosing $\zeta$ as the piecewise-constant function $\zeta(x) = -f_\star(x) + \int_M f_\theta_0(x') \, d\mu(x')$. We have seen the rigorous justifications for these approximations in Chapter 3, in particular Theorem A.2.2, Lemma A.3.8, and Section A.4.3.
Figure 3.1: The $\otimes$-number—theory and practice. Left: We generate a parametric family of space curves with fixed maximum curvature and length, but decreasing $\otimes$-number, by reflecting ‘petals’ of a clover about a circumscribing square. We set $M_+$ to be a fixed circle with large radius that crosses the center of the configurations, then rescale and project the entire geometry onto the sphere to create a two curve problem instance. In the insets, we show a two-dimensional projection of each of the blue $M_-$ curves as well as a base point $x \in M_+$ at the center (also highlighted in the three-dimensional plots). The intersection of $M_-$ with the neighborhood of $x$ denoted in orange represents the set whose covering number gives the $\otimes$-number of the configuration (see (3.2)). Top right: We numerically generate a certificate for each of the four geometries at left and plot its norm as a function of $\otimes$-number. The trend demonstrates that increasing $\otimes$-number correlates with increasing classification difficulty, measured through the certificate problem: this is in line with the intuition we have discussed. Bottom right: t-SNE projection of MNIST images (top: a “four” digit; bottom: a “one” digit) subject to rotations. Due to the approximate symmetry of the one digit under rotation by an angle $\pi$, the projection appears to nearly intersect itself. This may lead to a higher $\otimes$-number compared to the embedding of the less-symmetric four digit. For experimental details for all panels, see Section B.1.

Theorem 3.3.1 (Approximate Certificates for Curves). Let $M$ be two disjoint smooth, regular, simple closed curves, satisfying $\angle(x, x') \leq \pi/2$ for all $x, x' \in M$. There exist absolute constants $C, C', C'', C'''$ and a polynomial $P = \text{poly}(M_3, M_4, M_5, \text{len}(M), \Delta^{-1})$ of degree at most 36, with degree at most 12 in $(M_3, M_4, M_5, \text{len}(M))$ and degree at most 24 in $\Delta^{-1}$, such that when

$$L \geq \max \left\{ \exp(C\text{len}(M)\kappa), \left(\Delta\sqrt{1 + \kappa^2}\right)^{-C''\otimes(M)}, C'''\kappa^{10}, P, \rho_{\text{max}}^{12} \right\},$$
there exists a certificate $g$ with $\|g\|_{L^2_{\mu}} \leq \frac{C\|\zeta\|_{L^2_{\mu}}}{\rho_{\text{min}} n \log L}$ such that $\|\Theta_\mu [g] - \zeta\|_{L^2_{\mu}} \leq \frac{\|\zeta\|_{L^\infty}}{L}.$

Theorem 3.3.1 is our main technical result in this chapter: it provides a sufficient condition on the network depth $L$ to resolve the approximate certificate problem for the class of geometries we consider, with the required resources depending only on the geometric properties we introduce in Section 3.2.2. Given the connection between certificates and gradient descent, Theorem 3.3.1 demonstrates that deeper networks fit more complex geometries, which shows that the network depth plays the role of a fitting resource in classifying the two curves. We provide a numerical corroboration of the interaction between the network depth, the geometry, and the size of the certificate in Figure 3.2. For any family of geometries with bounded $\mathcal{R}$-number, Theorem 3.3.1 implies a polynomial dependence of the depth on the angle injectivity radius $\Delta$, whereas we are unable to avoid an exponential dependence of the depth on the curvature $\kappa$. Nevertheless, these dependences may seem overly pessimistic in light of the existence of ‘easy’ two curve problem instances—say, linearly-separable classes, each of which is a highly nonlinear manifold—for which one would expect gradient descent to succeed without needing an unduly large depth. In fact, such geometries will not admit a small certificate norm in general unless the depth is sufficiently large: intuitively, this is a consequence of the operator $\Theta_\mu$ being ill-conditioned for such geometries.\footnote{Again, the equivalence between the difficulty of the certificate problem and the progress of gradient descent on decreasing the error is a consequence of our analysis proceeding in the kernel regime with the square loss—using alternate techniques to analyze the dynamics can allow one to prove that neural networks continue to fit such ‘easy’ classification problems efficiently (e.g. [63]).}

The proof of Theorem 3.3.1 is novel, both in the context of kernel regression on manifolds and in the context of NTK-regime neural network training. We detail the key intuitions for the proof in Section 3.4. As suggested above, applying Theorem 3.3.1 to construct a certificate is straightforward: given a suitable setting of $L$ for a two curve problem instance, we obtain an approximate certificate $g$ via Theorem 3.3.1. Then with the triangle inequality and the Schwarz inequality, we can bound

$$\|\Theta_\mu^{\text{NTK}} [g] - \zeta_0\|_{L^2_{\mu}} \leq \|\Theta_\mu^{\text{NTK}} - \Theta_\mu\|_{L^2_{\mu} \to L^2_{\mu}} \|g\|_{L^2_{\mu}} + \|\zeta_0 - \zeta\|_{L^2_{\mu}} + \|\Theta_\mu [g] - \zeta\|_{L^2_{\mu}}.$$
Figure 3.2: The effect of geometry and depth on the certificate. Left: The certificate $g$ computed numerically from the kernel $\Theta$ for depth $L = 50$ (defined in (3.3)) and the geometry from Figure 2.1 with a uniform density, graphed over the manifolds. Control of the norm of the certificate implies rapid progress of gradient descent, as reflected in Theorem 3.3.2. Comparing to Section 2.1, we note that the certificate has large magnitude near the point of minimum distance between the two curves—this is suggestive of the way the geometry sets the difficulty of the fitting problem. Right: To visualize the certificate norm more precisely, we graph the log-magnitude of the certificate for kernels $\Theta$ of varying depth $L$, viewing them through the arc-length parameterizations $x_\sigma$ for the curves (left: $M_+$; right: $M_-$). At a coarse scale, the maximum magnitude decreases as the depth increases; at a finer scale, curvature-associated defects are ‘smoothed out’. This indicates the role of depth as a fitting resource. See Section B.1 for further experimental details.

and leveraging suitable probabilistic control (see Section B.6) of the approximation errors in the previous expression, as well as on $\|\zeta\|_{L^2_{\mu}}$, then yields bounds for the certificate problem. Applying the reductions from gradient descent dynamics in the NTK regime to certificates we have developed in Chapter 2, we then obtain an end-to-end guarantee for the two curve problem.

**Theorem 3.3.2 (Generalization).** Let $M$ be two disjoint smooth, regular, simple closed curves, satisfying $\angle(x, x') \leq \pi/2$ for all $x, x' \in M$. For any $0 < \delta \leq 1/e$, choose $L$ so that

$$L \geq K \max \left\{ \frac{1}{\Delta \sqrt{1 + \kappa^2}} e^{C' \max\{\text{len}(M)\xi, \log(\hat{\kappa})\}}, P \right\}$$

$$n = K' L^{99} \log^9(1/\delta) \log^{18}(L n_0)$$

$$N \geq L^{10},$$

and fix $\tau > 0$ such that $\frac{C'}{n L^2} \leq \tau \leq \frac{e}{n L}$. Then with probability at least $1 - \delta$, the parameters obtained at iteration $\lfloor L^{39/44}/(n \tau) \rfloor$ of gradient descent on the finite sample loss yield a classifier
that separates the two manifolds.

The constants $c, C, C', C'', K, K' > 0$ are absolute, the constant $C_{\mu} = \frac{\max\{\rho_{\min}^{19}, \rho_{\min}^{19}\}(1+\rho_{\max})^{12}}{(\min \{\mu(M_\nu), \mu(M_\lambda)\})^{1/2}}$, $\mathcal{P}$ is a polynomial $\text{poly}\{M_3, M_4, M_5, \text{len}(M), \Delta^{-1}\}$ of degree at most 36, with degree at most 12 when viewed as a polynomial in $M_3, M_4, M_5$ and $\text{len}(M)$, and of degree at most 24 as a polynomial in $\Delta^{-1}$.

Theorem 3.3.2 represents the first end-to-end guarantee for training a deep neural network to classify a nontrivial class of low-dimensional nonlinear manifolds. We call attention to the fact that the hypotheses of Theorem 3.3.2 are completely self-contained, making reference only to intrinsic properties of the data and the architectural hyperparameters of the neural network (as well as $\text{poly}(\log n_0)$), and that the result is algorithmic, as it applies to training the network via constant-stepping gradient descent on the empirical square loss and guarantees generalization within $L^2$ iterations. Furthermore, Theorem 3.3.2 can be readily extended to the more general setting of regression on curves, given that we have focused on training with the square loss.

3.4 Proof Sketch

In this section, we provide an overview of the key elements of the proof of Theorem 3.3.1, where we show that the equation $\Theta_\mu[g] \approx \zeta$ admits a solution $g$ (the certificate) of small norm. To solve the certificate problem for $M$, we require a fine-grained understanding of the kernel $\Theta$. The most natural approach is to formally set $g = \sum_{i=1}^{\infty} \lambda_i^{-1} \langle \zeta, v_i \rangle_{L_\mu^2} v_i$ using the eigendecomposition of $\Theta_\mu$ (just as constructed in Section 3.2.1 for $\Theta_\mu^{\text{NTK}}$), and then argue that this formal expression converges by studying the rate of decay of $\lambda_i$ and the alignment of $\zeta$ with eigenvectors of $\Theta_\mu$; this is the standard approach in the literature [62, 64]. However, as discussed in Section 3.2.1, the nonlinear structure of $M$ makes obtaining a full diagonalization for $\Theta_\mu$ intractable, and simple asymptotic characterizations of its spectrum are insufficient to prove that the solution $g$ has small norm. Our approach will therefore be more direct: we will study the ‘spatial’ properties of the kernel $\Theta$ itself, in particular its rate of decay away from $x = x'$, and thereby use the network depth $L$ as a resource to reduce the study of the operator $\Theta_\mu$ to a simpler, localized operator whose
invertibility can be proved using harmonic analysis. We will then use differentiability properties of $\Theta$ to transfer the solution obtained by inverting this auxiliary operator back to the operator $\Theta_\mu$. We refer readers to Section B.4 for the full proof.

We simplify the proceedings using two basic reductions. First, with a small amount of auxiliary argumentation, we can reduce from the study of the operator-with-density $\Theta_\mu$ to the density-free operator $\Theta$. Second, the kernel $\Theta(x, x')$ is a function of the angle $\angle(x, x')$, and hence is rotationally invariant. This kernel is maximized at $\angle(x, x') = 0$ and decreases monotonically as the angle increases, reaching its minimum value at $\angle(x, x') = \pi$. If we subtract this minimum value, it should not affect our ability to fit functions, and we obtain a rotationally invariant kernel $\Theta^\circ(x, x') = \psi^\circ(\angle(x, x'))$ that is concentrated around angle 0. In the following, we focus on certificate construction for the kernel $\Theta^\circ$. Both simplifications are justified [114, Section E.3].

3.4.1 The Importance of Depth: Localization of the Neural Tangent Kernel

The first problem one encounters when attempting to directly establish (a property like) invertibility of the operator $\Theta^\circ$ is its action across connected components of $M$: the operator $\Theta^\circ$ acts by integrating against functions defined on $M = M_+ \cup M_-$, and although it is intuitive that most of its image’s values on each component will be due to integration of the input over the same component, there will always be some ‘cross-talk’ corresponding to integration over the opposite component that interferes with our ability to apply harmonic analysis tools. To work around this basic issue (as well as others we will see below), our argument proceeds via a localization approach: we will exploit the fact that as the depth $L$ increases, the kernel $\Theta^\circ$ sharpens and concentrates around its value at $x = x'$, to the extent that we can neglect its action across components of $M$ and even pass to the analysis of an auxiliary localized operator. This reduction is enabled by new sharp estimates for the decay of the angle function $\psi^\circ$ that we establish in Section B.5.3. Moreover, the perspective of using the network depth as a resource to localize the kernel $\Theta^\circ$ and exploiting this to solve the classification problem appears to be new: this localization is typically presented as a deficiency in the literature (e.g. [115]).
At a more formal level, when the network is deep enough compared to geometric properties of the curves, for each point \( x \), the majority of the mass of the kernel \( \Theta^\circ(x, x') \) is taken within a small neighborhood \( d_M(x, x') \leq r \) of \( x \). When \( d_M(x, x') \) is small relative to \( \kappa \), we have \( d_M(x, x') \approx \angle(x, x') \). This allows us to approximate the local component by the following invariant operator:

\[
\tilde{M}[f](x_{\sigma}(s)) = \int_{s' = s - r}^{s + r} \psi^\circ(|s - s'|) f(x_{\sigma}(s')) ds'.
\] (3.4)

This approximation has two main benefits: (i) the operator \( \tilde{M} \) is defined by intrinsic distance \( s' - s \), and (ii) it is highly localized. In fact, (3.4) takes the form of a convolution over the arc length parameter \( s \). This implies that \( \tilde{M} \) diagonalizes in the Fourier basis, giving an explicit characterization of its eigenvalues and eigenvectors. Moreover, because \( \tilde{M} \) is localized, the eigenvalues corresponding to slowly oscillating Fourier basis functions are large, and \( \tilde{M} \) is stably invertible over such functions. Both of these benefits can be seen as consequences of depth: depth leads to localization, which facilitates approximation by \( \tilde{M} \), and renders that approximation invertible over low-frequency functions. In our proofs, we will work with a subspace \( S \) spanned by low-frequency basis functions that are nearly constant over a length \( 2r \) interval (this subspace ends up having dimension proportional to \( 1/r \); see Section B.2.3 for a formal definition), and use Fourier arguments to prove invertibility of \( \tilde{M} \) over \( S \) (see Lemma B.4.6).

3.5 Discussion

A role for depth. In the setting of fitting functions on the sphere \( \mathbb{S}^{m-1} \) in the NTK regime with unstructured (e.g., uniformly random) data, it is well-known that there is very little marginal benefit to using a deeper network: for example, [62, 84, 116] show that the risk lower bound for RKHS methods is nearly met by kernel regression with a 2-layer network’s NTK in an asymptotic \( (n_0 \to \infty) \) setting, and results for fitting degree-1 functions in the nonasymptotic setting [86] are suggestive of a similar phenomenon. In a similar vein, fitting in the NTK regime with a deeper network does not change the kernel’s RKHS [117, 118, 119], and in a certain “infinite-depth”
limit, the corresponding NTK for networks with ReLU activations, as we consider here, is a spike, guaranteeing that it fails to generalize [115, 120]. Our results are certainly not in contradiction to these facts—we consider a setting where the data are highly structured, and our proofs only show that an appropriate choice of the depth relative to this structure is sufficient to guarantee generalization, not necessary—but they nonetheless highlight an important role for the network depth in the NTK regime that has not been explored in the existing literature. In particular, the localization phenomenon exhibited by the deep NTK is completely inaccessible by fixed-depth networks, and simultaneously essential to our arguments to proving Theorem 3.3.2, as we have described in Section 3.4. It is an interesting open problem to determine whether there exist low-dimensional geometries that cannot be efficiently separated without a deep NTK, or whether the essential sufficiency of the depth-two NTK persists.

**Closing the gap to real networks and data.** Theorem 3.3.2 represents an initial step towards understanding the interaction between neural networks and data with low-dimensional structure, and identifying network resource requirements sufficient to guarantee generalization. There are several important avenues for future work. First, although the resource requirements in Theorem 3.3.1, and by extension Theorem 3.3.2, reflect only intrinsic properties of the data, the rates are far from optimal—improvements here will demand a more refined harmonic analysis argument beyond the localization approach we take in Section 3.4.1. A more fundamental advance would consist of extending the analysis to the setting of a model for image data, such as cartoon articulation manifolds, and the NTK of a convolutional neural network with architectural settings that impose translation invariance [121, 122]—recent results show asymptotic statistical efficiency guarantees with the NTK of a simple convolutional architecture, but only in the context of generic data [123]. The approach to certificate construction we develop in Theorem 3.3.1 will be of use in establishing guarantees analogous to Theorem 3.3.2 here, as our approach does not require an explicit diagonalization of the NTK. In addition, extending our certificate construction approach to smooth manifolds of dimension larger than one is a natural next step. We believe our localization
argument generalizes to this setting: as our bounds for the kernel $\psi$ are sharp with respect to depth and independent of the manifold dimension, one could seek to prove guarantees analogous to Theorem 3.3.1 with a similar subspace-restriction argument for sufficiently regular manifolds, such as manifolds diffeomorphic to spheres, where the geometric parameters of Section 3.2.2 have natural extensions. Such a generalization would incur at best an exponential dependence of the network on the manifold dimension for localization in high dimensions.
Chapter 4: Designing Network Architectures from Low-Dimensional Structure: Resource-Efficient Invariant Networks

Our results on the multiple manifold problem in Chapters 2 and 3 illustrate how studying data with low-dimensional structure leads to new mathematical insights into generic deep learning architectures. A careful consideration of structure in data has important computational benefits, as well: it allows the researcher to encode known geometric priors into the network architecture, potentially leading to improved efficiency and interpretability in specific applications. In this chapter, we will discuss a framework revolving around these ideas that implies resource-efficient neural network architectures for a broad array of applications in invariant computer vision. The proposed approach will use *optimization* rather than exhaustive sampling of nuisance transformations (i.e., as in standard data augmentation [24, 49, 50]) as the primitive to build networks, leading to a potential *exponential improvement in efficiency* while maintaining invariance.

4.1 Introduction

In computing with any kind of realistic visual data, one must contend with a dizzying array of complex variabilities: statistical variations due to appearance and shape, geometric variations due to pose and perspective, photometric variations due to illumination and cast shadows, and more. Practical systems cope with these variations by a data-driven approach, with deep neural network architectures trained on massive datasets. This approach is especially successful at coping with variations in texture and appearance. For invariance to geometric transformations of the input (e.g., translations, rotations, and scaling, as in Figure 4.1(a-d)), the predominant approach in practice is also data-driven: the ‘standard pipeline’ is to deploy an architecture that is structurally invariant to translations (say, by virtue of convolution and pooling), and improve its stability with
Figure 4.1: Comparing the complexity of covering-based and optimization-based methods for invariant recognition of a template embedded in visual clutter. (a-d): We consider four different classes of deformations that generate the observation of the template, ranging across shifts, rotations, scale, and skew. The dimension \( d \) of the family of transformations increases from left to right. (e): A geometric illustration of the covering and optimization approaches to global invariance: in certifying that a query (labeled with a star) is a transformed instance of the template (at the base point of the solid red/blue lines), optimization can be vastly more efficient than covering, because it effectively covers the space at the scale of the basin of attraction of the optimization problem, which is always larger than the template’s associated \( \varepsilon_{\text{COVER}} \). (f): Plotting the average number of convolution-like operations necessary to reach a zero-normalized cross-correlation (ZNCC) of 0.9 between the template and a randomly-transformed query across the different deformation classes. Optimization leads to an efficiency gain of several orders of magnitude as the dimensionality of the family of transformations grows. Precise experimental details are recorded in Section C.1.3.

respect to other types of transformations by data augmentation. Data augmentation generates large numbers of synthetic training samples by applying various transformations to the available training data, and demonstrably contributes to the performance of state-of-the-art systems [49, 50, 124]. However, it runs into a basic resource efficiency barrier associated with the dimensionality of the set of nuisances: learning over a \( d \)-dimensional group of transformations requires both data and architectural resources that are exponential in \( d \) [34, 35, 36, 37, 38]. This is a major obstacle to achieving invariance to large, structured deformations such as 3D rigid body motion (\( d = 6 \)), homography (\( d = 8 \)), and linked rigid body motion [125] and even nonrigid deformations [126].
It is no surprise, then, that systems trained in this fashion remain vulnerable to adversarial transformations of domain [51, 52, 53, 54, 55, 56, 57]—it simply is not possible to generate enough artificial training data to learn transformation manifolds of even moderate dimension. Moreover, this approach is fundamentally wasteful: learning nuisances known to be present in the input data wastes architectural capacity that would be better spent coping with statistical variability in the input, or learning to perform complex tasks.

These limitations of the standard pipeline are well-established, and they have inspired a range of alternative architectural approaches to achieving invariance, where each layer of the network incorporates computational operations that reflect the variabilities present in the data. Nevertheless, as we will survey in detail in Section 4.2, all known approaches are subject to some form of exponential complexity barrier: the computational primitives demand either a filter count that grows as \( \exp(d) \) or integration over a \( d \)-dimensional space, again incurring complexity exponential in \( d \). Like data augmentation, these approaches can be seen as obtaining invariance by exhaustively sampling transformations from the \( d \)-dimensional space of nuisances, which seems fundamentally inefficient: in many concrete high-dimensional signal recovery problems, optimization provides a significant advantage over naive grid searching when exploring a high-dimensional space [127, 128, 129], as in Figure 4.1(e). This motivates us to ask:

**Can we break the barrier between resource-efficiency and invariance using optimization as the architectural primitive, rather than sampling?**

In Figure 4.1, we conduct a simple experiment that suggests a promising avenue to answer this question in the affirmative. Given a known synthetic textured motif subject to an unknown structured transformation and embedded in a background, we calculate the number of computations (convolutions and interpolations) required to certify with high confidence that the motif appears in the image. Our baseline approach is template matching, which enumerates as many transformations of the input as are necessary to certify the motif’s occurrence (analogous to existing architectural approaches with sampling/integration as the computational primitive)—each enumeration requires one interpolation and one convolution. We compare to a gradient-based optimization ap-
proach that attempts to match the appearance of the test image to the motif, which uses three inter-
polations and several convolutions per iteration (and on the order of $10^2$ iterations). As the dimen-
sionality of the space of transformations grows, the optimization-based approach demonstrates an
increasingly-significant efficiency advantage over brute-force enumeration of templates—at affine
transformations, for which $d = 6$, it becomes challenging to even obtain a suitable transformation
of the template by sampling.

To build from the promising optimization-based approach to local invariance in this experiment
to a full invariant neural network architecture capable of computing with realistic visual data, one
needs a general method to incorporate prior information about the specific visual data, observ-
able transformations, and target task into the design of the network. We take the first steps towards
realizing this goal: inspired by classical methods for image registration in the computer vision liter-
ature, we propose an optimization formulation for seeking a structured transformation of an input
image that matches previously-observed images, and we show how combining this formulation
with *unrolled optimization* [130, 131, 132], which converts an iterative solver for an optimization
problem into a neural network, implies resource-efficient and principled invariant neural archi-
tectural primitives. In addition to providing network architectures incorporating ‘invariance by
design’, this is a principled approach that leads to networks amenable to theoretical analysis, and
in particular we provide convergence guarantees for specific instances of our optimization formul-
ations that transfer to the corresponding unrolled networks. On the practical side, we illustrate how
these architectural primitives can be combined into a task-specific neural network by designing
and evaluating an invariant network architecture for an idealized single-template hierarchical ob-
ject detection task, and present an experimental corroboration of the soundness of the formulation
for invariant visual motif recognition used in the experiment in Figure 4.1. Taken altogether, these
results demonstrate a promising new direction to obtain theoretically-principled, resource-efficient
neural networks that achieve guaranteed invariance to structured deformations of image data.

The remainder of the chapter is organized as follows: Section 4.2 surveys the broad range of
architectural approaches to invariance that have appeared in the literature; Section 4.3 describes
our proposed optimization formulations and the unrolling approach to network design; Section 4.4 describes the hierarchical invariant object detection task and a corresponding invariant network architecture; Section 4.5 establishes convergence guarantees for our optimization approach under a data model inspired by the hierarchical invariant object detection task; and Section 4.6 provides a more detailed look at the invariance capabilities of the formulation used in Figure 4.1.

4.2 Related Work

Augmentation-based invariance approaches in deep learning. The ‘standard pipeline’ for invariance in deep learning described in Section 4.1 occupies, in a certain sense, a minimal point on the tradeoff curve between a purely data-driven approach and incorporating prior knowledge about the data into the architecture: by using a convolutional neural network with pooling, invariance to translations of the input image (a two-dimensional group of nuisances) is (in principle) conferred, and invariance to more complex transformations is left up to a combination of learning from large datasets and data augmentation. A number of architectural proposals in the literature build from a similar perspective, but occupy different points on this tradeoff curve. Parallel channel networks [133] generate all possible transformations of the input and process them in parallel, and have been applied for invariance to rotations [134, 135] and scale [136]. Other architectures confer invariance by pooling over transformations at the feature level [137], similarly for rotation [138] and scale [139]. Evidently these approaches become impracticable for moderate-dimensional families of transformations, as they suffer from the same sampling-based bottleneck as the standard pipeline.

To avoid explicitly covering the space of transformations, one can instead incorporate learned deformation offsets into the network, as in deformable CNNs [140] and spatial transformer networks [141]. At a further level of generality, capsule networks [142, 143, 144, 145] allow more flexible deformations among distinct parts of an object to be modeled. The improved empirical performance observed with these architectures in certain tasks illustrates the value of explicitly modeling deformations in the network architecture. At the same time, when it comes to guaranteed
invariance to specific families of structured deformations, they suffer from the same exponential inefficiencies as the aforementioned approaches.

**Invariance-by-construction architectures in deep learning.** The fundamental efficiency bottleneck encountered by the preceding approaches has motivated the development of alternate networks that are invariant simply by virtue of their constituent computational building blocks. Scattering networks [146] are an especially principled and elegant approach: they repeatedly iterate layers that convolve an input signal with wavelet filters, take the modulus, and pool spatially. These networks provably obtain translation invariance in the limit of large depth, with feature representations that are Lipschitz-stable to general deformations [147]; moreover, the construction and provable invariance/stability guarantees generalize to feature extractors beyond wavelet scattering networks [148]. Nevertheless, these networks suffer from a similar exponential resource inefficiency to those that plague the augmentation-based approaches: each layer takes a wavelet transform of every feature map at the previous layer, resulting in a network of width growing exponentially with depth. Numerous mitigation strategies have been proposed for this limitation [146, 149, 150], and combinations of relatively-shallow scattering networks with standard learning machines have demonstrated competitive empirical performance on certain benchmark datasets [151]. However, the resulting hybrid networks still suffer from an inability to handle large, structured transformations of domain such as pose and perspective changes.

Group scattering networks attempt to remedy this deficiency by replacing the spatial convolution operation with a group convolution $w, x \mapsto [w \ast x](g) = \int_G x(g') w(g^{-1}g') \, d\mu(g')$ [147, 152, 153, 154]. In this formula, $G$ is a group with sufficient topological structure, $\mu$ is Haar measure on $G$, and $w$ and $x$ are the filter and signal (resp.), defined on $G$ (or a homogeneous space for $G$, as in spherical CNNs [155]). Spatial convolution of natural images coincides with the special case $G = \mathbb{Z}^2$ in this construction; for more general groups such as 3D rotation, networks constructed by iterated group convolutions yield feature representations equivariant to the group action, and intermixing pooling operations yields invariance, just as with 2D convolutional neural networks. At a
conceptual level, this basic construction implies invariant network architectures for an extremely broad class of groups and spaces admitting group actions [156], and has been especially successful in graph-structured tasks such as molecular prediction where there is an advantage to enforcing symmetries [154]. However, its application to visual data has been hindered by exponential inefficiencies in computing the group convolution—integration over a \( d \)-dimensional group \( G \) costs resources exponential in \( d \)—and more fundamentally by the fact that discrete images are defined on the image plane \( \mathbb{Z}^2 \), whereas group convolutions require the signal to be defined over the group \( G \) one seeks invariance to. In this sense, the ‘reflexivity’ of spatial convolution and discrete images seems to be the exception rather than the rule, and there remains a need for resource-efficient architectural primitives for invariance with visual data.

"Unrolling" iterative optimization algorithms. First introduced by Gregor and LeCun in the context of sparse coding [130], unrolled optimization provides a general method to convert an iterative algorithm for solving an optimization problem into a neural network (we will provide a concrete demonstration in the present context in Section 4.3), offering the possibility to combine the statistical learning capabilities of modern neural networks with very specific prior information about the problem at hand [131]. It has found broad use in scientific imaging and engineering applications [132, 157, 158, 159, 160], and most state-of-the-art methods for learned MRI reconstruction are based on this approach [161]. In many cases, the resulting networks are amenable to theoretical analysis [162, 163], leading to a mathematically-principled neural network construction.

4.3 Invariant Architecture Primitives: Optimization and Unrolling

Notation. We write \( \mathbb{R} \) for the reals, \( \mathbb{Z} \) for the integers, and \( \mathbb{N} \) for the positive integers. For positive integers \( m, n, \) and \( c \), we let \( \mathbb{R}^m, \mathbb{R}^{m \times n}, \) and \( \mathbb{R}^{m \times n \times c} \) denote the spaces of real-valued \( m \)-dimensional vectors, \( m \)-by-\( n \) matrices, and \( c \)-channel \( m \)-by-\( n \) images (resp.). We write \( e_i, e_{ij}, \) etc. to denote the elements of the canonical basis of these spaces, and \( \mathbf{1}_m \) and \( \mathbf{0}_{m,n} \) (etc.) to denote their all-
ones and all-zeros elements (resp.). We write $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_F$ to denote the euclidean inner product and associated norm of these spaces. We identify $m$ by $n$ images $x$ with functions on the integer grid $\{0, 1, \ldots, m - 1\} \times \{0, 1, \ldots, n - 1\}$, and therefore index images starting from 0; when applying operations such as filtering, we will assume that an implementation takes the necessary zero padding, shifting, and truncation steps to avoid boundary effects. For a subset $\Omega \subset \mathbb{Z}^2$, we write $P_\Omega$ for the orthogonal projection onto the space of images with support $\Omega$.

Given a deformation vector field $\tau \in \mathbb{R}^{m' \times n' \times 2}$ and an image $x \in \mathbb{R}^{m \times n \times c}$, we define the transformed image $x \circ \tau$ by $(x \circ \tau)_{ij} = \sum_{(k,l) \in \mathbb{Z}^2} x_{kl} \phi(\tau_{ij0} - k) \phi(\tau_{ij1} - l)$, where $\phi : \mathbb{R} \to \mathbb{R}$ is the cubic convolution interpolation kernel [164]. For parametric transformations of the image plane, we write $\tau_{A,b}$ to denote the vector field representation of the transformation parameterized by $(A, b)$, where $A \in \mathbb{R}^{2 \times 2}$ is nonsingular and $b \in \mathbb{R}^2$ (see Section C.1.1 for specific ‘implementation’ details). For two grayscale images $x \in \mathbb{R}^{m \times n}$ and $u \in \mathbb{R}^{m' \times n'}$, we write their linear convolution as $(x * u)_{ij} = \sum_{(k,l) \in \mathbb{Z}^2} x_{kl} u_{i-k,j-l}$. We write $g_{\sigma^2} \in \mathbb{R}^{Z \times Z}$ to denote a (sampled) gaussian with zero mean and variance $\sigma^2$. When $x \in \mathbb{R}^{m \times n}$ and $u \in \mathbb{R}^c$, we write $x \otimes u \in \mathbb{R}^{m \times n \times c}$ to denote the ‘tensor product’ of these elements, with $(x \otimes u)_{ijk} = x_{ij} u_k$. We use $x \odot u$ to denote elementwise multiplication of images.

4.3.1 Conceptual Framework

Given an input image $y \in \mathbb{R}^{m \times n \times c}$ (e.g., $c = 3$ for RGB images), we consider the following general optimization formulation for seeking a structured transformation of the input that explains it in terms of prior observations:

$$\min_{\tau} \varphi(y \circ \tau) + \lambda \rho(\tau).$$  \hspace{1cm} (4.1)

Here, $\tau \in \mathbb{R}^{m' \times n' \times 2}$ gives a vector field representation of transformations of the image plane, and $\lambda > 0$ is a regularization tradeoff parameter. Minimization of the function $\varphi$ encourages the transformed input image $y \circ \tau$ to be similar to previously-observed images, whereas minimization

\footnote{The function $\phi$ is compactly supported on the interval $[-2, 2]$, and differentiable with absolutely continuous derivative.}
of $\rho$ regularizes the complexity of the learned transformation $\tau$. Both terms allow to incorporate significant prior information about the visual data and task at hand, and an optimal solution $\tau$ to (4.1) furnishes an invariant representation of the input $y$.

### 4.3.2 Computational Primitive: Optimization for Domain Transformations

We illustrate the flexibility of the general formulation (4.1) by instantiating it for a variety of classes of visual data. In the most basic setting, we may consider the registration of the input image $y$ to a known motif $x_o$ assumed to be present in the image, and constrain the transformation $\tau$ to lie in a parametric family of transformations $T$, which yields the optimization formulation

$$\min_{\tau} \frac{1}{2}\|\mathcal{P}_\Omega [g_{\sigma^2} \ast (y \circ \tau - x_o)]\|_F^2 + \chi_T(\tau).$$  \hspace{1cm} (4.2)

Here, $\Omega$ denotes a subset of the image plane corresponding to the pixels on which the motif $x_o$ is supported, $g_{\sigma^2}$ is a gaussian filter with variance $\sigma^2$ applied individually to each channel, and $\chi_T(\tau)$ denotes the characteristic function for the set $T$ (zero if $\tau \in T$, $+\infty$ otherwise). The parameters in (4.2) are illustrated in Figure 4.2(a-d). We do not directly implement the basic formulation (4.2) in our experiments, but as a simple model for the more elaborate instantiations of (4.1) that follow later it furnishes several useful intuitions. For instance, although (4.2) is a nonconvex optimization problem with a ‘rough’ global landscape, well-known results suggest that under idealized conditions (e.g., when $y = x_o \circ \tau_o$ for some $\tau_o \in T$), multiscale solvers that repeatedly solve (4.2) with a smoothing level $\sigma^2_k$ then re-solve initialized at the previous optimal solution with a finer level of smoothing $\sigma^2_{k+1} < \sigma^2_k$ converge in a neighborhood of the true transformation [165, 166, 167, 168]. This basic fact underpins many classical computer vision methods for image registration and stitching [169, 170, 171, 172], active appearance models for objects [173], and optical flow estimation [174, 175, 176], and suggests that (4.2) is a suitable base for constructing invariant networks.

For our experiments on textured visual data in Figure 4.1 and Section 4.4, we will need two
Figure 4.2: Motif registration with the formulation (4.1), and an unrolled solver. (a-d): Visualization of components of a registration problem, such as (4.2). We model observations \( y \) as comprising an object involving the motif of interest (here, the body of the crab template we experiment with in Section 4.4) on a black background, as in (a), embedded in visual clutter (here, the beach background) and subject to a deformation, which leads to a novel pose. A mask \( \Omega \) for the nonzero pixels of the motif, as in (b), is used to avoid having pixels corresponding to clutter enter the registration cost. After solving this optimization problem, we obtain a transformation \( \mathbf{t} \) that registers the observation to the motif, as in (d). In (c-d), we set the red and blue pixels corresponding to the mask \( \Omega \) to 1 in order to visualize the relative location of the motif. (e): Optimization formulations imply network architectures, via unrolled optimization. Here we show two iterations of an unrolled solver for (4.2), as we detail in Section 4.3.3; parameters that could be learned from data, à la unrolled optimization, are highlighted with red text. The operations comprising this unrolled network consist of linear maps, convolutions, and interpolations, leading to efficient implementation on standard hardware accelerators.

elaborations of (4.2). The first arises due to the problem of obtaining invariant representations for images containing motifs \( x_o \) appearing in general backgrounds: in such a scenario, the input image \( y \) may contain the motif \( x_o \) in a completely novel scene (as in Figure 4.2(c-d)), which makes it inappropriate to smooth the entire motif with the filter \( g_{\sigma^2} \). In these scenarios, we consider instead a cost-smoothed formulation of (4.2):

\[
\min_{\mathbf{t}} \frac{1}{2} \sum_{\Delta \in \mathbb{Z} \times \mathbb{Z}} (g_{\sigma^2})_{\Delta} \| P_{\Omega} [ y \circ (\mathbf{t} + \mathbf{t}_0,\Delta) - x_o ] \|^2_F + \chi^T(\mathbf{t}).
\]  

(4.3)

In practice, we take the sum over a finite subset of shifts \( \Delta \) on which most of the mass of the gaussian filter lies. This formulation is inspired by more general cost-smoothing registration proposals in the literature [166], and it guarantees that pixels of \( y \circ \mathbf{t} \) corresponding to the background \( \Omega^c \) are never compared to pixels of \( x_o \) while incorporating the basin-expanding benefits of smoothing. Second, we consider a more general formulation which also incorporates a low-frequency model
for the image background:

$$\min_{\tau, \beta} \frac{1}{2} \left\| \mathcal{P}_{\tilde{\Omega}} \left[ g_{\sigma^2} * (y \circ \tau - x_o - \mathcal{P}_{\Omega'} [g_{C\sigma^2} * \beta]) \right] \right\|_F^2 + \chi_{T}(\tau).$$  \hspace{1cm} (4.4)

Here, $\beta \in \mathbb{R}^{m \times n \times c}$ acts as a learnable model for the image background, and $C > 1$ is a fixed constant that guarantees that the background model is at a coarser scale than the motif and image content. The set $\tilde{\Omega}$ represents a dilation by $\sigma$ of the motif support $\Omega$, and penalizing pixels in this dilated support ensures that an optimal $\tau$ accounts for both foreground and background agreement. We find background modeling essential in computing with scale-changing transformations, such as affine transforms in Figure 4.1.

### 4.3.3 Invariant Networks from Unrolled Optimization

The technique of unrolled optimization allows us to obtain principled network architectures from the optimization formulations developed in Section 4.3.2. We describe the basic approach using the abstract formulation (4.1). For a broad class of regularizers $\rho$, the proximal gradient method [17] can be used to attempt to solve the nonconvex problem (4.1): it defines a sequence of iterates

$$\tau^{(t+1)} = \text{prox}_{\lambda \nu_\tau \rho}(\tau^{(t)} - \nu_t \nabla_{\tau} \varphi(y \circ \tau^{(t)}))$$  \hspace{1cm} (4.5)

from a fixed initialization $\tau^{(0)}$, where $\nu_t > 0$ is a step size sequence and $\text{prox}_{\rho}(\tau) = \arg \min_{\tau'} \frac{1}{2} \| \tau - \tau' \|^2_F + \rho(\tau')$ is well-defined if $\rho$ is a proper convex function. Unrolled optimization suggests to truncate this iteration after $T$ steps, and treat the iterate $\tau^{(T)}$ at that iteration as the output of a neural network. One can then learn certain parameters of the neural network from datasets, as a principled approach to combining the structural priors of the original optimization problem with the benefits of a data-driven approach.

In Figure 4.2(e), we show an architectural diagram for a neural network unrolled from a proximal gradient descent solver for the registration formulation (4.2). We always initialize our networks with $\tau^{(0)}$ as the identity transformation field, and in this context we have $\text{prox}_{\lambda \nu_\tau \rho}(\tau) = \text{proj}_{\mathbb{T}}(\tau)$ as
the nearest point in \( \mathbb{T} \) to \( \tau \), which can be computed efficiently (computational details are provided in Section C.1.1). The cost (4.2) is differentiable; calculating its gradient as in Section C.1.2, (4.5) becomes

\[
\tau^{(t+1)} = \text{proj}_T \left( \tau^{(t)} - \nu_t \sum_{k=0}^{c-1} \left( g_{\sigma^2} * P_\Omega \left[ y \circ \tau^{(t)} - x_o \right] \otimes 1_2 \right) \odot (dy_k \circ \tau^{(t)}) \right),
\]

as we represent visually in Figure 4.2(e), where a subscript of \( k \) denotes the \( k \)-th channel of the image and \( dy \in \mathbb{R}^{m \times n \times c \times 2} \) is the Jacobian matrix of \( y \). The constituent operations in this network are convolutions, pointwise nonlinearities and linear maps, which lend themselves ideally to implementation in standard deep learning software packages and on hardware accelerators; and because the cubic convolution interpolation kernel \( \phi \) is twice continuously differentiable except at four points of \( \mathbb{R} \), these networks are end-to-end differentiable and can be backpropagated through efficiently. The calculations necessary to instantiate unrolled network architectures for other optimization formulations used in our experiments are deferred to Section C.1.2. A further advantage of the unrolled approach to network design is that hyperparameter selection becomes directly connected to convergence properties of the optimization formulation (4.1): we demonstrate how theory influences these selections in Section 4.5, and provide practical guidance for registration and detection problems through our experiments in Sections 4.4, 4.5.2 and 4.6.

### 4.4 Invariant Networks for Hierarchical Object Detection

The unrolled networks in Section 4.3 are architectural primitives for building deformation-invariant neural networks: they are effective at producing invariant representations for input images containing local motifs. In this section, we illustrate how these local modules can be combined into a network that performs invariant processing of nonlocally-structured visual data, via an invariant hierarchical object detection task with a fixed template. For simplicity, in this section we will focus on the setting where \( \mathbb{T} \) is the set of rigid motions of the image plane (i.e., translations and rotations), which we will write as SE(2).
4.4.1 Data Model and Problem Formulation

We consider an object detection task, where the objective is to locate a fixed template with independently-articulating parts (according to a SE(2) motion model) in visual clutter. More precisely, we assume the template is decomposable into a hierarchy of deformable parts, as in Figure 4.3(a): at the top level of the hierarchy is the template itself, with concrete visual motifs at the lowest levels that correspond to specific pixel subsets of the template, which constitute independent parts. Because these constituent parts deform independently of one another, detecting this template efficiently demands an approach to detection that captures the specific hierarchical structure of the template.\(^2\) Compared to existing methods for parts-based object detection that are formulated to work with objects subject to complicated variations in appearance [177, 178, 179, 180], focusing on the simpler setting of template detection allows us to develop a network that guarantees invariant detection under the motion model, and can incorporate large-scale statistical learning techniques by virtue of its unrolled construction (although we leave this latter direction for future work). We note that other approaches are possible, such as hierarchical sparse modeling [181, 182] or learning a graphical model [183].

More formally, we write \(y_o \in \mathbb{R}^{m_o \times n_o \times 3}\) for the RGB image corresponding to a canonized view of the template to be detected (e.g., the crab at the top of the hierarchy in Figure 4.3(a) left) embedded on a black background. For a \(K\)-motif object (e.g., \(K = 4\) for the crab template), we let \(x_k \in \mathbb{R}^{m_k \times n_k \times 3}\) denote the \(k\) distinct (canonized, black-background-embedded) transforming motifs in the object, each with non-overlapping occurrence coordinates \((i_k, j_k) \in \{0, \ldots, m_o\} \times \{0, \ldots, n_o\}\). The template \(y_o\) decomposes as

\[
y_o = \sum_{k=1}^{K} x_k * e_{i_k, j_k} + y_o - \sum_{k=1}^{K} x_k * e_{i_k, j_k}.
\]

(4.7)

\(^2\)Reasoning as in Section 4.1, the effective dimension of the space of all observable transformations of the object is the product of the dimension of the motion model and the number of articulating parts. A detector that exploits the hierarchical structure of the object effectively reduces the dimensionality to \(\text{dim(motion model)} + \log(\text{number of parts})\), yielding a serious advantage for moderate-dimensional families of deformations.
For example, the four transforming motifs for the crab template in Figure 4.3(a) are the two claws and two eyes. In our experiments with the crab template, we will consider detection of transformed templates $y_{\text{obs}}$ of the following form:

$$y_{\text{obs}} = \left[ \sum_{k=1}^{K} (x_k * e_{i,k,j,k}) \circ \tau_k + \left( y_o - \sum_{k=1}^{K} x_k * e_{i,k,j,k} \right) \right] \circ \tau_0, \quad (4.8)$$

where $\tau_0 \in \text{SE}(2)$, and $\tau_k \in \text{SO}(2)$ is sufficiently close to the identity transformation (which represents the physical constraints of the template). The detection task is then to decide, given an input scene $y \in \mathbb{R}^{m \times n \times 3}$ containing visual clutter (and, in practice, $m \gg m_o$ and $n \gg n_o$), whether or not a transformed instance $y_{\text{obs}}$ appears in $y$ or not, and to output estimates of its transformation parameters $\tau_k$.

Although our experiments will pertain to the observation model (4.8), as it agrees with our decomposition of the crab template in Figure 4.3(a), the networks we construct in Section 4.4.3 will be amenable to more complex observation models where parts at intermediate levels of the hierarchy also transform.\(^3\) To this end, we introduce additional notation that captures the hierarchical structure of the template $y_o$. Concretely, we identify a hierarchically-structured template with a directed rooted tree $G = (V, E)$, with 0 denoting the root node, and 1, \ldots, $K$ denoting the $K$ leaf nodes. Our networks will treat observations of the form

$$y_{\text{obs}} = \sum_{k=1}^{K} (\cdots ((x_k * e_{i,k,j}) \circ \tau_k) \circ \tau_{v_{d(k)-1}} \circ \cdots) \circ \tau_{v_1} \circ \tau_0 + \left( y_o - \sum_{k=1}^{K} x_k * e_{i,k,j,k} \right) \circ \tau_0, \quad (4.9)$$

where $d(k)$ is the depth of node $k$, and $v_1, \ldots, v_{d(k)-1} \in V$ with $0 \rightarrow v_1 \rightarrow \cdots \rightarrow v_{d(k)-1} \rightarrow k$ specifying the path from the root node to node $k$ in $G$. To motivate the observation model (4.9), consider the crab example of Figure 4.3(a), where in addition we imagine the coordinate frame of the eye pair motif transforms independently with a transformation $\tau_5$: in this case, the observation

---

\(^3\)For example, consider a simple extension of the crab template in Figure 4.3(a), where the left and right claw motifs are further decomposed into two pairs of pincers plus the left and right arms, with opening and closing motions for the pincers, and the same $\text{SO}(2)$ articulation model for the arms (which naturally moves the pincers in accordance with the rotational motion).
A model (4.9) can be written in an equivalent ‘hierarchical’ form

\[ y_{\text{obs}} = [(x_1 \ast e_{i_1 j_1}) \circ \tau_1 + (x_2 \ast e_{i_2 j_2}) \circ \tau_2 + [(x_3 \ast e_{i_3 j_3}) \circ \tau_3 + (x_4 \ast e_{i_4 j_4}) \circ \tau_4] \circ \tau_5] \circ \tau_0 + y_{\text{body}} \circ \tau_0, \]

by linearity of the interpolation operation \( x \mapsto x \circ \tau \) (with \( y_{\text{body}} = y_o - \sum_k x_k \ast e_{i_k j_k} \)).

4.4.2 Aside: Optimization Formulations for Registration of “Spiky” Motifs

To efficiently perform hierarchical invariant detection of templates following the model (4.9), the networks we design will build from the following basic paradigm, given an input scene \( y \):

1. **Visual motif detection**: First, perform detection of all of the lowest-level motifs \( x_1, \ldots, x_K \) in \( y \). The output of this process is an occurrence map for each of the \( K \) transforming motifs, i.e. an \( m \times n \) image taking (ideally) value 1 at the coordinates where detections occur and 0 elsewhere.

2. **Spiky motif detection for hierarchical motifs**: Detect intermediate-level abstractions using the occurrence maps in \( y \) obtained in the previous step. For example, if \( k = 3 \) and \( k = 4 \) index the left and right eye motifs in the crab template of Figure 4.3(a), detection of the eye pair motif corresponds to registration of the canonized eye pair’s occurrence map against the two-channel image corresponding to the concatenation of \( x_3 \) and \( x_4 \)’s occurrence maps in \( y \).

3. **Continue until the top of the hierarchy**: This occurrence map detection process is iterated until the top level of the hierarchy. For example, in Figure 4.3(a), a detection of the crab template occurs when the multichannel image corresponding to the occurrence maps for the left and right claws and the eye pair motif is matched.

To instantiate this paradigm, we find it necessary to develop a separate registration formulation for registration of occurrence maps, beyond the formulations we have introduced in Section 4.3. Indeed, occurrence maps contain no texture information and are maximally localized, motivating a
formulation that spreads out gradient information and avoids interpolation artifacts—and although there is still a need to cope with clutter in general, the fact that the occurrence maps are generated on a black background obviates the need for extensive background modeling, as in (4.4). We therefore consider the following “complementary smoothing” formulation for spike registration: for a $c$-channel occurrence map $y$ and canonized occurrence map $x_o$, we optimize over the affine group $\text{Aff}(2) = \text{GL}(2) \rtimes \mathbb{R}^2$ as

$$
\min_{A,b} \frac{1}{2c} \left\| g_{\sigma^2 I - \sigma_0^2 A^*} \left( \det^{-1/2}(AA^*) \left(g_{\sigma_0^2 I} * y \right) \circ \tau_{A^{-1},-A^{-1} b} \right) - g_{\sigma^2 I} * x_o \right\|_F^2 + \chi_{\text{Aff}(2)}(A,b),
$$

(4.10)

where $g_M$ denotes a single-channel centered gaussian filter with positive definite covariance matrix $M > 0$, and correlations are broadcast across channels. Here, $\sigma > 0$ is the main smoothing parameter to propagate gradient information, and $\sigma_0 > 0$ is an additional smoothing hyperparameter to mitigate interpolation artifacts.

In essence, the key modifications in (4.10) that make it amenable to registration of occurrence maps are the compensatory effects for scaling that it introduces: transformations that scale the image correspondingly reduce the amplitude of the (smoothed) spikes, which is essential given the discrete, single-pixel spike images we will register. Of course, since we consider only euclidean transformations in our experiments in this section, we always have $AA^* = I$, and the problem (4.10) can be implemented in a simpler form. However, these modifications lead the problem (4.10) to work excellently for scale-changing transformations as well: we explore the reasons behind this from both theoretical and practical perspectives in Section 4.5.

4.4.3 Invariant Network Architecture

The networks we design to detect under the observation model (4.9) consist of a configuration of unrolled motif registration networks, as in Figure 4.2(e), arranged in a ‘bottom-up’ hierarchical fashion, following the hierarchical structure in the example shown in Figure 4.3(a). The configuration for each motif registration sub-network is a ‘GLOM-style’ [145] collection of the networks
Figure 4.3: An example of a hierarchically-structured template, and the results of an implementation of our detection network. (a): Structure of the crab template, described in Section 4.4.1, and its interaction with our network architecture for detection, described in Section 4.4.3. *Left: top-down decomposition of the template into motifs.* A template of interest $y_o$ (here, the crab at top left) is decomposed into a hierarchy of abstractions. The hierarchical structure is captured by a tree $G = (V, E)$: nodes represent parts or aggregations of parts, and edges represent their relationships. *Right: bottom-up detection of the template in a novel scene.* To detect the template in a novel scene and pose, the network described in Section 4.4.3 first localizes each of the lowest-level visual motifs at left and their transformation parameters in the input scene $y$ (bottom right). Motifs and the derived occurrence maps are labeled in agreement with the notation we introduce in Section 4.4.3. The output of each round of optimization is an occurrence map $\omega_v$ for nodes $v \in V$; these occurrence maps then become the inputs for detection of the next level of concepts, following the connectivity structure of $G$, until the top-level template is reached (top right). (b-e): Evaluation of the hierarchical invariant object detection network implemented in Section 4.4.4: the learned transformation at the minimum-error stride for each motif is used to draw the motifs’ transformed bounding boxes. Relative to the canonized template, each observation has the template’s eye motifs independently rotated by 5 degrees, and each claw motif by 15 degrees. Insets at the bottom right corner of each result panel visualize the quality of the final detection trace $\omega_0$ for the template, with a value of 1 marked in dashed red (an ideal detection has unit $\ell^1$ norm). In this example, detection succeeds until a multiple instance issue between the two claw motifs occurs in panel (e), causing detection of the left claw motif to fail.

sketched in Figure 4.2(e), oriented at different pixel locations in the input scene $y$; the transformation parameters predicted of each of these configurations are aggregated across the image, weighted by the final optimization cost (as a measure of quality of the final solution), in order to determine detections. These detections are then used as feature maps for the next level of occurrence motifs, which in turn undergo the same registration-detection process until reaching the top-level object’s occurrence map, which we use as a solution to the detection problem. A suitable unrolled implementation of the registration and detection process leads to a network that is end-to-end differentiable and amenable to implementation on standard hardware acceleration platforms.
We now describe this construction formally, following notation introduced in Section 4.4.1. The network input is an RGB image $y \in \mathbb{R}^{m \times n \times 3}$. We shall assume that the canonized template $y_o$ is given, as are as the canonized visual motifs $x_1, \ldots, x_K$ and their masks $\Omega_1, \ldots, \Omega_K$; we also assume that for every $v \in V$ with $v \notin \{1, \ldots, K\}$, we are given the canonized occurrence map $x_v \in \mathbb{R}^{m_v \times n_v \times c_v}$ of the hierarchical feature $v$ in $y_o$. In practice, one obtains these occurrence maps through a process of “extraction”, using $y_o$ as an input to the network, which we describe in Section C.1.4. The network construction can be separated into three distinct steps:

**Traversal.** The network topology is determined by a simple traversal of the graph $G$. For each $v \in V$, let $d(v)$ denote the shortest-path distance from 0 to $v$, with unit weights for edges in $E$ (the “depth” of $v$ in $G$). We will process motifs in a deepest-first order for convenience, although this is not strictly necessary in all cases (e.g. for efficiency, it might be preferable to process all leaf nodes $1, \ldots, K$ first). Let $\text{diam}(G) = \max_{v \in V} d(v)$, and for an integer $\ell$ no larger than $\text{diam}(G)$, we let $D(\ell) \in \mathbb{N}$ denote the number of nodes in $V$ that are at depth $\ell$.

**Motif detection at one depth.** Take any integer $0 \leq \ell \leq \text{diam}(G)$, and let $v_1, \ldots, v_{D(\ell)}$ denote the nodes in $G$ at depth $\ell$, enumerated in increasing order (say). For each such vertex $v_k$, perform the following steps:

1. **Is this a leaf?** If the neighborhood $\{v' \mid (v_k, v') \in E\}$ is empty, this node is a leaf; otherwise it is not. Subsequent steps depend on this distinction. We phrase the condition more generally, although we have defined $1, \ldots, K$ as the leaf vertices here, to facilitate some implementation-independence.
2. Occurrence map aggregation for non-leaves: If \( v_k \) is not a leaf, construct its detection feature map from lower-level detections: concretely, let

\[
y_{v_k} = \sum_{v': \{(v_k, v') \in E\}} \omega_{v'} \otimes e_{\pi_{v_k}(v')},
\]  
(4.11)

where \( \pi_{v_k}(v') \) denotes a vertex-increasing-order enumeration of the set \( \{v' : (v_k, v') \in E\} \) starting from 0. By construction (see the fourth step below), \( y_{v_k} \) has the same width and height as the input scene \( y \), but \( c_{v_k} = |\{v' : (v_k, v') \in E\}| \) channels (one for each child node) instead of 3 RGB channels.

3. Perform strided registration: Because the motif \( x_{v_k} \) is in general much smaller in size than the scene \( y_{v_k} \), and because the optimization formulation (4.1) is generally nonconvex with a finite-radius basin of attraction around the true transformation parameters in the model (4.9), the detection process consists of a search for \( x_{v_k} \) anchored at a grid of points in \( y_{v_k} \). Concretely, let

\[
\Lambda_{v_k} = \{(i\Delta_H, v_k, f\Delta_W, v_k) | (i, j) \in \{0, \ldots, m-1\} \times \{0, \ldots, n-1\} \} \cap (\{0, \ldots, m-1\} \times \{0, \ldots, n-1\})
\]
denote the grid for the \( v_k \)-th motif; here \( \Delta_{H,v_k} \) and \( \Delta_{W,v_k} \) define the vertical and horizontal stride lengths of the grid (we discuss choices of these and other hyperparameters introduced below in Section C.1.4). When \( v_k \) is a leaf, for each \( \lambda \in \Lambda_{v_k} \), we let \( (U(v_k, \lambda), b(v_k, \lambda)) \in SE(2) \) denote the parameters obtained after running an unrolled solver for the cost-smoothed visual motif registration problem

\[
\min_{\tau} \frac{1}{2} \sum_{\Delta \in \mathbb{Z} \otimes \mathbb{Z}} (g_{\sigma'^2} \epsilon_{\Delta})_\lambda \left\| P_{\Omega_{v_k}} \left( (g_{\sigma'^2} \star y) \circ (\tau + \tau_0, \Delta, \lambda) - x_{v_k} \right) \right\|^2_F + \chi_{SE(2)}(\tau),
\]  
(4.12)

for \( T_{v_k} \) iterations, with step size \( \nu_{v_k} \). In addition, we employ a two-step multiscale smoothing strategy, which involves initializing an unrolled solver for (4.12) with a much smaller smoothing parameter \( (\sigma'^2) \) at \( (U(v_k, \lambda), b(v_k, \lambda)) \) and running it for an additional fixed number of iterations;
we let \( \text{loss}(v_k, \lambda) \) denote the final objective function value after this multiscale process, and abusing notation, we let \((U(v_k, \lambda), b(v_k, \lambda))\) denote the updated final parameters. Precise implementation details are discussed in Section C.1.4. When \( v_k \) is not a leaf, we instead define the same fields on the grid \( \Lambda_{v_k} \) via a solver for the spike registration problem

\[
\min_{\tau} \frac{1}{2c_{v_k}} \left\| P_{\Omega_{v_k}} \left[ g_{\sigma_{v_k}^2} \ast (y_{v_k} \circ (\tau + \tau_{0,\lambda}) - x_{v_k}) \right] \right\|^2_F + \chi_{SE(2)}(\tau), \tag{4.13}
\]

with \( \Omega_{v_k} \) denoting a dilated bounding box for \( x_{v_k} \), and otherwise the same notation and hyperparameters. We do not use multiscale smoothing for non-leaf motifs.

4. Aggregate registration outputs into detections (occurrence maps): We convert the registration fields into detection maps, by computing

\[
\omega_{v_k} = \sum_{\lambda \in \Lambda_{v_k}} \left( g_{\sigma_{v_k}^2} \ast e_{A+b(v_k, \lambda)} \right) \exp \left( -\alpha_{v_k} \max \{0, \text{loss}(v_k, \lambda) - \gamma_{v_k}\} \right), \tag{4.14}
\]

where each summand \( g_{\sigma_{v_k}^2} \ast e_{A+b(v_k, \lambda)} \) is truncated to be size \( m \times n \). The scale and threshold parameters \( \alpha_{v_k} \) and \( \gamma_{v_k} \) appearing in this formula are calibrated to achieve a specified level of performance under an assumed maximum level of visual clutter and transformation for the observations (4.9), as discussed in Section C.1.4.

We prefer to embed detections as occurrence maps and use these as inputs for higher-level detections using optimization, rather than a possible alternate approach (e.g. extracting landmarks and processing these using group synchronization), in order to have each occurrence map \( \omega_v \) for \( v \in V \) be differentiable with respect to the various filters and hyperparameters.

**Template detection.** To perform detection given an input \( y \), we repeat the four steps in the previous section for each motif depth, starting from depth \( \ell = \text{diam}(G) \), and each motif at each depth.

---

This convolutional notation is of course an abuse of notation, to avoid having to define a gaussian filter with a general mean parameter. In practice, this latter technique is both more efficient to implement and leads to a stably-differentiable occurrence map.
Algorithm 1 Invariant Hierarchical Motif Detection Network, Summarizing Section 4.4.3

**input** scene $y$, graph $G = (V, E)$, motifs $(x_v, \Omega_v)_{v \in V}$, hyperparameters $(\nu_v, T_v, \Delta_{H,v}, \Delta_{W,v}, \sigma^2_v, \sigma^2_{0,v}, \alpha_v, \gamma_v)_{v \in V}$

set $\text{diam}(G)$ and node enumerations by depth-first traversal of $G$

for all depths $\ell = \text{diam}(G), \text{diam}(G) - 1, \ldots, 0$ do

for all nodes $v$ at depth $\ell$ do

set $N_v = \{v' \mid (v, v') \in E\}$ and $c_v = |N_v|$

if $c_v > 0$ then

concatenate occurrence maps into $y_v = \sum_{v' \in N_v} \omega_v \otimes e_{\pi_v(v')}$

for all $\lambda \in \Lambda_v(\Delta_{H,v}, \Delta_{W,v})$ do

if $c_v > 0$ then

set $U(v, \lambda), b(v, \lambda) = \arg \min_\tau \frac{1}{2} \sum_{v} \|g_{\sigma^2_v - \sigma^2_{0,v}} \ast (y_v \circ (\tau + \tau_{0,v}) - x_v)\|_F^2 + \chi_{\text{SE}(2)}(\tau)$

set loss$(v, \lambda) = \min_\tau \frac{1}{2} \sum_{v} \|g_{\sigma^2_v - \sigma^2_{0,v}} \ast (y_v \circ (\tau + \tau_{0,v}) - x_v)\|_F^2 + \chi_{\text{SE}(2)}(\tau)$

(both with a $T_v$-layer unrolled solver)

else

set $U(v, \lambda), b(v, \lambda) = \arg \min_\tau \frac{1}{2} \sum_\lambda \|g_{\sigma^2_v} \Delta \|_{\mathcal{P}_\Omega_v} [(g_{\sigma^2_v} \ast y) \circ (\tau + \tau_{0,v}) - x_v]\|_F^2 + \chi_{\text{SE}(2)}(\tau)$

set loss$(v, \lambda) = \min_\tau \frac{1}{2} \sum_\lambda \|g_{\sigma^2_v} \Delta \|_{\mathcal{P}_\Omega_v} [(g_{\sigma^2_v} \ast y) \circ (\tau + \tau_{0,v}) - x_v]\|_F^2 + \chi_{\text{SE}(2)}(\tau)$

(both with a $T_v$-layer unrolled solver, with two-round multiscale smoothing)

construct the occurrence map $\omega_v = \sum_{\lambda \in \Lambda_v} (g_{\sigma^2_{0,v}} \ast e_{\lambda b(v, \lambda)}) \exp(-\alpha_v \max\{0, \text{loss}(v, \lambda) - \gamma_v\})$

output template occurrence map $\omega_0$

After processing depth $\ell = 0$, the output occurrence map $\omega_0$ can be thresholded to achieve a desired level of detection performance for observations of the form (4.9). The detection process is summarized as Algorithm 1.

By construction, this output $\omega_0$ can be differentiated with respect to each node $v \in V$’s hyperparameters or filters, and the unrolled structure of the sub-networks and $G$’s topology can be used to efficiently calculate such gradients via backpropagation. In addition, although we do not use the full transformation parameters $U(\lambda, v)$ calculated in the registration operations (4.12) and (4.13), these can be leveraged for various purposes (e.g. drawing detection bounding boxes, as in our experimental evaluations in Section 4.4.4).
4.4.4 Implementation and Evaluation

We implement the hierarchical invariant object detection network described in Section 4.4.3 in PyTorch [184], and test it for detection of the crab template from Figure 4.3(a) subject to a global rotation (i.e., $\tau_0$ in the model (4.9)) of varying size (Figure 4.3(b-e)), with independent rotations of the eye and claw motifs by 5 and 15 degrees, respectively. In $512 \times 384$ pixel scenes on a “beach” background, a calibrated detector perfectly detects the crab from its constituent parts up to rotations of 30 degrees—at rotations around 40 degrees, a multiple-instance issue due to similarity between the two claw motifs begins to hinder the detection performance. Traces in each panel of Figure 4.3, right demonstrate the precise localization of the template.

For hyperparameters, we set $T_v = 32$ and $\Delta_{H,v} = \Delta_{W,v} = 20$ for all $v \in V$, and calibrate detection parameters as described in Section C.1.4; for visual motifs, we calibrate the remaining registration hyperparameters as described in Section C.1.4 on a per-motif basis, and for spike motifs, we find the prescriptions for $\sigma_v^2$ and the step sizes $\nu_v$ implied by theory (Section 4.5) to work excellently without any fine-tuning. We also implement selective filtering of strides for spiky motif alignment that are unlikely to succeed: due to the common background, this type of screening is particularly effective here. The strided registration formulations (4.12) and (4.13) afford efficient batched implementation on a hardware accelerator, given that the motifs $x_v$ for $v \in V$ are significantly smaller than the full input scene $y$, and the costs only depend on pixels near to the motifs $x_v$. On a single NVIDIA TITAN X Pascal GPU accelerator (12 GB memory), it takes approximately forty seconds to complete a full detection. Our implementation is available at https://github.com/sdbuch/refine.

4.5 Guaranteed, Efficient Detection of Occurrence Maps

In Section 4.4, we described how invariant processing of hierarchically-structured visual data naturally leads to problems of registering ‘spiky’ occurrence maps, and we introduced the formulation (4.10) for this purpose. In this section, we provide a theoretical analysis of a continuum model
for the proximal gradient descent method applied to the optimization formula (4.10). A byproduct of our analysis is a concrete prescription for the step size and rate of smoothing—in Section 4.5.2, we demonstrate experimentally that these prescriptions work excellently for the discrete formulation (4.10), leading to rapid registration of the input scene.

4.5.1 Multichannel Spike Model

We consider continuous signals defined on $\mathbb{R}^2$ in this section, as an ‘infinite resolution’ idealization of discrete images, free of interpolation artifacts. We refer to Section C.2 for full technical details. Consider a target signal

$$X_o = \sum_{i=1}^{c} \delta_{v_i} \otimes e_i,$$

where $\delta_{v_i}$ is a Dirac distribution centered at the point $v_i$, and an observation

$$X = \sum_{i=1}^{c} \delta_{u_i} \otimes e_i,$$

satisfying

$$v_i = A_\ast u_i + b_\ast.$$

In words, the observed signal is an affine transformation of the spike signal $X_o$, as in Figure 4.4(a, b). This model is directly motivated by the occurrence maps (4.11) arising in our hierarchical detection networks. Following (4.10), consider the objective function

$$\varphi_{L^2,\sigma}(A, b) \equiv \frac{1}{2c} \sum_{i=1}^{c} \left\| g_{\sigma^2 I - \sigma_0^2 (A^* A)^{-1}} \left( \det^{1/2}(A^* A) \left( g_{\sigma^2 I} \ast X_i \right) \circ \tau_{A,b} \right) - g_{0,\sigma^2 I} \ast (X_o) \right\|_{L^2}^2.$$

We study the following “inverse parameterization” of this function:

$$\varphi_{L^2,\sigma}^{\text{inv}}(A, b) \equiv \varphi_{L^2,\sigma}(A^{-1}, -A^{-1} b). \quad (4.15)$$
We analyze the performance of gradient descent for solving the optimization problem

$$\min_{A, b} \varphi^{\text{inv}}_{L^2, \sigma}(A, b).$$

Under mild conditions, local minimizers of this problem are global. Moreover, if $\sigma$ is set appropriately, the method exhibits linear convergence to the truth:

**Theorem 4.5.1** (Multichannel Spike Model, Affine Transforms, $L^2$). Consider an instance of the multichannel spike model, with $U = [u_1, \ldots, u_c] \in \mathbb{R}^{2 \times c}$. Assume that the spikes $U$ are centered and nondegenerate, so that $U1 = 0$ and $\text{rank}(U) = 2$. Then gradient descent

$$A_{k+1} = A_k - t_A \nabla A \varphi^{\text{inv}}_{L^2, \sigma}(A_k, b_k),$$

$$b_{k+1} = b_k - t_b \nabla b \varphi^{\text{inv}}_{L^2, \sigma}(A_k, b_k)$$

with smoothing

$$\sigma^2 \geq 2 \frac{\max_i \|u_i\|^2}{s_{\text{min}}(U)^2} \left( s_{\text{max}}(U)^2 \|A_* - I\|^2_F + c\|b_*\|^2_2 \right)$$

(4.16)

and step sizes

$$t_A = \frac{8\pi c \sigma^4}{s_{\text{max}}(U)^2},$$

$$t_b = 8\pi \sigma^4,$$

(4.17)

from initialization $A_0 = I, b_0 = 0$ satisfies

$$\frac{8\pi \sigma^4}{t_A} \|A_k - A_*\|^2_F + \|b_k - b_*\|^2_2 \leq \left( 1 - \frac{1}{2k} \right)^{2k} \left( \frac{8\pi \sigma^4}{t_A} \|I - A_*\|^2_F + \|b_*\|^2_2 \right),$$

(4.18)

where

$$k = \frac{s_{\text{max}}(U)^2}{s_{\text{min}}(U)^2},$$

with, $s_{\text{min}}(U)$ and $s_{\text{max}}(U)$ denoting the minimum and maximum singular values of the matrix $U$. 
Theorem 4.5.1 establishes a global linear rate of convergence for the continuum occurrence map registration formulation (4.15) in the relevant product norm, where the rate depends on the condition number of the matrix of observed spike locations $U$. This dependence arises from the intuitive fact that recovery of the transformation parameters $(A_\ast, b_\ast)$ is a more challenging problem than registering the observation to the motif—in practice, we do not observe significant degradation of the ability to rapidly register the observed scene as the condition number increases. The proof of Theorem 4.5.1 reveals that the use of inverse parameterization in (4.15) dramatically improves the landscape of optimization: the problem becomes strongly convex around the true parameters when the smoothing level is set appropriately. In particular, (4.16) suggests a level of smoothing commensurate with the maximum distance the spikes need to travel for a successful registration, and (4.17) suggests larger step sizes for larger smoothing levels, with appropriate scaling of the step size on the $A$ parameters to account for the larger motions experienced by objects further from the origin. In the proof, the ‘centered locations’ assumption $U1 = 0$ allows us to obtain a global linear rate of convergence in both the $A$ and $b$ parameters. This is not a restrictive assumption, as in practice it is always possible to center the spike scene (e.g., by computing its center of mass and subtracting), and we also find it to accelerate convergence empirically when it is applied.

4.5.2 Experimental Verification

To verify the practical implications of Theorem 4.5.1, which is formulated in the continuum, we conduct numerical experiments on registering affine-transformed multichannel spike images using the discrete formulation (4.10). We implement a proximal gradient descent solver for (4.10), and use it to register randomly-transformed occurrence maps, as visualized in Figure 4.4(a-b). We set the step sizes and level of smoothing in accordance with (4.16) and (4.17), with a complementary smoothing value of $\sigma_0 = 3$. Figure 4.4 shows representative results taken from one such run: the objective value rapidly converges to near working precision, and the normalized cross-correlation between the transformed scene and the motif rapidly reaches a value of 0.972. This rapid convergence implies the formulation (4.10) is a suitable base for an unrolled architecture with mild
Figure 4.4: Numerical verification of Theorem 4.5.1. (a): A multichannel spike motif containing 5 spikes. (b): A scene generated by applying a random affine transformation to the motif. (c): The solution to (4.10) with these data. The skewing apparent here is undone by the compensated external gaussian filter, which enables accurate localization in spite of these artifacts. (d): Change in objective value of (4.10) across iterations of proximal gradient descent. Convergence occurs in tens of iterations. (e): Change in normalized cross correlation across iterations (see Section C.1.3). We observe that the method successfully registers the multichannel spike scene. (f): Comparison between the left and right-hand side of equation (4.18) with gradient descent iterates from (4.10) (labeled as \( \varphi \) here). After an initial faster-than-predicted linear rate, the discretized solver saturates at a sub-optimal level. This is because accurate estimation of the transformation parameters \((A, b)\) requires subpixel-level preciseness, which is affected by discretization and interpolation artifacts. It does not hinder correct localization of the scene, as (e) shows.

depth, and is a direct consequence of the robust step size prescription offered by Theorem 4.5.1. Figure 4.4(f) plots the left-hand and right-hand sides of the parameter error bound (4.18) to evaluate its applicability to the discretized formulation: we observe an initial faster-than-predicted linear rate, followed by saturation at a suboptimal value. This gap is due to the difference between the continuum theory of Theorem 4.5.1 and practice: in the discretized setting, interpolation errors and finite-resolution artifacts prevent subpixel-perfect registration of the parameters, and hence exact recovery of the transformation \((A_*, b_*)\). In practice, successful registration of the spike scene, as demonstrated by Figure 4.4(e), is sufficient for applications, as in the networks we develop for hierarchical detection in Section 4.4.
4.6 Basin of Attraction for Textured Motif Registration with (4.4)

The theory and experiments we have presented in Section 4.5 justify the use of local optimization for alignment of spiky motifs. In this section, we provide additional corroboration beyond the experiment of Figure 4.1 of the efficacy of our textured motif registration formulation (4.4), under euclidean and similarity motion models. To this end, in Figure 4.5 we empirically evaluate the basin of attraction of a suitably-configured solver for registration of the crab body motif from Figure 4.2 with this formulation. Two-dimensional search grids are generated for each of the two setups as shown in the figure. For each given pair of transformation parameters, a similar multi-scale scheme over $\sigma$ as in the above complexity experiment is used, starting at $\sigma = 10$ and step size 0.05, and halved every 50 iterations. The process terminates after a total of 250 iterations. The final ZNCC calculated over the motif support is reported, and the figure plots the average over 10 independent runs, where the background image is randomly generated for each pair of parameters in each run. The ZNCC ranges from 0 to 1, with a value of 1 implying equality of the channel-mean-subtracted motif and transformed image content over the corresponding support (up to scale).

Panels (a) and (b) of Figure 4.5 show that the optimization method tends to succeed unconditionally up to moderate amounts of transformation. For larger sets of transformations, it is important to first appropriately center the image, which will significantly improve the optimization performance. In practice, one may use a combination of optimization and a small number of covering, so that the entire transformation space is covered by the union of the basins of attractions. We note that irregularity near the edges, especially in panel (a), can be attributed in part due to the randomness in the background embedding, and in this sense the size of the basin in these results conveys a level of performance across a range of simulated operating conditions. In general, these basins are also motif-dependent: we would expect these results to change if we were testing with the eye motif from Figure 4.3(a), for example. A notable phenomenon in Figure 4.5(b), where translation is varied against scale, is the lack of a clear-cut boundary of the basin at small scales.
Figure 4.5: Plotting a basin of attraction for the textured motif registration formulation (4.4). (a): Heatmap of the ZNCC at convergence (see Section C.1.3), for translation versus rotation. Optimization conducted with SE(2) motion model. (b): Heatmap of the ZNCC at convergence, for translation versus scale. Optimization conducted in ‘similarity mode’, a SE(2) motion model with an extra global scale parameter. In both experiments, each reported data point is averaged over 10 independent runs. (c-d): Notably, when the registration target $y$ is zoomed out relative to the motif $x_0$, resolution is lost in the detection target, so recovering it will cause interpolation artifacts and blur the image. This prevents the ZNCC value from converging to 1 despite correct alignment with the motif, and accounts for the results shown in (b) at small scales.

This is due to the effect illustrated in Figure 4.5(c-d), where interpolation artifacts corrupt the motif when it is ‘zoomed out’ by optimization over deformations, and hence registration can never achieve a ZNCC close to 1. For applications where perfect reconstruction is not required, such as the hierarchical detection task studied in Section 4.4, these interpolation artifacts will not hinder the ability to localize the motif in the scene at intermediate scales, and if the basin were generated with a success metric other than ZNCC, a better-defined boundary to the basin would emerge.
4.7 Discussion

In this chapter, we have taken initial steps towards realizing the potential of optimization over transformations of domain as an approach to achieve resource-efficient invariance with visual data. Below, we discuss several important future directions for the basic framework we have developed.

Statistical variability and complex tasks. To build invariant networks for complex visual tasks and real-world data beyond matching against fixed templates $x_o$, it will be necessary to incorporate more refined appearance models for objects, such as a sparse dictionary model or a deep generative model [185, 186, 187], and train the resulting hybrid networks in an end-to-end fashion. The invariant architectures we have designed in this work naturally plug into such a framework, and will allow for investigations similar to what we have developed in Section 4.4 into challenging tasks with additional structure (e.g., temporal or 3D data). Coping with the more complex motion models in these applications will demand regularizers $\rho$ for our general optimization formulation (4.1) that go beyond parametric constraints.

Theory for registration of textured motifs in visual clutter. Our experiments in Section 4.5 have demonstrated the value that theoretical studies of optimization formulations have with respect to the design of the corresponding unrolled networks. Extending our theory for spiky motif registration to more general textured motifs will enable similar insights into the roles played by the various problem parameters in a formulation like (4.4) with respect to texture and shape properties of visual data and the clutter present, and allow for similarly resource-efficient architectures to be derived in applications like the hierarchical template detection task we have developed in Section 4.4.3.

Hierarchical detection networks in real-time. The above directions will enable the networks we have demoed for hierarchical detection in Section 4.4 to scale to more general data models. At the same time, there are promising directions to improve the efficiency of the networks we design
for a task like this one at the modeling level. For example, the networks we design in Section 4.4.3 essentially operate in a ‘sliding window’ fashion, without sharing information across the strides $\lambda$, and they perform registration and detection separately. An architecture developed around an integrated approach to registration and detection, possibly building off advances in convolutional sparse modeling [188, 189, 190], may lead to further efficiency gains and push the resulting network closer to real-time operation capability.
Conclusion

In Chapter 1, we set ourselves the objective of translating three achievements of the compressed sensing research paradigm into the realm of theory and practice of deep learning; we recall their content at a high level:

1. Deep insights into the relationship between data geometry and efficient algorithms;
2. Theoretical insights guiding practical application;
3. A flexible modeling language robust to nonidealities and the real world.

In Chapters 2 and 3, inspired by the nonlinear low-dimensional structures ubiquitous in data in successful applications of deep learning, we studied resource tradeoffs for classifying data with geometric structure with deep feedforward neural networks, breaking new ground in the theoretical understanding of deep learning. We understood the role of the network depth in classifying nonlinear data—it acts as a fitting resource, with deeper networks fitting more complex geometries—as well as the role of the network width as a statistical resource, granting concentration of the initial random network and its gradients around certain well-structured objects. In a certain sense, these results represent the first stages of a vast generalization of our understanding of resource requirements for computing with piecewise linear data within the compressed sensing paradigm to the modern, nonlinear setting. In Chapter 4, we began translating some of these insights into practice by using the low-dimensional structure of image transformation manifolds to design resource-efficient and invariant deep neural network
architectures for processing image data in certain applications. Our networks use regularized optimization of a cost function for invariant registration as a basis, leading to a principled construction. We demonstrated these networks on simple toy tasks that display large improvements over brute-force template matching, while still guaranteeing invariance.

Still, we have only scratched the surface of the potential gains of exploiting low-dimensional structures in the practice of deep learning; and there remain deep theoretical questions raised by our studies in this thesis to answer, as well. To conclude this thesis, we briefly discuss three open directions to extend the results we have presented here.

The Role of the Network Depth in Optimization and Generalization. Although the modern “revolution” of deep learning came about through breakthroughs in the training of multi-layer neural networks [22], our theoretical understanding of the power of deep versus shallow networks is largely limited to training-independent approximation separations [191, 192, 193, 194, 195, 196], rather than how these networks can be efficiently learned in practice. In Chapters 2 and 3, we have established a concrete generalization benefit of depth for learning with one-dimensional nonlinear data: in a slogan, deeper networks fit more complicated geometries. As future work, it is of interest to extend this insight into the practical role of depth through two distinct investigations. First, an important direction is to extend our results for one-dimensional data manifolds to general, higher-dimensional datasets. Existing work on high-dimensional asymptotics for (neural tangent) kernel methods either does not treat the case where the depth grows with the number of data samples and network width [84, 119, 197], or obtains an incomplete picture of depth’s role by studying asymptotics with a finite training set [198]. Second, an interesting direction is to build off the technical tools we have introduced to investigate the emergence of neural collapse in network training, wherein $\ell^2$ regularization induces a certain predictable distribution of the logits in well-trained networks [199, 200]. Preliminary experiments, as well as published theoretical results [201], suggest that neural collapse may be a phenomenon unique to non-shallow models. Thus, in addition to providing the first theoretical characterization of the emergence of neural
collapse in neural network training,\textsuperscript{5} this investigation offers a chance to understand an optimization-centric difference between deep and shallow network training.

**Invariant Computing with Visual Data: A Unified Approach to Appearance and Deformation?** In Chapter 4, we have designed resource-efficient neural network architectures for computing with visual data subject to deformations of the image plane. These architectures offer a mathematically-principled approach to designing guaranteed invariant architectures with respect to known, structured groups of deformations, but a key challenge in deploying them on complex real-world tasks is simultaneously coping with hard-to-model data variabilities in appearance, shape, illumination, and so on. A promising direction to incorporate appearance modeling into these invariant architectures is to replace components that rely on prior observations of visual data (e.g. fixed, known templates for motifs) with flexible generative models for visual data, such as those constructed through the MCR\textsuperscript{2} loss or the LDR framework [187, 209, 210]. The resulting hybrid networks retain end-to-end trainability, by virtue of their unrolled construction, and can potentially be applied across a wide variety of downstream tasks, due to the MCR\textsuperscript{2} framework’s ability to be applied in a supervised or unsupervised fashion. The main research challenge consists of properly designing the training scheme for the hybrid network so that the appearance and deformation sub-networks can be jointly trained to produce a disentangled representation of appearance and deformation in the input.

**Sharp Gradient Concentration for General Neural Network Architectures.** Currently, the “neural tangent kernel” (NTK) theory represents one of the most general theoretical tools for proving theorems about training and generalization with realistic neural networks and datasets (in spite of its well-discussed limitations in terms of, at least, sample efficiency [211]). However, surprisingly little is known about the precise implications of the NTK theory on the choice of network architecture in practical regimes. Indeed, although NTK ideas are broadly used as

\textsuperscript{5}Existing theoretical studies of neural collapse apply to unconstrained feature models [202, 203, 204, 205, 206], which have a matrix factorization geometry rather than a neural network training geometry, or to properties of global minimizers, rather than to the ability of an efficient algorithm to produce such a minimizer [207, 208].
heuristics for trainability of general feedforward networks [212, 213], existing theorems do not apply to sharp width-depth scalings, making their connections to practical networks unclear; and recent results that do connect to sharp width-depth scalings for broad classes of network architectures (e.g., feedforward networks with smooth activations) are written non-algorithmically and challenging to apply to practical settings [214]. In Chapter 2, we presented the first result on uniform feature and gradient concentration in feedforward ReLU networks applicable to the optimal width-depth scaling regime. The technical tools developed to prove these theorems can be generalized to general activation functions and to architectures beyond feedforward networks, such as residual networks [215]. As future research, one might aim to carry out these generalizations, and explore the implications on trainability and generalization in networks of practical widths and depths. A core technical challenge is articulating the proper weight initialization schemes for different architectures [213, 214, 216]: a likely byproduct of this investigation is an improved understanding of the advantages and disadvantages of the popular ReLU activation function versus other smooth activations, as manifested through their NTKs.
References


Appendix A: Proofs for the Multiple Manifold Problem

A.1 Extended Problem Formulation

A.1.1 Regarding the Algorithm

We analyze a gradient-like method for the minimization of the empirical loss $L^\mu_N$. After randomly initializing the parameters $\theta_0^N$ as $W^\ell \sim_{\text{i.i.d.}} \mathcal{N}(0, 2/n)$ if $\ell \in [L]$ and $W^{L+1} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$, independently of the samples $x_1, \ldots, x_N$, we consider the sequence of iterates

$$
\theta_{k+1}^N = \theta_k^N - \tau \nabla L^{\mu_N}(\theta_k^N),
$$

(A.1)

where $\tau > 0$ is a step size, and $\nabla L^{\mu_N}$ represents a ‘formal gradient’ of the loss $L^{\mu_N}$, which we define as follows: first, we define formal gradients of the network output by

$$
\nabla_{W^\ell} f_\theta(x) = \beta_\theta^\ell(x) \alpha_\theta^{\ell-1}(x)^*,
$$

for $\ell \in [L]$ and $x \in \mathcal{M}$, where we have introduced the definitions

$$
\beta_\theta^\ell(x) = \left( W^{L+1} P_{l_L(x)} W^L P_{l_{L-1}(x)} \cdots W^{\ell+2} P_{l_{\ell+1}(x)} \right)^*,
$$

for $\ell = 0, 1, \ldots, L - 1$, and where we additionally define

$$
I_\ell(x) = \text{supp} \left( 1_{\alpha_\theta^\ell(x) > 0} \right), \quad P_{l_\ell(x)} = \sum_{i \in I_\ell(x)} e_i e_i^*
$$

for the orthogonal projection onto the set of coordinates where the $\ell$-th activation at input $x$ is positive. We call the vectors $\beta_\theta^\ell(x)$ the backward features or backward activations—they correspond
to the backward pass of our neural network. We also define

$$\nabla_{W_{L+1}} f_\theta(x) = \alpha_\theta^L(x)^*.$$ 

We then define the formal gradient of the loss $L_{\mu_N}$ by

$$\nabla L_{\mu_N}(\theta) = \int_M \nabla f_\theta(x) \zeta_\theta(x) \, d\mu_N(x).$$

Let us emphasize again that the expressions above are definitions, not gradients in the analytical sense: we introduce these definitions to cope with nonsmoothness of the ReLU $[\cdot]_+$. On the other hand, our formal gradient definitions coincide with the expressions one obtains by applying the chain rule to differentiate $L_{\mu_N}$ at points where the ReLU is differentiable, and we will make use of this fact to proceed with these formal gradients in a manner almost identical to the differentiable setting.

We reiterate here our notational conventions for quantities evaluated at these iterates: we denote evaluation of quantities such as the features and prediction error at parameters along the gradient descent trajectory using a subscript $k$, with an omitted subscript denoting evaluation at the initial $k = 0$ parameters, and we add a superscript $N$ to parameters such as the prediction error to emphasize that they are evaluated at the parameters generated by (A.1). For example, in this notation we express $\zeta_\theta^N_k$ as $\zeta_k^N$. In addition, we use $\theta_0$ to denote the initial parameters $\theta_0^N$. We emphasize the dependence of certain quantities on these random initial parameters notationally, including the initial network function $f_{\theta_0}$.

A.1.2 Regarding the Data Manifolds

We now provide additional details regarding our assumptions on the data manifolds. For background on curves and more broadly Riemannian manifolds, we refer the reader to [217, 218].
assume that $M = M_+ \cup M_-$, where $M_+$ and $M_-$ are two disjoint complete connected\(^1\) Riemannian submanifolds of the unit sphere $S^{n_0-1}$, with $n_0 \geq 3$. In particular, $M_\pm$ are compact. We take as metric on these manifolds the metric induced by that of the sphere, which we take in turn as that induced by the euclidean metric on $\mathbb{R}^{n_0}$. We write $\mu^\infty_+$ and $\mu^\infty_-$ for the measures on $M_+$ and $M_-$ (respectively) induced by the data measure $\mu^\infty$, and we assume that $\mu^\infty$ admits a density $\rho$ with respect to the Riemannian measure on $M$, writing $\rho_+$ and $\rho_-$ for the densities on $M_\pm$ induced by the density $\rho$. When $d_0 = 1$, we add additional structural assumptions to the above: we assume that $M_\pm$ are smooth, simple, regular curves.

Concretely, that $M$ admits a density $\rho$ with respect to the Riemannian measure means that

$$1 = \int_M d\mu^\infty(x) = \int_{M_+} \rho_+(x) \, dV_+(x) + \int_{M_-} \rho_-(x) \, dV_-(x).$$

When $d_0 = 1$, because $M_\pm$ are smooth regular curves, they admit global unit-speed parameterizations with respect to arc length $\gamma_\pm : I_\pm \to S^{n_0-1}$, where $I_\pm$ are intervals of the form $[0, \text{len}(M_\pm)]$. In this setting, the curvature constraint is expressed as

$$\max\left\{ \sup_{s \in I_+} \|\gamma''_+(s)\|_2, \sup_{s \in I_-} \|\gamma''_-(s)\|_2 \right\} \leq \kappa,$$

and we observe that the fact that $M_\pm$ are sphere curves implies $\kappa \geq 1$.\(^2\) Exploiting the coordinate representation of the Riemannian measure and the fixed inherited metric from $\mathbb{R}^{n_0}$, we thus have

$$\int_{M_\pm} \rho_\pm(x) \, dV_\pm(x) = \int_{I_\pm} \rho_\pm \circ \gamma_\pm(t) \|\gamma'_\pm(t)\|_2 \, dt = \int_{I_\pm} \rho_\pm \circ \gamma_\pm(t) \, dt.$$

We will exploit this formula in the sequel to compare between $L_\mu^p(M)$ and $L^p(M)$ norms of functions defined on the manifold. More generally, we will frequently make use of similar reasoning

\(^1\)Certain parts of our argument, such as the concentration result Theorem A.2.2, are naturally applicable to cases where $M_\pm$ themselves have a finite number of connected components with a mild dependence on this number, and we state them as such. We skip this extra generality in our dynamics arguments to avoid an additional ‘juggling act’ that would obscure the main ideas.

\(^2\)We point out that the curvature of the manifolds does not enter into the proof of the concentration result Theorem A.2.2, so there is no ambiguity in discussing curvature only in the context of curves.
that leverages the existence of unit-speed parameterizations for the curves.

For clarity we rewrite the global regularity condition: we assume there exist constants $0 < c_L \leq 1$, $K_L \geq 1$ such that

$$\forall s \in (0, c_L/k], (x, x') \in M_\star \times M_\star, \star \in \{+, -\} : \angle(x, x') \leq s \Rightarrow \text{dist}_M(x, x') \leq K_L s,$$

(A.2)

where $\text{dist}_M$ denotes the Riemannian distance between points in a common connected component, and we define $C_L = K_L^2/c_L^2$. Because $M_\pm$ are simple curves, they do not self-intersect; the assumption (A.2) gives a quantitative characterization of how far the curves are from self-intersecting. We illustrate how the associated constants can be obtained from the assumption that the manifolds are simple curves: for either $\star \in \{+, -\}$, consider a connected component $M_\star \subset M$, and for any $0 < s \leq \text{len}(M_\star)$, define

$$r_\star(s) = \inf_{x, x' \in M_\star \times M_\star, \text{dist}_M(x, x') > s} \angle(x, x').$$

If $r_\star(s) = 0$, by compactness we can construct a sequence of pairs of points that converges to $r_\star(s)$, but this would imply that $M_\star$ is self-intersecting, contradicting our assumption that it is simple. It follows that $r_\star(s) > 0$ for any value of $s$. If we now define $\tilde{K}_s = r_\star(s)/s$, it follows that for any $(x, x') \in M_\star \times M_\star$,

$$\angle(x, x') \leq s \Rightarrow \text{dist}_M(x, x') \leq \tilde{K}_s s.$$

Our regularity assumption implies that a single such constant holds for a range of scales below the curvature scale, which is a mild assumption since $\tilde{K}_s$ approaches 1 as $s$ approaches 0.

A.1.3 Regarding the Initialization

The manner in which we have defined our initial random neural network $f_{\theta_0}$ is sometimes referred to as “fan-out initialization” in the literature—it guarantees that feature norms are preserved from layer to layer in the network, and thereby avoids the vanishing and exploding gradient problems. The difference between this initialization and the so called “standard” or “fan-in” ini-
tialization is only in the first and last layer weights, yet in a sufficiently deep network trained in the NTK regime the effect of any single layer is negligible and the dynamics of our network will be essentially identical to one with standard initialization. On the other hand, following the work of Jacot et al. [75], it has become common in the theoretical literature to consider a different construction of the neural network called “NTK parameterization”, which is in some ways more convenient for theoretical analysis. In particular, Arora et al. [65] prove their results on NTK concentration using this parameterization; to facilitate a comparison between our concentration result (Theorem 2) and theirs, we discuss the connection between fan-out and NTK parameterization in this section. This material is well-known and no doubt can be found already in the literature, but we believe it may be helpful to translate it into our notation.

Recall our definitions for the weights and features in our neural network: we have $W^\ell \sim_{\text{i.i.d.}} \mathcal{N}(0, 2/n)$ if $\ell \in \{0, 1 \ldots, L\}$ and $W^{L+1} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$, with features defined for $\ell = 0, 1, \ldots, L$ by

$$
\alpha^\ell(x) = \begin{cases} 
  x & \ell = 0 \\
  [W^\ell \alpha^{\ell-1}(x)]_+ & \text{otherwise},
\end{cases}
$$

and output $f_{\theta_0}(x) = W^{L+1}\alpha^L(x)$. Within this section—and only within this section—we shall define auxiliary weights by $G^1 \in \mathbb{R}^{n \times n_0}$, $G^\ell \in \mathbb{R}^{n \times n}$ for integer $1 \leq \ell < L + 1$, and $G^{L+1} \in \mathbb{R}^{1 \times n}$, with distributions $G^\ell \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$ for $\ell \in \{0, 1 \ldots, L + 1\}$, independent of everything else in the problem. As before, for $\ell \in \{0, 1, \ldots, L\}$ we use $\alpha^{(\ell)}_{\text{NTK}}(x)$ to denote the layer-$\ell$ features:

$$
\alpha^{(\ell)}_{\text{NTK}}(x) = \begin{cases} 
  x & \ell = 0 \\
  [G^\ell \alpha^{(\ell-1)}_{\text{NTK}}(x)]_+ & \text{otherwise}.
\end{cases}
$$

This network’s output will be written

$$
f_{\text{NTK}}(x) = \left( \prod_{\ell=1}^L \sqrt{\frac{2}{n}} \right) G^{L+1}\alpha^{(L)}_{\text{NTK}}(x).
$$
By 1-homogeneity (absolute) of $\sigma$, it follows that $f_{\theta_0} \overset{d}{=} f_{\text{NTK}}$. As the notation suggests, the network $f_{\text{NTK}}$ corresponds to a “NTK parameterization” network—although this network and $f_{\theta_0}$ are equivalent in terms of predictions, their “gradients” are not equivalent. The NTK for the NTK parameterization network is obtained by differentiating (at points of differentiability): after calculating (as in Lemma A.2.7), we introduce notation as we did for the fan-out parameterization network in Section A.1.1, so that

$$\Theta_{\text{NTK}}(x, x') = \left( \nabla f_{\text{NTK}}(x), \nabla f_{\text{NTK}}(x') \right),$$

with (for $\ell = 1, \ldots, L + 1$)

$$\nabla \mathcal{G}^\ell f_{\text{NTK}}(x) = \left( \prod_{\ell=1}^{L} \sqrt{\frac{2}{n}} \beta_{\text{NTK}}^{(\ell-1)}(x) a_{\text{NTK}}^{(\ell-1)}(x) \right)^*$$

where

$$\beta_{\text{NTK}}^{(\ell)}(x) = \begin{cases} G^{L+1} P_{L}^{\text{NTK}}(x) G^{L} P_{L-1}^{\text{NTK}}(x) \cdots G^{\ell+2} P_{\ell+1}^{\text{NTK}}(x) \ast & \ell = 0, 1, \ldots, L - 1 \\ 1 & \ell = L, \end{cases}$$

and

$$I_{\ell}^{\text{NTK}}(x) = \text{supp} \left( \mathbb{1}_{a_{\text{NTK}}^{(\ell)}(x) > 0} \right).$$

We shall relate the NTK parameterization $\Theta_{\text{NTK}}$ to our fan-out parameterization $\Theta$ using homogeneity of the ReLU. First, let us observe that

$$\left\{ i \in [n] \right\} \left( a^\ell(x) \right)_i > 0 \overset{d}{=} \left\{ i \in [n] \right\} \left( a_{\text{NTK}}^{(\ell)}(x) \right)_i > 0,$$

because $[\cdot]_+$ is 1-homogeneous and we have $G^\ell \overset{d}{=} \sqrt{n/2} W^\ell$ when $\ell \leq L$. For $\ell \in \{0, 1, \ldots, L\}$, we note that $a^\ell(x)$ and $a_{\text{NTK}}^{(\ell)}(x)$ depend only on the parameters $(W^1, \ldots, W^\ell)$ and $(G^1, \ldots, G^\ell)$, respectively. If we write $\theta = (W^1, \ldots, W^L)$ and $\theta_{\text{NTK}} = (G^1, \ldots, G^L)$, then it follows that
\(\alpha^\ell(x)\) is a \(\ell\)-homogeneous function of \(\theta\) (and likewise for \(\alpha^{(L)}_{\text{NTK}}(x)\)). In addition, the projection matrices \(P_{\ell}(x)\) are 0-homogeneous functions of \(\theta\), and so taking \(\ell \in \{0, 1, \ldots, L-1\}\) and counting parameters in the definitions of \(\beta^\ell(x)\) and \(\beta^{(L)}_{\text{NTK}}(x)\) implies that these two functions are \((L - \ell - 1)\)-homogeneous functions of \(\theta\) and \(\theta_{\text{NTK}}\), respectively. Of course, for \(\ell = L\), they are 0-homogeneous. Thus, using that \(G^\ell d = \sqrt{n/2} W^\ell\) for \(\ell \leq L\), we obtain

\[
\nabla G^\ell f_{\text{NTK}}(x) = \begin{cases} \sqrt{2/n} W^\ell f_{\theta_0}(x) & \ell = 0, 1, \ldots, L \\ \nabla W^\ell f_{\theta_0}(x) & \ell = L + 1. \end{cases}
\]

Although we have argued equidistributionality above for each index \(\ell\) separately for simplicity, the elementary nature of our arguments (we are just moving scalars around) and the statistical dependencies across gradients allows us to apply the same argument ‘in parallel’ to the sum of inner products between gradients, yielding

\[
\Theta_{\text{NTK}}(x, x') \overset{d}{=} \langle \alpha^L(x), \alpha^L(x') \rangle + \frac{2}{n} \sum_{\ell=1}^{L} \langle \alpha^{\ell-1}(x), \alpha^{\ell-1}(x') \rangle \langle \beta^{\ell-1}(x), \beta^{\ell-1}(x') \rangle.
\]

This expression makes it immediately clear that our concentration framework proves sharp concentration of the NTK of a uniform-width NTK parameterization feedforward ReLU network that improves over the results of Arora et al. [65] when the data are on the sphere\(^3\) — a simple adaptation of the proof of Theorem A.2.3 will suffice.

A.1.4 Notation

General Notation

If \(n \in \mathbb{N}\), we write \([n] = \{1, \ldots, n\}\). We generally use bold notation \(x, A\) for vectors, matrices, and operators and non-bold notation for scalars and scalar-valued functions. For a vector \(x\) or a matrix \(A\), we will write entries as either \(x_j\) or \(A_{ij}\), or \((x)_j\) or \((A)_{ij}\); we will occasionally index

\(^3\) The results of Arora et al. [65] apply to data of norm no larger than 1, but it is straightforward to extend our results for spherical data to this setting, using the \(1\)-homogeneity of \(\Theta\) in each argument (as a kernel on the entire ambient space \(\mathbb{R}^n \times \mathbb{R}^n\)) to write \(\Theta(x, x') = \|x\|_2 \|x'\|_2 \Theta(x/\|x\|_2, x'/\|x'\|_2)\).
the rows or columns of $A$ similarly as $(A)_i$ or $(A)_j$, with the particular meaning made clear from context. We write $[x]_+ = \max\{x, 0\}$ for the ReLU activation function; if $x$ is a vector, we write $[x]_+$ to denote the vector given by the application of $[ \cdot ]_+$ to each coordinate of $x$, and we will generally adopt this convention for applying scalar functions to vectors. If $x, x' \in \mathbb{R}^n$ are nonzero, we write $\langle x, x' \rangle = \cos^{-1}(\langle x, x' \rangle / \|x\|_2\|x'\|_2)$ for the angle between $x$ and $x'$.

The vectors $(e_i)$ denote the canonical basis for $\mathbb{R}^n$. We write $\langle x, y \rangle = \sum_i x_i y_i$ for the euclidean inner product on $\mathbb{R}^n$, and if $0 < p < +\infty$ we write $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ for the $\ell^p$ norms (when $p \geq 1$) on $\mathbb{R}^n$. We also write $\|x\|_0 = \{|i \in [n] | x_i ≠ 0\}$ and $\|x\|_\infty = \max_{i \in [n]} |x_i|$. The unit ball in $\mathbb{R}^n$ is written $\mathbb{B}^n = \{x \in \mathbb{R}^n | \|x\|_2 \leq 1\}$, and we denote its (topological) boundary, the unit sphere, as $\mathbb{S}^{n-1}$. We reserve the notation $\| \cdot \|$ for the operator norm of a $m \times n$ matrix $A$, defined as $\|A\| = \sup_{\|x\|_2 \leq 1} \|Ax\|_2$; more generally, we write $\|A\|_{\ell^p \to \ell^q} = \sup_{\|x\|_p \leq 1} \|Ax\|_q$ for the corresponding induced matrix norm. For $m \times n$ matrices $A$ and $B$, we write $\langle A, B \rangle = \text{tr}(A^*B)$ for the standard inner product, where $A^*$ denotes the transpose of $A$, and $\|A\|_F = \sqrt{\langle A, A \rangle}$ for the Frobenius norm of $A$.

The Banach space of (equivalence classes of) real-valued measurable functions on a measure space $(X, \mu)$ satisfying $(\int_X |f|^p \, d\mu)^{1/p} < +\infty$ is written $L^p(\mu)$ or simply $L^p$ if the space and/or measure is clear from context; we write $\| \cdot \|_{L^p}$ for the associated norm, and $\langle \cdot, \cdot \rangle_{L^2}$ for the associated inner product when $p = 2$, with the adjoint operation denoted by $^*$. For an operator $\mathcal{T} : L^p_\mu \to L^q_\nu$, we write $\mathcal{T}[f]$ to denote the image of $f$ under $\mathcal{T}$, $\mathcal{T}^i$ to denote the operator that applies $\mathcal{T}$ $i$ times, and $\|\mathcal{T}\|_{L^p_\mu \to L^q_\nu} = \sup_{\|f\|_{L^p_\mu} \leq 1} \|\mathcal{T}[f]\|_{L^q_\nu}$. We use $\text{Id}$ to denote the identity operator, i.e. $\text{Id}[g] = g$ for every $g \in L^p_\mu$. We say that $\mathcal{T}$ is positive if $\langle f, \mathcal{T}[f]\rangle_{L^2} \geq 0$ for all $f \in L^2$; for example, the identity operator is positive.

For an event $\mathcal{E}$ in a probability space, we write $1_\mathcal{E}$ to denote the indicator random variable that takes the value 1 if $\omega \in \mathcal{E}$ and 0 otherwise. If $\sigma > 0$, by $g \sim \mathcal{N}(0, \sigma^2 I)$ we mean that $g \in \mathbb{R}^n$ is distributed according to the standard i.i.d. gaussian law with variance $\sigma^2$, i.e., it admits the density $(2\pi\sigma^2)^{-n/2} \exp(-\|x\|_2^2/(2\sigma^2))$ with respect to Lebesgue measure on $\mathbb{R}^n$; we will occasionally write this equivalently as $g \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$. We use $\overset{d}{=} \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$ to denote the “identically-distributed”
equivalence relation.

We use “numerical constant” and “absolute constant” interchangeably for numbers that are independent of all problem parameters. Throughout the text, unless specified otherwise we use $c, c', c'', C, C', C'', K, K', K''$, and so on to refer to numerical constants whose value may change from line to line within a proof. Numerical constants with numbered subscripts $C_1, C_2, \ldots$ and so on will have values fixed at the scope of the proof of a single result, unless otherwise specified. We generally use lower-case letters to refer to numerical constants whose value should be small, and upper case for those that should be large; we will generally use $K, K'$ and so on to denote numerical constants involved in lower bounds on the size of parameters required for results to be valid. If $f$ and $g$ are two functions, the notation $f \preceq g$ means that there exists a numerical constant $C > 0$ such that $f \leq Cg$; the notation $f \succeq g$ means that there exists a numerical constant $C > 0$ such that $f \geq Cg$; and when both are true simultaneously we write $f \asymp g$. If $f$ is a real-valued function with sufficient differentiability properties, we will write both $f'$ and $\hat{f}$ for the derivative of $f$, and when higher derivatives are available we will occasionally denote them by $f^{(n)}$, with this usage specifically made clear in context. For a metric space $X$ and a Lipschitz function $f : X \to \mathbb{R}$, we write $\|f\|_{\text{Lip}}$ to denote the minimal Lipschitz constant of $f$.

**Summary of Operator and Error Definitions**

We collect some of the important definitions that appear throughout the main text and the appendices in this section. We begin with the NTK-type operators that appear in our analysis. Recall from Section A.1.1 our definition for the backward features: we have

$$p_{\ell}^\ell(x) = \left(W^{L+1}P_{L}(x)W^{L}P_{L-1}(x)\cdots W^{\ell+2}P_{\ell+1}(x)\right)^*$$

for $\ell = 0, 1, \ldots, L - 1$, and where we additionally define

$$I_{\ell}(x) = \text{supp} \left(1_{a_{\ell}^\ell(x)>0}\right), \quad P_{\ell}(x) = \sum_{i \in I_{\ell}(x)} e_ie_i^*$$

106
for the orthogonal projection onto the set of coordinates where the \( \ell \)-th activation at input \( x \) is positive. "The" neural tangent kernel is defined as

\[
\Theta(x, x') = \left\langle \nabla f_{\theta_0}(x), \nabla f_{\theta_0}(x') \right\rangle = \langle \alpha^L(x), \alpha^L(x') \rangle + \sum_{\ell=0}^{L-1} \langle \alpha^\ell(x), \alpha^\ell(x') \rangle \langle \beta^\ell(x), \beta^\ell(x') \rangle,
\]

with corresponding operator on \( L^2_{\mu^{\infty}}(M) \)

\[
\Theta[g](x) = \int_M \Theta(x, x')g(x') \, d\mu^{\infty}(x').
\]

As shown in Lemma A.2.7, this is not exactly the kernel that governs the dynamics of gradient descent: the relevant kernels in this context are defined as

\[
\Theta^N_k(x, x') = \int_0^1 \left\langle \nabla f_{\theta^N_k}(x'), \nabla f_{\theta^N_{k-1}}(x) \right\rangle \, dt.
\]

We define operators \( \Theta^N_k \) on \( L^2_{\mu^{N}}(M) \) corresponding to integration against these kernel in a manner analogous to the definition of \( \Theta \):

\[
\Theta^N_k[g](x) = \int_M \Theta^N_k(x, x')g(x') \, d\mu^{N}(x').
\]

We then move to the deterministic approximations for \( \Theta \) that we develop: we define

\[
\varphi(\nu) = \cos^{-1}\left( (1 - \nu/\pi) \cos \nu + (1/\pi) \sin \nu \right),
\]

which governs the angle evolution process in the initial random network, as studied in Section A.5, and write \( \varphi^{(\ell)} \) to denote \( \ell \)-fold composition of \( \varphi \) with itself. We define

\[
\psi_1(\nu) = \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(\nu) \right) \prod_{\ell'=\ell}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(\nu)}{\pi} \right),
\]

107
which is the “output” of our main result on concentration, Theorem A.2.2, and

\[ \psi(\nu) = \frac{n}{2} \sum_{\ell=0}^{L-1} \prod_{\nu' = \ell}^{L-1} \left( 1 - \frac{\varphi(\nu')}{\pi} \right), \]

which is at the core of the certificate construction problem. We think of \( \psi \) as an analytically-simpler version of \( \psi_1 \), with an approximation guarantee given in Lemma A.3.8. Throughout these appendices, we will make use of basic properties of \( \psi_1 \) and \( \psi \) that follow from properties of \( \varphi \) without explicit reference; the source material for these types of claims is Lemma A.5.5, which gives elementary properties of \( \varphi \) (for example, that it takes values in \([0, \pi/2]\), which implies that \( \psi \) and \( \psi_1 \) are no larger than \( nL/2 \)). For derived estimates, we call the reader’s attention to the contents of Section A.3.1; we will make explicit reference to these results when we need them, however. Although we have mentioned approximations \( \hat{\Theta} \) and \( \hat{\Theta} \) in the main text, we will prefer in these appendices to explicitly reference \( \psi \) and \( \psi_1 \) to avoid confusion; as an exception, we will use the \( \hat{\Theta} \) notation in Section A.3 as discussed there. Our approximation for the initial prediction error is

\[ \hat{\zeta}(x) = -f_*(x) + \int_M f_{\theta_0}(x') d\mu^\infty(x'), \quad (A.3) \]

where we recall \( f_{\theta_0} \) denotes the network function with the initial (random) weights. In particular, this approximates the network function with a constant, and the error as a piecewise constant function on \( M_{\pm} \). This approximation is justified in Lemma A.4.11.

A.2 Proofs of the Main Results

A.2.1 Main Results

**Theorem A.2.1.** Let \( M \) be a one-dimensional Riemannian manifold satisfying our regularity assumptions. For any \( 0 < \delta \leq 1/e \), choose \( L \) so that

\[ L \geq C_1 \max\{C_{\mu}\log^9(1/\delta) \log^{24}(C_{\mu}n_0 \log(1/\delta)), \kappa^2 C_A\}, \]
let $N \geq L^{10}$, set $n = C_2 L^{99} \log^9 (1/\delta) \log^{18} (L n_0)$, and fix $\tau > 0$ such that

$$\frac{C_3}{n L^2} \leq \tau \leq \frac{C_4}{n L}.$$ 

Then if there exists a function $g \in L^2_{\mu^\infty}(M)$ such that

$$\|\Theta[g] - \zeta\|_{L^2_{\mu^\infty}(M)} \leq C_5 \frac{\sqrt{\log(1/\delta) \log(n n_0)}}{L \min\{\rho^q_{\min}, \rho^{-q}_{\min}\}}; \quad \|g\|_{L^2_{\mu^\infty}(M)} \leq C_6 \frac{\sqrt{\log(1/\delta) \log(n n_0)}}{n \rho^q_{\min}},$$

(A.4)

with probability at least $1 - \delta$ over the random initialization of the network and the i.i.d. sample from $\mu^\infty$, the parameters obtained at iteration $\lfloor L^{39/44} / (n \tau) \rfloor$ of gradient descent on the finite sample loss $L_{\mu^N}$ yield a classifier that separates the two manifolds.

The constants $C_1, \ldots, C_4 > 0$ depend only on the constants $q_{\text{cert}}, C_5, C_6 > 0$, the constants $\kappa, C_\lambda$ are respectively the extrinsic curvature constant and the global regularity constant defined in Section A.1, and the constant $C_{\mu^\infty} = \max\{\rho^q_{\min}, \rho^{-q}_{\min}\} (1 + \rho_{\max})^6 (\min \{\mu^\infty(M_+), \mu^\infty(M_-)\})^{-11/2}$, where $q = 11 + 8q_{\text{cert}}$.

Proof. The proof is an application of Lemma A.2.6, with suitable instantiations of the parameters of that result; to avoid clashing with the probability parameter $\delta$ in this theorem, we use $\epsilon$ for the parameter $\delta$ appearing in Lemma A.2.6. Define $C_\rho = \max\{\rho_{\min}, \rho_{\min}^{-1}\}$. We will pick $q = 39/44$ and $\epsilon = 5/47$, so that the relevant hypotheses of Lemma A.2.6 become (after worst-casing in the bound on $N$ somewhat for readability)

$$d \geq K \log(n n_0 C_M)$$

$$n \geq K' \max\left\{L^{99} d^9 \log^9 L, \kappa^{2/5}, \left(\frac{\kappa}{C_\lambda}\right)^{1/3}\right\}$$

$$L \geq K'' \max\{C_\rho^{2q_{\text{cert}}} d, \kappa^2 C_\lambda\}$$

$$N \geq K'''' \frac{C_\rho^{133/18 + (152/27)q_{\text{cert}}} (1 + \rho_{\max})^{133/54}}{\min \{\mu^\infty(M_+), \mu^\infty(M_-)\}^{19/18}} d^{8/3} L^9 \log^3 L,$$
and the conclusion we will appeal to becomes

\[
\mathbb{P}\left[ \left\| \zeta_N^{[L^{39/44} / (n\tau)]} \right\|_{L^\infty(M)} \leq \frac{CC_p^{1+2q_{\text{cert}}/3} (1 + \rho_{\text{max}})^{1/2}}{\min \{\mu^\infty(M_+), \mu^\infty(M_-)\}^{1/2}} \frac{d^{3/4} \log^{4/3} L}{L^{1/11}} \right] \geq 1 - \frac{C' L e^{-c'd}}{n\tau}.
\]

Under our choice of \( \tau \) and enforcing

\[
L \geq \frac{(2C)^{11} C_p^{11+22q_{\text{cert}}/3} (1 + \rho_{\text{max}})^{11/2} d^{33/4} \log^{44/3} L}{(\min \{\mu^\infty(M_+), \mu^\infty(M_-)\})^{11/2}},
\]

we have the equivalent result

\[
\mathbb{P}\left[ \left\| \zeta_N^{[L^{39/44} / (n\tau)]} \right\|_{L^\infty(M)} \leq \frac{1}{2} \right] \geq 1 - L^3 e^{-c'd}
\]
\[
\geq 1 - e^{-c'd},
\]

where the last bound holds when \( d \geq K \log L \), which is redundant with the hypotheses on \( n \) and \( d \) required to use Lemma A.2.6. Thus, when in addition \( d \geq (1/c') \log(1/\delta) \), we obtain

\[
\mathbb{P}\left[ \left\| \zeta_N^{[L^{39/44} / (n\tau)]} \right\|_{L^\infty(M)} \leq \frac{1}{2} \right] \geq 1 - \delta.
\]

(A.6)

Therefore to conclude, we need only argue that our choices of \( n, N, L, d, \) and \( \delta \) in the theorem statement suffice to satisfy the hypotheses of Lemma A.2.6. We have already satisfied the conditions on \( \varepsilon \) and \( q \). We notice that (A.5) implies that it suffices to enforce simply \( N \geq L^{10} \), and following Lemma A.3.1, we can bound \( C_M \) as in (A.63) in the proof of Lemma A.2.6 by

\[
C_M \leq 1 + \frac{\text{len}(M_+)}{\mu^\infty(M_+)} + \frac{\text{len}(M_-)}{\mu^\infty(M_-)} \leq 2 \frac{1 + \rho_{\text{max}}}{\rho_{\text{min}}}
\]

Because \( n \geq L^{99} \) and \( L \geq C_p (1 + \rho_{\text{max}}) \), we can eliminate \( C_M \) from the lower bound on \( d \) while paying only an extra factor of 2 in the constant. In addition, because \( \kappa \geq 1 \) and \( C_{\lambda} \geq \max\{1, 1/c_{\lambda}\} \), we can remove the \( \kappa^{2/5} \) and \( \left( \frac{\kappa}{c_{\lambda}} \right)^{1/3} \) lower bounds on \( n \), since they are enforced through \( L \) already via the bound \( L \geq K' \kappa^2 C_{\lambda} \), worsening the absolute constant if needed. These simplifications lead
us to the sufficient conditions (plus the certificate existence hypotheses)

\[ d \geq K \max \{ \log(1/\delta), \log(nn_0) \} \]

\[ n \geq K'L^{99} d^9 \log^9 L \]

\[ L \geq K'' \max \left\{ \frac{C_\rho^{11+22q_{\text{cert}}/3} (1 + \rho_{\max})^{11/2} d^{33/4} \log^{44/3} L}{\left( \min \{\mu_\infty(M_+), \mu_\infty(M_-)\} \right)^{11/2}}, \kappa^2 C_\lambda \right\} \]

\[ N \geq L^{10}. \]

We ignore the condition on \( N \) below, since it matches with the theorem statement. When \( \delta \leq 1/e \), given that \( n_0 \geq 3 \) we have \( nn_0 \geq e \) and \( \max \{ \log(1/\delta), \log(nn_0) \} \leq \log(1/\delta) \log(nn_0) \). For the sake of simplicity, we can also round up the fractional constants in the lower bound on \( L \). We can eliminate \( d \) from these sufficient conditions by substituting the lower bound into the conditions on \( n \) and \( L \), and this also implies that our conditions on certificate existence in the theorem statement suffice for the certificate existence hypothesis for Lemma A.2.6. Thus, we have the remaining sufficient conditions

\[ n \geq K L^{99} \log^9 (1/\delta) \log^9 (nn_0) \log^9 L \]

\[ L \geq K' \max \left\{ \frac{C_\rho^{11+8q_{\text{cert}}/3} (1 + \rho_{\max})^{6} \log^9 (1/\delta) \log^9 (nn_0) \log^{15} L}{\left( \min \{\mu_\infty(M_+), \mu_\infty(M_-)\} \right)^{11/2}}, \kappa^2 C_\lambda \right\}. \]

Using Lemma A.2.14 and choosing \( L \) larger than a sufficiently large absolute constant and larger than \( \log(1/\delta) \), we obtain that it suffices to enforce for \( n \)

\[ n \geq K L^{99} \log^9 (1/\delta) \log^{18} (Ln_0). \]

In the hypotheses of the theorem, we have chosen the equality \( n = K L^{99} \log^9 (1/\delta) \log^{18} (Ln_0) \) in the last bound. This implies \( \log(nn_0) \leq C \log(Ln_0) \), so it suffices to enforce the \( L \) lower bound

\[ L \geq K' \max \left\{ \frac{C_\rho^{11+8q_{\text{cert}}/3} (1 + \rho_{\max})^{6} \log^9 (1/\delta) \log^{24} (Ln_0)}{\left( \min \{\mu_\infty(M_+), \mu_\infty(M_-)\} \right)^{11/2}}, \kappa^2 C_\lambda \right\}. \]
Defining, as in the theorem
\[
C_{\mu} = \frac{C_1^{11 + 8q_{\text{cert}}}}{(\min \{\mu^\infty(M_+), \mu^\infty(M_-)\})^{11/2}},
\]
and using \(C_{\mu} \geq 1\), we can worsen the absolute constant \(K'\) in order to apply Lemma A.2.14 once again, obtaining the simplified condition
\[
L \geq CK' \max\{C_{\mu} \log^9(1/\delta) \log^{24} \left(C_{\mu} n_0 \log(1/\delta)\right), \kappa^2 C_1\}.
\]
These conditions reflect what is stated in the lemma.

Theorem A.2.2. Let \(M\) be a \(d_0\)-dimensional Riemannian submanifold of \(S^{n_0-1}\). For any \(d \geq Kd_0 \log(nn_0 C_M)\), if \(n \geq K'd^4 L\) then one has on an event of probability at least \(1 - e^{-cd}\)
\[
\sup_{(x, x') \in M \times M} \left| \Theta(x, x') - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(v) \right) \prod_{\ell'=\ell}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(v)}{\pi} \right) \right| \leq \sqrt{d^4 n L^3},
\]
where we write \(v = \angle(x, x')\) in context with an abuse of notation, \(c, K, K' > 0\) are absolute constants, and \(C_M > 0\) depends only on the number of connected components of \(M\) and their diameters and curvatures (Lemma A.3.1).

Proof. We have by the definition of \(\Theta\)
\[
\Theta(x, x') = \langle \alpha^{L}(x), \alpha^{L}(x') \rangle + \sum_{\ell=0}^{L-1} \langle \alpha^{\ell}(x), \alpha^{\ell}(x') \rangle \langle \beta^{\ell}(x), \beta^{\ell}(x') \rangle.
\] (A.7)
Under the stated hypotheses, Lemmas A.4.10 and A.4.13 give uniform control of each of the terms appearing in this expression with suitable probability to tolerate \(2L + 1\) union bounds, which gives simultaneous uniform control of the factors on an event \(\mathcal{E}\) with probability at least \(1 - e^{-cd}\). Starting
from (A.7), we can write with the triangle inequality
\[
\Theta(x, x') - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(y) \right) \prod_{l' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(l')}(y)}{\pi} \right) \leq \left| \langle \alpha^{L}(x), \alpha^{L}(x') \rangle \right|
\]
\[
+ \sum_{\ell=0}^{L-1} \left| \langle \alpha^{\ell}(x), \alpha^{\ell}(x') \rangle \right| \langle \beta^{\ell}(x), \beta^{\ell}(x') \rangle - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(y) \right) \prod_{l' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(l')}(y)}{\pi} \right) \right|.
\] (A.8)

By the triangle inequality, we have
\[
\left| \langle \alpha^{\ell}(x), \alpha^{\ell}(x') \rangle \right| \langle \beta^{\ell}(x), \beta^{\ell}(x') \rangle - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(y) \right) \prod_{l' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(l')}(y)}{\pi} \right) \right|
\]
\[
\leq \left| \langle \alpha^{\ell}(x), \alpha^{\ell}(x') \rangle \right| \langle \beta^{\ell}(x), \beta^{\ell}(x') \rangle - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(y) \right) \prod_{l' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(l')}(y)}{\pi} \right) \right|
\]
\[
+ \frac{n}{2} \sum_{\ell=0}^{L-1} \left( 1 - \frac{\varphi^{(\ell)}(y)}{\pi} \right) \left| \langle \alpha^{\ell}(x), \alpha^{\ell}(x') \rangle - \cos \left( \varphi^{(\ell)}(y) \right) \right|.
\]

Under the conditions on \( n, L, \) and \( d \), we have on the event \( \mathcal{E} \) that for each \( \ell \)
\[
\sup_{(x, x') \in \mathcal{M}} \left| \langle \alpha^{\ell}(x), \alpha^{\ell}(x') \rangle \right| \leq 2,
\]
so we can conclude that on \( \mathcal{E} \)
\[
\left| \langle \alpha^{\ell}(x), \alpha^{\ell}(x') \rangle \right| \langle \beta^{\ell}(x), \beta^{\ell}(x') \rangle - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(y) \right) \prod_{l' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(l')}(y)}{\pi} \right) \right| \leq 3\sqrt{d}^4 n L.
\]
The conditions on \( n, d, \) and \( L \) imply that this residual is larger than that incurred by the level-\( L \) features, which is no larger than 2. Returning to (A.8), we have shown that on \( \mathcal{E} \)
\[
\Theta(x, x') - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \left( \varphi^{(\ell)}(y) \right) \prod_{l' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(l')}(y)}{\pi} \right) \right| \leq C\sqrt{d}^4 n L^3.
\]
After adjusting the other absolute constants to absorb \( C \) into \( d \), this gives the claim. \( \square \)

**Theorem A.2.3** (Pointwise Version of Theorem A.2.2). Let \( \mathcal{M} \) be a \( d_0 \)-dimensional Riemannian
submanifold of $\mathbb{S}^{n_0-1}$. For any $d \geq K \log n$, if $n \geq K' \max\{1, d^4L\}$ then one has for any $(x, x') \in \mathcal{M} \times \mathcal{M}$

$$\mathbb{P}\left[ \left\| \Theta(x, x') - \frac{n}{2} \sum_{\ell=0}^{L-1} \cos(\varphi^{(\ell)}(v)) \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(v)}{\pi} \right) \right\| \leq \sqrt{d^4nL^3} \right] \geq 1 - e^{-cd},$$

where we write $v = \angle(x, x')$ in context with an abuse of notation, and $c, K, K' > 0$ are absolute constants.

**Proof.** Follow the proof of Theorem A.2.2, but invoke the pointwise versions of the uniform concentration results used there (i.e., Lemmas A.4.1 and A.4.4) after rescaling $d$ to relocate the $\log n$ terms. \hfill \Box

**Theorem A.2.4.** There exist absolute constants $c, C, K, K' > 0$ such that for any $d \geq K_n \log n$, if $n \geq K'd^4L$, then on an event of probability at least $1 - e^{-cd}$ the natural extension of $f_{\theta_0}$ to $\mathbb{R}^{n_0}$ is $3\sqrt{d}$-Lipschitz.

**Proof.** The proof is a simple application of Lemma A.2.16, which (because $f_{\theta_0}$ is 1-nonnegatively homogeneous and so are all its intermediate feature maps $\alpha^{\ell}_{\theta_0}(x)$) implies that it suffices to control the Lipschitz constants of the maps and bound them on the unit sphere, together with Lemmas A.4.11 and A.4.12. In particular, for any $d \geq K_n \log(n)$ and any $n \geq K'd^4L$, we have that there exists an event of probability at least $1 - e^{-cd}$ on which

$$\|f_{\theta}\|_{L^\infty(S^{n_0-1})} \leq \sqrt{d},$$

and

$$\|f_{\theta}|_{S^{n_0-1}}\|_{\text{Lip}} \leq \sqrt{d}.$$  

Applying Lemma A.2.16, it follows that $f_{\theta_0} : \mathbb{R}^{n_0} \to \mathbb{R}$ is $3\sqrt{d}$-Lipschitz on an event of probability at least $1 - e^{-cd}$. \hfill \Box

114
A.2.2 Supporting Results on Dynamics

Lemma A.2.5 (Nominal). Suppose $C_{\text{err}}, C_{\text{cert}}, q_{\text{cert}} > 0$ are absolute constants. Then there exist absolute constants $c, c', C', C'', C''' > 0$ and absolute constants $K, K', K'' > 0$ such that for any $d \geq Kd_0 \log(n \alpha_0 C_M)$ and any $1/2 \leq q \leq 1$, if $n \geq K'd^4 L^5$, if $L \geq K''d C_{\rho}^{2q_{\text{cert}}}$, and if additionally there exists $g \in L^2_{\mu^\infty}(M)$ satisfying

$$\|\Theta[g] - \zeta\|_{L^2_{\mu^\infty}(M)} \leq C_{\text{err}} C_{\rho}^{q_{\text{cert}}} \frac{\sqrt{d}}{L}; \quad \|g\|_{L^2_{\mu^\infty}(M)} \leq C_{\text{cert}} \rho_{\min}^{-q_{\text{cert}}} \frac{\sqrt{d}}{n}$$

and $\tau > 0$ is chosen such that

$$\tau \leq \frac{c'}{nL},$$

then one has

$$\mathbb{P}\left[ \bigcap_{0 \leq k \leq L^q/(n\tau)} \left\{ \|\zeta_k\|_{L^2_{\mu^\infty}(M)} \leq \sqrt{d} \right\} \right] \geq 1 - e^{-cd},$$

and in addition

$$\mathbb{P}\left[ \bigcap_{C'\sqrt{d}/(n\tau \rho_{\min}^{-q_{\text{cert}}}) \leq k \leq L^q/(n\tau)} \left\{ \|\epsilon_k\|_{L^2_{\mu^\infty}(M)} \leq \frac{C'' C_{\rho}^{q_{\text{cert}}} \sqrt{d} \log L}{nk\tau} \right\} \right] \geq 1 - e^{-cd}.$$

Moreover, one has

$$\mathbb{P}\left[ \bigcap_{0 \leq k \leq L^q/(n\tau)} \left\{ \sum_{s=0}^{k} \|\epsilon_s\|_{L^2_{\mu^\infty}(M)} \leq C_{\rho}^{2q_{\text{cert}}} C'''' d \log^2 L \right\} \right] \geq 1 - e^{-cd}.$$

The constant $C_{\rho} = \max\{\rho_{\min}, \rho_{\min}^{-1}\}$.

Proof. We will combine Lemma A.2.11 with various probabilistic results to obtain a simple final form for the bound from this result.
Invoking Lemma A.2.11, we can assert that for any step size \( \tau > 0 \) satisfying
\[
\tau < \frac{1}{\| \Theta \|_{L^\infty_{\mu_\infty}(M) \to L^2_{\mu_\infty}(M)}}, \tag{A.9}
\]
and for any \( k \) satisfying
\[
k \tau \geq \sqrt{\frac{3e \| g \|_{L^2_{\mu_\infty}(M)}}{2 \| \zeta \|_{L^\infty(M)}}}, \tag{A.10}
\]
the population dynamics satisfy
\[
\| \xi_k \|_{L^2_{\mu_\infty}(M)} \leq \sqrt{3} \| \Theta [g] - \zeta \|_{L^2_{\mu_\infty}(M)} - \frac{3 \| g \|_{L^2_{\mu_\infty}(M)}}{k \tau} \log \left( \sqrt{\frac{3}{2 \| \zeta \|_{L^\infty(M)} k \tau}} \right). \tag{A.11}
\]

We state the bounds we will apply to simplify this expression. An application of Lemma A.4.11 gives
\[
\mathbb{P} \left[ \| \hat{\zeta} - \zeta \|_{L^\infty(M)} \leq \frac{\sqrt{2d}}{L} \right] \geq 1 - e^{-cd} \tag{A.12}
\]
and
\[
\mathbb{P} \left[ \| \zeta \|_{L^\infty(M)} \leq \sqrt{d} \right] \geq 1 - e^{-cd} \tag{A.13}
\]
as long as \( n \geq K d^4 L^5 \) and \( d \geq K^* d_0 \log(n n_0 C_M) \), where we use these conditions to simplify the residual that appears in the version of (A.12) quoted in Lemma A.4.11. In particular, combining (A.12) and (A.13) with the triangle inequality and a union bound and then rescaling \( d \), which worsens the constant \( c \) and the absolute constants in the preceding conditions, gives
\[
\mathbb{P} \left[ \| \hat{\zeta} \|_{L^\infty(M)} \leq \sqrt{d} \right] \geq 1 - 2e^{-cd}. \tag{A.14}
\]

In addition, we can write using the triangle inequality
\[
\| \zeta \|_{L^\infty(M)} \geq \| \hat{\zeta} \|_{L^\infty(M)} - \| \zeta - \hat{\zeta} \|_{L^\infty(M)},
\]
and

\[
\left\| \zeta \right\|_{L^\infty(M)} = \sup_{x \in M} \left| f_*(x) - \int_M f_{\theta_0}(x') \, d\mu^\infty(x') \right|
\]

\[
= \max \left\{ \left| \int_M f_{\theta_0}(x') \, d\mu^\infty(x') - 1 \right|, \left| \int_M f_{\theta_0}(x') \, d\mu^\infty(x') + 1 \right| \right\}
\]

\[
\geq 1,
\]

so that, by (A.12), we have if \( L \geq 2 \sqrt{d} \)

\[
\mathbb{P} \left[ \left\| \zeta \right\|_{L^\infty(M)} \geq \frac{1}{2} \right] \geq 1 - e^{-cd}.
\] (A.15)

Because \( \mu^\infty \) is a probability measure, Jensen’s inequality, the Schwarz inequality, and the triangle inequality give

\[
\left\| \Theta \right\|_{L^2_{\mu^\infty}(M) \rightarrow L^2_{\mu^\infty}(M)} \leq \sup_{(x,x') \in M \times M} |\Theta(x,x')|
\]

\[
\leq \sup_{(x,x') \in M \times M} |\Theta(x,x') - \psi_1 \circ \angle(x,x')|
\]

\[
+ \sup_{(x,x') \in M \times M} |\psi_1 \circ \angle(x,x')|,
\]

and an application of Theorem A.2.2 and Lemma A.5.5 then gives that on an event of probability at least \( 1 - e^{-cd} \)

\[
\left\| \Theta \right\|_{L^2_{\mu^\infty}(M) \rightarrow L^2_{\mu^\infty}(M)} \leq CnL
\] (A.16)

provided \( d \geq Kd_0 \log(nn_0C_M) \) and \( n \geq K'd^4L \). We will write \( E \) for the event consisting of the union of the events invoked for the bounds (A.12), (A.13), (A.14), (A.15) and (A.16), which has probability at least \( 1 - e^{-cd} \) by a union bound and a choice of \( d \geq K \). We will conclude by simplifying (A.11) on \( E \). First, we note that by (A.16), the step size condition (A.9) is satisfied on \( E \) provided

\[
\tau \leq \frac{c}{nL},
\] (A.17)
which holds under our hypotheses. Next, on $E$, we write using decreasingness of $x \mapsto -\log x$ and (A.13)

$$-\frac{3\|g\|_{L^2_{\mu_\infty}(M)}}{k\tau} \log \left( \sqrt{\frac{3}{2}} \frac{\|g\|_{L^2_{\mu_\infty}(M)}}{\|\zeta\|_{L^\infty(M)}k\tau} \right) \leq -\frac{3\|g\|_{L^2_{\mu_\infty}(M)}}{k\tau} \log \left( \sqrt{\frac{3}{2}} \frac{\|g\|_{L^2_{\mu_\infty}(M)}}{k\tau \sqrt{d}} \right)$$

$$= -\sqrt{6d} \frac{\|g\|_{L^2_{\mu_\infty}(M)}}{\sqrt{2k\tau \sqrt{d}}} \log \left( \sqrt{\frac{3}{2}} \frac{\|g\|_{L^2_{\mu_\infty}(M)}}{k\tau \sqrt{d}} \right). \quad (A.18)$$

By the hypothesis on $g$, we have on $E$

$$\|g\|_{L^2_{\mu_\infty}(M)} \leq C \sqrt{d} \frac{\rho_{\min}^{-q_{\text{cert}}}}{n}, \quad (A.19)$$

and so it follows that on $E$

$$\sqrt{\frac{3}{2}} \frac{\|g\|_{L^2_{\mu_\infty}(M)}}{k\tau \sqrt{d}} \leq \frac{C}{nk\tau \rho_{\min}^{q_{\text{cert}}}}.$$

The function $x \mapsto -x \log x$ is a strictly increasing function on $[0, e^{-1}]$, so when $k$ is chosen such that

$$\frac{Ce}{n\tau \rho_{\min}^{q_{\text{cert}}}} \leq k, \quad (A.20)$$

we have on $E$ by (A.18)

$$-\frac{3\|g\|_{L^2_{\mu_\infty}(M)}}{k\tau} \log \left( \sqrt{\frac{3}{2}} \frac{\|g\|_{L^2_{\mu_\infty}(M)}}{\|\zeta\|_{L^\infty(M)}k\tau} \right) \leq \frac{C \sqrt{6d}}{nk\tau \rho_{\min}^{q_{\text{cert}}}} \log \left( C^{-1} nk\tau \rho_{\min}^{q_{\text{cert}}} \right). \quad (A.21)$$

Additionally, in the context of the condition (A.10), notice that by (A.15) and (A.19), on $E$ we have

$$\sqrt{\frac{3e}{2}} \frac{\|g\|_{L^2_{\mu_\infty}(M)}}{\tau \|\zeta\|_{L^\infty(M)}} \leq \frac{Ce \sqrt{d}}{n\tau \rho_{\min}^{q_{\text{cert}}}} ,$$

so that given $d \geq 1$, we have that the choice

$$k \geq \frac{Ce \sqrt{d}}{n\tau \rho_{\min}^{q_{\text{cert}}}} \quad (A.22)$$
implies both conditions (A.10) and (A.20). We can simplify (A.21) using the hypothesis $k\tau \leq L^q/n$ with $1/2 \leq q \leq 1$: we get

$$\frac{nk\tau \rho_{\text{min}}^{q_{\text{cert}}}}{C} \leq \frac{L^q \rho_{\text{min}}^{q_{\text{cert}}}}{C} \leq L^{1+q},$$

where the last inequality requires $L \geq \rho_{\text{min}}^{q_{\text{cert}}}/C$, which implies

$$\frac{3\|g\|_{L^2_{\rho_{\text{min}}}(M)}}{k\tau} \log \left( \sqrt{\frac{3}{2}} \frac{\|g\|_{L^2_{\rho_{\text{min}}}(M)}}{\|\xi\|_{L^\infty(M)} k\tau} \right) \leq \frac{C' \sqrt{d} \log L}{nk\tau \rho_{\text{min}}^{q_{\text{cert}}}}. \quad (A.23)$$

The conditions we need to satisfy on $k\tau$ can be stated together as

$$\frac{Ce \sqrt{d}}{np_{\text{min}}^{q_{\text{cert}}}} \leq k\tau \leq L^q/n,$$

and it is possible to satisfy these conditions simultaneously as long as

$$L \geq \left( \frac{Ce \sqrt{d}}{\rho_{\text{min}}^{q_{\text{cert}}}} \right)^{1/q}.$$

We obtain an upper bound $\frac{C e^2 d}{L^{1/2}}$ for the quantity on the RHS of this inequality from $q \geq 1/2$; it suffices to choose $L$ larger than this upper bound instead. The other simplifications are easier: using the assumption on the norm of $\Theta[g] - \zeta$, we have

$$\|\Theta[g] - \zeta\|_{L^2_{\rho_{\text{min}}}(\mathcal{M})} \leq C_{\rho}^{q_{\text{cert}}} \frac{C \sqrt{d}}{L \rho_{\text{min}}^{1/2}}.$$

Worst-casing terms using our hypotheses on $d$ and $L$ to obtain a simplified bound, on $E$, we have thus shown that when (A.22) is satisfied, we have

$$\|\phi_k\|_{L^2_{\rho_{\text{min}}}(\mathcal{M})} \leq CC_{\rho}^{q_{\text{cert}}} \sqrt{d} \left( \frac{1}{L} \right. + \frac{\log L}{nk\tau} \left. \right).$$
We have
\[
\frac{1}{L} \leq \frac{\log L}{nk\tau} \iff L \log L \geq \frac{k\tau}{n},
\]
which is implied by the hypothesis \(k\tau \leq L^q/n\) as long as \(L \geq e\). So we can simplify to
\[
\|\xi^\infty_k\|_{L^2_{\mu^\infty}(M)} \leq \frac{CC^q_{\text{cert}} \sqrt{d} \log L}{nk\tau}.
\]

We also need a bound that works for \(k\) that do not satisfy (A.22). From the update equation for the dynamics in the proof of Lemma A.2.11 and the choice of \(\tau\), we also have
\[
\|\xi^\infty_k\|_{L^2_{\mu^\infty}(M)} \leq \|\xi\|_{L^2_{\mu^\infty}(M)} \leq \sqrt{d},
\]
where the last bound is valid on \(E\). Finally, we can obtain the claimed sum bound by calculating using our ‘small-\(k\)’ and ‘large-\(k\)’ bounds:
\[
\sum_{s=0}^{k} \|\xi^\infty_s\|_{L^2_{\mu^\infty}(M)} = \sum_{s=0}^{[C\sqrt{d}/(n\tau\rho_{\text{cert}}^{\text{min}})]} \|\xi^\infty_s\|_{L^2_{\mu^\infty}(M)} + \sum_{s=[C\sqrt{d}/(n\tau\rho_{\text{cert}}^{\text{min}})]}^{k} \|\xi^\infty_s\|_{L^2_{\mu^\infty}(M)}
\]
\[
\leq \sqrt{d} \left(1 + \frac{C'\sqrt{d}}{n\tau\rho_{\text{cert}}^{\text{min}}} + \frac{C''C^q_{\text{cert}} \sqrt{d} \log L}{n\tau} \sum_{s=[C\sqrt{d}/(n\tau\rho_{\text{cert}}^{\text{min}})]}^{k} \frac{1}{s}\right)
\]
\[
\leq \frac{C'd}{n\tau\rho_{\text{cert}}^{\text{min}}} + C''C^q_{\text{cert}} \sqrt{d} \log L \left(\frac{n\tau\rho_{\text{cert}}^{\text{min}}}{C\sqrt{d}} + \int_{C\sqrt{d}/(n\tau\rho_{\text{cert}}^{\text{min}})}^{k} \frac{dx}{x}\right)
\]
\[
\leq \frac{Cd}{n\tau\rho_{\text{cert}}^{\text{min}}} + C' \max\left\{\rho_{\text{cert}}^{2q_{\text{min}}}, 1\right\} \log L + \frac{C''C^q_{\text{cert}} \sqrt{d} \log^2 L}{n\tau},
\]
where the second inequality uses standard estimates for the harmonic numbers and the fact that \(C'\sqrt{d}/(n\tau\rho_{\text{cert}}^{\text{min}}) \geq 1\), which follows from \(\tau \leq c'(nL), d \geq 1\) and \(L \geq K\rho_{\text{min}}^{q_{\text{cert}}}\) for a suitable absolute constant \(K\); and the third inequality integrates and simplifies, using \(k\tau \leq L/n\) and again \(d \geq 1\) and \(L \geq C\rho_{\text{min}}^{q_{\text{cert}}}\). Worst-casing constants and using \(n\tau \leq 1\), we simplify this last bound to
\[
\sum_{s=0}^{k} \|\xi^\infty_s\|_{L^2_{\mu^\infty}(M)} \leq \max\left\{\rho_{\text{cert}}^{2q_{\text{min}}}, \frac{1}{\rho_{\text{min}}^{2q_{\text{cert}}}}\right\} \frac{Cd \log^2 L}{n\tau}.
\]
To see that the conditions on \( L \) in the statement of the result suffice, note that we have to satisfy (say) \( L \geq K \rho_{\text{min}}^{q_{\text{cert}}} \) and \( L \geq K' \rho_{\text{min}}^{-q_{\text{cert}}} \); the first of these lower bounds is tighter when \( \rho_{\text{min}} \geq 1 \), and the second when \( \rho_{\text{min}} < 1 \), and so it suffices to require \( L \geq K \rho_{\text{min}}^{2q_{\text{cert}}} \) and \( L \geq K' \rho_{\text{min}}^{-2q_{\text{cert}}} \) instead. \( \square \)

**Lemma A.2.6** (Nominal to Finite). Let \( d_0 = 1 \), and suppose \( C_{\text{err}}, C_{\text{cert}}, q_{\text{cert}} > 0 \) are absolute constants. Then there exist absolute constants \( c, c', C', C'' > 0 \) such that for any \( \rho_{\text{min}} \geq \frac{1}{2} \) and \( \rho < 1 \), if \( L \geq K \max \{ C \rho_{\text{min}}^{2q_{\text{cert}}} d, \kappa^2 \} \), if

\[
n \geq K'' \max \left\{ e \frac{252}{\delta} L^{60 + 44q} d^9 \log^9 L, \kappa^{2/5} \left( \frac{k}{\ell} \right)^{1/3} \right\},
\]

and if

\[
N^{1/(2+\delta)} \geq K'' C_0^{7/2 + 8q_{\text{cert}}/3} (1 + \rho_{\text{max}})^{7/6} e^{119/(3\delta)} \min \{ \mu^\infty (\mathcal{M}^+)^{1/2}, \mu^\infty (\mathcal{M}^-)^{1/2} \} d^{5/4} L^{5/2 + 2q} \log L,
\]

and if additionally there exists \( g \in L_2^\infty (\mathcal{M}) \) satisfying

\[
\| \Theta [g] - \zeta \|_{L_2^\infty (\mathcal{M})} \leq C_{\text{err}} C_0 \rho_{\text{min}} \sqrt{d} L, \quad \| g \|_{L_2^\infty (\mathcal{M})} \leq C_{\text{cert}} \rho_{\text{min}}^{-q_{\text{cert}}} \sqrt{d} \frac{n}{n}
\]

and \( \tau > 0 \) is chosen such that

\[
\tau \leq \frac{c'}{n L},
\]

then one has generalization in \( L_2^\infty (\mathcal{M}) \):

\[
P \left[ \| \mathbf{Z}[L^{d/(n \tau)}] \|_{L_2^\infty (\mathcal{M})} \leq \frac{C' C_0^{q_{\text{cert}}} \sqrt{d} \log L}{L^q} \right] \geq 1 - \frac{C'' L e^{-c d}}{n \tau},
\]

and in addition, one has generalization in \( L^\infty (\mathcal{M}) \):

\[
P \left[ \| \mathbf{Z}[L^{d/(n \tau)}] \|_{L^\infty (\mathcal{M})} \leq \frac{C' C_0^{1+2q_{\text{cert}}/3} (1 + \rho_{\text{max}})^{1/2} e^{14/(3\delta)} d^{3/4} \log^{4/3} L}{\min \{ \mu^\infty (\mathcal{M}^+), \mu^\infty (\mathcal{M}^-) \}^{1/2} L^{(4q-3)/6}} \right] \geq 1 - \frac{C'' L e^{-c d}}{n \tau}.
\]

The constant \( C_0 = \max \{ \rho_{\text{min}}, \rho_{\text{min}}^{-1} \} \).
Proof. The proof controls the $L^\infty$ norm of the error evaluated along the finite sample dynamics using an interpolation inequality for Lipschitz functions on an interval (Lemma A.2.13), which relates the $L^\infty$ norm to a certain combination of the predictor’s Lipschitz constant and its $L^2_{\mu^\infty}$ norm. We can control these two quantities at time zero using our measure concentration results; to control them for larger times $0 < k \leq L^q/\left(n\tau\right)$, we set up a system of coupled ‘discrete integral equations’ for the generalization error of the finite sample predictor and the Lipschitz constant of the finite sample predictor, and use the fact that $k\tau$ is not large to argue by induction that not much blow-up can occur. Along the way, we control the generalization error of the finite sample predictor by linking it to the generalization error of the nominal predictor as controlled in Lemma A.2.5; the residual that arises is shown to be small by applying Corollary A.2.10 and applying basic results from optimal transport theory adapted to our setting, encapsulated in Lemmas A.2.12 and A.2.15.

To begin, we will lay out the probabilistic bounds we will rely on for simplifications, so that the rest of the proof can proceed without interruption. We will want to satisfy

$$\tau < \frac{1}{\max \left\{ \left\| \Theta^{\mu^N} \right\|_{L^2_{\mu^N}(M) \to L^2_{\mu^N}(M)^*}, \left\| \Theta^{\mu^\infty} \right\|_{L^2_{\mu^\infty}(M) \to L^2_{\mu^\infty}(M)} \right\}},$$

(A.24)

following the notation of Lemma A.2.9. Using Jensen’s inequality, the Schwarz inequality, and the triangle inequality, we have for $\star \in \{N, \infty\}$

$$\left\| \Theta^{\mu^\star} \right\|_{L^2_{\mu^\star}(M) \to L^2_{\mu^\star}(M)} = \sup_{\left\| g \right\|_{L^2_{\mu^\star}(M)} \leq 1} \left\| \int_M \Theta(x, x') g(x') \, d\mu^\star(x') \right\|_{L^2_{\mu^\star}(M)}$$

$$\leq \left\| g \right\|_{L^1_{\mu^\star}(M)} \sup_{(x, x') \in M \times M} |\Theta(x, x')|$$

$$\leq \sup_{(x, x') \in M \times M} |\Theta(x, x') - \psi_1 \circ \angle(x, x')|$$

$$+ \sup_{(x, x') \in M \times M} |\psi_1 \circ \angle(x, x')|,$$

(A.25)

where the notation $\psi_1$ follows the definition in Section A.3.1. The first term in (A.25) can be
controlled using Theorem A.2.2: we obtain that on an event of probability at least $1 - e^{-cd}$

$$
\|\Theta - \psi_1 \circ L\|_{L^{\infty}(M, M)} \leq \sqrt{d^4 nL^3}
$$

(A.26)

if $d \geq Kd_0 \log(nM_C)$ and $n \geq K'd^4L$. The second term in (A.25) can be controlled using the triangle inequality, Lemma A.5.5, and the definition of $\psi_1$: we obtain that it is no larger than $nL/2$.

Combining these two bounds, we have on an event of probability at least $1 - e^{-cd}$

$$
\max \left\{ \|\Theta_{\mu}^\infty\|_{L^2_{\mu^N}(M) \to L^2_{\mu^N}(M)}, \|\Theta_{\mu}^\infty\|_{L^2_{\mu^N}(M) \to L^2_{\mu^N}(M)} \right\} \leq CnL
$$

(A.27)

provided $d \geq Kd_0 \log(nM_C)$ and $n \geq K'd^4L$. Thus, with probability at least $1 - e^{-cd}$, our choice of step size $\tau \leq c/(nL)$ satisfies (A.24). Under our hypotheses on the function $g$ in the statement of the result and taking a union bound with the event in (A.27), we can invoke Lemma A.2.5 to obtain

$$
P \left[ \bigcap_{C\sqrt{d}/(n\rho_{\min}^{\rho_{\text{cert}}}) \leq k \leq L^q/(n\tau)} \left\{ \|\xi_{k}^{\infty}\|_{L^2_{\mu^N}(M)} \leq \frac{C'C_{\rho}^{\rho_{\text{cert}}} \sqrt{d \log L}}{nk\tau} \right\} \right] \geq 1 - \frac{C''L e^{-cd}}{n\tau} 
$$

(A.28)

and

$$
P \left[ \bigcap_{0 \leq k \leq L^q/(n\tau)} \left\{ \sum_{s=0}^{k} \|\xi_{s}^{\infty}\|_{L^2_{\mu^N}(M)} \leq C_{\rho}^{2\rho_{\text{cert}}} C'' d^2 \log^2 L / n\tau \right\} \right] \geq 1 - \frac{C''L e^{-cd}}{n\tau} 
$$

(A.29)

provided $d \geq Kd_0 \log(nM_C)$, $1/2 \leq q < 1$, $n \geq K'd^4L^5$, and $L \geq K' C_{\rho}^{2\rho_{\text{cert}}} d$. We have by Lemmas A.2.5 and A.2.9, a union bound with (A.27), and our condition on $\tau$ that

$$
P \left[ \bigcap_{0 \leq k \leq L^q/(n\tau)} \left\{ \|\xi_{k}^{\infty}\|_{L^2_{\mu^N}(M)} \leq \sqrt{d} \right\} \cap \bigcap_{0 \leq k \leq L^q/(n\tau)} \left\{ \|\xi_{k}^{\infty}\|_{L^2_{\mu^N}(M)} \leq \sqrt{d} \right\} \right] \geq 1 - \frac{CLe^{-cd}}{n\tau} 
$$

(A.30)

as long as $d \geq Kd_0 \log(nM_C)$ and $n \geq K'L^{48+20q}d^9 \log^9 L$, and where we used our conditions on $\tau$ and $q$ to obtain that $L^q/n\tau \geq 1$ and simplify the probability bound; and, following the notation of Corollary A.2.10, we have by this result (again under our condition on $\tau$ and a union bound)
that there is an event of probability at least $1 - CLe^{-cd}/(n\tau)$ on which

$$\Delta^N_{[L^q/(n\tau)]-1} \leq \left(n^{11}L^{4g+8g}d^9\log^9 L\right)^{1/12} \tag{A.31}$$

under the previous conditions on $n$ and $d$. In addition, applying Lemma A.4.12 and a union bound gives that on an event of probability at least $1 - Ce^{-cd}$

$$\max\left\{\|\xi|_{M_t}\|_{\text{Lip}}, \|\xi|_{M_N}\|_{\text{Lip}}\right\} \leq \sqrt{d} \tag{A.32}$$

provided $d \geq Kd_0\log(nn_0C_M)$ and $n \geq K'\max\{d^4L, (\kappa/c_\lambda)^{1/3}, \kappa^{2/5}\}$. Finally, we have by Lemma A.2.12 that for any $0 < \delta \leq 1$

$$\mathbb{P}\left[ \bigcap_{f \in \text{Lip}(M)} \left\{ \frac{\left| \int_M f(x) \, d\mu^\infty(x) - \int_M f(x) \, d\mu^N(x) \right|}{\|f\|_{\text{Lip}(M)}\sqrt{d}} \leq 2\|f\|_{\text{Lip}(M)} + \frac{e^{14/\delta}C_{\mu^\infty,M_N}\max_{e \in \mathbb{R}^+} \|f\|_{\text{Lip}(M)} \sqrt{d}}{N^{1/12+\delta}} \right\} \right] \geq 1 - 8e^{-d}, \tag{A.33}$$

as long as $d \geq 1$ and $N \geq 2\sqrt{d}/\min\{\mu^\infty(M_+), \mu^\infty(M_-)\}$. We let $E(q, \delta)$ denote the event consisting of the union of the events appearing in the bounds (A.26), (A.27), (A.28), (A.29), (A.30), (A.31), (A.32) and (A.33) hold; by a union bound and the previous observation that $L^q/n\tau \geq 1$, we have

$$\mathbb{P}[E] \geq 1 - \frac{C'Le^{-cd}}{n\tau}. \tag{A.34}$$

In the sequel, we will use the events defining $E$ to simplify our residuals without explicitly referencing that our bounds hold only on $E$ to save time.

We start from the dynamics update equations given by Lemma A.2.7, which we use to write

$$\zeta^\infty_k - \zeta^N_k = (\text{Id} - \tau\Theta) \left[ \zeta^\infty_{k-1} - \zeta^N_{k-1} \right] + \tau\Theta^N_K \left[ \zeta^N_{k-1} - \tau\Theta \left[ \zeta^N_{k-1} \right] \right],$$

where $\Theta$ is defined as in Lemma A.2.11. Under the choice of $\tau$ and positivity of $\Theta$ (Lemma A.2.8), we apply the triangle inequality and a telescoping series with the common initial conditions to
obtain
\[ \| \xi_k - \xi_k^N \|_{L_\infty^2(\mathcal{M})} \leq \tau \sum_{s=0}^{k-1} \| \Theta_s^N [ \xi_s^N ] - \Theta [ \xi_s ] \|_{L_\infty^2(\mathcal{M})}. \] (A.34)

We can write
\[
\Theta_s^N [ \xi_s^N ] (x) = \int_M \Theta_s^N (x, x') \xi_s^N (x') \, d\mu^N (x') \\
= \int_M (\Theta_s^N (x, x') - \psi_1 \circ \angle (x, x')) \xi_s^N (x') \, d\mu^N (x') \\
+ \int_M \psi_1 \circ \angle (x, x') \xi_s^N (x') \, d\mu^N (x'),
\]
and analogously
\[
\Theta [ \xi_s^N ] (x) = \int_M (\Theta (x, x') - \psi_1 \circ \angle (x, x')) \xi_s^N (x') \, d\mu^\infty (x') \\
+ \int_M \psi_1 \circ \angle (x, x') \xi_s^N (x') \, d\mu^\infty (x').
\]

Using Jensen’s inequality and the Schwarz inequality, we have
\[
\left\| \int_M (\Theta_s^N (x, x') - \psi_1 \circ \angle (x, x')) \xi_s^N (x') \, d\mu^N (x') \right\|_{L_\infty^2(\mathcal{M})} \\
\leq \int_M \| \Theta_s^N (\cdot, x') - \psi_1 \circ \angle (\cdot, x') \|_{L_\infty^2(\mathcal{M})} | \xi_s^N (x') | \, d\mu^N (x') \\
\leq \| \Theta_s^N - \psi_1 \circ \angle \|_{L_\infty (\mathcal{M} \times \mathcal{M})} \| \xi_s^N \|_{L_1^\infty (\mathcal{M})} \\
\leq \| \Theta_s^N - \psi_1 \circ \angle \|_{L_\infty (\mathcal{M} \times \mathcal{M})} \| \xi_s^N \|_{L_2^\infty (\mathcal{M})},
\]

since \( \mu^N \) is a probability measure. Repeating an analogous calculation with \( \mu^\infty \) for the other term.
and applying the triangle inequality, we have

\[ \| \Theta^N \left[ \zeta^N_s \right] - \Theta \left[ \xi^N_s \right] \|_{L^2_{\mu\infty}(M)} \leq \| \Theta - \psi_1 \circ \angle \|_{L^\infty(M \times M)} \left( \| \zeta^\infty_s - \zeta^N_s \|_{L^2_{\mu\infty}(M)} + \| \xi^\infty_s \|_{L^2_{\mu\infty}(M)} + \| \xi^N_s \|_{L^2_{\mu\infty}(M)} \right) \\
\]

+ \| \Theta^N - \Theta \|_{L^\infty(M \times M)} \| \xi^N_s \|_{L^2_{\mu\infty}(M)}

+ \left\| \int_M \psi_1 \circ \angle (\cdot, x') \xi^N_s(x') \left( d\mu^\infty(x') - d\mu^N(x') \right) \right\|_{L^2_{\mu\infty}(M)}. \tag{A.35} \\

We detour briefly to simplify residuals appearing in (A.35) before using the result to update (A.34). Using (A.26) and (A.31), we get

\[ \| \Theta - \psi_1 \circ \angle \|_{L^\infty(M \times M)} \left( \| \zeta^\infty_s - \zeta^N_s \|_{L^2_{\mu\infty}(M)} + \| \xi^\infty_s \|_{L^2_{\mu\infty}(M)} + \| \xi^N_s \|_{L^2_{\mu\infty}(M)} \right) \\
\]

+ \| \Theta^N - \Theta \|_{L^\infty(M \times M)} \| \xi^N_s \|_{L^2_{\mu\infty}(M)}

\[ \leq \sqrt{d^4 n L^3} \left( \| \zeta^\infty_s - \zeta^N_s \|_{L^2_{\mu\infty}(M)} + \| \xi^\infty_s \|_{L^2_{\mu\infty}(M)} + \| \xi^N_s \|_{L^2_{\mu\infty}(M)} \right) \]

+ \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12} \| \xi^N_s \|_{L^2_{\mu\infty}(M)}

\[ \leq \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12} \left( \| \zeta^\infty_s - \zeta^N_s \|_{L^2_{\mu\infty}(M)} + \| \xi^\infty_s \|_{L^2_{\mu\infty}(M)} + 2 \| \xi^N_s \|_{L^2_{\mu\infty}(M)} \right). \tag{A.36} \]

where the final bound holds when \( n \geq d^3 \). Using (A.30), we can further simplify the RHS of the last bound above to

\[ \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12} \left( \| \zeta^\infty_s - \zeta^N_s \|_{L^2_{\mu\infty}(M)} + \| \xi^\infty_s \|_{L^2_{\mu\infty}(M)} + 2 \| \xi^N_s \|_{L^2_{\mu\infty}(M)} \right) \]

\[ \leq 2 \left( n^{11} L^{48+8q} d^{15} \log^9 L \right)^{1/12} + \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12} \| \zeta^\infty_s - \zeta^N_s \|_{L^2_{\mu\infty}(M)}. \]
With this last bound and (A.35), we can use $k \tau \leq L^q / n$ to simplify (A.34) to

$$
\| \zeta^N_k - \zeta^N \|_{L^2_{\mu^N}(M)} \leq C \left( \frac{L^{48+20} d^{15} \log^9 L}{n} \right)^{1/12}
+ \tau \left( \sum_{s=0}^{k-1} \| \zeta^N_s \|_{L^2_{\mu^N}(M)} \right)^{1/12}
+ \tau \sum_{s=0}^{k-1} \left\| \int_M \psi_1 \circ \angle (\cdot, x') \zeta^N_s (x') \left( d\mu^\infty (x') - d\mu^N (x') \right) \right\|_{L^2_{\mu^N}(M)}.
$$

(A.37)

To control the remaining term in (A.37), we split the error $\zeta^N_s$ into a Lipschitz component whose evolution is governed by the nominal kernel $\psi_1 \circ \angle$ and a nonsmooth component which is small in $L^\infty$. Formally, we define $\Theta^{\text{nom}} : L^2_{\mu_N} (M) \to L^2_{\mu_N^N} (M)$ by

$$
\Theta^{\text{nom}} [g] (x) = \int_M \psi_1 \circ \angle (x, x') g(x') d\mu^N (x'),
$$

and use the update equation from Lemma A.2.7 to write

$$
\zeta^N_s = \zeta - \tau \sum_{i=0}^{s-1} \Theta^{\text{nom}} [\zeta^N_i]
= \zeta - \tau \sum_{i=0}^{s-1} \Theta^{\text{nom}} [\zeta^N_i] + \tau \sum_{i=0}^{s-1} \left( \Theta^{\text{nom}} - \Theta^{\text{nom}}_i \right) [\zeta^N_i],
$$

so that $\zeta^N_s = \zeta, \delta^N_s = \zeta, \delta^N_0 = 0$. It is straightforward to control $\delta^N_s$ in $L^\infty$: we have (as usual) by the triangle inequality, Jensen’s inequality, and the Schwarz inequality

$$
\| \delta^N_s \|_{L^\infty (M)} \leq \tau \sum_{i=0}^{s-1} \int_M \| \psi_1 \circ \angle (\cdot, x') - \Theta^N_i (\cdot, x') \|_{L^\infty (M)} \| \zeta^N_i (x') \|_{L^2_{\mu_N^N} (M)} d\mu^N (x')
\leq \tau \sum_{i=0}^{s-1} \| \psi_1 \circ \angle - \Theta^N_i \|_{L^\infty (M \times M)} \| \zeta^N_i \|_{L^2_{\mu_N^N} (M)},
$$

\[127\]
and then the triangle inequality together with (A.26), (A.30) and (A.31) yield
\[
\|\delta_s^N\|_{L^\infty(M)} \leq s\tau \sqrt{d} \left( \sqrt{d^4 n L^2} + \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12} \right)
\]
\[
\leq s\tau \sqrt{d} \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12},
\]
(A.38)
where the second line applies the same simplifications that led us to (A.36). The triangle inequality gives
\[
\left\| \int_M \psi_1 \circ \angle(\cdot, x') \delta_s^N(x') \left( d\mu^\infty(x') - d\mu^N(x') \right) \right\|_{L^2_{\mu^\infty}(M)}
\leq \sum_{s \in \{N, \infty\}} \left\| \int_M \psi_1 \circ \angle(\cdot, x') \delta_s^N(x') \right\|_{L^2_{\mu^\infty}(M)},
\]
and simplifying as usual using Jensen’s inequality and the Hölder inequality, we obtain
\[
\left\| \int_M \psi_1 \circ \angle(\cdot, x') \delta_s^N(x') \left( d\mu^\infty(x') - d\mu^N(x') \right) \right\|_{L^2_{\mu^\infty}(M)} \leq nL \|\delta_s^N\|_{L^\infty(M)}
\leq s\tau \left( n^{23} L^{60+8q} d^{15} \log^9 L \right)^{1/12},
\]
where the last bound uses (A.38). Then using the triangle inequality and \( k\tau \leq L^q/n \) to simplify in (A.37), we obtain
\[
\left\| \xi_n^\infty - \xi_k^N \right\|_{L^2_{\mu^\infty}(M)} \leq C \left( \frac{L^{60+32q} d^{15} \log^9 L}{n} \right)^{1/12}
+ \tau \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12} \sum_{s=0}^{k-1} \left\| \xi_s^\infty - \xi_s^N \right\|_{L^2_{\mu^\infty}(M)}
+ \tau \sum_{s=0}^{k-1} \left\| \int_M \psi_1 \circ \angle(\cdot, x') \zeta_s^{N, \text{Lip}}(x') \left( d\mu^\infty(x') - d\mu^N(x') \right) \right\|_{L^2_{\mu^\infty}(M)}.
\]
(A.39)
To simplify the remaining term in (A.39), we aim to apply (A.33); to do this we will need to justify the notation and establish that \( \zeta_s^{N, \text{Lip}} \in \text{Lip}(M) \) regardless of the random sample from \( \mu^\infty \) and the random instance of the weights. Because \( \zeta_s^{N, \text{Lip}} \) is a sum of functions, we can bound its minimal
Lipschitz constant by the sum of bounds on the Lipschitz constants of each summand. We always have for either \( \star \in \{+,-\} \)

\[
\left\| \xi^{N,\text{Lip}}_{\star} \right\|_{\text{Lip} | M_{\star}} \leq \left\| \xi | M_{\star} \right\|_{\text{Lip}} + \tau \sum_{i=0}^{s-1} \int_{M} \psi_{1} \circ \angle (\cdot, x') \xi_{i}^{N}(x') \, d\mu^{N}(x').
\]

(A.40)

We note that because the ReLU \([\cdot ]_{+}\) is 1-Lipschitz as a map on \(\mathbb{R}^{n}\), we have

\[
\left\| \xi | M_{\star} \right\|_{\text{Lip}} \leq \| W^{L+1} \|_{2} \frac{L}{\ell = 1} \| W^{\ell} \| < +\infty,
\]

so we need only develop a Lipschitz property for the summands in the second term of (A.40). To do this, we will start by showing that \( t \mapsto \psi_{1} \circ \cos^{-1}(\gamma_{\star}(t), x') \) is absolutely continuous for each \( x' \). Continuity is immediate. The only obstruction to differentiability comes from the inverse cosine, which fails to be differentiable at \( \pm 1 \), and because \( M \subset \mathbb{S}^{n-1} \) we have \( \langle \gamma_{\star}(t), x' \rangle = \pm 1 \) only if \( \gamma_{\star}(t) = \pm x' \); because \( \gamma_{\star} \) are simple curves, this shows that there are at most two points of nondifferentiability in \([0, \text{len}(M_{\star})]\). At points of differentiability, we calculate using the chain rule the derivative

\[
t \mapsto -\left( \psi_{1} \circ \cos^{-1}(\gamma_{\star}(t), x') \right) \left( \frac{\gamma'_{\star}(t)}{\sqrt{1 - \langle \gamma_{\star}(t), x' \rangle^{2}}} x' \right).
\]

and because \( \gamma_{\star} \) is a sphere curve, it holds \( (I - \gamma_{\star}(t) \gamma_{\star}^{*}(t)) \gamma'_{\star}(t) = \gamma'_{\star}(t) \) for all \( t \), whence by Cauchy-Schwarz

\[
\left\| \frac{\gamma'_{\star}(t)}{\sqrt{1 - \langle \gamma_{\star}(t), x' \rangle^{2}}} x' \right\| \leq \left\| \frac{(I - \gamma_{\star}(t) \gamma_{\star}^{*}(t)) x'}{\sqrt{1 - \langle \gamma_{\star}(t), x' \rangle^{2}}} \right\| \leq 1,
\]

(A.41)

where we also used that \( \gamma_{\star} \) are unit-speed curves. In particular, the derivative is bounded, hence integrable on \([0, \text{len}(M_{\star})]\), and so an application of [219, Theorem 6.3.11] establishes that \( t \mapsto
\(\psi_1 \circ \cos^{-1}\langle \mathbf{y}_*(t), \mathbf{x}' \rangle\) is absolutely continuous, with the expansion

\[
\begin{align*}
|\psi_1 \circ \cos^{-1}\langle \mathbf{y}_*(t), \mathbf{x}' \rangle - \psi_1 \circ \cos^{-1}\langle \mathbf{y}_*(t'), \mathbf{x}' \rangle| &= \left| \int_t^{t'} \left( \psi_1' \circ \cos^{-1}\langle \mathbf{y}_*(t''), \mathbf{x}' \rangle \right) \left( \frac{\mathbf{y}_*'(t'')}{\sqrt{1 - \langle \mathbf{y}_*(t''), \mathbf{x}' \rangle^2}} \right) \, dt'' \right|. 
\end{align*}
\]

which gives an avenue to establish Lipschitz estimates for \(t \mapsto \psi_1 \circ \cos^{-1}\langle \mathbf{y}_*(t), \mathbf{x}' \rangle\). Because \(\mathbf{x}' \mapsto \zeta_i^N(\mathbf{x}')\) is continuous and \(s \leq k \leq L^q/(n\tau) < +\infty\), an application of Fubini’s theorem enables us to also use this result to obtain Lipschitz estimates for the summands examined in (A.40), to wit

\[
\begin{align*}
\left\| \int_M \psi_1 \circ \angle(\cdot, \mathbf{x}') \zeta_i^N(\mathbf{x}') \, d\mu^N(\mathbf{x}') \right\|_{\text{Lip}} &\leq \sup_{x \in M_*} \int_M \left| \psi_1' \circ \angle(x, \mathbf{x}') \right| \zeta_i^N(\mathbf{x}') \, d\mu^N(\mathbf{x}') \\
&\leq \|\zeta_i^N\|_{L^2(\mu^N(M))} \sup_{x \in M_*} \left( \int_M \left( \psi_1' \circ \angle(x, \mathbf{x}') \right)^2 \, d\mu^N(\mathbf{x}') \right)^{1/2} \quad \text{(A.42)}
\end{align*}
\]

after using the bound (A.41) in the first inequality and the Schwarz inequality for the second.

Before proceeding with further simplifications, we note that the \(C^2\) property of \(\psi_1\), continuity of \(\zeta_i^N\), boundedness of \(i\), and compactness of \(M\) let us assert using (A.40) and (A.42) that \(\zeta_i^{N,\text{Lip}} \in \text{Lip}(M)\) whether or not we are working on the event \(E\). Continuing, we develop a bound for the RHS of (A.42) that is valid on \(E\). Using the triangle inequality and the Minkowski inequality, we have for the second term on the RHS of the last bound in (A.42)

\[
\begin{align*}
\sup_{x \in M_*} \left( \int_M \left( \psi_1' \circ \angle(x, \mathbf{x}') \right)^2 \, d\mu^N(\mathbf{x}') \right)^{1/2} &\leq \sup_{x \in M_*} \left( \left( \int_M \left( \psi_1' \circ \angle(x, \mathbf{x}') \right)^2 \left( d\mu^N(\mathbf{x}') - d\mu^\infty(\mathbf{x}') \right) \right)^{1/2} \right) \\
& \quad + \sup_{x \in M_*} \left( \int_M \left( \psi_1' \circ \angle(x, \mathbf{x}') \right)^2 \, d\mu^\infty(\mathbf{x}') \right)^{1/2} \quad \text{(A.43)}
\end{align*}
\]

For the first term in (A.43), we use Lemmas A.3.4, A.3.19 and A.3.21 to obtain that \(\mathbf{x}' \mapsto (\psi_1' \circ \angle(x, \mathbf{x}') \right)^2 \, d\mu^\infty(\mathbf{x}') \right)^{1/2} \).
\( \mathcal{L}(x, x')^2 \) is bounded by \( Cn^2L^4 \) and \( C'n^2L^5 \)-Lipschitz for every \( x \), and then applying (A.33) gives

\[
\sup_{x \in M_*} \left( \left( \int_{\mathcal{M}} (\psi_1' \circ \mathcal{L}(x, x'))^2 \left( d\mu^N(x') - d\mu^\infty(x') \right) \right) \right)^{1/2} \\
\leq \left( \frac{Cn^2L^4 \sqrt{d}}{N} + \frac{e^{14/\delta} C_{\mu^\infty,M} C'n^2L^5 \sqrt{d}}{N^{1/(2+\delta)}} \right)^{1/2} \\
\leq C \left( 1 + C_{\mu^\infty,M} \right)^{1/2} e^{7/\delta} nL^{5/2}d^{1/4}.
\] (A.44)

For the second term in (A.43), we apply Lemmas A.3.5 and A.3.19 together with the choice \( L \geq K\alpha^2C_\lambda \) to get

\[
\sup_{x \in M_*} \left( \int_{\mathcal{M}} (\psi_1' \circ \mathcal{L}(x, x'))^2 \, d\mu^\infty(x') \right)^{1/2} \\
\leq C nL^2 \sup_{x \in M_*} \left( \int_{\mathcal{M}} \left( \frac{d\mu^\infty(x')}{1 + (L/\pi) \mathcal{L}(x, x')} \right)^2 \right)^{1/2} \\
\leq C nL^{3/2} \rho_{\max}^{1/2} \left( \text{len}(M_+) + \text{len}(M_-) \right)^{1/2} \\
\leq C \rho_{\max}^{1/2} C_{\mu^\infty,M} nL^{3/2}.
\] (A.45)

Combining (A.44) and (A.45) to control the RHS of (A.43), we obtain from (A.42)

\[
\left\| \int_{\mathcal{M}} \psi_1 \circ \mathcal{L}(\cdot, x') \xi_i^N(x') \, d\mu^N(x') \right\|_{\text{Lip}} \\
\leq C \left\| \xi_i^N \right\|_{L^2_{\mu^N(M)}} \left( \frac{\left( 1 + C_{\mu^\infty,M} \right)^{1/2} e^{7/\delta}}{N^{1/(4+2\delta)}} nL^{5/2}d^{1/4} + \rho_{\max} C_{\mu^\infty,M} nL^{3/2} \right) \\
\leq C \left\| \xi_i^N \right\|_{L^2_{\mu^N(M)}} \left( 1 + C_{\mu^\infty,M} \right)^{1/2} e^{7/\delta} \left( 1 + \rho_{\max} \right)^{1/2} d^{1/4} nL^{3/2},
\] (A.46)

where in the second line we used \( N \geq L^{4+2\delta} \). Plugging (A.46) into (A.40) and applying in addition (A.32), we get

\[
\left\| \int_{\mathcal{M}} \xi_{\lambda}^N \right\|_{L^2_{\mu^N(M)}} \left\| \xi_{\lambda}^N \right\|_{L^2_{\mu^N(M)}} \\
\leq \sqrt{d} + C \tau e^{7/\delta} \left( 1 + C_{\mu^\infty,M} \right)^{1/2} \left( 1 + \rho_{\max} \right)^{1/2} d^{1/4} nL^{3/2} \sum_{i=0}^{s-1} \left\| \xi_i^N \right\|_{L^2_{\mu^N(M)}}.
\] (A.47)

Let us briefly pause to reorient ourselves. We do not have control of the empirical losses
appearing in (A.47) by an outside result, so we need to make some further simplifications to this bound. We will control the sum of empirical losses term in (A.47) by linking it to the difference population error, which we last saw in (A.39), and the population error using the triangle inequality and a change of measure inequality. Meanwhile, with the Lipschitz property of $\zeta_i^{N,\text{Lip}}$ we have shown, we will be able to obtain a bound in terms of simpler quantities for the last term on the RHS of (A.39) using (A.33). The two resulting bounds will give us a system of two coupled ‘discrete integral equations’ for the difference population error and the Lipschitz constants of $\zeta_i^{N,\text{Lip}}$, which we will solve inductively.

First, we continue simplifying (A.47). The triangle inequality and the fact that $\mu^N$ is a probability measure give
\[
\left\| \zeta_i^N \right\|_{L^2_{\mu^N}(M)} \leq \left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} + \left\| \delta_i^N \right\|_{L^\infty(M)}, \tag{A.48}
\]
and we have by the triangle inequality and Hölder-$\frac{1}{2}$ continuity of $x \mapsto \sqrt{x}$
\[
\left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} \leq \left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} + \left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} - \left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} \leq \left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} + \sqrt{\int_M \left( \zeta_i^{N,\text{Lip}}(x) \right)^2 \left( d\mu^\infty(x) - d\mu^N(x) \right)} \right). \tag{A.49}
\]
We have shown that $\zeta_i^{N,\text{Lip}} \in \text{Lip}(M)$ and $\zeta_i^{N,\text{Lip}} \in L^\infty(M)$ above, and so $\left( \zeta_i^{N,\text{Lip}} \right)^2 \in \text{Lip}(M)$ as well, with
\[
\left\| \left( \zeta_i^{N,\text{Lip}} \right)^2 \right\|_{\text{Lip}} \leq 2 \left\| \zeta_i^{N,\text{Lip}} \right\|_{L^\infty(M)} \left\| \zeta_i^{N,\text{Lip}} \right\|_{\text{Lip}}.
\]
Applying the previous equation with (A.33) to control (A.49), we get
\[
\left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} \leq \left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} \leq \left\| \zeta_i^{N,\text{Lip}} \right\|_{L^2_{\mu^N}(M)} + e^{14/\delta} C \mu^\infty(M) \max_{\epsilon \in \{+,-\}} \frac{\left\| \zeta_i^{N,\text{Lip}} \right\|_{L^\infty(M)}}{N^{1/(2+\delta)}},
\]
132
where the second line applies the Minkowski inequality. Using the triangle inequality and that \( \mu^\infty \) is a probability measure, we have

\[
\| \zeta_i \|_{L^2_{\mu^\infty}(M)} \leq \| \zeta_i - \zeta_i^\infty \|_{L^2_{\mu^\infty}(M)} + \| \zeta_i^\infty \|_{L^2_{\mu^\infty}(M)}.
\]

(A.50)

Substituting (A.50) into (A.48) and using (A.38) to simplify gives

\[
\left( \frac{\| \zeta_i \|_{L^2_{\mu^\infty}(M)}}{\sqrt{N}} + C d^{1/4} \left( \left\| \zeta_i \right\|_{L^\infty(M)} + \frac{\| \zeta_i \|_{L^\infty(M)}}{N^{1/4}} \right) \right)^{1/2} + C d^{1/4} \left( \frac{\| \zeta_i \|_{L^\infty(M)}}{\sqrt{N}} + \frac{\| \zeta_i \|_{L^\infty(M)}}{N^{1/4}} \right)^{1/2}.
\]

(A.51)

Following (A.47), we need to sum the previous bound over \( i \). To simplify residuals, we use (A.29) to get

\[
C s^2 \tau \sqrt{d} \left( n L^{48+8q} d^9 \log^9 L \right)^{1/12} + \sum_{i=0}^{s-1} \| \zeta_i \|_{L^2_{\mu^\infty}(M)} + \frac{C^2 q \log^2 L}{n \tau},
\]

where the second bound uses the control \( s \tau \leq k \tau \leq L^q / n \) and holds under the condition
\[ n \geq (C/C')^{12}L^{48+32d}d^3. \] Summing in (A.51) and using the previous bound, it follows

\[
\sum_{i=0}^{s-1} \left\| \xi_i^N \right\|^2_{\mu^N(M)} + \frac{CC^2\rho \sigma^2 e \log L}{n\tau} + \sum_{i=0}^{s-1} \left\| \xi_i^N - \xi_i^\infty \right\|^2_{\mu^\infty(M)}
\]

\[
+ C d^{1/4} \left( \left\| \xi_i^N, \text{Lip} \right\|_{L^\infty(M)} + \frac{e^{7/\delta} C_{1/2}^{1/2} \max_{\star \in \{+,-\}} \left\| \xi_i^N, \text{Lip} \right\|_{L^\infty(M)}^{1/2}}{N^{1/(4+2\delta)}} \right).
\]

Plugging (A.52) into (A.47), we obtain

\[
\left\| \xi_s^N, \text{Lip} \right\|_{\mu_s^*} \leq C_1 d^{1/4} L^{3/2}
\]

\[
+ C_1 d^{1/4} \sum_{i=0}^{s-1} \frac{\left\| \xi_i^N, \text{Lip} \right\|_{L^\infty(M)}}{\sqrt{N}} + \frac{\max_{\star \in \{+,-\}} \left\| \xi_i^N, \text{Lip} \right\|_{\mu_s^*}^{1/2}}{N^{1/(4+2\delta)}},
\]

where for concision we have defined

\[
C_1(\delta, \mu^\infty) = CC^2\rho \sigma^2 e^{14/\delta} \left( 1 + C_{1/2} \right) \left( 1 + \rho_{\max} \right)^{1/2}
\]

and simplified the \(\sqrt{d}\) residual in (A.47) by worst-casing with the larger residual from the population error term in (A.52), and made other simplifications by worst-casing some constants. We simplify (A.39) next: we have shown that \(\xi_s^N, \text{Lip} \in \text{Lip}(M)\) and \(\xi_s^N, \text{Lip} \in L^\infty(M)\) above, and so for every \(x \in M\), we have

\[
\psi_1 \circ \angle (x, \cdot) \xi_s^N, \text{Lip} \in \text{Lip}(M)
\]

as well, with

\[
\left\| \psi_1 \circ \angle (x, \cdot) \xi_s^N, \text{Lip} \right\|_{\mu_s^*} \leq CnL \max_{\star \in \{+,-\}} \left\| \xi_s^N, \text{Lip} \right\|_{\mu_s^*} + C' nL^2 \left\| \xi_s^N, \text{Lip} \right\|_{L^\infty(M)}
\]
using the definition of $\psi_1$, Lemmas A.3.4, A.5.5 and A.3.19, and

$$\left\| \psi_1 \circ \angle(x, \cdot) \xi_s^{N, \text{Lip}} \right\|_{L^\infty(M)} \leq CnL \left\| \xi_s^{N, \text{Lip}} \right\|_{L^\infty(M)}.$$  \hfill (A.56)

The bounds (A.55) and (A.56) retain no $x$ dependence. Applying (A.33) and integrating over $x$, we obtain from (A.55) and (A.56)

$$\left\| \int_M \psi_1 \circ \angle(\cdot, x') \xi_s^{N, \text{Lip}}(x') \left( d\mu^\infty(x') - d\mu^N(x') \right) \right\|_{L^2(\mu, M)} \leq \frac{Cn \sqrt{d} \left\| \xi_s^{N, \text{Lip}} \right\|_{L^\infty(M)}}{N},$$

$$+ \frac{CnLe^{14/i}C_{\mu^\infty, M} \sqrt{d} \max_{* \in \{+, -\}} \left\| \xi_s^{N, \text{Lip}} \right\|_{M_*} \right\|_{\text{Lip}}}{N^{1/(2+i)}},$$

$$+ \frac{CnL^2e^{14/i}C_{\mu^\infty, M} \sqrt{d} \left\| \xi_s^{N, \text{Lip}} \right\|_{L^\infty(M)}}{N^{1/(2+i)}},$$

and we can combine the first and third terms on the RHS of the previous bound by worst-casing, giving

$$\left\| \int_M \psi_1 \circ \angle(\cdot, x') \xi_s^{N, \text{Lip}}(x') \left( d\mu^\infty(x') - d\mu^N(x') \right) \right\|_{L^2(\mu, M)} \leq \frac{C \sqrt{d} \mu^{14/i} (1 + C_{\mu^\infty, M}) (1 + C_{\mu^\infty, M}) \left( \sum_{* \in \{+, -\}} \left\| \xi_s^{N, \text{Lip}} \right\|_{M_*} \right\|_{\text{Lip}} + L \left\| \xi_s^{N, \text{Lip}} \right\|_{L^\infty(M)}}{N^{1/(2+i)}}.$$

Plugging the previous bound into (A.39), we obtain

$$\left\| \xi_k^\infty - \xi_k^N \right\|_{L^2(\mu, M)} \leq C \left( \frac{L_6^{60+32q} d^{15} \log^9 L}{n} \right)^{1/12} + \tau \left( \frac{n^{11} L_6^{48+8q} d^{9} \log^9 L}{L} \right)^{1/12} \sum_{s=0}^{k-1} \left\| \xi_s^\infty - \xi_s^N \right\|_{L^2(\mu, M)}$$

$$+ \frac{C \tau \sqrt{d} \mu^{14/i} (1 + C_{\mu^\infty, M}) \left( \sum_{* \in \{+, -\}} \left\| \xi_s^{N, \text{Lip}} \right\|_{M_*} \right\|_{\text{Lip}} + L \left\| \xi_s^{N, \text{Lip}} \right\|_{L^\infty(M)}}{N^{1/(2+i)}}.$$

(A.57)
To finish coupling (A.53) and (A.57), we need to remove the $L^\infty(M)$ terms. We accomplish this using Lemma A.2.13, which gives

$$
\| \kappa^N_{s,Lip} \|_{L^\infty(M)} \leq CC_2^{1/2} \| \kappa^N_{s,Lip} \|_{L^2(M)}^2 + \frac{C}{\rho_{\min}^{1/3}} \| \kappa^N_{s,Lip} \|_{L^2(M)}^{2/3} \max_{* \in \{+, -\}} \| \kappa^N_{s,Lip} \|_{M_*}^{1/3},
$$
(A.58)

where we have defined

$$
C_2(\mu^\infty) = \frac{\rho_{\max}}{\rho_{\min} \min \{\mu^\infty(M_+), \mu^\infty(M_-)\}}.
$$
(A.59)

For coupling purposes, it will suffice to use a version of (A.58) obtained by simplifying with some coarse estimates. Using (A.30), (A.50) and (A.38), we have

$$
\left\| \kappa^N_{i,Lip} \right\|_{L^2(M)} \leq \sqrt{d} + i \tau \sqrt{d} \left( n^{11} L^{48 + 8q} d^9 \log^9 L \right)^{1/12} + \left\| \kappa^N_{i} - \zeta^\infty_{i} \right\|_{L^2(M)}
$$
\[\leq 2 \sqrt{d} + \left\| \kappa^N_{i} - \zeta^\infty_{i} \right\|_{L^2(M)},\]

using $i \tau \leq L^q / n$ and $n \geq L^{48 + 20q} d^9 \log^9 L$ in the second line, and plugging this into (A.58) and using the Minkowski inequality gives

$$
\left\| \kappa^N_{s,Lip} \right\|_{L^\infty(M)} \leq CC_2^{1/2} \sqrt{d} + CC_2^{1/2} \left\| \kappa^N_{i} - \zeta^\infty_{i} \right\|_{L^2(M)} + C d^{1/3} \max_{* \in \{+, -\}} \left\| \kappa^N_{s,Lip} \right\|_{M_*}^{1/3} \left\| \kappa^N_{s,Lip} \right\|_{L^2(M)}^{2/3} \max_{* \in \{+, -\}} \left\| \kappa^N_{s,Lip} \right\|_{M_*}^{1/3},
$$
(A.60)

To make some of the subsequent bounds more concise, we introduce additional notation

$$
\Lambda_s = \max_{* \in \{+, -\}} \left\| \kappa^N_{s,Lip} \right\|_{M_*}^{1/3}.
$$
Plugging (A.60) into (A.53) and using the Minkowski inequality, we obtain

\[ \Lambda_s \leq CC_1d^{1/4}L^{3/2}\left( d \log^2 L + \frac{C_2^{1/2}d^{3/4}n\tau}{\sqrt{N}} + n\tau \left( 1 + \frac{C_2^{1/2}d^{1/4}}{\sqrt{N}} \right) \sum_{i=0}^{s-1} \left\| \xi_i^N - \xi_i^\infty \right\|_{L^2_{\mu^\infty}(M)} \right) \]

\[ + \frac{n\tau d^{7/12}}{\rho^{1/3}_{\min} N^{1/(4+2\delta)}} \sum_{i=0}^{s-1} \Lambda_i^{1/3} + \frac{C_2^{1/4}n\tau d^{1/2}}{N^{1/(4+2\delta)}} \sum_{i=0}^{s-1} \Lambda_i^{1/2} \]

\[ + \frac{n\tau d^{5/12}}{\rho^{1/6}_{\min} N^{1/(4+2\delta)}} \sum_{i=0}^{s-1} \Lambda_i^{1/3} + \frac{n\tau d^{1/4}}{\rho^{1/3}_{\min} N^{1/(4+2\delta)}} \sum_{i=0}^{s-1} \left\| \xi_i^N - \xi_i^\infty \right\|_{L^2_{\mu^\infty}(M)} \Lambda_i^{1/2} \]

\[ + \frac{n\tau d^{1/4}}{\rho^{1/6}_{\min} N^{1/(4+2\delta)}} \sum_{i=0}^{s-1} \left\| \xi_i^N - \xi_i^\infty \right\|_{L^2_{\mu^\infty}(M)} \Lambda_i^{1/2} \] (A.61)

To simplify (A.61), we use \( s\tau \leq L^q/n \), \( C_2 \geq 1 \), and \( q \leq 1 \), and so if additionally we choose \( N \geq C_2 \max\{ \sqrt{d}, L^2 \} \) we obtain

\[ \Lambda_s \leq CC_1d^{1/4}L^{3/2}\left( d \log^2 L + n\tau \sum_{i=0}^{s-1} \left\| \xi_i^N - \xi_i^\infty \right\|_{L^2_{\mu^\infty}(M)} \right) \]

\[ + \frac{n\tau d^{1/4}}{\rho^{1/6}_{\min} N^{1/(4+2\delta)}} \sum_{i=0}^{s-1} \left\| \xi_i^N - \xi_i^\infty \right\|_{L^2_{\mu^\infty}(M)} \Lambda_i^{1/2} \]

\[ + \frac{n\tau d^{1/4}}{\rho^{1/6}_{\min} N^{1/(4+2\delta)}} \sum_{i=0}^{s-1} \Lambda_i^{1/2} \]

\[ + \frac{n\tau d^{7/12}}{\rho^{1/3}_{\min} N^{1/(4+2\delta)}} \sum_{i=0}^{s-1} \Lambda_i^{1/3} \] (A.62)

Meanwhile, we recall

\[ C_{\mu^\infty,M} = \frac{\text{len}(M_+)}{\mu^\infty(M_+)} + \frac{\text{len}(M_-)}{\mu^\infty(M_-)} \]

\[ \text{len}(M_+) = \sum_{i=0}^{s-1} \left\| \xi_i^N - \xi_i^\infty \right\|_{L^2_{\mu^\infty}(M)} \Lambda_i^{1/3} \]

\[ \text{len}(M_-) = \sum_{i=0}^{s-1} \left\| \xi_i^N - \xi_i^\infty \right\|_{L^2_{\mu^\infty}(M)} \Lambda_i^{1/2} \]
and an integration in coordinates gives

\[ \mu^\infty(M_\pm) = \int_0^{\text{len}(M_\pm)} \rho_\pm \circ \gamma_\pm(t) \, dt \geq \rho_{\min} \text{len}(M_\pm), \]

so that

\[ C_{\mu^\infty,M} \leq \frac{2}{\rho_{\min}}. \quad (A.63) \]

Using (A.63) and plugging (A.60) into (A.57), we obtain

\[ \left\| \xi_k - \xi_N \right\|_{L_{\mu^\infty}^2(M)} \leq C \left( \frac{L^{60+32q} d^{15} \log^9 L}{n} \right)^{1/12} \]

\[ + \tau \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12} + \frac{C' C_1^{1/2} \sqrt{d} n L^2 e^{14/\delta}}{\rho_{\min} N^{1/2+\delta}} \sum_{s=0}^{k-1} \left\| \xi_s - \xi_N \right\|_{L_{\mu^\infty}^2(M)} \]

\[ + \frac{C' \sqrt{d} n L e^{14/\delta}}{\rho_{\min} N^{1/2+\delta}} \sum_{s=0}^{k-1} \left( \Lambda_s + L C_2^{1/2} \sqrt{d} \right)^{1/3} \left( L d^{1/3} \rho_{\min}^{-1/3} \Lambda_s^{1/3} + L \rho_{\min}^{-1/3} \left\| \xi_s - \xi_N \right\|_{L_{\mu^\infty}^2(M)}^{2/3} \right) \Lambda_s^{1/3}. \quad (A.64) \]

In (A.62) and (A.64), we now have a suitable system of coupled discrete integral equations for \( \left\| \xi_k - \xi_N \right\|_{L_{\mu^\infty}^2(M)} \) and \( \Lambda_k \). We will solve these equations by positing bounds for each parameter that are valid for all indices \( 0 \leq k \leq \lfloor L^q/(n\tau) \rfloor \) based on inspection of (A.62) and (A.64), then proving the bounds hold by induction on \( k \). Positing the bounds is not too hard, because each term in (A.62) and (A.64) with a factor of \( N \) in its denominator can be forced to be small by requiring \( N \) to be large enough. For all \( 0 \leq k \leq \lfloor L^q/(n\tau) \rfloor \), we claim

\[ \left\| \xi_k - \xi_N \right\|_{L_{\mu^\infty}^2(M)} \leq C_{\text{diff}} \max \left\{ \frac{C_{\text{lip}} C_1 C_2^{1/2} C_\rho^{4/3} d^{7/4} L^{5/2+q} \log^2 L}{n^{1/2+\delta}} \right\} \]

\[ \Lambda_k \leq C_{\text{lip}} C_1 d^{5/4} L^{3/2} \log^2 L, \quad (A.66) \]

where \( C_{\text{diff}} \) and \( C_{\text{lip}} \) are two absolute constants that we will specify in our arguments below. We
prove (A.65) and (A.66) by induction on $k$. The case of $k = 0$ is immediate, since $\zeta_0^\infty = \zeta_0^N$ for (A.65); and by construction $\zeta_0^{N,\text{lip}} = \zeta$, and (A.32) and $d \geq 1$ then gives (A.66) if $L \geq e$.

We therefore move to the induction step, assuming that (A.65) and (A.66) hold for $k - 1$ and showing that this implies the bounds for $k$. We begin by verifying (A.65). Applying the induction hypothesis for $k - 1$ via (A.66), we can write

$$\Lambda_s + LC_2^{1/2} \sqrt{d} + L \left( \frac{d \Lambda_s}{\rho_{\min}} \right)^{1/3}$$

$$\leq \text{lip}C_1 d^{5/4} L^{3/2} \log^2 L + LC_2^{1/2} \sqrt{d} + \left( \frac{\text{lip}C_1}{\rho_{\min}} \right)^{1/3} d^{3/4} L^{3/2} \log^{2/3} L$$

$$\leq \text{lip}C_1C_2^{1/2} C_{\rho}^{1/3} d^{5/4} L^{3/2} \log^2 L,$$

where we worst-cased in the second line using $\text{lip} \geq 1$ and $C_1 \geq 1$, $C_2 \geq 1$, which follow from (A.54) and (A.59). We use $k \tau \leq L^n/n$ with the last bound to note that

$$\frac{C' \sqrt{d} \ln L e^{14/\delta}}{\rho_{\min} N^{1/(2+\delta)}} \sum_{s=0}^{k-1} \text{lip}C_1 C_2^{1/2} C_{\rho}^{1/3} d^{5/4} L^{3/2} \log^2 L \leq C'' \text{lip}C_1 C_2^{1/2} C_{\rho}^{4/3} d^{7/4} L^{5/2+q} \log^2 L,$$

where $C'' \geq 1$. Using this bound and (A.66) once more, we can simplify (A.64) to

$$\left\| \xi_k - \xi_k^N \right\|_{L^2_{\mu^\infty}(\mathcal{M})}$$

$$\leq C'' \text{lip}C_1 C_2^{1/2} C_{\rho}^{4/3} d^{7/4} L^{5/2+q} \log^2 L + \frac{C (L^{60+32q} d^{15} \log^9 L)}{n} \right)^{1/12}$$

$$+ \tau \left( \left( n^{11} L^{48+8q} d^9 \log^9 L \right)^{1/12} + \frac{C'C_2^{1/2} \sqrt{d} \ln L e^{14/\delta}}{\rho_{\min} N^{1/(2+\delta)}} \right) \sum_{s=0}^{k-1} \left\| \xi_s^\infty - \xi_s^N \right\|_{L^2_{\mu^\infty}(\mathcal{M})}$$

$$+ \frac{C'C_1^{1/3} C_{\rho}^{14/3}}{\rho_{\min}^{4/3}} \tau d^{11/12} n L^{5/2} \log^{2/3} L \sum_{s=0}^{k-1} \left\| \xi_s^\infty - \xi_s^N \right\|_{L^2_{\mu^\infty}(\mathcal{M})}^{2/3}.$$  

(A.67)

Noticing that the RHS of the bound (A.65) does not depend on $k$, let us momentarily denote it by $C_{\text{diff}} M$ (i.e., the part of the RHS of this bound that does not involve $C_{\text{diff}}$ is denoted as $M$).
Plugging into (A.67) and using \( k \tau \leq L^9/n \), we obtain

\[
\| \xi^\infty_k - \xi^N_k \|_{L^2_{\mu^\infty}(M)} \leq C'' C_{\text{lip}} C_1^{1/2} C_2^{4/3} d^{7/4} L^{5/2+q} \log^2 L + C \left( \frac{L^{60+32q} d^{15} \log^9 L}{n} \right)^{1/12} + C_{\text{diff}} \left( \frac{L^{48+20q} d^9 \log^9 L}{n} \right)^{1/12} + C' C_2^{1/2} \sqrt{d} L^{2+q} e^{14/\delta} M + C' C_{\text{diff}}^{2/3} C_{\text{lip}}^{1/3} C_1^{1/3} \frac{e^{14/\delta}}{\rho_{\min}^{4/3}} d^{11/12} L^{5/2+q} \log^{2/3} L M^{2/3}.
\]

In particular, if \( C_{\text{diff}} = 6 \max\{C, C''\} \) (for the constants in the first line of the previous bound), we can bound the RHS of the previous bound and obtain

\[
\| \xi^\infty_k - \xi^N_k \|_{L^2_{\mu^\infty}(M)} \leq \frac{C_{\text{diff}} M}{3} + C_{\text{diff}} \left( \frac{L^{48+20q} d^9 \log^9 L}{n} \right)^{1/12} + C' C_2^{1/2} \sqrt{d} L^{2+q} e^{14/\delta} M + C' C_{\text{diff}}^{2/3} C_{\text{lip}}^{1/3} C_1^{1/3} \frac{e^{14/\delta}}{\rho_{\min}^{4/3}} d^{11/12} L^{5/2+q} \log^{2/3} L M^{2/3}.
\] (A.68)

We can conclude (A.65) from (A.68) provided we can show the second and third terms are no larger than \( C_{\text{diff}} M/3 \). For the second term in (A.68), if we choose \( N \) such that

\[
N^{1/(2+\delta)} \geq 6 C' C_2^{1/2} \rho_{\min}^{-1} e^{14/\delta} d^{1/2} L^{2+q}
\]

and \( n \) such that

\[
n \geq 6^{12} L^{48+20q} d^9 \log^9 L
\]

then we have

\[
C_{\text{diff}} \left( \frac{L^{48+20q} d^9 \log^9 L}{n} \right)^{1/12} + C' C_2^{1/2} \sqrt{d} L^{2+q} e^{14/\delta} M \leq \frac{C_{\text{diff}} M}{3}.
\]

For the third term in (A.68), we proceed in cases: first, when

\[
C_{\text{lip}} C_1^{1/2} C_2^{4/3} d^{7/4} L^{5/2+q} \log^2 L \leq \left( \frac{L^{60+32q} d^{15} \log^9 L}{n} \right)^{1/12}, \quad \text{(A.69)}
\]
we have by (A.65)

$$M = \left( \frac{L^{60+32q} d^{15} \log^9 L}{n} \right)^{1/12},$$

and if we require additionally $C_{\text{diff}} \geq 1$, it follows that

$$\frac{C' C_{\text{diff}}^{2/3} C_{\text{lip}}^{1/3} C_1^{1/3} e^{14/\delta}}{\rho_{\min}^{4/3}} \frac{d^{11/12} L^{5/2+q} \log^{2/3} L M^{2/3}}{N^{1/(2+\delta)}} \leq C' C_{\text{diff}} C_1 C_2^{1/2} C_{\rho}^{4/3} e^{14/\delta} \frac{d^{7/4} L^{5/2+q} \log^2 L M^{2/3}}{N^{1/(2+\delta)}}$$

$$\leq C' C_{\text{diff}}^{14/\delta} M^{1+2/3},$$

using $C_1 \geq 1$, $C_2 \geq 1$, and $C_{\text{lip}} \geq 1$ and worst-casing exponents on $d$ and $\log L$ in the first line, and (A.69) in the second line. In particular, by the value of $M$ in this regime, if

$$n \geq (3C' e^{14/\delta})^{18} L^{60+32q} d^{15} \log^9 L$$

then we obtain for the third term in (A.68)

$$\frac{C' C_{\text{diff}}^{2/3} C_{\text{lip}}^{4/3} C_1^{1/3} e^{14/\delta}}{\rho_{\min}^{4/3}} \frac{d^{11/12} L^{5/2+q} \log^{2/3} L M^{2/3}}{N^{1/(2+\delta)}} \leq \frac{C_{\text{diff}} M}{3},$$

as desired. Next, we consider the remaining case

$$C_{\text{lip}} C_1 C_2^{1/2} C_{\rho}^{4/3} d^{7/4} L^{5/2+q} \log^2 L \geq \left( \frac{L^{60+32q} d^{15} \log^9 L}{n} \right)^{1/12},$$

which by (A.65) implies

$$M = C_{\text{lip}} C_1 C_2^{1/2} C_{\rho}^{4/3} d^{7/4} L^{5/2+q} \log^2 L \frac{N^{1/(2+\delta)}}{n}.$$
With this setting of $M$, the third term in (A.68) can be bounded as

\[
\frac{C' C_{\text{lip}}^{2/3} C_1^{1/3} e^{14/\delta}}{\rho^{4/3}_{\min}} \cdot \frac{d^{11/12} L^{5/2+q} \log^{2/3} L}{N^{1/(2+\delta)}} M^{2/3} = C' C_{\text{lip}}^{2/3} C_1^{1/3} C_2^{4/3} e^{14/\delta} \cdot \frac{d^{7/4+1/3} L^{5/3+2q/3}}{N^{1/(2+\delta)+2/(2+3\delta)}} \log^2 L
\leq C' C_{\text{lip}}^{2/3} C_1^{1/3} e^{14/\delta} \cdot \frac{d^{1/3} L^{5/3+2q/3}}{N^{2/(2+3\delta)}} M,
\]

and using the RHS of the final bound in the previous expression, we see that if we choose

\[
N^{1/(2+\delta)} \geq (3C')^{3/2} C_\rho^{4/3} e^{21/\delta} d^{1/2} L^{5/2+q},
\]

then we have for the case (A.70)

\[
\frac{C' C_{\text{lip}}^{2/3} C_1^{1/3} e^{14/\delta}}{\rho^{4/3}_{\min}} \cdot \frac{d^{11/12} L^{5/2+q} \log^{2/3} L}{N^{1/(2+\delta)}} M^{2/3} \leq \frac{C_{\text{diff}} M}{3}.
\]

Combining the bounds on the third term in (A.68) over both cases (A.69) and (A.70), we have shown

\[
\|\tilde{\xi}_k - \xi_k\|_{L^2_{\infty}(M)} \leq C_{\text{diff}} M,
\]

which proves (A.65). Next, to verify (A.66), we proceed with a similar idea: the bound claimed in (A.66) corresponds to a constant multiple of the first term in parentheses in (A.62), so to establish (A.66) it suffices to show that each of the other terms in (A.62) is no larger than a certain constant. To work with the maximum operation in (A.65), we will again split the analysis into two cases. First, we consider the case where (A.70) holds, so that the maximum in (A.65) is achieved by the
and (to combine the first and second terms)

\[
\frac{d^{5/3} L^{11/6+4q/3} \log^2 L}{N^{6/(10+5\delta)}} \leq \frac{d^{7/4} L^{5/2+2q} \log^2 L}{N^{1/(2+\delta)}},
\]

and because \(0 < \delta \leq 1\), we have \(1/(2+\delta) \leq 1/2\) and \((5+\delta)/(4+2\delta) \geq 1\), and if \(N \geq d^{1/12}\) this implies (to combine the first and second-to-last terms)

\[
\frac{d^{11/6} L^{13/6+5q/3} \log^2 L}{N^{(5+\delta)/(4+2\delta)}} \leq \frac{d^{7/4} L^{5/2+2q} \log^2 L}{N^{1/(2+\delta)}}.
\]
We can worst-case the remaining three terms, and we thus obtain

\[ \Lambda_k \leq CC_1 d^{1/4} L^{3/2} \left( d \log^2 L + 4C_{\text{diff}} C_{\text{lip}} C_1 C_2^{1/2} C_\rho^{5/3} d^{1+3/2} L^{5/2+2q} \log^2 L \right) \]

We can then pick \( C_{\text{lip}} = 3C \), and if

\[ N^{1/(4+2\delta)} \geq 3(3C)^{2/3} C_1 C_2^{1/2} C_\rho^{1/3} d^{1/4} L^{1+q}, \]

and

\[ N^{1/(2+\delta)} \geq 12CC_{\text{diff}} C_1 C_2^{1/2} C_\rho^{5/3} d^{3/4} L^{5/2+2q}, \]

then it follows from the previous bound

\[ \Lambda_k \leq 3CC_1 d^{5/4} L^{3/2} \log^2 L, \]

which establishes (A.66) in the first case, where (A.70) holds. Next, we consider the remaining case where (A.69) holds, so that the maximum in (A.65) is saturated by the first argument. We start by grouping some terms in (A.62) so that it will be slightly easier to simplify later: we can write

\[ \Lambda_k \leq CC_1 d^{1/4} L^{3/2} \left( d \log^2 L + n\tau \sum_{s=0}^{k-1} \| \xi_s^N - \xi_s^\infty \|_{L^2_{\mu^\infty}(\mathcal{M})} + \frac{n\tau d^{1/4}}{\rho_{\min}^{1/6} N^{1/(4+2\delta)}} \sum_{s=0}^{k-1} \left( \| \xi_s^N - \xi_s^\infty \|_{L^2_{\mu^\infty}(\mathcal{M})} + d^{1/6} \right) \Lambda_s^{2/3} \right) \]

\[ + \frac{n\tau d^{1/4}}{\rho_{\min}^{1/3} \sqrt{N}} \sum_{s=0}^{k-1} \left( \| \xi_s^N - \xi_s^\infty \|_{L^2_{\mu^\infty}(\mathcal{M})} + d^{1/3} \right) \Lambda_s^{1/3} \]

\[ + \frac{C_2^{1/4} n\tau d^{1/4}}{N^{1/(4+2\delta)}} \sum_{s=0}^{k-1} \left( \| \xi_s^N - \xi_s^\infty \|_{L^2_{\mu^\infty}(\mathcal{M})} + d^{1/4} \right) \Lambda_s^{1/2} \right). \]

(A.71)
By the case-defining condition (A.69) and (A.65), enforcing

\[ n \geq C_{\text{diff}}^{12} L^{60+32q} d^9 \log^9 L \]

implies

\[ \| \zeta_N^N - \xi^\infty \|_{L^2_{\mu^\infty}(M)} + d^{1/2} \leq 2d^{1/2}, \]

and we can use this to simplify (A.71), obtaining

\[
\Lambda_k \leq CC_1 d^{1/4} L^{3/2} \left( d \log^2 L + n \tau \sum_{s=0}^{k-1} \| \zeta_N^N - \xi_s^\infty \|_{L^2_{\mu^\infty}(M)} + \frac{2n \tau d^{5/12}}{\rho_{\min}^{1/6} N^{1/(4+2\delta)}} \sum_{s=0}^{k-1} \Lambda_s^{2/3} + \frac{2n \tau d^{7/12}}{\rho_{\min}^{1/3} \sqrt{N}} \sum_{s=0}^{k-1} \Lambda_s^{1/3} + \frac{2C_1^{1/4} n \tau d^{1/2}}{N^{1/(4+2\delta)}} \sum_{s=0}^{k-1} \Lambda_s^{1/2} \right). \tag{A.72}
\]

Plugging (A.65) and (A.66) into (A.72) and using \( k \tau \leq L^q/n \) and (A.69), we get the bound

\[
\Lambda_k \leq CC_1 d^{1/4} L^{3/2} \left( d \log^2 L + C_{\text{diff}} \left( \frac{L^{60+44q} d^{15} \log^9 L}{n} \right)^{1/12} + 2C_{\text{lip}}^{2/3} C_1^{2/3} C_\rho^{1/6} \frac{d^{1+1/4} L^{1+q} \log^{2/3} L}{N^{1/(4+2\delta)}} \right) \tag{A.73}
\]

From (A.73), we see that if we choose \( n \) such that

\[ n \geq (2C_{\text{diff}})^{12} L^{60+44q} d^3 \]

and we choose \( N \) such that

\[ N^{1/(2+\delta)} \geq 16C_{\text{lip}}^{4/3} C_1^{4/3} C_2^{1/2} C_\rho^{1/3} d^{1/2} L^{2+2q} \]
then (A.73) implies the bound
\[ \Lambda_k \leq 3CC_1 d^{5/4} L^{3/2} \log^2 L, \]
which agrees with the previous choice \( C_{\text{lip}} = 3C \) and thus proves (A.66) in the remaining case of (A.72). By induction, then, we have proved that (A.65) and (A.66) hold for each index \( 0 \leq k \leq [L^q/(n\tau)] \).

We can now wrap up the proof: we will obtain the desired conclusion by plugging the results we have developed into (A.58) and simplifying. Plugging (A.28), (A.65) and (A.38) into (A.50) and bounding the maximum by the sum, we get
\[
\left\| \xi_{[L^q/(n\tau)]}^N \right\|_{L^2_{\mu^\infty}(M)} \leq \left\| \xi_{[L^q/(n\tau)]}^\infty \right\|_{L^2_{\mu^\infty}(M)} + \left\| \xi_{[L^q/(n\tau)]}^N - \xi_{[L^q/(n\tau)]}^\infty \right\|_{L^2_{\mu^\infty}(M)} + \left\| \delta_{[L^q/(n\tau)]}^N \right\|_{L^\infty(M)} \leq \frac{CC_{\text{cert}}^q \sqrt{d} \log L}{L^q} \ 
\]
where in the third inequality we apply \([L^q/(n\tau)] \geq L^q/(2n\tau)\), which follows from our choice of step size, and in the fourth inequality we simplify residuals using \( n \geq (C'/C)^{12} d^9 L^{60+44q} \) and \( N^{1/(2+\delta)} \geq C'' C_1 C_2^{1/2} C_\rho^{4/3} d^{7/4} L^{5/2+2q} \log L \). Applying (A.74), the triangle inequality (with (A.38) and the fact that \( \mu^\infty \) is a probability measure) and our previous choice of large \( n \), we get
\[
\left\| \xi_{[L^q/(n\tau)]}^N \right\|_{L^2_{\mu^\infty}(M)} \leq \frac{CC_{\text{cert}}^q \sqrt{d} \log L}{L^q},
\]
i.e. generalization in \( L^2_{\mu^\infty}(M) \). We can bootstrap generalization in \( L^\infty(M) \) from (A.74) using the
triangle inequality and (A.58): we get

\[
\left\| \mathbf{s}_{[L^q/(\sigma n)]} \right\|_{L^\infty(M)}^N 
\leq C C_2^{1/2} \left\| \xi_{[L^q/(\sigma n)]} \right\|_{L^2(M)}^{N, \text{Lip}} + \frac{C}{\rho_{\min}^{1/3}} \left\| \xi_{[L^q/(\sigma n)]} \right\|_{L^2(M)}^{N/3, \text{Lip}}^{2/3} \Lambda_{[L^q/(\sigma n)]}^{1/3} + \left\| \delta_{[L^q/(\sigma n)]}^N \right\|_{L^\infty(M)}
\]

\[
\leq \frac{C C_2^{1/2} C_{q_{\text{cert}}}}{L} \sqrt{d \log L} + \frac{C' C_1^{1/3} d^{3/4} L^{(3-4q)/6} \log^{4/3} L}{\min\left\{ \rho_{\min}^{1/3} \left(1 + \rho_{\max}\right)^{1/2} \right\}},
\]

where in the second line we apply (A.38) and our previous choice of large \(n\) to absorb the residual from \(\delta_{[L^q/(\sigma n)]}^N\), and apply (A.66) to bound the \(\Lambda_{[L^q/(\sigma n)]}^{1/3}\) term. Worst-casing the errors in the previous bound, we obtain

\[
\left\| \mathbf{s}_{[L^q/(\sigma n)]} \right\|_{L^\infty(M)}^N \leq C \left( C_{q_{\text{cert}}^N} C_2^{1/2} + C_1^{1/3} C_{\rho}^{2/3} \right) \frac{d^{3/4} \log^{4/3} L}{L^{(4q-3)/6}}.
\]

To conclude, we will tally dependencies and make some simplifications to show the conditions stated in the result suffice. Recalling (A.54) and (A.59) and using (A.63), we have

\[
C_1 \leq C C_{\rho}^{-2q_{\text{cert}}+1} \left(1 + \rho_{\max}\right)^{1/2} e^{14/3\delta},
\]

so we can simplify to

\[
C_{\rho}^{1/2} C_2^{1/2} + C_1^{1/3} C_{\rho}^{2/3} \leq C_{\rho} \left( \frac{\rho_{\max}}{\min \{ \mu_{\infty}(M_+) + \mu_{\infty}(M_-) \}} \right)^{1/2}
+ C C_{\rho}^{1+2q_{\text{cert}}/3} \left(1 + \rho_{\max}\right)^{1/6} e^{14/(3\delta)}
\leq C C_{\rho}^{1+2q_{\text{cert}}/3} \left(1 + \rho_{\max}\right)^{1/2} e^{14/(3\delta)}
\leq \frac{C C_{\rho}^{1+2q_{\text{cert}}/3} \left(1 + \rho_{\max}\right)^{1/2} e^{14/(3\delta)}}{\min \{ \mu_{\infty}(M_+), \mu_{\infty}(M_-) \}^{1/2}}.
\]

We can use this to obtain a simplified generalization in \(L^\infty(M)\) bound from our previous expres-
tion: it becomes
\[
\left\| \phi^N \right\|_{L^\infty(M)} \leq \frac{CC_\rho^{1+2q_{\text{cert}}/3}(1 + \rho_{\text{max}})^{1/2}e^{14/(3\delta)}d^{3/4}\log^{4/3}L}{\min \{\mu^\infty(M_+), \mu^\infty(M_-)\}^{1/2}} \frac{d^{3/4}L}{L^{(4q-3)/6}},
\] (A.76)

which can be made nonvacuous when \( q > 3/4 \). Tallying dependencies, we find after worst-casing (and using \( q \geq 1/2 \) and some interdependences between parameters to simplify) that it suffices to choose \( N \) such that
\[
N^{1/(2+\delta)} \geq CC_1^{4/3}C_2^{1/2}C_\rho^{5/3}e^{21/\delta}d^{5/4}L^{5/2+2q} \log L,
\]
the depth \( L \) such that
\[
L \geq C \max \{C_\rho^{2q_{\text{cert}}}d, \kappa^2C_1\},
\]
the width \( n \) such that
\[
n \geq C \max \left\{ e^{252/\delta}L^{60+44q}d^{9/2} \log^{9/2}L, \kappa^{2/5} \left( \frac{k}{C_1} \right)^{1/3} \right\},
\]
and \( d \) such that \( d \geq Cd_0 \log(nn_0C_M) \). Unpacking the constants in the condition on \( N \), we see that it suffices to choose \( N \) such that
\[
N^{1/(2+\delta)} \geq \frac{CC_\rho^{7/2+8q_{\text{cert}}/3}(1 + \rho_{\text{max}})^{7/6}e^{119/(3\delta)}d^{5/4}L^{5/2+2d} \log L.}
\]

\[\square\]

A.2.3 Auxiliary Results

**Lemma A.2.7.** Defining a kernel

\[
\Theta^N_k(x, x') = \int_0^1 \left\langle \tilde{\nabla} f_{\theta^N_k}(x'), \tilde{\nabla} f_{\theta^N_{k-t}}(x) \right\rangle dt
\]
and corresponding operator on $L^2_{\mu^N}(M)$

$$
\Theta_k^N[g](x) = \int_M \Theta_k^N(x,x') g(x') \, d\mu^N(x'),
$$

we have that $\Theta_k^N$ is bounded, and

$$
\xi_{k+1}^N = \left(\text{Id} - \tau \Theta_k^N\right) \left[\xi_k^N\right].
$$

**Proof.** By the definition of the gradient iteration, we have that

$$
\xi_{k+1}^N - \xi_k^N = f_{\theta_k^N - \tau \nabla L_{\mu^N}(\theta_k^N)} - f_{\theta_k^N}.
$$

The total number of trainable parameters in the network is $M = n(n(L - 1) + n_0 + 1)$, and the euclidean space in which $\theta$ lies is isomorphic to $\mathbb{R}^M$. For $k \in \mathbb{N}_0$, define paths $\gamma_k^N : [0, 1] \to \mathbb{R}^M$ by

$$
\gamma_k^N(t) = \theta_k^N - t \tau \nabla L_{\mu^N}(\theta_k^N),
$$

so that

$$
\xi_{k+1}^N - \xi_k^N = f_{\gamma_k^N(1)} - f_{\gamma_k^N(0)}.
$$

We will justify a first-order Taylor representation in integral form based on the previous expression by arguing that for every $x \in M$, $t \mapsto f_{\gamma_k^N(t)}(x)$ is absolutely continuous on $[0, 1]$, by checking the hypotheses of [219, Theorem 6.3.11]. Because $\gamma_k^N$ is smooth and $f(\cdot)(x)$ is continuous, $f_{\gamma_k^N(t)}$ is also continuous. Continuity of the features as a function of the parameters and of $\gamma_k^N$ implies that for every $\ell \geq 0$, the image of $[0, 1]$ under the map

$$
t \mapsto \alpha_{\gamma_k^N(t)}^\ell(x)
$$

is compact. By repeated application of Lemma A.5.21, we conclude that $t \mapsto f_{\gamma_k^N(t)}(x)$ is differentiable at all but countably many points of $[0, 1]$. Following the proof of Lemma A.5.21, we see that
the points of nondifferentiability of \( t \mapsto f_{\gamma_{k}^N(t)}(x) \) are contained in the set of points of \([0, 1]\) where there exists a layer \( \ell \) at which at least one of the coordinates of \( \alpha_{\gamma_{k}^N(t)}^{\ell}(x) \) vanishes. Applying the chain rule at points of differentiability of the ReLU \( \lfloor \cdot \rfloor_{+} \) and assigning 0 otherwise, it follows that the derivative of \( t \mapsto f_{\gamma_{k}^N(t)}(x) \) at \( t \in [0, 1] \) is equal to

\[
-\tau \left\langle \nabla L_{\mu}^{N}(\theta_{k}^N), \nabla f_{\gamma_{k}^N(t)}(x) \right\rangle
\]

at all but countably many points \( t \in [0, 1] \). We finally need to check integrability of this derivative on \([0, 1]\). We have by linearity

\[
-\tau \left\langle \nabla L_{\mu}^{N}(\theta_{k}^N), \nabla f_{\gamma_{k}^N(t)}(x) \right\rangle = -\tau \int_{M} \zeta_{\theta_{k}^N}(x') \left\langle \nabla f_{\theta_{k}^N}(x'), \nabla f_{\gamma_{k}^N(t)}(x) \right\rangle d\mu^{N}(x'),
\]

and by definition

\[
\left\langle \nabla f_{\gamma_{k}^N(t)}(x), \nabla f_{\theta_{k}^N}(x') \right\rangle = \left\langle \alpha_{\gamma_{k}^N(t)}^{\ell}(x), \alpha_{\theta_{k}^N}^{\ell}(x') \right\rangle + \sum_{\ell=0}^{L-1} \left\langle \alpha_{\gamma_{k}^N(t)}^{\ell}(x), \alpha_{\theta_{k}^N}^{\ell}(x') \right\rangle \left\langle \beta_{\gamma_{k}^N(t)}^{\ell}(x), \beta_{\theta_{k}^N}^{\ell}(x') \right\rangle.
\]

By construction of the network, the feature maps \((t, x) \mapsto \alpha_{\gamma_{k}^N(t)}^{\ell}(x)\) are continuous. For the backward feature maps, we can write for any \( \theta_1 = (W_1^1, \ldots, W_{L+1}^1) \) and any \( \theta_2 = (W_1^2, \ldots, W_{L+1}^2) \) using Cauchy-Schwarz

\[
\left| \left\langle \beta_{\theta_1}^{\ell}(x), \beta_{\theta_2}^{\ell}(x') \right\rangle \right| \leq \prod_{\ell' = \ell+1}^{L} \| W_{\ell'+1}^1 \| \| W_{\ell'+1}^2 \|,
\]

and the RHS of this bound is a continuous function of \((\theta, x)\). Because our domain of interest \([0, 1] \times M\) is compact, we have from the triangle inequality, the previous bound on the backward feature inner products and the Weierstrass theorem

\[
\sup_{t \in [0, 1], x \in M} \left| \left\langle \nabla f_{\gamma_{k}^N(t)}(x), \nabla f_{\theta_{k}^N}(x') \right\rangle \right| < +\infty,
\]
so that in particular, we can bound our expression for the derivative of \( t \mapsto f_{\gamma_N(t)}(x) \) using the triangle inequality as

\[
\left| -\tau \left( \nabla L_{\mu_N}(\theta_k^N), \nabla f_{\gamma_N(t)}(x) \right) \right| \leq C \tau \int_M \left| \zeta_{\theta_k^N}(x') \right| \, d\mu_N(x')
\]

for some constant \( C > 0 \). The RHS of the previous bound does not depend on \( t \), so by an application of [219, Theorem 6.3.11], it follows that \( t \mapsto f_{\gamma_N(t)}(x) \) is absolutely continuous, and we have the representation

\[
\zeta_{k+1}^N(x) - \zeta_k^N(x) = -\tau \int_0^1 \left( \int_M \zeta_{\theta_k^N}(x') \left( \nabla f_{\theta_k^N}(x'), \nabla f_{\gamma_N(t)}(x) \right) \, d\mu_N(x') \right) \, dt.
\]

Using (A.77), we can express this as

\[
\zeta_{k+1}^N(x) - \zeta_k^N(x) = -\tau \int_0^1 \left( \int_M \zeta_{\theta_k^N}(x') \left( \nabla f_{\theta_k^N}(x'), \nabla f_{\gamma_N(t)}(x) \right) \, d\mu_N(x') \right) \, dt.
\]

To conclude, it will be convenient to switch the order of integration appearing in the previous expression. Applying (A.78), we have

\[
\left| \zeta_{\theta_k^N}(x') \left( \nabla f_{\theta_k^N}(x'), \nabla f_{\gamma_N(t)}(x) \right) \right| \leq C \left| \zeta_{\theta_k^N}(x') \right|,
\]

and the RHS of this bound is integrable over \([0, 1] \times M\) because the network is a continuous function of the input. By Fubini’s theorem, it follows

\[
\zeta_{k+1}^N(x) - \zeta_k^N(x) = -\tau \int_M \left( \int_0^1 \left( \nabla f_{\theta_k^N}(x'), \nabla f_{\gamma_N(t)}(x) \right) \, dt \right) \zeta_{\theta_k^N}(x') \, d\mu_N(x') \quad (A.79)
\]

Defining

\[
\Theta_k^N(x, x') = \int_0^1 \left( \nabla f_{\theta_k^N}(x'), \nabla f_{\gamma_N(t)}(x) \right) \, dt
\]
and using (A.78), we can define bounded operators \( \Theta^N_k : L^2_{\mu^N}(\mathcal{M}) \to L^2_{\mu^N}(\mathcal{M}) \) by

\[
\Theta^N_k[g](x) = \int_{\mathcal{M}} \Theta^N_k(x, x') g(x') \, d\mu^N(x'),
\]

and with this definition, (A.79) becomes

\[
\zeta^N_{k+1} - \zeta^N_k = -\tau \Theta^N_k \begin{bmatrix} \zeta^N_k \end{bmatrix},
\]

as claimed. \( \square \)

**Lemma A.2.8.** For any network parameters \( \theta \), define kernels

\[
\Theta_{\theta}(x, x') = \left\langle \nabla f_{\theta}(x'), \nabla f_{\theta}(x) \right\rangle,
\]

and for \( \star \in \{N, \infty\} \), define corresponding bounded operators on \( L^2_{\mu^\star}(\mathcal{M}) \) by

\[
\Theta_{\theta,\mu^\star}[g](x) = \int_{\mathcal{M}} \Theta_{\theta}(x, x') g(x') \, d\mu^\star(x').
\]

For any settings of the parameters \( \theta \), the operators \( \Theta_{\theta,\mu^\star} \) are self-adjoint, positive, and compact. In particular, they diagonalize in a countable orthonormal basis of \( L^2_{\mu^\star}(\mathcal{M}) \) functions with corresponding nonnegative eigenvalues.

**Proof.** When \( \star = N \), an identification reduces \( \Theta_{\theta,\mu^\star} \) to operators on finite-dimensional vector spaces, and the claims follow immediately from general principles and the finite-dimensional spectral theorem. We therefore only work out the details for the case \( \star = \infty \). Boundedness follows from an argument identical to the one developed in the proof of Lemma A.2.7, in particular to develop an estimate analogous to (A.78). This estimate, together with separability and compactness of \( \mathcal{M} \), also establishes that \( \Theta_{\theta,\infty} \) is compact, by standard results for Hilbert-Schmidt operators [220, §B].

152
In addition, this estimate allows us to apply Fubini’s theorem to write for any \(g_1, g_2 \in L^2_{\mu^\infty}(M)\)

\[
\langle g_1, \Theta_{\theta, \infty}[g_2] \rangle_{L^2_{\mu^\infty}(M)} = \iint_{M \times M} \Theta_{\theta}(x, x') g_1(x) g_2(x') \, d\mu^\infty(x) \, d\mu^\infty(x') = \langle g_2, \Theta_{\theta, \infty}[g_1] \rangle
\]

since \(\Theta_{\theta}(x, x') = \Theta_{\theta}(x', x)\). A similar calculation establishes positivity: we have for any \(g \in L^2_{\mu^\infty}(M)\)

\[
\langle g, \Theta_{\theta, \infty}[g] \rangle_{L^2_{\mu^\infty}(M)} = \iint_{M \times M} \left( \nabla f_{\theta}(x'), \nabla f_{\theta}(x) \right) g(x) g(x') \, d\mu^\infty(x) \, d\mu^\infty(x')
\]

\[
= \left( \int_M \nabla f_{\theta}(x) g(x) \, d\mu^\infty(x), \int_M \nabla f_{\theta}(x) g(x) \, d\mu^\infty(x) \right) \geq 0,
\]

where we applied Fubini’s theorem and linearity of the integral. These facts and the spectral theorem for self-adjoint compact operators on a Hilbert space imply in particular that the operator \(\Theta_{\theta, \infty}\) can be diagonalized in a countable orthonormal basis of eigenfunctions \((v_i)_{i \in \mathbb{N}} \subset L^2_{\mu^\infty}(M)\) with corresponding nonnegative eigenvalues \((\lambda_i)_{i \in \mathbb{N}} \subset [0, +\infty)\).

\[\square\]

**Lemma A.2.9.** Write \(\Theta_{\mu^N}\) for the operator defined in Lemma A.2.8, with the parameters \(\theta\) set to the initial random network weights and the measure set to \(\mu^N\). There exist absolute constants \(c, K, K' > 0\) such that for any \(q \geq 0\) and any \(d \geq Kd_0 \log(nn_0C_M)\), if

\[
\tau < \frac{1}{\|\Theta_{\mu^N}\|_{\mu^N}(M) \to \mu^N(M)}
\]

and if in addition \(n \geq K'L^{48+20q}d^9 \log^9 L\), then one has

\[
\mathbb{P} \left[ \bigcap_{0 \leq k \leq L^q/(n\tau)} \left\{ \|\xi_k^N\|_{\mu^N(M)} \leq \sqrt{d} \right\} \right] \geq 1 - \left( 1 + \frac{2L^q}{n\tau} \right) e^{-cd}.
\]

**Proof.** Consider the nominal error evolution \(\xi_k^{N, \text{nom}}\), defined iteratively as

\[
\xi_{k+1}^{N, \text{nom}} = \xi_k^{N, \text{nom}} - \tau \Theta_{\mu^N} \left[ \xi_k^{N, \text{nom}} \right];
\]

\[\text{153}\]
\[ \zeta_0^{N,\text{nom}} = \zeta \]

for a step size \( \tau > 0 \), which satisfies

\[ \tau < \frac{1}{\| \Theta_{\mu^N} \|_{L^2_{\mu^N}(M) \to L^2_{\mu^N}(M)}}. \]

We will prove the claim by showing that this auxiliary iteration is monotone decreasing in the loss, and close enough to the gradient-like iteration of interest that we can prove that the gradient-like iteration also retains a controlled loss. These dynamics satisfy the ‘update equation’

\[ \zeta_k^{N,\text{nom}} = \left( \text{Id} - \tau \Theta_{\mu^N} \right)^k [\zeta]. \]

Because \( M \) is compact and \( \zeta \) is a continuous function of the input, we have \( \zeta \in L^\infty(M) \) for all values of the random weights. Because \( \mu^N \) is a probability measure, this means \( \zeta \) has finite \( L^p_{\mu^N}(M) \) norm for every \( p > 0 \). Meanwhile, the choice of \( \tau \) and positivity of the operator (by Lemma A.2.8) guarantees

\[ \| \text{Id} - \tau \Theta_{\mu^N} \|_{L^2_{\mu^N}(M) \to L^2_{\mu^N}(M)} \leq 1, \]

from which it follows from the update equation

\[ \| \zeta_k^{N,\text{nom}} \|_{L^2_{\mu^N}(M)} \leq \| \zeta \|_{L^2_{\mu^N}(M)} \leq \| \zeta \|_{L^\infty(M)}, \quad \text{(A.80)} \]

where the last inequality uses that \( \mu^N \) is a probability measure. In particular, this nominal error evolution is nonincreasing in the relevant loss. Now, we recall the update equation for the finite-sample dynamics

\[ \zeta_{k+1}^N = \left( \text{Id} - \tau \Theta^N \right) [\zeta_k^N], \]

which follows from Lemma A.2.7. Subtracting and rearranging, this gives an update equation for
the difference:

\[
\zeta_{N, k+1} - \zeta_{N, \text{nom}, k+1} = (\text{Id} - \tau \Theta_{\mu^N}) \left( \zeta_{N, k} - \zeta_{N, \text{nom}, k} \right) - \tau \left( \Theta_{k}^{N} - \Theta_{\mu^N}^{N} \right) \left[ \zeta_{k}^{N} \right]. \tag{A.81}
\]

Under our hypothesis on \( \tau \), (A.81) and the triangle inequality imply the bound

\[
\left\| \zeta_{N, k+1} - \zeta_{N, \text{nom}, k+1} \right\|_{L^2_{\mu^N}(M)} \leq \left\| \zeta_{k}^{N} - \zeta_{k}^{N, \text{nom}} \right\|_{L^2_{\mu^N}(M)} + \tau \left\| \zeta_{k}^{N} \right\|_{L^2_{\mu^N}(M)} \left\| \Theta_{k}^{N} - \Theta_{\mu^N}^{N} \right\|_{L^2_{\mu^N}(M) \to L^2_{\mu^N}(M)}.
\]

Using Jensen’s inequality and the Schwarz inequality, we have

\[
\left\| \Theta_{k}^{N} - \Theta_{\mu^N}^{N} \right\|_{L^2_{\mu^N}(M) \to L^2_{\mu^N}(M)} \leq \sup_{\|\|g\|\|_{L^1_{\mu^N}(M)} \leq 1} \int_{M} \left| \Theta_{k}^{N}(\cdot, x') - \Theta(\cdot, x') \right| |g(x')| d\mu^N(x')
\]

\[
\leq \sup_{\|\|g\|\|_{L^1_{\mu^N}(M)} \leq 1} \left\| \Theta_{k}^{N} - \Theta \right\|_{L^\infty(M \times M)} \|g\|_{L^1_{\mu^N}(M)}
\]

\[
\leq \left\| \Theta_{k}^{N} - \Theta \right\|_{L^\infty(M \times M)},
\]

since \( \mu^N \) is a probability measure. Defining

\[
\Delta_{k}^{N} = \max_{i \in \{0, 1, \ldots, k\}} \left\| \Theta_{i}^{N} - \Theta \right\|_{L^\infty(M \times M)},
\]

by a telescoping series and the identical initial conditions, we thus obtain

\[
\left\| \zeta_{N, k+1} - \zeta_{N, \text{nom}, k+1} \right\|_{L^2_{\mu^N}(M)} \leq \tau \Delta_{k}^{N} \sum_{i=0}^{k} \left\| \zeta_{i}^{N} \right\|_{L^2_{\mu^N}(M)},
\]

155
and the triangle inequality and (A.80) then yield

\[ \left\| \xi_{k+1}^N \right\|_{L^2_{\mu_N}(M)} \leq \left\| \xi \right\|_{L^\infty(M)} + \tau \Delta_k^N \sum_{i=0}^{k} \left\| \xi_i^N \right\|_{L^2_{\mu_N}(M)} \cdot \]

Using a discrete version of (the standard) Gronwall’s inequality, the previous bound implies

\[ \left\| \xi_k^N \right\|_{L^2_{\mu_N}(M)} \leq \left\| \xi \right\|_{L^\infty(M)} + \left\| \xi \right\|_{L^\infty(M)} \sum_{i=0}^{k-1} \tau \Delta_{k-1}^N \exp \left( \sum_{j=i+1}^{k-1} \tau \Delta_{k-1}^N \right) \]

\[ \leq \left\| \xi \right\|_{L^\infty(M)} \left( 1 + k \tau \Delta_{k-1}^N \exp \left( k \tau \Delta_{k-1}^N \right) \right) \]  \hspace{1cm} (A.82)

To conclude, we will use [221, Lemma F.5] and an inductive argument based on (A.82). Let us first observe that by Lemma A.4.11 (with a rescaling of \( d \), which worsens the absolute constants), we have

\[ \mathbb{P} \left[ \left\| \xi \right\|_{L^\infty(M)} \leq \frac{\sqrt{d}}{2} \right] \geq 1 - e^{-cd} \]  \hspace{1cm} (A.83)

as long as \( n \geq K d^4 L \) and \( d \geq K'd_0 \log(nn_0C_M) \). Define events \( \mathcal{E}_k^N \) by

\[ \mathcal{E}_k^N = \left\{ \left\| \xi_k^N \right\|_{L^2_{\mu_N}(M)} > \sqrt{d} \right\} \]

where \( d > 0 \) is sufficiently large to satisfy the conditions on \( d \) given above. We are interested in controlling the probability of \( \bigcup_{i=0}^{k} \mathcal{E}_i^N \) for \( k \in \mathbb{N}_0 \). We can write

\[ \mathbb{P} \left[ \bigcup_{i=0}^{k} \mathcal{E}_i^N \right] = \mathbb{P} \left[ \bigcup_{i=0}^{k-1} \mathcal{E}_i^N \right] + \mathbb{P} \left[ \mathcal{E}_k^N \cap \bigcap_{i=0}^{k-1} \left( \mathcal{E}_i^N \right)^c \right] \]

and unraveling, we obtain

\[ \mathbb{P} \left[ \bigcup_{i=0}^{k} \mathcal{E}_i^N \right] = \sum_{i=0}^{k} \mathbb{P} \left[ \mathcal{E}_i^N \cap \bigcap_{j=0}^{i-1} \left( \mathcal{E}_j^N \right)^c \right] \]

In words, it is enough to control the sum of the measures of the parts of \( \mathcal{E}_k^N \) that are common with
the part of the space where none of the past events occurs. First, note that (A.83) implies

$$\mathbb{P} [ \mathcal{E}_0^N ] \leq e^{-cd},$$

and so assume \( i > 0 \) below. For any \( q > 0 \), if \( k \tau \leq L^q / n, n \geq KL^{36+8q} d^9 \) and additionally \( d \geq K'd_0 \log(nn_0C_M), [221, \text{Lemma F.5}] \) gives that there are events \( \mathcal{B}_i^N \) that respectively contain the sets

$$\{ \Delta_{i-1}^N > CL^{4+2q/3} d^{3/4} n^{11/12} \log^{3/4} L \},$$

and which satisfy in addition

$$\mathbb{P} \left[ \mathcal{B}_i^N \cap \bigcap_{j=0}^{i-1} \left( \mathcal{E}_j^N \right)^c \right] \leq e^{-cd}.$$

We thus have by this last bound, a partition, and intersection monotonicity

$$\mathbb{P} \left[ \mathcal{E}_i^N \cap \bigcap_{j=0}^{i-1} \left( \mathcal{E}_j^N \right)^c \right] \leq e^{-cd} + \mathbb{P} \left[ \mathcal{E}_i^N \cap \mathcal{B}_i^N \right],$$

and by construction, one has \( \Delta_{i-1}^N \leq CL^{4+2q/3} d^{3/4} n^{11/12} \log^{3/4} L \) on \( (\mathcal{B}_i^N)^c \). Another partition and (A.83) give

$$\mathbb{P} \left[ \mathcal{E}_i^N \cap \left( \mathcal{B}_i^N \right)^c \right] \leq e^{-cd} + \mathbb{P} \left[ \mathcal{E}_i^N \cap \left( \mathcal{B}_i^N \right)^c \cap \left\{ \| \zeta \|_{L^\infty(M)} \leq \frac{\sqrt{d}}{2} \right\} \right].$$

When the two events on the RHS of the last bound are active, we can obtain using (A.82)

$$\| \zeta_k^N \|_{L^2_{\nu_N^1}(M)} \leq \frac{\sqrt{d}}{2} \left( 1 + k \tau CL^{4+2q/3} d^{3/4} n^{11/12} \log^{3/4} L \exp \left( k \tau CL^{4+2q/3} d^{3/4} n^{11/12} \log^{3/4} L \right) \right).$$
Given that \( k \tau \leq L^q / n \), we have

\[
k \tau C L^{4+2q/3} d^{3/4} n^{11/12} \log^{3/4} L \leq \left( \frac{C^{12} L^{48+20q} d^9 \log^9 L}{n} \right)^{1/12} \leq 1/e,
\]

where the last bound holds provided \( n \geq K L^{48+20q} d^9 \log^9 L \). Thus, on the event

\[
\left( \mathcal{B}_i^N \right)^c \cap \left\{ \| \zeta_N \|_{L^\infty(M)} \leq \frac{\sqrt{d}}{2} \right\},
\]

we have

\[
\| \zeta_k^N \|_{L^2(M)} \leq \sqrt{d},
\]

and thus

\[
P \left[ \mathcal{E}_i^N \cap \left( \mathcal{B}_i^N \right)^c \cap \left\{ \| \zeta_N \|_{L^\infty(M)} \leq \frac{\sqrt{d}}{2} \right\} \right] = 0.
\]

By our previous reductions, we conclude

\[
P \left[ \mathcal{E}_i^N \cap \bigcap_{j=0}^{i-1} \left( \mathcal{E}_j^N \right)^c \right] \leq 2e^{-cd},
\]

and in particular

\[
P \left[ \bigcup_{i=0}^{k} \mathcal{E}_i^N \right] \leq (2k + 1)e^{-cd}.
\]

The claim is then established by taking \( k \) as large as \( L^q / (n\tau) \). \( \square \)

**Corollary A.2.10.** Write \( \Theta_{\mu^N} \) for the operator defined in Lemma A.2.8, with the parameters \( \theta \) set to the initial random network weights \( \theta_0 \) and the measure set to \( \mu^N \), and define for \( k \in \mathbb{N}_0 \)

\[
\Delta_k^N = \max_{i \in \{0, 1, \ldots, k\}} \| \Theta_i^N - \Theta \|_{L^\infty(M \times M)}.
\]

There exist absolute constants \( c, C, C', K, K' > 0 \) such that for any \( q \geq 0 \) and any \( d \) satisfying...
\[ d \geq K_0 d_0 \log(n_0 C_M), \]

if

\[ \tau < \frac{1}{\|\Theta^N\|_{L^2_{\mu^N} (M) \to L^2_{\mu^N} (M)}}. \]

and if in addition

\[ n \geq K' L^4 + 9 d^9 \log^9 L, \]

then one has on an event of probability at least

\[ 1 - C' (1 + L^9 / (n \tau)) e^{-cd}, \]

\[ \Delta^N_{[L^9 / (n \tau)] - 1} \leq C \log^{3/4} (L) d^{3/4} L^{4 + 4q^3} n^{11/12}. \]

**Proof.** Use Lemma A.2.9 to remove the hypothesis about boundedness of the errors from [221, Lemma F.5], then apply this result together with a union bound. \(\square\)

**Lemma A.2.11.** Write \( \Theta \) for the operator defined in Lemma A.2.8, with the parameters \( \theta \) set to the initial random network weights and the measure set to \( \mu^\infty \). Consider the (population) nominal error evolution \( \zeta^\infty_k \), defined iteratively as

\[
\zeta^\infty_{k+1} = \zeta^\infty_k - \tau \Theta [\zeta^\infty_k]; \\
\zeta^\infty_0 = \zeta
\]

for a step size \( \tau > 0 \), which satisfies

\[ \tau < \frac{1}{\|\Theta\|_{L^2_{\mu^\infty} (M) \to L^2_{\mu^\infty} (M)}}. \]

Then for any \( g \in L^2_{\mu^\infty} (M) \) and any \( k \) satisfying

\[ k \tau \geq \sqrt{\frac{3e}{2} \|g\|_{L^2_{\mu^\infty} (M)}}, \]

we have

\[
\|\zeta^\infty_k\|_{L^2_{\mu^\infty} (M)} \leq \sqrt{3} \|\Theta[g] - \zeta\|_{L^2_{\mu^\infty} (M)} - \frac{3\|g\|_{L^2_{\mu^\infty} (M)}}{k \tau} \log \left( \sqrt{\frac{3}{2} \|g\|_{L^2_{\mu^\infty} (M)} k \tau} \right). \]
Proof. The dynamics satisfy the ‘update equation’

\[ \zeta_k^\infty = (\text{Id} - \tau \Theta)^k [\zeta] . \]

Because \( M \) is compact and \( \zeta \) is a continuous function of the input, we have \( \zeta \in L^\infty(M) \) for all values of the random weights. Because \( \mu^\infty \) is a probability measure, this means \( \zeta \) has finite \( L_p^\mu(M) \) norm for every \( p > 0 \). Using the eigendecomposition of \( \Theta \) as developed in Lemma A.2.8, we can therefore write

\[ \zeta = \sum_{i=0}^\infty \langle v_i, \zeta \rangle_{L_{\mu^\infty}^2(M)} v_i \]

in the sense of \( L_{\mu^\infty}^2(M) \). Because \( \Theta \) and \( \text{Id} - \tau \Theta \) diagonalize simultaneously, we obtain

\[ \| \zeta_k^\infty \|^2_{L_{\mu^\infty}^2(M)} = \sum_{i=1}^\infty (1 - \tau \lambda_i)^{2k} \langle v_i, \zeta \rangle^2_{L_{\mu^\infty}^2(M)} \leq \sum_{i=1}^\infty e^{-2k \tau \lambda_i} \langle v_i, \zeta \rangle^2_{L_{\mu^\infty}^2(M)}, \]

where the inequality follows from the elementary estimate \( 1 - x \leq e^{-x} \) for \( x \geq 0 \) and our choice of \( \tau \), which guarantees that \( 1 - \tau \lambda_i > 0 \) for all \( i \in \mathbb{N} \) so that the elementary estimate is valid after squaring. We can split this last sum into two parts: for any \( \lambda \in \mathbb{R} \), we have

\[ \| \zeta_k^\infty \|^2_{L_{\mu^\infty}^2(M)} = \sum_{i : \lambda_i \geq \lambda} e^{-2k \tau \lambda_i} \langle v_i, \zeta \rangle^2_{L_{\mu^\infty}^2(M)} + \sum_{i : \lambda_i < \lambda} e^{-2k \tau \lambda_i} \langle v_i, \zeta \rangle^2_{L_{\mu^\infty}^2(M)}. \]

Because \( \Theta \) is positive, we have further that \( \lambda_i \geq 0 \) for all \( i \), so we can take \( \lambda \geq 0 \). The first sum consists of large eigenvalues: we use \( \exp(-2k \tau \lambda_i) \leq \exp(-2k \tau \lambda) \) to preserve their effect, and then upper bound the remainder of the sum by the squared \( L_{\mu^\infty}^2 \) norm of \( \zeta \). The second sum consists of small eigenvalues: we replace \( \exp(-2k \tau \lambda_i) \leq 1 \), and then plug in \( \zeta = \Theta[g] - (\Theta[g] - \zeta) \) and use bilinearity, self-adjointness of \( \Theta \), and the triangle inequality to get

\[ \left| \langle v_i, \zeta \rangle_{L_{\mu^\infty}^2(M)} \right| \leq \lambda \left| \langle v_i, g \rangle_{L_{\mu^\infty}^2(M)} \right| + \left| \langle v_i, \Theta[g] - \zeta \rangle_{L_{\mu^\infty}^2(M)} \right|. \]

We then square both (nonnegative) sides of the inequality and use Cauchy-Schwarz to replace the
squared sum with the sum of squares times a constant, obtaining

\[ \left\| \xi^\infty \right\|_{L^2_{\mu^\infty}(M)}^2 \leq e^{-2k\tau_1} \left\| \xi \right\|_{L^2_{\mu^\infty}(M)}^2 + 3t^2 \left\| g \right\|_{L^2_{\mu^\infty}(M)}^2 + 3\left\| \Theta [g] - \xi \right\|_{L^2_{\mu^\infty}(M)}^2 \]

after re-adding indices \( i \) to the sum to obtain the third residual. We will choose \( \lambda \geq 0 \) to minimize the sum of the first and second terms. Differentiating and setting to zero gives the critical point equation

\[ \frac{2}{3} \left( \frac{\left\| \xi \right\|_{L^2_{\mu^\infty}(M)}^2}{\left\| g \right\|_{L^2_{\mu^\infty}(M)}^2} \right) (k\tau)^2 = (2t\lambda) e^{2k\tau_1}, \]

which can be inverted to give the unique critical point

\[ \lambda = \frac{1}{2k\tau} W \left( \frac{2}{3} \frac{\left\| \xi \right\|_{L^2_{\mu^\infty}(M)}^2}{\left\| g \right\|_{L^2_{\mu^\infty}(M)}^2} (k\tau)^2 \right), \]

where \( W \) is the Lambert \( W \) function, defined as the principal branch of the inverse of \( z \mapsto ze^z \); we know that this critical point is a minimizer because the function of \( \lambda \) we differentiated diverges as \( \lambda \to \infty \). Plugging this point into the sum of the first two terms gives

\[ \left\| \xi^\infty \right\|_{L^2_{\mu^\infty}(M)}^2 \leq \left( \frac{3}{2} \right) \left( \frac{1}{2} \right) W \left( \frac{2}{3} \frac{\left\| \xi \right\|_{L^2_{\mu^\infty}(M)}^2}{\left\| g \right\|_{L^2_{\mu^\infty}(M)}^2} (k\tau)^2 \right) \frac{\left\| \xi \right\|_{L^2_{\mu^\infty}(M)}^2}{\left\| g \right\|_{L^2_{\mu^\infty}(M)}^2} \left( \frac{1}{2} \right) \frac{\left\| \xi \right\|_{L^2_{\mu^\infty}(M)}^2}{\left\| g \right\|_{L^2_{\mu^\infty}(M)}^2} \left( \frac{3}{2} \right). \]

For \( x \geq 0 \), the function \( x \mapsto W(x) \) is strictly increasing, as the inverse of \( y \mapsto ye^y \); by definition \( W(e) = 1 \); and we have the representation \( W(z) + \log W(z) = \log z \) \( \{222\} \), whence \( W(x) \leq \log x \) if \( x \geq e \). Because \( \mu^\infty \) is a probability measure, we have

\[ \left\| \xi \right\|_{L^2_{\mu^\infty}(M)}^2 \leq \left\| \xi \right\|_{L^\infty}^2, \]
and therefore if
\[ k \tau \geq \sqrt{\frac{3e}{2}} \frac{\|g\|_{L^2_{\mu^\infty(M)}}}{\|\xi\|_{L^\infty(M)}}, \]
we can simplify the previous bound to
\[ \|\xi_k\|_{L^2_{\mu^\infty(M)}}^2 \leq 3\|\Theta[g] - \xi\|_{L^2_{\mu^\infty(M)}}^2 + \frac{9\|g\|_{L^2_{\mu^\infty(M)}}^2}{4(k\tau)^2} \log^2 \left( \frac{3}{2} \frac{\|g\|_{L^2_{\mu^\infty(M)}}}{\|\xi\|_{L^\infty(M)}} (k\tau)^2 \right), \]
using also properties of the logarithm. Taking square roots and using the Minkowski inequality then yields
\[ \|\xi_k\|_{L^2_{\mu^\infty(M)}} \leq \sqrt{3}\|\Theta[g] - \xi\|_{L^2_{\mu^\infty(M)}} - \frac{3\|g\|_{L^2_{\mu^\infty(M)}}}{k\tau} \log \left( \sqrt{3} \frac{\|g\|_{L^2_{\mu^\infty(M)}}}{\|\xi\|_{L^\infty(M)}} (k\tau) \right), \]
where we used the previous lower bound on $k \tau$ to determine the sign that the absolute value of the logarithm takes. This gives the claim.

\[ \square \]

**Lemma A.2.12 (Kantorovich-Rubinstein Duality).** Let $\text{Lip}(M)$ denote the class of functions $f : M \to \mathbb{R}$ such that both $f \big|_{M_\pm}$ are Lipschitz with respect to the Riemannian distances on $M_\pm$. For any $d \geq 1$, any $0 < \delta \leq 1$ and any $N \geq 2\sqrt{d}/\min\{\mu^\infty(M_+), \mu^\infty(M_-)\}$, one has that on an event of probability at least $1 - 8e^{-d}$, simultaneously for all $f \in \text{Lip}(M)$
\[ \left| \int_M f(x) \, d\mu^\infty(x) - \int_M f(x) \, d\mu^N(x) \right| \leq 2\frac{\|f\|_{L^\infty(M)}}{N} \sqrt{d} + \frac{e^{14/\delta} C_{\mu^\infty,M} \sqrt{d} \max_{* \in \{+,-\}} \|f|_{M_*}\|_{\text{Lip}}}{N^{1/(2+\delta)}}, \]
where
\[ C_{\mu^\infty,M} = \frac{\text{len}(M_+)}{\mu^\infty(M_+)} + \frac{\text{len}(M_-)}{\mu^\infty(M_-)}. \]

**Proof.** The proof is an application of Kantorovich-Rubinstein duality for the 1-Wasserstein distance [93, eq. (1)], which states that for any two Borel probability measures $\mu, \nu$ on $M_\pm$, one...
has
\[ W(\mu, \nu) = \sup_{\|f\|_{\text{Lip}} \leq 1} \left| \int_{M_\pm} f(x) \, d\mu(x) - \int_{M_\pm} f(x) \, d\nu(x) \right|, \]
where \( M_\pm \) denotes either of \( M_+ \) or \( M_- \), and \( \| \cdot \|_{\text{Lip}} \) is the minimal Lipschitz constant with respect to the Riemannian distance on \( M_\pm \). Therefore for any \( f : M_\pm \to \mathbb{R} \) Lipschitz, we have
\[ \left| \int_{M_\pm} f(x) \, d\mu(x) - \int_{M_\pm} f(x) \, d\nu(x) \right| \leq \|f\|_{\text{Lip}} W(\mu, \nu), \tag{A.84} \]
where one checks separately the case where \( \|f\|_{\text{Lip}} = 0 \) to see that this bound holds there as well.

To go from (A.84) to the desired conclusion, we need to pass from the measures \( \mu_\infty \) and \( \mu_N \), both supported on \( M_\star \) (with \( \star \in \{N, \infty\} \)), supported on the manifolds \( M_\pm \) (which we will define in detail below); the challenge here is that the number of ‘hits’ of each manifold \( M_\pm \) that show up in the finite sample measure \( \mu_N \) is a random variable, which requires a small detour to control. Let us define random variables \( N_+, N_- \) by
\[ N_+ = N\mu_N(M_+) \quad \text{and} \quad N_- = N\mu_N(M_-), \]
so that \( N_\pm \) have support in \( \{0, 1, \ldots, N\} \), and \( N_+ + N_- = N \). Define in addition
\[ p_+ = \mu_\infty(M_+) \quad \text{and} \quad p_- = \mu_\infty(M_-), \]
which represent the degree of imbalance between the positive and negative classes in the data. By definition of the i.i.d. sample, we have that \( N_+ \sim \text{Binom}(N, p_+) \). Using \( N_+ \) and \( N_- \), we can define ‘conditional’ finite sample measures \( \mu^N_+ \) and \( \mu^N_- \) by
\[ \mu^N_+ = \frac{1}{\max\{1, N_+\}} \sum_{i \in [N] : x_i \in M_+} \delta_{(x_i)}; \quad \mu^N_- = \frac{1}{\max\{1, N_-\}} \sum_{i \in [N] : x_i \in M_-} \delta_{(x_i)}, \]
so that \( (N_+/N)\mu^N_+ + (N_-/N)\mu^N_- = \mu^N \), and \( \mu^N_+ \) and \( \mu^N_- \) are both probability measures except when
\[ ^4 \text{Here we treat the empty sum as the appropriate ‘zero element’ of the space of finite signed Borel measures on} \]

163
\( N_+ \in \{0, N\} \), in which case exactly one is a probability measure. By the triangle inequality, we have for any continuous \( f : \mathcal{M} \to \mathbb{R} \)

\[
\left| \int_{\mathcal{M}} f(x) \, d\mu^\infty(x) - \int_{\mathcal{M}} f(x) \, d\mu^N(x) \right| \\
\leq \sum_{* \in \{+,-\}} \left| p_* \int_{\mathcal{M}_*} f(x) \, d\mu^\infty_*(x) - \frac{N_*}{N} \int_{\mathcal{M}_*} f(x) \, d\mu^N_*(x) \right| \\
\leq \sum_{* \in \{+,-\}} \| f \|_{L^\infty(\mathcal{M})} \left| \frac{N_*}{N} - p_* \right| + \left| \int_{\mathcal{M}_*} f(x) \, d\mu^\infty_*(x) \right| - \left| \int_{\mathcal{M}_*} f(x) \, d\mu^N_*(x) \right|. \tag{A.85}
\]

By Lemma A.6.1, we have

\[
\mathbb{P} \left[ \left| \frac{N_*}{N} - p_* \right| \leq \frac{\sqrt{d}}{N} \right] \geq 1 - 2e^{-2d}. \tag{A.86}
\]

Using that \( N - N_+ = N_- \) and \( 1 - p_+ = p_- \), the bound (A.86) implies if \( N \geq 2\sqrt{d}/\min\{p_+, p_-\} \)

\[
\mathbb{P} \left[ \frac{p_*}{2} \leq \frac{N_*}{N} \leq \frac{1 - p_*}{2} \right] \geq 1 - 2e^{-2d}. \tag{A.87}
\]

Now fix an arbitrary \( f \in \operatorname{Lip}(\mathcal{M}) \). For either \( * \in \{+,-\} \), we can write

\[
\int_{\mathcal{M}_*} f(x) \, d\mu_*(x) = \frac{1}{\max\{1, N_*\}} \sum_{i: x_i \in \mathcal{M}_*} f(x_i) = \frac{1}{\max\{1, \sum_{i=1}^N 1_{x_i \in \mathcal{M}_*}\}} \sum_{i=1}^N 1_{x_i \in \mathcal{M}_*} f(x_i),
\]

and since \( \mathcal{M}_+ \) and \( \mathcal{M}_- \) are separated by a positive distance \( \Delta > 0 \), we have that \( x_i \mapsto 1_{x_i \in \mathcal{M}_*} \) are continuous functions on \( \mathcal{M} \). Since \( f \) is continuous on \( \mathcal{M} \) by the same reasoning and the fact that \( \mathcal{M} \) is compact, it follows that the functions \( (x_1, \ldots, x_N) \mapsto \int_{\mathcal{M}_*} f(x) \, d\mu_*(x) \) are continuous on \( \mathcal{M} \times \cdots \times \mathcal{M} \) as well, and in particular for any \( t > 0 \) the sets

\[
\left\{ \left| \int_{\mathcal{M}_*} f(x) \frac{d\mu^\infty_*(x)}{p_*} - \int_{\mathcal{M}_*} f(x) \, d\mu^N_*(x) \right| > t \right\}
\]

are open in \( \mathcal{M} \), and so is their union over all \( f \in \operatorname{Lip}(\mathcal{M}) \). By conditioning, we can then apply \( \mathcal{M}_d \), namely the trivial measure that assigns zero to every Borel subset of \( \mathcal{M}_d \).
(A.87) to write

\[
\mathbb{P} \left[ \bigcup_{f \in \text{Lip}(M)} \left\{ \left| \int_{M} f(x) \frac{d\mu^{\infty}_{\star}(x)}{p_{\star}} - \int_{M} f(x) d\mu_{\star}^{N}(x) \right| > t \right\} \right]
\leq 2e^{-2d} + \sum_{k=[Np_{\star}/2]}^{[N(1-p_{\star})/2]} \mathbb{P} \left[ \bigcup_{f \in \text{Lip}(M)} \left\{ \left| \int_{M} f(x) \frac{d\mu^{\infty}_{\star}(x)}{p_{\star}} - \int_{M} f(x) d\mu_{\star}^{N}(x) \right| > t \right\} \bigg| N_{\star} = k \right] \mathbb{P}[N_{\star} = k].
\]  

(A.88)

Conditioned on \{N_{\star} = k\} with 0 < k < N, the measure \(\mu_{\star}^{N}\) is distributed as an empirical measure of sample size \(k\) from the probability measure \(\mu^{\infty}_{\star}/p_{\star}\) supported on \(M_{\star}\). For \([Np_{\star}/2] \leq k \leq [N(1-p_{\star})/2]\), any \(\delta > 0\) and any \(d \geq 1\) we have for both possible values of \(\star\)

\[
\sqrt{d}e^{14/\delta} \frac{\text{len}(M_{\star})}{k^{1/(2+\delta)}} \leq \frac{\sqrt{d}e^{14/\delta} \text{len}(M_{\star})}{\left(\left\lfloor Np_{\star}/2 \right\rfloor \right)^{1/(2+\delta)}} \leq \frac{\sqrt{2d}e^{14/\delta} \text{len}(M_{\star})}{(Np_{\star})^{1/(2+\delta)}},
\]

and so an application of Lemma A.2.15 thus gives for any \(0 < \delta \leq 1\) and any \(d \geq 2\)

\[
\mathbb{P} \left[ \mathcal{W} \left( \frac{d\mu^{\infty}_{\star}(x)}{p_{\star}}, d\mu_{\star}^{N} \right) > \sqrt{d}e^{14/\delta} \frac{\text{len}(M_{\star})}{(Np_{\star})^{1/(2+\delta)}} \bigg| N_{\star} = k \right] \leq e^{-d}.
\]

Combining this last bound with (A.84) and (A.88) gives

\[
\mathbb{P} \left[ \bigcup_{f \in \text{Lip}(M)} \left\{ \left| \int_{M} f(x) \frac{d\mu^{\infty}_{\star}(x)}{p_{\star}} - \int_{M} f(x) d\mu_{\star}^{N}(x) \right| > \sqrt{d}e^{14/\delta} \frac{\left\| f_{\star} \right\|_{\text{Lip}(M_{\star})}}{N^{1/(2+\delta)}} \frac{\text{len}(M_{\star})}{p_{\star}} \right\} \right] \leq 3e^{-d}.
\]

where we used \(\max\{p_{+}, p_{-}\} \leq 1\) to remove the exponent of \(1/(2 + \delta)\) on these terms. Taking a max over the Lipschitz constants and combining this bound with (A.86) and (A.85) and a union
bound, we obtain

$$\mathbb{P} \left[ \bigcup_{f \in \text{Lip}(M)} \left\{ \frac{1}{M} \left( \int_M f(x) \, d\mu^\infty - \int_M f(x) \, d\mu^N \right) \right\} > \frac{2\|f\|_{L^\infty(M)} \sqrt{d}}{N} + e^{14/\delta} C_{\mu^\infty, M} \sqrt{d} \max_{\star \in \{+, -\}} \|f|_{M^\star}\|_{\text{Lip}} \right \} \leq 8e^{-d},$$

where the constant is defined as in the statement of the lemma. □

**Lemma A.2.13.** Let $d_0 = 1$. There is an absolute constant $C > 0$ such that for any function $f : M \to \mathbb{R}$ with $f|_{M^\pm}$ Lipschitz with respect to the Riemannian distances on $M^\pm$, one has

$$\|f\|_{L^\infty} \leq C \max \left\{ \frac{\rho^{1/2}\|f\|_{L^2_{\mu^\infty}(M)}}{\rho_{\min}^{1/2}(\min \{\mu^\infty(M^+), \mu^\infty(M^-)\})^{1/2}}, \frac{\|f\|_{L^2_{\mu^\infty}(M)}^{2/3} \max_{\star \in \{+, -\}} \|f|_{M^\star}\|_{\text{Lip}}^{1/3}}{\rho_{\min}^{1/3}} \right\}. \tag{A.89}$$

**Proof.** For any $T > 0$ and a nonconstant Lipschitz function $f : [0, T] \to \mathbb{R}$, we will establish the inequality

$$\|f\|_{L^\infty} \leq C \max \left\{ \frac{\|f\|_{L^2}}{\sqrt{T}}, \frac{\|f\|_{L^2}^{2/3} \|f\|_{\text{Lip}}^{1/3}}{\rho_{\min}^{1/3}} \right\}, \tag{A.89}$$

where the constant $C > 0$ is absolute. We can use this result to establish the claim. We start by writing

$$\|f\|_{L^\infty} = \max_{\star \in \{+, -\}} \|f|_{M^\star}\|_{L^\infty};$$

and for $\star \in \{+, -, \}$, we have

$$\|f|_{M^\star}\|_{L^\infty} = \|f \circ \gamma_\star\|_{L^\infty}, \tag{A.90}$$

where $\gamma_\star : [0, \text{len}(M^\star)] \to M^\star$ are the smooth unit-speed curves parameterized with respect to arc length parameterizing the manifolds. Similarly, the curves’ parameterization with respect to arc length implies

$$\|f \circ \gamma_\star\|_{\text{Lip}} \leq \|f|_{M^\star}\|_{\text{Lip}}, \tag{A.91}$$

166
Applying (A.89) with (A.90) and (A.91), we obtain

\[
\|f\|_{L^\infty} \leq C \max \left\{ \frac{\|f \circ \gamma\|_{L^2}}{\sqrt{\text{len}(M_\star)}}, \|f\|_{L^2}^{2/3} \|f\|_{L^\infty}^{1/3}, \|\gamma\|_{L^2}^{2/3} \|\gamma\|_{L^\infty}^{1/3} \right\}.
\]

For the first term in the max, we have

\[
\frac{\|f \circ \gamma\|_{L^2}}{\text{len}(M_\star)} = \left| \frac{1}{\text{len}(M_\star)} \int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 \, dt \right|
\leq \int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 \rho_\star \circ \gamma_\star(t) \, dt
\leq \int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 \rho_\star \circ \gamma_\star(t) \, dt + \int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 \rho_\star \circ \gamma_\star(t) \, dt
\leq \|f\|_{L^2}^2 \mu^\infty(M),
\]

using the triangle inequality. For the second term in the last bound, we note that

\[
\int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 \rho_\star \circ \gamma_\star(t) \, dt
\leq \int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 \rho_\star \circ \gamma_\star(t) \, dt + \int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 \rho_\star \circ \gamma_\star(t) \, dt
\leq \|f\|_{L^2}^2 \mu^\infty(M),
\]

and for the first term, we have

\[
\max_{t \in [0, \text{len}(M_\star)]} \left| \frac{\rho_\star \circ \gamma_\star(t) - \frac{1}{\text{len}(M_\star)}}{\rho_\star \circ \gamma_\star(t)} \right|
\leq \max_{t \in [0, \text{len}(M_\star)]} \left| \frac{\rho_\star \circ \gamma_\star(t) - \frac{\mu^\infty(M_\star)}{\mu^\infty(M_\star)}}{\mu^\infty(M_\star)} \right|
\leq \frac{1 - \mu^\infty(M_\star)}{\mu^\infty(M_\star)} + \frac{\rho_\max}{\mu^\infty(M_\star)} \mu^\infty(M_\star)
\leq \frac{2\rho_\max}{\mu^\infty(M_\star) \rho_\min},
\]

(A.92)
where in the first inequality we used the triangle inequality, and for the second we used that $\rho_\star \circ \gamma_\star$ integrates to $\mu^\infty(M_\star)$ over $[0, \text{len}(M_\star)]$, which implies that there exists at least one $t \in [0, \text{len}(M_\star)]$ at which $\rho_\star \circ \gamma_\star(t) \geq \mu^\infty(M_\star)/\text{len}(M_\star)$, so that the maximum of the difference in the second term on the RHS of the first inequality is bounded by the maximum of the density term. Thus, by Hölder’s inequality and (A.92) and (A.93), we have

$$
\left| \int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 \rho_\star \circ \gamma_\star(t) - \frac{1}{\text{len}(M_\star)} \right| \rho_* \circ \gamma_\star(t) dt 
\leq \frac{3\rho_{\max}}{\rho_{\min} \min \{\mu^\infty(M_+, \mu^\infty(M_-)\}} \|f\|_{L^2_{\mu^\infty}(M)}^2.
$$

Similarly, for the second term in the max, we have

$$
\|f \circ \gamma_\star\|_{L^2_{\mu^\infty}(M)}^{2/3} 
= \left( \int_0^{\text{len}(M_\star)} f \circ \gamma_\star(t)^2 dt \right)^{1/3} 
\leq \left( \int_0^{\text{len}(M_+)} f \circ \gamma_\star(t)^2 dt + \int_0^{\text{len}(M_-)} f \circ \gamma_\star(t)^2 dt \right)^{1/3} 
\leq \frac{1}{\rho_{\max}^{1/3}} \left( \int_0^{\text{len}(M_+)} f \circ \gamma_\star(t)^2 \rho_+ \circ \gamma_\star(t) dt + \int_0^{\text{len}(M_-)} f \circ \gamma_\star(t)^2 \rho_- \circ \gamma_\star(t) dt \right)^{1/3} 
\leq \frac{\|f\|_{L^2_{\mu^\infty}(M)}^{2/3}}{\rho_{\min}^{1/3}}.
$$

Thus, we have

$$
\|f\|_{L^\infty} \leq C \max \left\{ \rho_{\max}^{1/2} \|f\|_{L^2_{\mu^\infty}(M)} \rho_{\min}^{1/2} \min \{\mu^\infty(M_+, \mu^\infty(M_-)\})^{1/2}, \rho_{\max} \|f\|_{L^2_{\mu^\infty}(M)} \|f\|_{L^\infty} \rho_{\min}^{1/3} \right\},
$$

and taking a maximum over $\star \in \{+, -, \}$ establishes the claim.

To prove (A.89), consider first the trivial case where $\|f\|_{L^\infty} = 0$: here the LHS and RHS of (A.89) are identical, and the proof is immediate. When $\|f\|_{L^\infty} > 0$, the Weierstrass theorem
implies that there exists $t \in [0, T]$ such that $|f(t)| = \|f\|_{L^\infty}$; we consider the case $\text{sign}(f(t)) > 0$. For any $t' \in [0, T]$, we can write by the Lipschitz property

$$f(t') \geq \|f\|_{L^\infty} - \|f\|_{\text{Lip}} |t - t'|,$$

and the RHS of the previous bound is nonnegative on the intersection of the interval

$$[t - \|f\|_{L^\infty}\|f\|_{\text{Lip}}^{-1}, t + \|f\|_{L^\infty}\|f\|_{\text{Lip}}^{-1}]$$

with the domain $[0, T]$ (with standard extended-valued arithmetic conventions when $\|f\|_{\text{Lip}} = 0$). This gives the bound

$$\|f\|_{L^2}^2 \geq \int \min \left\{ t + \frac{\|f\|_{L^\infty}}{\|f\|_{\text{Lip}}}, T \right\} \left( \|f\|_{L^\infty} - \|f\|_{\text{Lip}} |t - t'| \right)^2 \, dt'$$

$$= \int \min \left\{ \frac{\|f\|_{L^\infty}}{\|f\|_{\text{Lip}}}, T - t \right\} \left( \|f\|_{L^\infty} - \|f\|_{\text{Lip}} |t'| \right)^2 \, dt',$$

where the second line follows from the changes of variables $t' \mapsto t' + t$. The integrand on the RHS of the second line in the previous bound is even-symmetric, and $\max \left\{ -\frac{\|f\|_{L^\infty}}{\|f\|_{\text{Lip}}}, -t \right\} = -\min \left\{ \frac{\|f\|_{L^\infty}}{\|f\|_{\text{Lip}}}, t \right\}$, so we can discard one side of the interval of integration to get

$$\int \min \left\{ \frac{\|f\|_{L^\infty}}{\|f\|_{\text{Lip}}}, T - t \right\} \left( \|f\|_{L^\infty} - \|f\|_{\text{Lip}} |t'| \right)^2 \, dt' \quad \text{(A.94)}$$

$$\geq \int_{0}^{\min \{ \frac{\|f\|_{L^\infty}}{\|f\|_{\text{Lip}}}, \max \{ t, T - t \} \}} \left( \|f\|_{L^\infty} - \|f\|_{\text{Lip}} |t'| \right)^2 \, dt'. \quad \text{(A.95)}$$

We proceed analyzing two distinct cases. First, if $\|f\|_{L^\infty} \leq \max \{ t, T - t \} \|f\|_{\text{Lip}}$, then we must have $\|f\|_{\text{Lip}} > 0$; integrating the RHS of (A.95), we obtain

$$\|f\|_{L^2}^2 \geq \frac{\|f\|_{L^\infty}^3}{3\|f\|_{\text{Lip}}}.$$
or equivalently
\[ \|f\|_{L^\infty} \leq 3^{1/3} \|f\|_{L^2}^{2/3} \|f\|_{\text{Lip}}^{1/3}. \]  
(A.96)

Next, we consider the case \( \|f\|_{L^\infty} > \max\{t, T - t\} \|f\|_{\text{Lip}} \). We split on two sub-cases: when \( \|f\|_{\text{Lip}} = 0 \), integrating (A.95) gives
\[ \|f\|_{L^2}^2 \geq \|f\|_{L^\infty}^2 \max\{t, T - t\} \geq \frac{T \|f\|_{L^\infty}^2}{2}, \]  
(A.97)

where we used \( \max\{t, T - t\} \geq T/2 \). When \( \|f\|_{\text{Lip}} > 0 \), integrating (A.95) gives
\[ \|f\|_{L^2}^2 \geq \frac{1}{3\|f\|_{\text{Lip}}} \left( \|f\|_{L^\infty}^3 - (\|f\|_{L^\infty} - \|f\|_{\text{Lip}} \max\{t, T - t\})^3 \right) \]
\[ = \frac{\max\{t, T - t\}}{3} \sum_{k=0}^{2} \|f\|_{L^\infty}^{2-k} (\|f\|_{L^\infty} - \|f\|_{\text{Lip}} \max\{t, T - t\})^k \]
\[ \geq \frac{T \|f\|_{L^\infty}^2}{6}, \]  
(A.98)

where the second line uses a standard algebraic identity, and the third line uses \( \max\{t, T - t\} \geq T/2 \) together with the definition of the case to get that \( \|f\|_{L^\infty} - \|f\|_{\text{Lip}} \max\{t, T - t\} > 0 \) in order to discard all but the \( k = 0 \) summand. Combining (A.97) and (A.98), we obtain for this case
\[ \|f\|_{L^\infty} \leq \frac{\sqrt{6}\|f\|_{L^2}}{\sqrt{T}}, \]  
(A.99)

and combining (A.96) and (A.99) gives unconditionally
\[ \|f\|_{L^\infty} \leq \max \left\{ \frac{\sqrt{6}\|f\|_{L^2}}{\sqrt{T}}, 3^{1/3} \|f\|_{L^2}^{2/3} \|f\|_{\text{Lip}}^{1/3} \right\}, \]

which establishes (A.89). For the case \( \text{sign}(f(t)) < 0 \), apply the preceding argument to \(-f\) to conclude. See [223, Exercise 8.15] for a sketch of a proof that leads to more general versions of (A.89).
Lemma A.2.14. For any \( p \in \mathbb{N} \), if \( C \geq (4p)^{4p} \), then one has

\[
n \geq C \log^p n \quad \text{if} \quad n \geq 2^p C \log^p (2^p C).
\]

Proof. We first give a proof for \( p = 1 \), then build off this proof for the general case. Consider the function \( f(x) = cx - \log x \). We have \( f'(x) = c - 1/x \), which is nonnegative for every \( x \geq 1/c \), so in particular \( f \) is increasing under this condition. By concavity of the logarithm, we have

\[
\log x \leq \log(2/c) + (c/2)(x - 2/c),
\]

whence

\[
f(x) \geq 1 + cx/2 - \log(2/c).
\]

The RHS of this bound is equal to zero at \( x = (2/c)(\log(2/c) - 1) \), and

\[
\frac{2}{c} \left( \log \left( \frac{2}{c} \right) - 1 \right) \geq \frac{1}{c} \quad \iff \quad c \leq 2e^{-3/2}.
\]

In particular, we have \( f(x) \geq 0 \) for every \( x \geq (2/c) \log(2/c) \). Rearranging this bound, we can assert the desired conclusion that if \( C \geq 3 \), then \( n \geq C \log n \) for every \( n \geq 2C \log 2C \). Equivalently, we have for all such \( n \) that \( Cn^{-1} \log n \leq 1 \). Next, we consider the case of \( p > 1 \). We will show

\[
C \frac{\log^p n}{n} \leq 1
\]

under suitable conditions. Let us consider the choice \( n = KC \log^p KC \), where \( K > 0 \) is a constant we will specify below. Consider the function \( f(x) = Cx^{-1} \log^p x \), which satisfies

\[
f'(x) = C \frac{\log^{-1}(x)(p - \log^{-1}(x))}{x^2}.
\]

In particular, \( f \) is decreasing as soon as \( p \leq \log^{-1}(x) \). Now, we can calculate

\[
f(KC \log^p KC) = \frac{1}{K} \left( 1 + p \frac{\log \log KC}{\log KC} \right)^p.
\]
and by our result for the case $p = 1$, we have for all $p \geq 2$

$$\frac{p \log \log KC}{\log KC} \leq 1 \quad \text{if} \quad \log KC \geq 4p \log 4p.$$ 

This condition is satisfied for $KC \geq (4p)^4p$, so if we set $K = 2^p$, we obtain the above conclusion when $C \geq (4p)^4p$. Under these conditions, we then get

$$f(KC \log^p KC) \leq 1.$$ 

Similarly, we have $\log^{p-1}(KC \log KC) \geq \log^{p-1}((4p)^4p) = (4p)^{p-1} \log^{p-1}(4p)$, which is larger than $p$ because $4p \geq e$. It follows that $f(x) \leq 1$ for every $x \geq KC \log KC$, which completes the proof. \qed

**Lemma A.2.15** (Concentration of Empirical Measure in Wasserstein Distance [93]). Let $d_0 = 1$. For either $\star \in \{+, -\}$, let $\mu$ be a Borel probability measure on $M_\star$, and write $\mu^N$ for the empirical measure corresponding to $N$ i.i.d. samples from $\mu$. Then for any $d \geq 1$ and any $0 < \delta \leq 1$, one has

$$P \left[ W(\mu, \mu^N) \leq \frac{\sqrt{d} e^{14/\delta} \log(M_\star)}{N^{1/(2+\delta)}} \right] \geq 1 - e^{-2d},$$

where the 1-Wasserstein distance is taken with respect to the Riemannian distance.

**Proof.** The proof is a direct application of the results of [93] on concentration of empirical measures in Wasserstein distance. For the duration of the proof, we will work on the metric space $(M_\star, \log(M_\star)^{-1} \text{dist}_{M_\star}(\cdot, \cdot))$, i.e., the same metric space scaled to have unit diameter; we will then obtain the result in terms of the unscaled metric by the definition of the 1-Wasserstein distance.

Because $d_0 = 1$ and $M_\star$ can be given as a unit-speed curve parameterized with respect to arc length, we have for any Borel $S \subset [0, 1]$ and any $\varepsilon > 0$

$$N_\varepsilon(S) \leq \frac{1}{\varepsilon},$$

172
where \( \mathcal{N}_\varepsilon(S) \) denotes the \( \varepsilon \)-covering number of \( S \) by closed balls in the rescaled metric. Following the notation of [93, \S 4.1], we then obtain for any \( s > 2 \)

\[
d_\varepsilon(\mu, \varepsilon^{s/(s-2)}) = \log \inf \left\{ \mathcal{N}_\varepsilon(S) \left| \mu(S) \geq 1 - \varepsilon^{s/(s-2)} \right. \right\} - \log \varepsilon \leq 1.
\]

Invoking [93, Proposition 5], we obtain after some simplifications of the constants that for any \( 0 < \delta \leq 1 \) (putting \( s = \delta + 2 \) in the previous estimates)

\[
\mathbb{E} \left[ W(\mu, \mu^N) \right] \leq 3^{11/\delta} N^{-1/(2+\delta)} + 3^6 N^{-1/2} \leq e^{14/\delta} N^{-1/(2+\delta)},
\]

where the final inequality worst-cases constants for convenience. Using [93, Proposition 20], we have

\[
\mathbb{P} \left[ W(\mu, \mu^N) + \mathbb{E} \left[ W(\mu, \mu^N) \right] \geq \sqrt{d/N} \right] \leq e^{-2d},
\]

and hence

\[
\mathbb{P} \left[ W(\mu, \mu^N) \geq \frac{\sqrt{d/e^{14/\delta}}}{N^{1/(2+\delta)}} \right] \leq \mathbb{P} \left[ W(\mu, \mu^N) \geq \frac{e^{14/\delta}}{N^{1/(2+\delta)}} + \sqrt{d/N} \right] \leq e^{-2d}
\]

if \( d \geq 1 \).

**Lemma A.2.16.** Let \( n, m \in \mathbb{N} \). Let \( F : \mathbb{R}^n \to \mathbb{R}^m \) be 1-nonnegatively homogeneous, and suppose there exist \( M, L \geq 0 \) such that

1. \( \|F|_{\mathbb{R}^{n-1}}\|_2\|_\infty \leq M; \)

2. \( F|_{\mathbb{R}^{n-1}} \) is \( L \)-Lipschitz.

Then for any \( x, x' \in \mathbb{R}^n \), one has

\[
\|F(x) - F(x')\|_2 \leq (2L + M)\|x - x'\|_2,
\]

so that \( F \) is \((2L + M)\)-Lipschitz.

173
Proof. For any numbers \( a, b \geq 0 \) and any \( u, v \in \mathbb{R}^m \), one has by the triangle inequality

\[
\|au - bv\|_2 \leq \min\{a\|u - v\|_2 + |a - b|\|v\|_2, b\|u - v\|_2 + |a - b|\|u\|_2\}.
\]

Using an elementary property of the min and the max, we thus have

\[
\|au - bv\|_2 \leq \min\{a, b\}\|u - v\|_2 + \max\{\|u\|_2, \|v\|_2\}|a - b|.
\] (A.100)

Now we proceed to show the claim. Noting that the case where both \( x, x' \) are zero is trivial, first consider the case where \( x \) is nonzero and \( x' \) is zero. By nonnegative homogeneity, it suffices to proceed as

\[
\|F(x) - F(x')\|_2 = \|F(x)\|_2 = \|x\|_2 \left\| F\left(\frac{x}{\|x\|_2}\right) \right\|_2 \leq M\|x\|_2 = M\|x - x'\|_2
\]

to conclude; for the inequality we used the boundedness assumption on \( F \). Now fix \( x, x' \in \mathbb{R}^n \) nonzero. The inequality (A.100) can be applied to get

\[
\|F(x) - F(x')\|_2 = \left\| \|x\|_2 F\left(\frac{x}{\|x\|_2}\right) - \|x'\|_2 F\left(\frac{x'}{\|x'\|_2}\right) \right\|_2
\]
\[
\leq \min\{\|x\|_2, \|x'\|_2\} \left\| F\left(\frac{x}{\|x\|_2}\right) - F\left(\frac{x'}{\|x'\|_2}\right) \right\|_2
\]
\[
+ \max\left\{ \left\| F\left(\frac{x}{\|x\|_2}\right) \right\|_2, \left\| F\left(\frac{x'}{\|x'\|_2}\right) \right\|_2 \right\} \|x - x'\|_2,
\]

where in the inequality we also applied the \( \ell^2 \) triangle inequality. Using the assumed properties of \( F \), we thus have

\[
\|F(x) - F(x')\|_2 \leq L \min\{\|x\|_2, \|x'\|_2\} \left\| \frac{x}{\|x\|_2} - \frac{x'}{\|x'\|_2} \right\|_2 + M\|x - x'\|_2.
\]
By a classical inequality (e.g. proved in (A.189)), one has

\[ \left\| \frac{x}{\|x\|_2} - \frac{x'}{\|x'\|_2} \right\|_2 \leq 2 \frac{\|x - x'\|_2}{\max\{\|x\|_2, \|x'\|_2\}}. \]

whence

\[ \|F(x) - F(x')\|_2 \leq (2L + M)\|x - x'\|_2, \]

as was to be shown. \(\square\)

A.3 Geometric and Kernel Estimates

A.3.1 Auxiliary Results

Geometric Results

**Lemma A.3.1.** Let \( M \) be a complete Riemannian submanifold of the unit sphere \( S^{n_0-1} \) (with respect to the spherical metric induced by the euclidean metric on \( \mathbb{R}^{n_0} \)) with finitely many connected components \( K \). If \( d_0 = 1 \), assume moreover that each connected component of \( M \) is a smooth regular curve. Then for every \( 0 < \varepsilon \leq 1 \), there is a \( \varepsilon \)-net for \( M \) in the euclidean metric \( \|\cdot\|_2 \) having cardinality no larger than \( (C_M/\varepsilon)^{d_0} \), where \( C_M \geq 1 \) is a constant depending only on the diameters \( \sup_{x,x' \in M_i} \text{dist}_{M_i}(x,x') \) and, when \( d_0 \geq 2 \), additionally on the extremal Ricci curvatures of \( M_i \). Moreover, these nets have the property that if \( x \in M \) is given, there is a point in the net \( \bar{x} \) within euclidean distance \( \varepsilon \) of \( x \) such that \( \bar{x} \) lies in the same connected component of \( M \) as \( x \).

**Proof.** Consider a fixed connected component \( M_i \) with \( i \in [K] \). We write the Riemannian distance of \( M_i \) as \( \text{dist}_{M_i} \); because \( M_i \) is a Riemannian submanifold of \( \mathbb{R}^{n_0} \), we have \( \text{dist}_{M_i}(x,y) \geq \|x - y\|_2 \) for every \( x,y \) in \( M_i \). Because \( \text{dist}_{M_i}(x,y) \geq \|x - y\|_2 \), it suffices to estimate the covering number in terms of the Riemannian distance. We will consider distinctly the cases \( d_0 = 1 \) and \( d_0 \geq 2 \), starting with \( d_0 = 1 \).

When \( d_0 = 1 \), we have assumed that \( M_i \) are regular curves, so it is without loss of generality to assume they are moreover unit-speed curves parameterized by arc length, with lengths \( \text{len}(M_i) \). It
follows that we can obtain an $\varepsilon$-net for $M_i$ in terms of $\text{dist}_{M_i}$ having cardinality at most $\text{len}(M_i)/\varepsilon$ when $0 < \varepsilon \leq 1$, and by the submanifold property these sets also constitute $\varepsilon$-nets for $M_i$ in terms of the $\ell^2$ distance. Covering each connected component $M_i$ in this way gives a $\varepsilon$-net for $M$ by taking the union of each connected component’s net.

When $d_0 \geq 2$, we make use of standard results relating the covering number to the curvature and diameter of $M$. Let $\text{diam}(M_i) = \sup_{x,x' \in M_i} \text{dist}_{M_i}(x,x')$, and let $\text{Ric}_i$ denote the Ricci curvature tensor of $M_i$ (recall that we assume the metric on $M$ is the one induced by the euclidean metric). Then because $M$ is compact, (1) $\max_{i \in [K]} \text{diam}(M_i) < +\infty$; and (2) because $\text{Ric}_i$ is moreover continuous, there are constants $k_i > 0$ such that $\text{Ric}_i \geq -(d_0 - 1)k_i$ for each $i \in [K]$. Applying Lemma A.3.2, it follows that for any $\varepsilon > 0$, there is a $\varepsilon$-net for $M_i$ in terms of $\text{dist}_{M_i}$ with cardinality no larger than $(C_{M_i}/\varepsilon)^{d_0}$, where $C_{M_i} \leq \text{diam}(M_i)e^{2\text{diam}(M_i)\sqrt{k_i}}$.

Thus, for any $i \in [K]$, any $d_0 \geq 1$ and any $0 < \varepsilon \leq 1$, we can conclude that there is a $\varepsilon$-net for $M_i$ in the euclidean metric having cardinality no larger than $(C_{M_i}/\varepsilon)^{d_0}$, where

$$C_{M_i} = \begin{cases} \text{len}(M_i) & d_0 = 1 \\ 16 \text{diam}(M_i)e^{2\text{diam}(M_i)\sqrt{k_i}} & d_0 \geq 2. \end{cases}$$

Taking the union of these nets and applying Lemma A.6.10 for simplicity, we conclude that for any $d_0 \geq 1$ and any $0 < \varepsilon \leq 1$, there is a $\varepsilon$-net for $M$ in the euclidean metric having cardinality no larger than $(C_M/\varepsilon)^{d_0}$, where

$$C_M = \begin{cases} 1 + \sum_{i=1}^{K} \text{len}(M_i) & d_0 = 1 \\ 1 + 16 \sum_{i=1}^{K} \text{diam}(M_i)e^{2\text{diam}(M_i)\sqrt{k_i}} & d_0 \geq 2. \end{cases}$$

The additional property claimed is satisfied by our construction of the nets. □

**Lemma A.3.2.** Given $k > 0$ and integer $d \geq 2$, suppose that $M$ is a $d$-dimensional complete Riemannian manifold with Ricci curvature tensor satisfying $\text{Ric} \geq -(d - 1)k$. Then for any $r, \varepsilon > 0$ and any $p \in M$, there exists an $\varepsilon$-net (measured in the Riemannian distance $\text{dist}_M$) of
the metric ball \( \{ x \in M \mid \text{dist}_M(p, x) \leq r \} \) with cardinality at most \((C_M/\varepsilon)^d\), where \( C_M > 0 \) is a constant depending only on \( k \) and \( r \).

**Proof.** The proof is essentially an application of [224, Lemma 3.6] together with some calculations on volumes of geodesic balls in hyperbolic space that we record here for completeness, although they are classical. For any \( r > 0 \) and any \( p \in M \), write

\[
B_r(p) = \{ x \in M \mid \text{dist}_M(p, x) \leq r \}.
\]

Fix \( p \in M \) and \( r, \varepsilon > 0 \). The hypotheses of the lemma make [224, Lemma 3.6] applicable, whence

\[
\inf \left\{ \text{card}(S) \right| S \subset B_r(p), B_r(p) \subset \bigcup_{p' \in S} B_\varepsilon(p') \right\} \leq \frac{\text{vol}(B^k(2r))}{\text{vol}(B^k(\varepsilon/4))},
\]

where \( \text{card}(S) \) denotes the cardinality of a set \( S \), and for all \( \varepsilon > 0 \), \( \text{vol}(B^k(\varepsilon)) \) denotes the volume of a geodesic ball of radius \( r \) in the \( d \)-dimensional simply-connected hyperbolic space of constant sectional curvature \( -k \); these spaces are homogeneous and isotropic so the base point does not matter (c.f. [217, Proposition 3.9]). In particular, we can calculate these volumes in any model of hyperbolic space and anchored at any base point; we choose the Poincaré ball model and the base point \( 0 \), where the maximal unit-speed geodesics take the simple form

\[
\gamma(t) = k^{-1/2} v \tanh \frac{\sqrt{k} t}{2}
\]

for \( v \in S^d \) and \( t \in \mathbb{R} \) [217, Theorem 3.7, Proposition 5.28]. Integrating the volume form in coordinates, we then get for any \( \varepsilon > 0 \)

\[
\text{vol}(B^k(\varepsilon)) = \int_{(k^{-1/2} \tanh \sqrt{k} \varepsilon/2) \mathbb{B}^d} \left( \frac{2/k}{1/k - \|x\|^2} \right)^d dx
\]

\[
= k^{-d/2} \int_{(\tanh \sqrt{k} \varepsilon/2) \mathbb{B}^d} \left( \frac{2}{1 - \|x\|^2} \right)^d dx
\]
where the second line changes coordinates $x \mapsto k^{-1/2}x$. Changing to polar coordinates in the last expression, we get

$$\text{vol}(B^k(\varepsilon)) = k^{-d/2} \text{vol}(S^{d-1}) \int_{[0, \sinh \sqrt{k \varepsilon}/2]} x^{d-1} \left(\frac{2}{1 - x^2}\right)^d \, dx,$$

and then changing coordinates $x \mapsto \tanh x$, we obtain after applying several trigonometric identities

$$\text{vol}(B^k(\varepsilon)) = k^{-d/2} \text{vol}(S^{d-1}) \int_{[0, \sqrt{k \varepsilon}]} 2 \sinh^{d-1}(2x) \, dx$$

$$= k^{-d/2} \text{vol}(S^{d-1}) \int_{[0, \sqrt{k \varepsilon}]} \sinh^{d-1}(x) \, dx,$$

whence

$$\frac{\text{vol}(B^k(2r))}{\text{vol}(B^k(\varepsilon/4))} = \frac{\int_{[0, 2r \sqrt{k}]} \sinh^{d-1}(x) \, dx}{\int_{[0, \varepsilon \sqrt{k}/4]} \sinh^{d-1}(x) \, dx}.$$

We have bounds $x \leq \sinh x \leq xe^x$ for nonnegative $x$,\footnote{The lower bound is implied by $\cosh x \geq 1$; the upper bound follows from writing $\sinh x = 0.5e^x(1 - e^{-2x})$ and using $e^{-x} \geq 1 - x$.} which gives after integration

$$\frac{\text{vol}(B^k(2r))}{\text{vol}(B^k(\varepsilon/4))} \leq \frac{\int_{[0, 2r \sqrt{k}]} x^{d-1}e^{(d-1)x} \, dx}{\int_{[0, \varepsilon \sqrt{k}/4]} x^{d-1} \, dx}$$

$$\leq d \frac{(2r \sqrt{k})^d \int_{[0, 1]} x^{d-1}e^{2r \sqrt{k}(d-1)x} \, dx}{(\varepsilon \sqrt{k}/4)^d}$$

$$\leq \left(\frac{16re^{2r \sqrt{k}}}{\varepsilon}\right)^d,$$

where in the second line we change coordinates $x \mapsto (2r \sqrt{k})x$, and then use $L^\infty$ control of the (monotone increasing) integrand in the second line to move to the expression in the third line. $\square$

Remark A.3.3. The constant $C_M$ in Lemma A.3.2 can be sharpened if more is known about the curvature of $\mathcal{M}$: if $\text{Ric} \geq 0$, the exponential dependence on curvature and diameter can be removed (intuitively, taking $k \downarrow 0$ “recovers” this from the proved result), and if $\text{Ric} > 0$, the dependence
on diameter can be completely removed using Myers’ theorem [224, Theorem 3.4(1)].

**Lemma A.3.4.** For any \(x, x', \tilde{x}, \tilde{x}'\) in \(\mathbb{S}^{n_0-1}\), one has

\[
|\angle(x, x') - \angle(\tilde{x}, \tilde{x}')| \leq \sqrt{2}\|x - x'\|_2 - \|\tilde{x} - \tilde{x}'\|_2.
\]

**Proof.** Writing 
\[
\angle(x, x') = \arccos\left(\langle x, x' \rangle\right) = \arccos\left(1 - \frac{1}{2}\|x - x'\|_2^2\right),
\]
consider the function 
\[
f(x) = \frac{x}{\sqrt{1 - (1 - \frac{1}{2}x^2)^2}} = \frac{\text{sign } x}{\sqrt{1 - \frac{1}{4}x^2}},
\]
and taking limits at 0 shows that \(f\) admits left and right derivatives on all of \([-\sqrt{2}, \sqrt{2}]\). \(f'\) is even-symmetric, so by checking values at 0 and \(\sqrt{2}\) we conclude that \(|f'| \leq \sqrt{2}\), which shows that \(f\) is \(\sqrt{2}\)-Lipschitz. The claim follows. \(\square\)

**Lemma A.3.5.** Let \(d_0 = 1\). Choose \(L\) so that \(L \geq K\kappa^2 C_{\lambda}\), where \(\kappa\) and \(C_{\lambda}\) are respectively the curvature and global regularity constants defined in Section A.1.2, and \(K, K' > 0\) are absolute constants. Then

\[
\sup_{x \in M_\pm} \int_M \frac{d\mu^\infty(x')}{(1 + (L/\pi)\angle(x, x'))^2} \leq \frac{C\rho_{\max}(\text{len}(M_+) + \text{len}(M_-))}{L},
\]

where \(C\) is an absolute constant and \(M_\pm\) denotes either \(M_+\) or \(M_-\).

**Proof.** See [221, Lemma C.8]. \(\square\)

**Lemma A.3.6.** Given a smooth, simple open curve in \(\mathbb{R}^n\) of length \(S\) with unit-speed parametrization \(\gamma: [0, S] \rightarrow \mathbb{R}^n\) such that for some \(\kappa > 0\)

1. \(\|\dot{\gamma}\|_2 \leq \kappa\)
2. \(S \leq \frac{\pi}{\kappa}\)
define by \( \tilde{\gamma} \) an arc of any circle of radius \( \frac{1}{k} \) such that \( \tilde{\gamma}(0) = \gamma(0), \tilde{\gamma}(S) = \gamma(S), S \leq \frac{\pi}{k} \). We then have

\[
S \leq \tilde{S}
\]

**Proof.** See [221, Lemma C.9]. \( \square \)

**Analysis of the Skeleton**

**Notation.** Define \( \varphi^{(0)} = \text{Id}, \) and for \( \ell \in \mathbb{N} \) define \( \varphi^{(\ell)} \) as the \( \ell \)-fold composition of \( \varphi \) with itself, where

\[
\varphi(v) = \cos^{-1}\left((1 - \pi^{-1} v) \cos v + \pi^{-1} \sin v\right)
\]

is the heuristic angle evolution function. We will make use of basic properties of this function such as smoothness (established in Lemma A.5.5) below. In this section, we will study the skeleton

\[
\psi_1(v) = \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \varphi^{(\ell)}(v) \prod_{\ell'=\ell}^{L-1} \left(1 - \pi^{-1} \varphi^{(\ell')}(v)\right), \quad v \in [0, \pi],
\]

where we have not included the additive factor \( \cos \varphi^{(L)}(v) \), as it is easily removed along the lines of Theorem A.2.2. We define

\[
\xi^{(\ell)}(v) = \prod_{\ell'=\ell}^{L-1} \left(1 - \pi^{-1} \varphi^{(\ell')}(v)\right), \quad \ell = 0, \ldots, L - 1,
\]

so that

\[
\psi_1(v) = \frac{n}{2} \sum_{\ell=0}^{L-1} \cos \varphi^{(\ell)}(v) \xi^{(\ell)}(v). \quad \text{(A.101)}
\]

We will also establish a convenient approximation to the skeleton. Define

\[
\psi(v) = \frac{n}{2} \sum_{\ell=0}^{L-1} \xi^{(\ell)}(v).
\]

---

\(^{6}\)For any circle and choice of endpoints there will be two such arcs, and the last condition implies that we choose the shorter of the two.
Lemma A.3.7 implies that $\psi$ is convex; it is less trivial to obtain the same for $\psi_1$. We will prove several estimates below for the terms $\xi^{(\ell)}$ and their derivatives that can be used to immediately obtain useful estimates for $\psi$ and its derivatives.

**Lemma A.3.7.** For each $\ell = 0, 1, \ldots, L$, the functions $\varphi^{(\ell)}$ are nonnegative, strictly increasing, and concave (positive and strictly concave on $(0, \pi)$); if $0 \leq \ell < L$, the functions $\xi^{(\ell)}$, are nonnegative, strictly decreasing, and convex (positive and strictly convex on $(0, \pi)$).

**Proof.** These claims are a consequence of some general facts for smooth functions that we articulate here so that we can rely on them often in the sequel. First, we have for any smooth function $f : (0, \pi) \to \mathbb{R}$

$$
(f \circ f)' = (f' \circ f)f',
$$
$$
(f \circ f)'' = (f' \circ f)f'' + (f')^2(f'' \circ f).
$$

These equations show that if $f > 0$, $f' > 0$, and $f'' < 0$, then $f \circ f$ also satisfies these three properties. Lemma A.5.5 shows that $\varphi$ satisfies these three properties on $(0, \pi)$; we conclude from the mean value theorem and a simple induction the same for $\varphi^{(\ell)}$, as claimed. Meanwhile, if $f, g$ are smooth real-valued functions on $(0, \pi)$, we have

$$
(fg)' = f'g + g'f,
$$
$$
(fg)'' = f''g + g''f + 2f'g'.
$$

Thus, if $f$ and $g$ are both positive, strictly decreasing, strictly convex functions on $(0, \pi)$, then $fg$ also satisfies these three properties. Lemma A.5.5 implies that $0 < 1 - \pi^{-1}\varphi^{(\ell)} < 1$ on $(0, \pi)$, and the first and second derivatives are scaled and negated versions of those of $\varphi^{(\ell)}$; we conclude by another induction that the same three properties apply to the functions $\xi^{(\ell)}$. \qed

**Lemma A.3.8.** There is an absolute constant $C > 0$ such that if $L \geq 12$ and $n \geq L$, then one has

$$
\|\psi_1 - \psi\|_{L^\infty} \leq \frac{Cn}{L}.
$$
Proof. We have from the triangle inequality

\[ \|\psi_1 - \psi\|_{L^\infty} \leq \sup_{\nu \in [0,\pi]} \left( \frac{n}{2} \sum_{\ell=0}^{L-1} \left| \cos \varphi^{(\ell)}(\nu) - 1 \right| \right) \]

\[ \leq \frac{n}{2} \sum_{\ell=0}^{L-1} \sup_{\nu \in [0,\pi]} \left( \left| \cos \varphi^{(\ell)}(\nu) - 1 \right| \right), \]

where we use Lemma A.3.7 to take \( \xi^{(\ell)} \) outside the absolute value. Notice that \( (\cos \varphi^{(\ell)} - 1) \xi^{(\ell)} \leq 0 \), so to control the \( L^\infty \) norm of this term it suffices to bound it from below. We will show the monotonicity property

\[ (\cos \varphi^{(\ell)} - 1) \xi^{(\ell)} - (\cos \varphi^{(\ell+1)} - 1) \xi^{(\ell+1)} \geq 0, \]

(A.102)

from which it follows

\[ \|\psi_1 - \psi\|_{L^\infty} \leq \frac{nL}{2} \sup_{\nu \in [0,\pi]} \left| \cos \varphi^{(L-1)}(\nu) - 1 \right|, \]

using also \( \xi^{(L-1)}(\nu) \leq 1 \). Since \( \cos x \geq 1 - (1/2)x^2 \), and since Lemma A.3.9 gives that \( \varphi^{(L-1)} \leq C/(L - 1) \) (and also estimates the constant), we have as soon as \( L \geq 1 + C/\sqrt{2} \)

\[ \|\psi_1 - \psi\|_{L^\infty} \leq \frac{C^2 nL}{4(L - 1)^2} \]

which gives the claim provided \( L \geq 2 \) and \( n \geq L \). So to conclude, we need only establish (A.102). To this end, write the LHS of (A.102) as

\[ (\cos \varphi^{(\ell)} - 1) \xi^{(\ell)} - (\cos \varphi^{(\ell+1)} - 1) \xi^{(\ell+1)} = \left[ (\cos \varphi^{(\ell)} - \cos \varphi^{(\ell+1)}) - \frac{\varphi^{(\ell)}}{\pi} (\cos \varphi^{(\ell)} - 1) \right] \xi^{(\ell+1)} \]

to notice that it suffices to prove nonnegativity of the bracketed quantity. In addition, since \( \ell \geq 0 \) and \( \varphi(\nu) \leq \nu \) by Lemma A.5.5, we can instead prove the inequality

\[ (\cos x - \cos \varphi(x)) - \frac{x}{\pi} (\cos x - 1) \geq 0 \]
for all $x \in [0, \pi]$. Using the closed-form expression for $\cos \varphi(x)$ in Lemma A.5.2, we can plug into the previous inequality and cancel to get the equivalent inequality

$$x - \sin x \geq 0.$$  

But this is immediate from the concavity estimate $\sin x \leq x$, and (A.102) is proved. □

**Lemma A.3.9.** If $\ell \in \mathbb{N}_0$, one has the “fluid” estimate for the angle evolution function

$$\varphi^{(\ell)}(\nu) \leq \frac{\nu}{1 + c_1\nu},$$

where $c > 0$ is an absolute constant. In particular, if $\ell \in \mathbb{N}$ one has $\varphi^{(\ell)} \leq 1/c\ell$.

**Proof.** The second claim follows from the first claim and $1 + c\ell\nu \geq c\ell\nu$, so we will focus on establishing the first estimate. The proof is by induction on $\ell \in \mathbb{N}$, since the case of $\ell = 0$ is immediate. By Lemma A.5.5, there is a constant $c_1 > 0$ such that $\varphi(\nu) \leq \nu(1 - c_1\nu)$, and using the numerical inequality $x(1 - x) \leq x(1 + x)^{-1}$, valid for $x \geq 0$, we get

$$\varphi(\nu) \leq \frac{\nu}{1 + c_1\nu}, \tag{A.103}$$

which establishes the claim in the case $\ell = 1$. Assuming the claim holds for $\ell - 1$, we calculate

$$\varphi^{(\ell)}(\nu) \leq \frac{\varphi^{(\ell-1)}(\nu)}{1 + c_1\varphi^{(\ell-1)}(\nu)} \leq \frac{\nu}{1 + c_1(1 + \ell\nu/\pi)},$$

where the first inequality uses (A.103), and the second inequality uses the induction hypothesis and the relation $x(1 + x)^{-1} = 1 - (1 + x)^{-1}$ to see that $x \mapsto x(1 + c_1x)^{-1}$ is increasing. Clearing denominators in the numerator and denominator of the RHS of this last bound, we see that it is equal to $\nu/(1 + \ell\nu/\pi)$, and the claim follows by induction. □
Lemma A.3.10. If \( \ell \in \mathbb{N}_0 \), the iterated angle evolution function satisfies the estimate

\[
\varphi^{(\ell)}(\nu) \geq \frac{\nu}{1 + \ell \nu / \pi}.
\]

Proof. The proof is by induction on \( \ell \in \mathbb{N} \), since the case \( \ell = 0 \) is immediate. The case \( \ell = 1 \) follows from Lemma A.3.11. Assuming the claim holds for \( \ell - 1 \), we calculate

\[
\varphi^{(\ell)}(\nu) \geq \frac{\varphi^{(\ell-1)}(\nu)}{1 + \varphi^{(\ell-1)}(\nu) / \pi} \geq \frac{\nu}{1 + \frac{1}{\pi} \frac{\nu}{1 + (\ell-1) \nu / \pi}},
\]

where the first inequality applies Lemma A.3.11, and the second uses the fact that the RHS of the bound in Lemma A.3.11 is strictly increasing and the induction hypothesis. Clearing denominators in the numerator and denominator of the RHS of this last bound, we see that it is equal to \( \nu / (1 + \ell \nu / \pi) \), and the claim follows by induction. \( \square \)

Lemma A.3.11. It holds

\[
\varphi(\nu) \geq \frac{\nu}{1 + \nu / \pi}.
\]

Proof. After some rearranging using Lemma A.5.2, it suffices to prove

\[
\left(1 - \frac{\nu}{\pi}\right) \cos \nu + \frac{\sin \nu}{\pi} \leq \cos \left(\frac{\pi \nu}{\pi + \nu}\right). \tag{A.104}
\]

Using Lemma A.5.5, we see that both the LHS and RHS of this bound are nonincreasing. We will prove the estimate in three stages, using “small angle”, “large angle”, and “intermediate angle” estimates of the quantities on both sides of (A.104). Since \( \pi \nu / (\pi + \nu) \in [0, \pi/2] \), we can use standard estimates for \( \cos \) to get RHS estimates

\[
\cos \left(\frac{\pi \nu}{\pi + \nu}\right) \geq 1 - \frac{1}{2} \left(\frac{\pi \nu}{\pi + \nu}\right)^2 \tag{A.105}
\]

and

\[
\cos \left(\frac{\pi \nu}{\pi + \nu}\right) \geq \frac{\nu}{\pi + \nu}. \tag{A.106}
\]
As for the LHS, we can obtain an estimate near $\nu = \pi$ in a straightforward way. Transforming the domain by $\nu \mapsto \pi - \nu$, it suffices to get estimates on $\sin \nu - \nu \cos \nu$ near $\nu = 0$, then divide by $\pi$. Using $\cos \nu \geq 1 - (1/2)\nu^2$ and $\sin \nu \leq \nu$, it follows that $\sin \nu - \nu \cos \nu \leq (1/2)\nu^3$. We conclude

$$\left(1 - \frac{\nu}{\pi}\right) \cos \nu + \frac{\sin \nu}{\pi} \leq \frac{1}{2\pi} (\pi - \nu)^3.$$  \quad (A.107)

We will develop a second-order approximation to the LHS near 0 for the small-angle estimates. The first, second, and third derivatives of the LHS are $(1 - \nu/\pi) \sin \nu, (1/\pi) \sin \nu - (1 - \nu/\pi) \cos \nu,$ and $(2/\pi) \cos \nu + (1 - \nu/\pi) \sin \nu$, respectively. To bound the third derivative, we will use the estimate $\cos \nu \leq 1 - \nu^2/3$ on $[0, \pi/2]$. To prove this, note that Taylor’s formula implies the bound $\cos \nu \leq 1 - \nu^2/3$ on $[0, \cos^{-1}(2/3)]$; because $\cos$ is concave on $[0, \pi/2]$, we also have the tangent line bound $\cos(\nu) \leq -\nu \sqrt{5}/3 + (2/3 + \sqrt{5} \cos^{-1}(2/3)/3)$ on $[0, \pi/2]$. We can then solve for the zeros of the quadratic polynomial $1 - \nu^2/3 + (\sqrt{5}/3)\nu - (2/3 + \sqrt{5} \cos^{-1}(2/3)/3)$; a numerical evaluation shows that both roots are real and outside the interval $[\cos^{-1}(2/3), \pi/2]$. Since the tangent line touches the graph of $\cos$ at $\nu = \cos^{-1}(2/3)$, this proves that $\cos \nu \leq 1 - \nu^2/3$ on $[0, \pi/2]$. We can therefore write

$$2 \cos \nu + (\pi - \nu) \sin \nu \leq 2(1 - \nu^2/3) + \nu(\pi - \nu), \quad \nu \in [0, \pi/2].$$

The RHS of this inequality is a concave quadratic; we calculate its maximum analytically as $2 + 3\pi^2/20$. Meanwhile, if $\nu \in [\pi/2, \pi]$, we have $2 \cos \nu \leq 0$, and $(\pi - \nu) \sin \nu \leq \pi/2$. We conclude that $(2/\pi) \cos \nu + (1 - \nu/\pi) \sin \nu \leq 2 + 3\pi^2/20$ on $[0, \pi]$. Writing $c = 1/(3\pi) + \pi/40$, this implies an estimate

$$\left(1 - \frac{\nu}{\pi}\right) \cos \nu + \frac{\sin \nu}{\pi} \leq 1 - \frac{\nu^2}{2} + c\nu^3.$$  \quad (A.108)

Finally, we will need some estimates for interpolating the small and large angle regimes. We note that the second derivative $(1/\pi) \sin \nu - (1 - \nu/\pi) \cos \nu$ of the LHS of (A.104) is nonnegative if $\nu \geq \pi/2$, because $\cos \geq 0$ here; meanwhile, the third derivative $(2/\pi) \cos \nu + (1 - \nu/\pi) \sin \nu$ of the
LHS of (A.104) is nonnegative if $0 \leq \nu \leq \pi/2$, since $\cos \geq 0$ here, and it follows that the second derivative is increasing on $[0, \pi/2]$. Checking numerically that the value of the second derivative at 1.42 is positive, we conclude that the LHS of (A.104) is convex on $[1.42, \pi]$. In addition, we use calculus to evaluate the first and second derivative of the RHS of (A.105) as $-\nu \pi^3/(\pi + \nu)^3$ and $-\pi^3(\pi - 2\nu)/(\pi + \nu)^4$, respectively; this shows that the RHS of (A.105) is convex for $\nu \geq \pi/2$, and concave for $\nu \leq \pi/2$. Taking a tangent line to the graph of the RHS of (A.105) at $1.42$, it follows that the function

$$g(x) = \begin{cases} 
1 - (\pi^2/2)\nu^2/(\pi + \nu)^2 & x \leq \pi/2 \\
-(4\pi/27)\nu + (1 + \pi^2/54) & x \geq \pi/2
\end{cases}$$

(A.109)

is a concave lower bound for the RHS of (A.105) on $[0, \pi]$.

We proceed to using the estimates developed in the previous paragraph to prove (A.104). We first argue that for $\nu$ in a neighborhood of 0, we have

$$1 - \nu^2/2 + c\nu^3 \leq 1 - (\pi^2/2)\nu^2/(\pi + \nu)^2,$$

which will in turn prove (A.104) in the same neighborhood. Cancelling and rearranging, it is equivalent to show

$$(2/\pi - 2c) - (4c/\pi - 1/\pi^2)\nu - (2c/\pi^2)\nu^2 \geq 0.$$ 

The LHS is a concave quadratic, with value $2/\pi - 2c > 0$ at 0; we calculate its two distinct roots numerically as lying in the intervals $[-5.1, -5]$ and $[1.42, 1.43]$, respectively. It follows that (A.104) holds for $\nu \in [0, 1.42]$. Next, we argue that for $\nu$ in a neighborhood of $\pi$, we have

$$\frac{1}{2\pi}(\pi - \nu)^3 \leq \frac{\pi - \nu}{\pi + \nu},$$

which will in turn prove (A.104) in the same neighborhood. Transforming with $\nu \mapsto \pi - \nu$ and rearranging, it is equivalent to show $\nu^2(2\pi - \nu) \leq 4\pi^2$ in a neighborhood of 0. The LHS of this last inequality is 0 at 0, and nonnegative on $[0, \pi]$; its first and second derivatives are $\nu(4\pi - 3\nu)$ and
$4\pi - 6\nu$, respectively, which shows that it is a strictly increasing function of $\nu$ on $[0, \pi]$. Verifying numerically the three distinct real roots of $\nu^3 - 2\pi\nu^2 + 1 = 0$ and transferring the result back via another transformation $\nu \mapsto \pi - \nu$, we conclude that (A.104) holds on $[\pi - 1.1, \pi]$. To obtain that (A.104) holds on $[1.42, \pi - 1.1]$, we use that the function $g$ defined in (A.109) is a concave lower bound for the RHS of (A.105), so that it suffices to show that the LHS of (A.104) is upper bounded by $g$ on $[1.42, \pi - 1.1]$. The LHS of (A.104) is convex on $[1.42, \pi]$, so it follows that it is sufficient to show that the values of the LHS of (A.104) at $1.42$ and at $\pi - 1.1$ are upper bounded by those of $g$ at the same points. Confirming this numerically, we can conclude the proof.

Lemma A.3.12. If $\ell \in \mathbb{N}_0$, one has

\[
\left| \phi^{(\ell)}(\nu) \right| \leq \frac{1}{1 + (c/2)\ell\nu},
\]

where $c > 0$ is the absolute constant also appearing in Lemma A.5.5 (property 4), and in particular $c/2$ is equal to the absolute constant appearing in Lemma A.3.9. In particular, if $\ell \in \mathbb{N}$ and $\nu \in [0, \pi]$, we have the estimate

\[
\left| \nu \phi^{(\ell)}(\nu) \right| \leq \frac{2}{c\ell}.
\]

Proof. The case of $\ell = 0$ follows directly (as an equality) from $\phi^{(0)}(\nu) = \nu$. Now we assume $\ell \in \mathbb{N}$. Smoothness of $\phi^{(\ell)}$ follows from Lemma A.5.5. Applying the chain rule and an induction, we have

\[
\phi^{(\ell)} = \left( \phi \circ \phi^{(\ell-1)} \right) \phi^{(\ell-1)} = \prod_{\ell' = 0}^{\ell-1} \phi \circ \phi^{(\ell')}, \tag{A.110}
\]

and applying the chain rule also gives

\[
\phi^{(\ell)} = \left( \phi^{(\ell-1)} \right)^2 \left( \phi \circ \phi^{(\ell-1)} \right) + \left( \phi^{(\ell-1)} \right) \left( \phi \circ \phi^{(\ell-1)} \right). \tag{A.111}
\]

By Lemma A.5.5, we have $\phi > 0$ on $[0, \pi]$, and the formula (A.110) then implies that $\phi^{(\ell)} > 0$.
on \([0, \pi]\) as well. Considering only angles in this half-open interval and distributing, it follows

\[
\frac{\varphi^{(\ell)}}{(\varphi^{(\ell)})^2} = \frac{\varphi \circ \varphi^{(\ell-1)}}{(\varphi \circ \varphi^{(\ell-1)})^2} + \frac{1}{\varphi \circ \varphi^{(\ell-1)}} \frac{\varphi^{(\ell-1)}}{(\varphi^{(\ell-1)})^2}.
\]

Applying an induction using the previous formula and distributing in the result, we obtain

\[
\frac{\varphi^{(\ell)}}{(\varphi^{(\ell)})^2} = \sum_{\ell' = 0}^{\ell-1} \left( \frac{1}{\prod_{\ell'' = \ell' + 1}^{\ell-1} \varphi \circ \varphi^{(\ell'')}} \right) \frac{\varphi}{\varphi^2} \circ \varphi^{(\ell')}.
\]  

(A.112)

By Lemma A.5.5, we have \(0 < \varphi \leq 1\) on \([0, \pi]\) and \(\varphi \leq 0\). Thus

\[
-\frac{\varphi^{(\ell)}}{(\varphi^{(\ell)})^2} \geq -\sum_{\ell' = 0}^{\ell-1} \varphi \circ \varphi^{(\ell')}.
\]

When \(\ell' > 0\), we have \(\varphi^{(\ell')} \leq \pi/2\), and by Lemma A.5.5, we have \(\varphi \leq -c < 0\) on \([0, \pi/2]\); thus, 

\(-\varphi \circ \varphi^{(\ell')} \geq c\) if \(\ell' > 0\). When \(\ell' = 0\), we can use the fact that \(\varphi \leq 0\) on \([0, \pi]\) to get a bound

\(\varphi \leq -c\mathbb{I}_{[0, \pi/2]}\). We conclude

\[
-\frac{\varphi^{(\ell)}}{(\varphi^{(\ell)})^2} \geq c(\ell - 1) + c\mathbb{I}_{[0, \pi/2]}.
\]  

(A.113)

Next, we notice using the chain rule that

\[
\left( \frac{1}{\varphi^{(\ell)}} \right)' = -\frac{\varphi^{(\ell)}}{(\varphi^{(\ell)})^2},
\]

and using (A.110) and Lemma A.5.5, we have that \(\varphi^{(\ell)}(0) = 1\). For any \(\nu \in [0, \pi]\), we integrate
both sides of (A.113) from 0 to $\nu$ to obtain using the fundamental theorem of calculus

\[
\frac{1}{\phi^{(\ell)}(\nu)} - 1 \geq c(\ell - 1)\nu + c \int_0^\nu 1_{[0,\pi/2]}(t) \, dt = c(\ell - 1)\nu + c \min\{\nu, \pi/2\} \geq \frac{c\ell\nu}{2},
\]

where in the final inequality we use the inequality $\min\{\nu, \pi/2\} \geq \nu/2$, valid for $\nu \in [0, \pi]$. Rearranging, we conclude for any $0 \leq \nu < \pi$

\[
\phi^{(\ell)}(\nu) \leq \frac{1}{1 + (c/2)\ell\nu},
\]

and noting that the LHS of this bound is equal to 0 at $\nu = \pi$ and the RHS is positive, we conclude the claimed bound for every $\nu \in [0, \pi]$. The second estimate claimed follows by multiplying this bound by $\nu$ on both sides, and using $1 + (c/2)\ell\nu \geq (c/2)\ell\nu$. □

**Lemma A.3.13.** If $\ell \in \mathbb{N}$, one has

\[
|\phi^{(\ell)}(\nu)| \leq \frac{C}{1 + (c/8)\ell\nu} \left( 1 + \frac{1}{(c/8)\nu} \log \left( 1 + (c/8)(\ell - 1)\nu \right) \right),
\]

where $C > 0$ is an absolute constant, and $c > 0$ is the absolute constant also appearing in Lemma A.5.5 (property 4), and in particular $c/2$ is equal to the absolute constant appearing in Lemma A.3.9. If $\nu \in [0, \pi]$, the RHS of this upper bound is a decreasing function of $\nu$, and moreover we have the estimates

\[
|\phi^{(\ell)}| \leq C\ell, \quad |\nu^2\phi^{(\ell)}(\nu)| \leq \frac{C\pi\nu}{1 + (c/8)\ell\nu} \left( 1 + \frac{8\log \ell}{c\pi} \right) \leq \frac{8\pi C}{c\ell} + \frac{64C\log \ell}{c^2\ell}.
\]

*Proof.* Smoothness follows from Lemma A.5.5; we make use of some results from the proof of Lemma A.3.12, in particular (A.110) and (A.112). We treat the case of $\ell = 1$ first. By Lemma A.5.5, we have $|\phi| \leq C$ for an absolute constant $C > 0$, and since $1/(1 + (c/2)\nu) \geq$
1/(3/2) = 2/3 by the numerical estimate of the absolute constant \( c > 0 \) in Lemma A.5.5, it follows

\[
|\varphi(v)| \leq \frac{3C/2}{1 + (c/2)v},
\]

which establishes the claim when \( \ell = 1 \) (after worst-casing constants if necessary). Next, we assume \( \ell > 1 \). Multiplying both sides of (A.112) by \( (\varphi(\ell))^{2} \) and cancelling using (A.110), we obtain

\[
\varphi(\ell) = \sum_{\ell'=0}^{\ell-1} \prod_{\ell''=0}^{\ell'-1} \left( \varphi \circ \varphi^{(\ell'')} \right)^{2} \varphi \circ \varphi^{(\ell)}
= \sum_{\ell'=0}^{\ell-1} \left( \prod_{\ell''=0}^{\ell'-1} \left( \varphi \circ \varphi^{(\ell'')} \right)^{2} \left( \prod_{\ell''=\ell'+1}^{\ell-1} \varphi \circ \varphi^{(\ell'')} \right) \varphi \circ \varphi^{(\ell')}
\]

(A.114)

where the last equality holds at least on \([0, \pi)\), by Lemmas A.5.5 and A.3.12, and where empty products are defined to be 1. If \( \ell' > 0 \), we have \( \varphi^{(\ell')} \leq \pi/2 \), and by Lemma A.5.5 we have that \( |\varphi| \leq C \) and \( \varphi \geq c' > 0 \) on \([0, \pi/2]\) for absolute constants \( C, C' > 0 \). Separating the \( \ell' = 0 \) summand, this gives a bound

\[
|\varphi^{(\ell)}| \leq C \left( \prod_{\ell'=1}^{\ell-1} \varphi \circ \varphi^{(\ell')} \right) + \frac{C}{c'} \varphi^{(\ell)} \sum_{\ell'=1}^{\ell-1} \varphi^{(\ell')}.
\]

(A.115)

By Lemma A.3.12, we have \( \varphi(\nu) \leq 1/(1 + (c/2)\nu) \), and by Lemma A.5.5, we have \( \varphi(\nu) \leq \nu \), hence \( \varphi^{(\ell')}(\nu) \leq \nu \). Using concavity of \( \varphi \), nonincreasingness of \( \varphi \) and nondecreasingness of \( \varphi^{(\ell')} \)
(which follow from Lemma A.5.5) and a simple re-indexing, we can write

\[
\prod_{\ell' = 1}^{\ell - 1} \varphi \circ \varphi^{(\ell')} (v) = \prod_{\ell' = 0}^{\ell - 2} \varphi \circ \varphi^{(\ell' + 1)} (v) = \prod_{\ell' = 0}^{\ell - 2} \varphi \circ \varphi^{(\ell')} \circ \varphi (v)
\]

\[
\leq \prod_{\ell' = 0}^{\ell - 2} \varphi^{(\ell')} (v/2))
\]

\[
= \varphi^{(\ell - 1)} (v/2)
\]

\[
\leq \frac{1}{1 + (c/4)(\ell - 1)v}
\]

\[
\leq \frac{1}{1 + (c/8)\ell v}
\]

where the third-to-last line follows from (A.110), the second-to-last line from Lemma A.3.12, and the last line follows from the inequality \(\ell - 1 \geq \ell/2\) if \(\ell \geq 2\). Following on from (A.115), we conclude by an application of Lemma A.3.12

\[
\left| \varphi^{(\ell)} (v) \right| \leq \frac{C}{1 + (c/8)\ell v} + \frac{C/c'}{1 + (c/2)\ell v} \sum_{\ell' = 1}^{\ell - 1} \frac{1}{1 + (c/2)\ell' v}
\]

\[
\leq \frac{C}{c'} \left( \frac{1}{1 + (c/8)\ell v} \sum_{\ell' = 0}^{\ell - 1} \frac{1}{1 + (c/8)\ell' v} \right),
\]

where the last line simply worst-cases the constants. For any \(\ell' \in \mathbb{N}_0\), the function \(x \mapsto 1/(1 + (c/8)\ell' x)\) is nonincreasing, so we can estimate the sum in the previous statement using an integral, obtaining

\[
\left| \varphi^{(\ell)} (v) \right| \leq \frac{C/c'}{1 + (c/8)\ell v} \left( 1 + \int_0^{\ell - 1} \frac{1}{1 + cvx} \, dx \right)
\]

\[
\leq \frac{C/c'}{1 + (c/8)\ell v} \left( 1 + \frac{1}{(c/8)v} \log (1 + (c/8)(\ell - 1)v) \right)
\]

after evaluating the integral—we define the quantity inside the parentheses on the RHS of the final inequality to be \(\ell - 1\) when \(v = 0\), which agrees with the integral representation in the previous line and with the unique continuous extension of the function on \((0, \pi]\) to \([0, \pi]\)—which establishes the first claim.
We now move on to the study of the bound we have derived. For decreasingness, we note that
the functions
\[
\nu \mapsto \frac{C/c'}{1 + (c/8)\ell \nu}, \quad \nu \mapsto 1 + \frac{1}{(c/8)\nu} \log \left(1 + (c/8)\left(\ell - 1\right)\nu\right),
\]
(A.116)
whose product is equal to our upper bound, are evidently both smooth nonnegative functions of \(\nu\) at
least on \((0, \pi]\), so that by the product rule for differentiable functions it suffices to prove that these
two functions are themselves decreasing functions of \(\nu\). The first function is evidently decreasing
as an increasing affine reparameterization of \(\nu \mapsto 1/\nu\); for the second function, after multiplying
by the constant \(\ell - 1\) and rescaling by a positive number (when \(\ell = 1\), the function is identically
zero on \((0, \pi]\), and the function’s continuous extension as defined above equals 0 at 0 as well), we
observe that it suffices to prove that \(x \mapsto x^{-1} \log(1 + x)\) is a decreasing function of \(\nu\) on \((0, \infty)\).
The derivative of this function is \(x \mapsto (x - (1 + x) \log(1 + x))/(x^2(1 + x))\), so it suffices to show that
\(x - (1 + x) \log(1 + x) \leq 0\). Noting that the function \(x \mapsto x \log x\) is convex (its second derivative is
1/\(x\)), it follows that \(x - (1 + x) \log(1 + x)\) is concave as a sum of concave functions, and is therefore
has its graph majorized by its supporting hyperplanes; its derivative is equal to \(-\log(1 + x)\), which
equals 0 at 0, and we therefore conclude from our previous reduction that the second function in
(A.116) is decreasing, and that our composite upper bound is as well. For the remaining estimates,
we use the concavity estimate \(\log(1 + x) \leq x\) to obtain from our previous result
\[
\left|\bar{\psi}^{(\ell)}(\nu)\right| \leq \frac{C\ell}{1 + (c/8)\ell \nu} \leq C\ell,
\]
since the function \(x \mapsto C/(1 + cx)\) is nonincreasing for any choice of the constants. Next, we use
the expression we have derived in the first claim to obtain
\[
\left|\nu^2 \bar{\psi}^{(\ell)}(\nu)\right| \leq \frac{CV}{1 + (c/8)\ell \nu} \left(\nu + \frac{1}{(c/8)} \log \left(1 + (c/8)\left(\ell - 1\right)\nu\right)\right).
\]
For any \(K > 0\), the function \(x \mapsto x/(1 + Kx)\) is nondecreasing, and using the numerical estimate
\( \pi(c/8) < 1 \) that follows from Lemma A.5.5, we obtain in addition \( 1 + \pi(c/8)(\ell - 1) \leq \ell \) for \( \ell \in \mathbb{N} \).

Thus

\[
\left| v^2 \varphi^{(\ell)}(v) \right| \leq \frac{C\pi^2}{1 + (c/8)\ell \pi} \left( 1 + \frac{\log \ell}{c\pi/8} \right)
\leq \frac{8\pi C}{c\ell} + \frac{64\pi \log \ell}{c^2 \ell},
\]

as claimed. \( \Box \)

**Lemma A.3.14.** One has for every \( \ell \in \{0, 1, \ldots, L\} \)

\[
\varphi^{(\ell)}(0) = 0; \quad \varphi^{(\ell)}(0) = 1; \quad \varphi^{(\ell)}(0) = -\frac{2\ell}{3\pi},
\]

and for every \( \ell \in [L] \)

\[
\varphi^{(\ell)}(\pi) = \varphi^{(\ell)}(\pi) = 0.
\]

Finally, we have \( \varphi^{(0)}(\pi) = 1 \) and \( \varphi^{(0)}(\pi) = 0. \)

**Proof.** The claims are consequences of Lemma A.5.5 when \( \ell = 1 \), and of \( \varphi^{(0)} = \text{Id} \) for smaller \( \ell \); assume \( \ell > 1 \) below. The claim for \( \varphi^{(\ell)}(0) \) follows from the fact that \( \varphi(0) = 0 \) and induction. For the claim about \( \varphi^{(\ell)}(0) \), we calculate using the chain rule

\[
\varphi^{(\ell)}(0) = \varphi(\varphi^{(\ell-1)}(0)) \varphi^{(\ell-1)}(0)
= \varphi(0) \varphi^{(\ell-1)}(0)
= \varphi^{(\ell-1)}(0).
\]

By induction and Lemma A.5.5, we obtain \( \varphi^{(\ell)}(0) = 1 \). The claim about \( \varphi^{(\ell)}(\pi) \) follows from the same argument. For the remaining claims about \( \varphi^{(\ell)} \), we calculate using the chain rule

\[
\varphi^{(\ell)} = (\varphi^{(\ell-1)})^2 \varphi \circ \varphi^{(\ell-1)} + (\varphi^{(\ell-1)}) \varphi \circ \varphi^{(\ell-1)},
\]

193
whence
\[ \psi^{(\ell)}(0) = \varphi(0) + \psi^{(\ell-1)}(0). \]

Using Lemma A.5.5 to get \( \varphi(0) = -2/(3\pi) \), this yields
\[ \psi^{(\ell)}(0) = \frac{2\ell}{3\pi}. \]

Similarly, since we have shown \( \varphi^{(\ell-1)}(\pi) = 0 \), we obtain \( \psi^{(\ell)}(\pi) = 0. \)

\[ \square \]

**Lemma A.3.15.** For first and second derivatives of \( \xi^{(\ell)} \), one has
\[ \xi^{(\ell)} = -\pi^{-1} \sum_{\ell' = \ell}^{L-1} \varphi^{(\ell')} \prod_{\ell'' = \ell}^{L-1} (1 - \pi^{-1} \varphi^{(\ell'')}), \] (A.117)

and
\[ \xi^{(\ell)} \] (A.118)

where empty sums are interpreted as zero, and empty products as 1. In particular, one calculates
\[ \xi^{(\ell)}(0) = 1; \quad \xi^{(\ell)}(0) = \frac{L - \ell}{\pi}; \quad \xi^{(\ell)}(0) = \frac{(L - \ell)(L - \ell - 1)}{\pi^2} + \frac{L(L - 1) - \ell(\ell - 1)}{3\pi^2}, \]

and
\[ \xi^{(\ell)}(\pi) = 0; \quad \xi^{(\ell)}(\pi) = \frac{1}{\pi} \xi^{(1)}(\pi) 1_{\ell=0}; \quad \xi^{(\ell)}(\pi) = 0. \]

**Proof.** The two derivative formulas are direct applications of the Leibniz rule to \( \xi^{(\ell)} \). The claims about values at 0 follow from plugging the results of Lemma A.3.14 into our derivative formulas and the definition of \( \xi^{(\ell)} \). For values at \( \pi \), we first note that \( \varphi^{(0)}(\pi) = \pi \), from which it follows \( \xi^{(0)}(\pi) = 0 \). Next, we use Lemma A.3.14 to get that \( \varphi^{(\ell)}(\pi) = 0 \) for all \( \ell \in \{0, 1, \ldots, L\} \) and
\( \varphi^{(\ell)}(\pi) = \mathbb{1}_{\ell=0} \) to get \( \tilde{\varphi}^{(\ell)}(\pi) = -\pi^{-1} \varphi^{(1)}(\pi) \mathbb{1}_{\ell=0} \). For \( \tilde{\varphi}^{(\ell)}(\pi) \), we have

\[
\tilde{\varphi}^{(\ell)} = \pi^{-2} \sum_{\ell' = 0}^{L-1} \varphi^{(\ell')} \left( \sum_{\ell'' \neq \ell'} (1 - \pi^{-1} \varphi^{(\ell'')}(\pi)) \right)
\]

If \( L = 1 \), the sum in the last expression is empty, and this quantity is 0. If \( L > 1 \), the sum is nonempty, and every summand is equal to zero by Lemma A.3.14. We conclude \( \tilde{\varphi}^{(\ell)}(\pi) = 0. \)

**Lemma A.3.16.** If \( L \geq 3 \), there exists an absolute constant \( 0 < C \leq \pi/2 \) such that on the interval \([0, C]\), one has for every \( \ell = 0, 1, \ldots, L - 1 \)

\[
\tilde{\varphi}^{(\ell)} \leq 0.
\]

**Proof.** We consider functions only on \([0, \pi/2]\) in this proof. Following the calculations in the proof of Lemma A.3.12, we have the expression

\[
(\varphi \circ \varphi^{(\ell-1)})^{(\ell)} = \left( \varphi \circ \varphi^{(\ell-1)} \right)^{\ell} + 3 \left( \varphi \circ \varphi^{(\ell-1)} \right) \varphi^{(\ell-1)} + \left( \varphi \circ \varphi^{(\ell-1)} \right)^3.
\]

Using as well Lemma A.5.5, we have first and second derivative estimates

\[
0 \leq \varphi^{(\ell)} \leq 1
\]

and

\[
-C_2 \ell \leq \varphi^{(\ell)} \leq -c_2 \ell.
\]

By Lemma A.5.5, \( \varphi \) extends to a continuous function on \([0, \pi/2]\), so in addition there exists a \( \delta > 0 \) such that on \([0, \delta]\) we have

\[
\varphi \geq -\frac{1}{2\pi^2}
\]
We lower bound (A.120) on $[0, \delta]$ using these estimates. For $\ell = 1$, we can do no better than (A.121). For $\ell > 1$, we can write

$$
\left(\varphi \circ \varphi^{(\ell-1)}\right)^{'''} \geq \left(\varphi \circ \varphi^{(\ell-1)}\right) \varphi^{(\ell-1)} + 3\varphi^{(\ell-1)} \left(\left(\varphi \circ \varphi^{(\ell-1)}\right) \varphi^{(\ell-1)} - \frac{1}{6\pi^2} \left(\varphi^{(\ell-1)}\right)^2\right) 
\geq \left(\varphi \circ \varphi^{(\ell-1)}\right) \varphi^{(\ell-1)} + 3\varphi^{(\ell-1)} \left(c_2^2(\ell - 1) - \frac{1}{6\pi^2}\right).
$$

We have the numerical estimate $c_2 = 0.14$ from Lemma A.5.5, and we check numerically that $(0.14)^2 > 1/6\pi^2$. This implies that on $[0, \delta]$ and for every $\ell \geq 2$, $\bar{\varphi}^{(\ell)}$ is lower bounded by a positive number plus a scaled version of $\varphi^{(\ell-1)}$. We check precisely using the original formula (A.120) and Lemma A.5.5 for $\ell = 2$

$$
\bar{\varphi}^{(2)}(0) = 2\varphi(0) + 3\varphi(0)^2 = \frac{2}{3\pi^2} > 0,
$$

so that in particular

$$
\bar{\varphi}^{(2)}(0) + \bar{\varphi}^{(1)}(0) = \frac{1}{3\pi^2} > 0.
$$

By continuity, it follows that there is a neighborhood $[0, \delta']$ on which we have $\bar{\varphi}^{(2)} + \bar{\varphi}^{(1)} > 0$. Thus, on $[0, \min\{\delta, \delta'\}]$, we guarantee that simultaneously

$$
\bar{\varphi}^{(\ell)} > 0 \text{ if } \ell \geq 2; \quad \bar{\varphi}^{(2)} + \bar{\varphi}^{(1)} > 0.
$$

Now we consider the third derivative of the skeleton summands $\xi^{(\ell)}$. Following the calculations of Lemmas A.3.7 and A.3.15, in particular applying the Leibniz rule, we observe that every term in the sum defining $\xi^{(\ell)}$ that does not involve a third derivative of one of the factors $(1 - (1/\pi)\varphi^{(\ell')})$ will be nonpositive, because $(1 - (1/\pi)\varphi^{(\ell')}) \geq 0$, $\varphi^{(\ell')} \geq 0$, and $\varphi^{(\ell')} \leq 0$. Meanwhile, by our calculations above, on the interval $[0, \min\{\delta, \delta'\}]$, the only terms that can be positive are those with $\ell = 0$ or $\ell = 1$ where we differentiate the $\ell' = 1$ factor three times, i.e., the $\ell' = 1$ term in the
sum
\[-1 \sum_{\ell'=\ell}^{L-1} \prod_{\ell''=\ell, \ell'' \neq \ell'} \left( 1 - \frac{\varphi(\ell'')}{\pi} \right) \]
with \( \ell = 0 \) or \( \ell = 1 \). We will compare the \( \ell' = 1 \) summand with the \( \ell' = 2 \) summand: we have that the sum of these two terms equals

\[-1 \sum_{\ell'=\ell}^{L-1} \prod_{\ell''=\ell, \ell'' \neq \ell'} \left( 1 - \frac{\varphi(\ell'')}{\pi} \right) \]

At 0, the quantity inside the right parentheses is equal to \( \bar{\varphi} + \bar{\varphi}^{(2)} > 0 \), by our calculations above. Thus, by continuity, there is a possibly smaller \( \delta'' > 0 \) such that on \([0, \delta'']\), the sum of terms (A.122) is negative. We conclude that on \([0, \min\{\delta, \delta', \delta''\}]\), we have for every \( \ell \geq 0 \)

\[\bar{\varphi}^{(\ell)} \leq 0,\]

and since we have chosen the neighborhood sizes \( \delta, \delta', \delta'' \) independently of the depth \( L \), we can conclude. □

**Lemma A.3.17.** For all \( \ell \in \{0, \ldots, L - 1\} \), one has

\[\xi^{(\ell)}(v) \leq \frac{1 + \ell v / \pi}{1 + L v / \pi}.\]

**Proof.** We have

\[\xi^{(\ell)}(v) = \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(v)}{\pi} \right) \leq \left( 1 - \frac{1}{\pi (L - \ell)} \sum_{\ell' = \ell}^{L-1} \varphi^{(\ell')}(v) \right)^{L-\ell} \leq \exp \left( -\frac{1}{\pi} \sum_{\ell' = \ell}^{L-1} \varphi^{(\ell')}(v) \right),\]

where the first inequality applies the AM-GM inequality, and the second uses the standard expo-
ential convexity estimate. Using Lemma A.3.10, we have

\[- \sum_{\ell' = \ell}^{L-1} \varphi^{(\ell')} (v) \leq - \sum_{\ell' = \ell}^{L-1} \frac{v}{1 + \ell' v / \pi} \leq - \int_{\ell}^{L} \frac{v}{1 + \ell' v / \pi} \, \mathrm{d} \ell',\]

where the last inequality uses the fact that \( \ell' \mapsto v/(1 + \ell' v / \pi) \) is nonincreasing for every \( v \in [0, \pi] \) together with a standard estimate from the integral test. We calculate

\[ \int_{\ell}^{L} \frac{v}{1 + \ell' v / \pi} \, \mathrm{d} \ell' = \pi \log \left( \frac{1 + L v / \pi}{1 + \ell v / \pi} \right), \]

which gives the claim after substituting into (A.123).

\[ \square \]

**Lemma A.3.18.** For all \( \ell \in \{0, \ldots, L - 1\} \), one has

\[ |\hat{\xi}^{(\ell)} (v)| \leq \frac{3}{1 + L v / \pi} \cdot \frac{L - \ell}{1 + L v / \pi}. \]

**Proof.** Using Lemma A.3.15, we have

\[ \hat{\xi}^{(\ell)} = -\frac{\xi^{(1)}}{\pi} \mathbb{1}_{\ell=0} - \frac{\xi^{(\ell)}}{\pi} \sum_{\ell' = \max\{\ell,1\}}^{L-1} \frac{\psi^{(\ell')}}{1 - \varphi^{(\ell')}}. \]

where we directly treat the case \( \ell = 0 \) to avoid dividing by zero at \( v = \pi \). The triangle inequality and Lemmas A.5.5 and A.3.17 then give

\[ |\hat{\xi}^{(\ell)} (v)| \leq \frac{2}{\pi} \left( \frac{1}{1 + L v / \pi} \mathbb{1}_{\ell=0} + \xi^{(\ell)} (v) \sum_{\ell' = \max\{\ell,1\}}^{L-1} \psi^{(\ell')} (v) \right). \]

Using Lemma A.3.12, we have

\[ \sum_{\ell' = \ell}^{L-1} \varphi^{(\ell')} (v) \leq \sum_{\ell' = \ell}^{L-1} \frac{1}{1 + c \ell' v} \leq \frac{1}{1 + c \ell v} + \int_{\ell}^{L-1} \frac{1}{1 + c \ell' v} \, \mathrm{d} \ell', \]

where the last inequality uses the fact that \( \ell' \mapsto 1/(1 + c \ell' v) \) is nonincreasing for every \( v \in [0, \pi] \).
together with a standard estimate from the integral test. Evaluating the integral, we obtain

\[ \sum_{\ell=\ell}^{L-1} \psi^{(\ell)}(\nu) \leq \frac{1}{1 + c \ell \nu} + \frac{1}{c \nu} \log \left( \frac{1 + c(L-1)\nu}{1 + c \ell \nu} \right), \]

where the second term on the RHS is defined at \( \nu = 0 \) by continuity. Using the standard concavity estimate \( \log(1 + x) \leq x \), we have

\[ \frac{1}{c \nu} \log \left( \frac{1 + c(L-1)\nu}{1 + c \ell \nu} \right) = \frac{1}{c \nu} \log \left( 1 + \frac{(L-\ell-1)c\nu}{1 + c \ell \nu} \right) \leq \frac{L-\ell-1}{1 + c \ell \nu}, \]

whence

\[ \sum_{\ell=\ell}^{L-1} \psi^{(\ell)}(\nu) \leq \frac{L-\ell}{1 + c \ell \nu}. \] (A.124)

Combined with the result of Lemma A.3.17, we conclude

\[ \xi^{(\ell)}(\nu) \sum_{\ell'=\ell}^{L-1} \psi^{(\ell')}(\nu) \leq \frac{1}{c \pi} \frac{L-\ell}{1 + L \nu / \pi}. \]

The numerical estimate \( c = 0.07 \) in Lemma A.5.5 then allows us to conclude

\[ |\xi^{(\ell)}(\nu)| \leq \frac{3}{1 + L \nu / \pi}, \]

as claimed. \( \square \)

**Lemma A.3.19.** One has

\[ |\psi'_1(\nu)| \leq \frac{5nL^2}{1 + L \nu / \pi}, \]

and

\[ |\psi'(\nu)| \leq \frac{(3/2)nL^2}{1 + L \nu / \pi}. \]

**Proof.** We calculate using the chain rule

\[ \psi'_1 = \frac{n}{2} \sum_{\ell=0}^{L-1} \xi^{(\ell)} \cos \varphi^{(\ell)} - \xi^{(\ell)} \phi^{(\ell)} \sin \varphi^{(\ell)}, \]
and the triangle inequality gives

\[ |\psi_1'| \leq \frac{n}{2} \sum_{\ell=0}^{L-1} |\xi^{(\ell)}| + x^{(\ell)} \phi^{(\ell)}. \]

Applying Lemmas A.3.12, A.3.17 and A.3.18 and Lemma A.5.5 to estimate the constant \( c \) in

Lemma A.3.12, we then obtain

\[
|\psi_1'(v)| \leq \frac{n}{2(1 + Lv/\pi)} \sum_{\ell=0}^{L-1} 3(L - \ell) + \frac{1 + \ell v/\pi}{1 + \ell v/(5\pi)} \\
\leq \frac{n}{2(1 + Lv/\pi)} \sum_{\ell=0}^{L-1} 3(L - \ell) + 1 + 4 \frac{\ell v/(5\pi)}{1 + \ell v/(5\pi)} \\
\leq \frac{n}{2(1 + Lv/\pi)} \left( \frac{3L^2}{2} + 5L \right) \\
\leq \frac{5nL^2}{1 + Lv/\pi}.
\]

The proof of the second claim is nearly identical, since in this case we need only use the bounds on \( |\xi^{(\ell)}| \). □

**Lemma A.3.20.** There are absolute constants \( c, C > 0 \) such that for all \( \ell \in \{0, \ldots, L - 1\} \), one has

\[ |\hat{\xi}^{(\ell)}| \leq C \frac{L(L - \ell)(1 + \ell v/\pi)}{(1 + cLv)^2} + C \frac{(L - \ell)^2}{(1 + cLv)(1 + c\ell v)}. \]

**Proof.** By Lemmas A.5.5 and A.3.15, we can write

\[
\hat{\xi}^{(\ell)} = -\frac{\xi^{(\ell)}}{\pi} \sum_{\ell' = \max\{1, \ell\}}^{L-1} \frac{\phi^{(\ell')}}{1 - \frac{\phi^{(\ell')}}{\pi}} \\
+ \frac{1}{\pi^2} \left( 2\xi^{(1)} I_{\ell=0} \sum_{\ell' = 1}^{L-1} \frac{\phi^{(\ell')}}{1 - \frac{\phi^{(\ell')}}{\pi}} + \xi^{(\ell)} \sum_{\ell' = \max\{1, \ell\}}^{L-1} \sum_{\ell'' = \max\{1, \ell\}}^{L-1} \frac{\phi^{(\ell')}}{1 - \frac{\phi^{(\ell')}}{\pi}} \left( 1 - \frac{\phi^{(\ell'')}}{\pi} \right) \left( 1 - \frac{\phi^{(\ell'')}}{\pi} \right) \right) .
\]
Focusing first on the second term, we have using Lemma A.5.5, (A.124) and Lemma A.3.17
\[
2\xi^{(1)} \sum_{\ell=0}^{L-1} \frac{\phi^{(\ell')}}{1 - \frac{\phi^{(\ell')}}{\pi}} + \xi^{(1)} \sum_{\ell=0}^{L-1} \sum_{\ell' = \max\{1, \ell\}}^{L-1} \sum_{\ell'' = \max\{1, \ell\}}^{L-1} \frac{\phi^{(\ell')}}{1 - \frac{\phi^{(\ell')}}{\pi}} \left(1 - \frac{\phi^{(\ell'')}}{\pi}ight)
\leq 4\xi^{(1)} \sum_{\ell=0}^{L-1} \phi^{(\ell')} + 4\xi^{(1)} \sum_{\ell=0}^{L-1} \sum_{\ell' = \max\{1, \ell\}}^{L-1} \sum_{\ell'' = \max\{1, \ell\}}^{L-1} \phi^{(\ell')} \phi^{(\ell'')}.
\]

We can then write using nonnegativity
\[
\sum_{\ell=0}^{L-1} \sum_{\ell' = \ell}^{L-1} \phi^{(\ell')} \phi^{(\ell'')} \leq \sum_{\ell=0}^{L-1} \sum_{\ell' = \ell}^{L-1} \phi^{(\ell')} \phi^{(\ell'')} = \left(\sum_{\ell=0}^{L-1} \phi^{(\ell')}ight)^2,
\]
and using (A.124) and Lemma A.3.17, we obtain thus
\[
\xi^{(1)} \sum_{\ell=0}^{L-1} \phi^{(\ell')} + \xi^{(1)} \sum_{\ell=0}^{L-1} \sum_{\ell' = \max\{1, \ell\}}^{L-1} \sum_{\ell'' = \max\{1, \ell\}}^{L-1} \phi^{(\ell')} \phi^{(\ell'')} \leq \frac{3}{c\pi} \left(\frac{(L - 1)^2}{1 + c\ell\nu}(1 + c\ell\nu) \right).
\]

Regarding the first term, we have using Lemma A.3.13
\[
\sum_{\ell=0}^{L-1} |\phi^{(\ell')}| \leq C \sum_{\ell=0}^{L-1} \frac{\ell'}{1 + (c/4)\ell\nu} \leq C \frac{L(L - \ell)}{1 + (c/4)L\nu},
\]
because the function \( \ell' \mapsto \ell' / (1 + c\ell\nu) \) is nondecreasing. Applying also Lemma A.3.17, we obtain using the triangle inequality and worst-casing constants
\[
\left|\tilde{\xi}(\ell')\right| \leq C_1 \frac{L(L - \ell)(1 + \ell\nu/\pi)}{(1 + cL\nu)^2} + C_2 \frac{(L - \ell)^2}{(1 + cL\nu)(1 + c\ell\nu)}.
\]

\[\square\]

**Lemma A.3.21.** One has
\[
|\psi_1^{(\nu)}(\nu)| \leq \frac{CnL^3}{1 + cL\nu},
\]

201
and

\[ |\psi''(\nu)| \leq \frac{Cn^3}{1 + cL\nu}, \]

where \( c, C > 0 \) are absolute constants.

**Proof.** We calculate using the chain rule

\[
\psi''_1 = \frac{n}{2} \sum_{\ell=0}^{L-1} \xi^{(\ell)} \cos \varphi^{(\ell)} - 2 \xi^{(\ell)} \varphi^{(\ell)} \sin \varphi^{(\ell)} - \xi^{(\ell)} \varphi^{(\ell)} \sin \varphi^{(\ell)} - \xi^{(\ell)} \left( \varphi^{(\ell)} \right)^2 \cos \varphi^{(\ell)},
\]

and the triangle inequality gives

\[
|\psi''_1| \leq \frac{n}{2} \sum_{\ell=0}^{L-1} \xi^{(\ell)} + 2 \left| \xi^{(\ell)} \right| |\varphi^{(\ell)}| + \xi^{(\ell)} \left( \varphi^{(\ell)} \right)^2.
\]

Using Lemmas A.3.12, A.3.13, A.3.17, A.3.18 and A.3.20 and worst-casing constants for convenience, we obtain from the last estimate

\[
|\psi''_1(\nu)| \leq Cn \sum_{\ell=0}^{L-1} \left( \frac{L(L-\ell)(1+\ell\nu/\pi)}{(1+cL\nu)^2} + \frac{(L-\ell)^2}{(1+cL\nu)(1+c\ell\nu)} \right) \leq \frac{Cn^3}{1 + cL\nu},
\]

where in the second line we made some estimates along the lines of the proof of Lemma A.3.19 and worsened the constant \( C \). The proof for \( \psi \) follows from the same argument, since in this case we have the same sum of \( \xi^{(\ell)} \) terms but none of the extra residuals. \( \square \)

### A.4 Concentration at Initialization

#### A.4.1 Notation and Framework

We recall the expression for the neural tangent kernel, as summarized in Section A.1.4:

\[
\Theta(x, x') = \left\langle \overline{\nabla f_{\theta_0}(x), \overline{f_{\theta_0}}(x')} \right\rangle
= \langle \alpha^L(x), \alpha^L(x') \rangle + \sum_{\ell=0}^{L-1} \langle \alpha^\ell(x), \alpha^\ell(x') \rangle \langle \beta^\ell(x), \beta^\ell(x') \rangle,
\]
The objective of this section is to establish supporting results for the proof of Theorem A.2.2, which gives uniform concentration of $\Theta(x, x')$ over $M \times M$ around the deterministic skeleton kernel. We take a pointwise-uniformize approach to proving this result: Section A.4.2 establishes concentration results for the constituents of $\Theta(x, x')$ when $x, x'$ are fixed, and Section A.4.3 develops results that control the number of local support changes near points in a discretization of $M \times M$ in order to provide a suitable stand-in for the continuity properties necessary to uniformize these pointwise results. We collect relevant technical results and their proofs in Section A.4.4.

A.4.2 Pointwise Concentration

We fix $(x, x')$ in this section, and generally suppress notation involving the specific points for concision. We separate our analysis into two distinct sub-problems: “forward concentration”, which consists of the study of the correlations $\langle \alpha^\ell(x), \alpha^\ell(x') \rangle$, and “backward concentration”, which consists of the study of the backward feature correlations $\langle \beta^\ell(x), \beta^\ell(x') \rangle$. Forward concentration is a prerequisite of our approach to backward concentration, so we begin there.

Forward Concentration

Notation. For $\ell = 0, 1, \ldots, L$, define random variables $z_1^\ell = ||\alpha^\ell(x)||_2$ and $z_2^\ell = ||\alpha^\ell(x')||_2$. With the convention $0 \cdot +\infty = 0$, we define for $\ell = 0, \ldots, L$, random variables $\nu^\ell$ by

$$\nu^\ell = \cos^{-1}\left(1_{z_1^\ell > 0}1_{z_2^\ell > 0} \frac{\alpha^\ell(x)}{||\alpha^\ell(x)||_2}, \frac{\alpha^\ell(x')}{||\alpha^\ell(x')||_2}\right) - 1_{\{z_1^\ell = 0\} \cup \{z_2^\ell = 0\}}.$$ 

These definitions guarantee that $\nu^\ell = \pi$ whenever either feature norm $z_i^\ell$ vanishes. These random variables are significant toward controlling $\Theta(x, x')$ because, for each $\ell$

$$\langle \alpha^\ell(x), \alpha^\ell(x') \rangle = z_1^\ell z_2^\ell \cos \nu^\ell.$$ 

Let us define pairs of gaussian vectors $g_1^\ell, g_2^\ell \sim \text{i.i.d. } N(0, (2/n)I)$ that are independent of everything else in the problem. For $\ell \geq 1$, we have by rotational invariance of the Gaussian distribution
and the probability chain rule
\[
z_1^\ell = \| W^\ell \alpha^{\ell-1}(x) \|_2 + \frac{d}{2} \| g_1^\ell \|_2 z_1^{\ell-1}.
\]
Since \( \alpha^0(x) = x \) and \( \|x\|_2 = 1 \), we have by an induction with analogous definitions
\[
z_1^\ell = \prod_{\ell' = 1}^{\ell} \| g_1^{\ell'} \|_2.
\]
Similarly, we have
\[
z_2^\ell = \prod_{\ell' = 1}^{\ell} \| g_1^{\ell'} \|_2.
\]
As for the angles, we have by rotational invariance
\[
z_1^\ell z_2^\ell = \| W^\ell \alpha^{\ell-1}(x) \|_2 \| W^\ell \alpha^{\ell-1}(x') \|_2 + \frac{d}{2} \| g_1^\ell \|_2 \| g_1^\ell \cos \nu^{\ell-1} + g_2^\ell \sin \nu^{\ell-1} \|_2 z_1^{\ell-1} z_2^{\ell-1},
\]
so that an inductive argument gives
\[
z_1^\ell z_2^\ell = \left( \prod_{\ell' = 1}^{\ell} \| g_1^{\ell'} \|_2 \right) \left( \prod_{\ell' = 1}^{\ell} \| g_1^{\ell'} \cos \nu^{\ell-1} + g_2^{\ell'} \sin \nu^{\ell-1} \|_2 \right) z_1^{\ell-1} z_2^{\ell-1}.
\]
We will write
\[
\bar{z}_1^\ell = \prod_{\ell' = 1}^{\ell} \| g_1^{\ell'} \|_2, \quad \bar{z}_2^\ell = \prod_{\ell' = 1}^{\ell} \| g_1^{\ell'} \cos \nu^{\ell-1} + g_2^{\ell'} \sin \nu^{\ell-1} \|_2,
\]
and similarly
\[
\bar{\nu}^\ell = \cos^{-1} \left( 1_{z_1^\ell z_2^\ell > 0} \left( \frac{\| g_1^\ell \|_2}{\| g_1^\ell \|_2} \| g_1^\ell \cos \nu^{\ell-1} + g_2^\ell \sin \nu^{\ell-1} \|_2 \| g_1^\ell \cos \nu^{\ell-1} + g_2^\ell \sin \nu^{\ell-1} \|_2 + 1_{z_1^\ell = 0 \cup z_2^\ell = 0} \right) \right),
\]
204
so that we obtain for the angles by a similar inductive argument

\[ \nu^\ell \overset{d}{=} \bar{\nu}^\ell. \quad (A.125) \]

For technical reasons, it will be convenient to consider an auxiliary angle process, defined for \( \ell \geq 1 \) as

\[ \hat{\nu}^\ell = \cos^{-1} \left( \mathbb{1}_E(\mathbf{g}_1^\ell, \mathbf{g}_2^\ell) \left( \left\| \mathbf{g}_1^\ell \right\|_2, \left\| \mathbf{g}_2^\ell \right\|_2 \right) \right), \quad (A.126) \]

where we define with notation from Section A.5.1

\[ \bar{E} = \bigcap_{i \in [n]} \{ (\mathbf{g}_1^i, \mathbf{g}_2^i) \mid \forall \nu \in [0, 2\pi], \frac{1}{2} \leq \left\| \mathbf{g}_1^i \cos \nu + \mathbf{g}_2^i \sin \nu \right\|_2 \leq 2 \}, \]

and \( \hat{\nu}^0 = \nu^0 = \langle x, x' \rangle \). We then observe

\[ \prod_{\ell' = 1}^\ell \mathbb{1}_{\bar{E}}(\mathbf{g}_1^{\ell'}, \mathbf{g}_2^{\ell'}) \leq \prod_{\ell' = 1}^\ell \mathbb{1}_{\bar{z}_i^L > 0}, \]

since the inductive structure of \( \bar{z}_i^L \) implies that all feature norms are nonvanishing if and only if the top-level feature norms \( \bar{z}_i^L \) are nonvanishing, and since the statement \( \prod_{\ell' = 1}^\ell \mathbb{1}_E(\mathbf{g}_1^{\ell'}, \mathbf{g}_2^{\ell'}) = 1 \) implies by definition that \( \bar{z}_1^L \geq 2^{-L} \) and \( \bar{z}_2^L \geq 2^{-L} \). By Lemma A.5.16, as long as \( n \geq 21 \) the event \( \bar{E} \) has overwhelming probability, and in particular a union bound implies

\[ \mathbb{P} \left[ \prod_{\ell' = 1}^\ell \mathbb{1}_{\bar{z}_i^L > 0} = 1 \right] \geq \mathbb{P} \left[ \prod_{\ell' = 1}^\ell \mathbb{1}_{\bar{E}}(\mathbf{g}_1^{\ell'}, \mathbf{g}_2^{\ell'}) = 1 \right] \geq 1 - CLe^{-cn}, \quad (A.127) \]

so that

\[ \mathbb{P} \left[ \forall \ell = 1, 2, \ldots, L, \hat{\nu}^\ell = \bar{\nu}^\ell \right] \geq 1 - CLe^{-cn}. \quad (A.128) \]

We can therefore pass from \( \bar{\nu}^\ell \) to \( \hat{\nu}^\ell \) with negligible error.

From the expression for \( \hat{\nu}^\ell \), we see that the angles \( \hat{\nu}^0 \to \hat{\nu}^1 \to \cdots \to \hat{\nu}^L \) form a Markov chain, and we will control them using martingale techniques. For \( \ell = 0, 1, \ldots, L \), we write \( \mathcal{F}_\ell \) to denote
the \( \sigma \)-algebra generated by the gaussian vectors \((g_1, g_2, g_1', g_2', \ldots, g_\ell, g_\ell')\), so that \((\mathcal{F}^0, \ldots, \mathcal{F}^L)\) is a filtration, and the sequences of random variables \((\check{v}^1, \ldots, \check{v}^L)\) and \((\check{v}^1, \ldots, \check{v}^L)\) are adapted to \((\mathcal{F}^1, \ldots, \mathcal{F}^L)\). Moreover, with these definitions we have

\[
\mathbb{E}[\check{v}^\ell | \mathcal{F}^{\ell-1}] = \check{\varphi}(\check{v}^{\ell-1}),
\]

where \(\check{\varphi}\) is the angle evolution function defined in Section A.5.1, which is well-approximated by the function

\[
\varphi(v) = \cos^{-1}\left(\left(1 - \frac{v}{\pi}\right) + \frac{\sin v}{\pi}\right)
\]

(see Lemmas A.5.1 and A.5.2). In the sequel, we will employ the notation \(\varphi^{(\ell)}\) to denote the \(\ell\)-fold composition of \(\varphi\) with itself. By Lemma A.5.5, the function \(\varphi\) is smooth, and the chain rule implies the same for \(\varphi^{(\ell)}\); we will employ the notation \(\check{\varphi}^{(\ell)}\) and \(\varphi^{(\ell)}\) for the first and second derivatives of \(\varphi^{(\ell)}\), respectively.

**Main results.**

**Lemma A.4.1.** There are absolute constants \(c, C, C' > 0\) and absolute constants \(K, K' > 0\) such that for any \(d \geq K\), if \(n \geq K' \max\{1, d^4 \log^4 n, d^3 L \log^3 n\}\) then one has for any \(\ell = 1, \ldots, L \)

\[
\mathbb{P}\left[ \left| \langle \alpha^{(\ell)}(x), \alpha^{(\ell)}(x') \rangle \right| - \cos \varphi^{(\ell)}(\angle(x, x')) \right| > C \sqrt{\frac{d^3 \ell \log^3 n}{n}} \right] \leq C'n^{-cd}.
\]

**Proof.** We have

\[
\langle \alpha^{(\ell)}(x), \alpha^{(\ell)}(x') \rangle = z_1^\ell z_2^\ell \cos \nu^\ell,
\]

and the triangle inequality (applied twice) then yields

\[
\left| \langle \alpha^{(\ell)}(x), \alpha^{(\ell)}(x') \rangle \right| - \cos \varphi^{(\ell)}(\nu^0) \leq |\cos \nu^\ell| |z_1^\ell z_2^\ell - 1| + |\cos \nu^\ell - \cos \varphi^{(\ell)}(\nu^0)|
\]

\[
\leq |z_2^\ell| |z_1^\ell - 1| + |z_1^\ell - 1| + |\nu^\ell - \varphi^{(\ell)}(\nu^0)|,
\]
where we also use $|\cos| \leq 1$ and that $\cos$ is 1-Lipschitz. Since $z_i^\ell = \tilde{z}_i^\ell$ for $i = 1, 2$, we obtain using Lemma A.4.2 and the choice $n \geq KdL$

$$P\left[|z_1^\ell - 1| > C \sqrt{\frac{d \ell}{n}}\right] \leq C' \ell e^{-d},$$

and as long as $n \geq C^2dL$, we obtain on one of the same events

$$P[z_2^\ell \leq 2] \geq 1 - C' \ell e^{-d}.$$  

By a union bound, we obtain

$$P\left[|z_2^\ell - z_1^\ell - 1| + |z_2^\ell - 1| \leq 3C \sqrt{\frac{d \ell}{n}}\right] \geq 1 - 2C' \ell e^{-d},$$

so that if we put $d' = d \log n$ and therefore choose $n \geq C^2dL \log n$, we have

$$P\left[|z_2^\ell - z_1^\ell - 1| + |z_2^\ell - 1| \leq 3C \sqrt{\frac{d \ell \log n}{n}}\right] \geq 1 - 2C' \ell n^{-d} \geq 1 - 2C' n^{-d},$$

with the second bound holding if $d \geq 1$ and $n \geq L$. For the remaining term, we have by the triangle inequality

$$|v^\ell - \varphi^{(i)}(v^0)| \leq |v^\ell - \hat{v}^\ell| + |\hat{v}^\ell - \varphi^{(i)}(v^0)|.$$  

By (A.128), the first term on the RHS of the previous expression is equal to zero with probability at least $1 - CLe^{-cn}$ as long as $n \geq 21$. The second term can be controlled with Lemma A.4.3 provided we select $n, L, d$ to satisfy the hypotheses of that lemma. We thus obtain via an additional union bound

$$P\left[|\alpha^\ell(x), \alpha^\ell(x') - \cos \varphi^{(i)}(v^0)| > 3C \sqrt{\frac{d \ell \log n}{n}} + C' \sqrt{\frac{d^3 \log n}{n \ell}}\right] \leq C'' n^{-cd} + C''' \ell e^{-c' n}.$$  

If $n \geq (2/c') \log L$ and $n \geq (2c/c')d \log n$, we have $C'' n^{-cd} + C''' \ell e^{-c' n} \leq (C'' + C''')n^{-cd}$. The
previous bound then becomes
\[
\mathbb{P} \left[ \left| \langle \alpha^\ell(x), \alpha^\ell(x') \rangle - \cos \varphi^{(l)}(\nu^0) \right| > 3C \sqrt{\frac{d \ell \log n}{n}} + C' \sqrt{\frac{d^3 \log^3 n}{n \ell}} \right] \leq (C'' + C''')n^{-c'd},
\]
and if we worst-case the dependence on \( \ell \) and \( d \) in the residual in the previous bound, we obtain
\[
\mathbb{P} \left[ \left| \langle \alpha^\ell(x), \alpha^\ell(x') \rangle - \cos \varphi^{(l)}(\nu^0) \right| > (3C + C') \sqrt{\frac{d^3 \log^3 n}{n}} \right] \leq (C'' + C''')n^{-c'd},
\]
as claimed. \( \square \)

**Lemma A.4.2.** There are absolute constants \( c, C, C' > 0 \) and an absolute constant \( K > 0 \) such that for \( i = 1, 2 \), every \( \ell = 1, \ldots, L \), and any \( d > 0 \), if \( n \geq \max \{ Kd\ell, 4 \} \), then one has
\[
\mathbb{P} \left[ |z_\ell^i - 1| > C' \sqrt{\frac{d \ell}{n}} \right] \leq C' \ell e^{-c'd}. \]

**Proof.** Because \( z_\ell^i \overset{d}{=\sim} z_1^\ell \), it suffices to show
\[
\mathbb{P} \left[ -1 + \prod_{\ell'=1}^\ell \left\| g_{1,\ell'}^{\ell'} \right\|_2 > C' \sqrt{\frac{d \ell}{n}} \right] \leq C' \ell e^{-c'd}. \tag{A.129}
\]
The proof will proceed by showing concentration of the squared quantity \( \prod_{\ell'=1}^\ell \left\| g_{1,\ell'}^{\ell'} \right\|_2^2 \) around 1, so that we can appeal to results like Lemma A.4.16, and then conclude by applying an inequality for the square root to pass to the actual quantity of interest. To enter the setting of Lemma A.4.16, it makes sense to normalize the factors in the product by their degree, but we must avoid dividing by zero. We have \( \prod_{\ell'=1}^\ell \left\| g_{1,\ell'}^{\ell'} \right\|_0 = 0 \) if and only if \( \prod_{\ell'=1}^\ell \left\| g_{1,\ell'}^{\ell'} \right\|_2 = 0 \), and whenever \( \prod_{\ell'=1}^\ell \left\| g_{1,\ell'}^{\ell'} \right\|_2 \neq 0 \), we can write
\[
\prod_{\ell'=1}^\ell \left\| g_{1,\ell'}^{\ell'} \right\|_2 = \left( \prod_{\ell'=1}^\ell \left\| g_{1,\ell'}^{\ell'} \right\|_2 \right)^{1/2} \tag{A.130}
\]
\[
\left( \prod_{\ell' = 1}^{\ell} \frac{2}{n} \left\| g_1^{\ell'} \right\|_0 \right)^{1/2} = \left( \prod_{\ell' = 1}^{\ell} \frac{1}{\sqrt{\alpha} \left\| g_1^{\ell'} \right\|_0} \right) \left( \frac{n}{2} \left\| g_1^{\ell'} \right\|_2^2 \right)^{1/2}, \tag{A.131}
\]

using 0-homogeneity of the \( \ell^0 \) “norm”. This leads to an extra product-of-degrees term; we will make use of Lemma A.4.17 to show that the product of degrees itself concentrates. We will also show that the event where a degree is zero is extremely unlikely and proceed with the degree-normalized main term by conditioning. By symmetry, the random variables \( \left\| g_1^{\ell'} \right\|_0 \) are i.i.d. sums of \( n \) Bernoulli random variables with rate \( \frac{1}{2} \). By Lemma A.6.1, we then have

\[
P \left[ \left\| g_1^{\ell'} \right\|_0 < n/2 - t \right] \leq e^{-2t^2/n},
\]

and so

\[
P \left[ \min_{\ell' = 1, \ldots, \ell} \left\| g_1^{\ell'} \right\|_0 < n/2 - t \right] = P \left[ \exists \ell' \in \{1, \ldots, \ell\} : \left\| g_1^{\ell'} \right\|_0 < n/2 - t \right] \\
\leq \ell P \left[ \left\| g_1^{\ell'} \right\|_0 < n/2 - t \right] \leq \ell e^{-2t^2/n},
\]

where the first inequality applies a union bound. Putting \( t = n/4 \), we conclude

\[
P \left[ \min_{\ell' = 1, \ldots, \ell} \left\| g_1^{\ell'} \right\|_0 < n/4 \right] \leq \ell e^{-n/8},
\]

so that whenever \( n \geq 16 \log \ell \), we have \( \left\| g_1^{\ell'} \right\|_0 \geq n/4 \) for every \( \ell' \leq \ell \) with probability at least \( 1 - e^{-n/16} \). This gives us enough to begin working on showing concentration of the squared version of (A.129): partitioning, we can use the previous simplified bound to write

\[
P \left[ -1 + \prod_{\ell' = 1}^{\ell} \left\| g_1^{\ell'} \right\|_2^2 > C \sqrt{\frac{d \ell}{n}} \right] \\
\leq e^{-n/16} + P \left[ \min_{\ell' = 1, \ldots, \ell} \left\| g_1^{\ell'} \right\|_0 \geq n/4, \ -1 + \prod_{\ell' = 1}^{\ell} \left\| g_1^{\ell'} \right\|_2^2 > C \sqrt{\frac{d \ell}{n}} \right]. \tag{A.132}
\]

\[
P \left[ \left| \prod_{\ell' = 1}^{\ell} \left\| g_1^{\ell'} \right\|_2^2 \right| > C \sqrt{\frac{d \ell}{n}} \right] \\
\leq e^{-n/16} + P \left[ \min_{\ell' = 1, \ldots, \ell} \left\| g_1^{\ell'} \right\|_0 \geq n/4, \ -1 + \prod_{\ell' = 1}^{\ell} \left\| g_1^{\ell'} \right\|_2^2 > C \sqrt{\frac{d \ell}{n}} \right]. \tag{A.133}
\]
Using (A.131) and the triangle inequality, we can write whenever no terms in the product vanish

\[
-1 + \prod_{\ell' = 1}^{\ell} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_2^2 = \left| \left( \prod_{\ell' = 1}^{\ell} \frac{2}{n} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_0 \right) \left( \prod_{\ell' = 1}^{\ell} \frac{1}{\left\| \sqrt{\frac{n}{2}} \left[ g_{1}^{\ell'} \right]_+ \right\|_0} \left\| \sqrt{\frac{n}{2}} \left[ g_{1}^{\ell'} \right]_+ \right\|_2^2 \right) - 1 \right| - 1 \leq \left( \prod_{\ell' = 1}^{\ell} \frac{2}{n} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_0 \right) \left( \prod_{\ell' = 1}^{\ell} \frac{1}{\left\| \sqrt{\frac{n}{2}} \left[ g_{1}^{\ell'} \right]_+ \right\|_0} \left\| \sqrt{\frac{n}{2}} \left[ g_{1}^{\ell'} \right]_+ \right\|_2^2 \right) - 1 \right| + \left( \prod_{\ell' = 1}^{\ell} \frac{2}{n} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_0 \right) - 1. \tag{A.134}
\]

Moreover, we have by Lemma A.4.17

\[
P\left( -1 + \prod_{\ell' = 1}^{\ell} \frac{2}{n} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_0 > 4 \sqrt{\frac{d\ell}{n}} \right) \leq 4\ell e^{-cd}
\]

as long as \( n \geq 128d\ell \). Choosing in addition \( n \geq 4d\ell \) and using nonnegativity, this implies

\[
P\left( \prod_{\ell' = 1}^{\ell} \frac{2}{n} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_0 > 2 \right) \leq 4\ell e^{-cd},
\]

occurring on the same event. Combining the previous two bounds with (A.133) and (A.134) via another partition, we get

\[
P\left( -1 + \prod_{\ell' = 1}^{\ell} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_2^2 > C \sqrt{\frac{d\ell}{n}} \right) \leq e^{-n/16} + 4\ell e^{-cd}
\]

\[
+ \left. \left. P\left( -1 + \prod_{\ell' = 1}^{\ell} \frac{1}{\left\| \sqrt{\frac{n}{2}} \left[ g_{1}^{\ell'} \right]_+ \right\|_0} \left\| \sqrt{\frac{n}{2}} \left[ g_{1}^{\ell'} \right]_+ \right\|_2^2 \right) > (C/2 + 2) \sqrt{\frac{d\ell}{n}} \right|_{\min_{\ell' = 1,\ldots,\ell} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_0 \geq n/4}, \tag{A.135}
\]

where we use here that on the event \( \{ \min_{\ell' = 1,\ldots,\ell} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_0 \geq n/4 \} \), the quantity \( \prod_{\ell' = 1}^{\ell} \left\| \left[ g_{1}^{\ell'} \right]_+ \right\|_2 \) is nonzero almost surely, which allowed us to invoke the identities (A.131). For \((k_1, \ldots, k_\ell) \in [n]^{\ell}\),
we define events $\mathcal{E}_1^{k_1}, \ldots, \mathcal{E}_\ell^{k_\ell}$ by

$$
\mathcal{E}_{k'} = \left\{ \left\| \sqrt{\frac{n}{2}} g^{(k')}_1 \right\|_0 = k' \right\}.
$$

Conditioning and then relaxing the bounds, we can write

$$
\Pr\left[ \min_{\ell' = 1, \ldots, \ell} \left\| g^{(\ell')}_1 \right\|_0 \geq n/4, \quad -1 + \prod_{\ell' = 1}^{\ell} \frac{1}{\sqrt{n/2} \left\| g^{(\ell')}_1 \right\|_0} \left\| \sqrt{\frac{n}{2}} g^{(\ell')}_1 \right\|_2 \geq C' \sqrt{\frac{d\ell}{n}} \right] \leq C'' \ell e^{-cd}
$$

as long as $n \geq K'' d\ell$, whence

$$
\Pr\left[ \min_{\ell' = 1, \ldots, \ell} \left\| g^{(\ell')}_1 \right\|_0 \geq n/4, \quad -1 + \prod_{\ell' = 1}^{\ell} \frac{1}{\sqrt{n/2} \left\| g^{(\ell')}_1 \right\|_0} \left\| \sqrt{\frac{n}{2}} g^{(\ell')}_1 \right\|_2 \geq C' \sqrt{\frac{d\ell}{n}} \right] \leq C'' \ell e^{-cd}.
$$

Combining this previous bound with (A.135) yields

$$
\Pr\left[ -1 + \prod_{\ell' = 1}^{\ell} \left\| g^{(\ell')}_1 \right\|_2 > C \sqrt{\frac{d\ell}{n}} \right] \leq e^{-n/16} + C' \ell e^{-cd},
$$

where we worst-cased constants in the probability bound. If we choose $n \geq 4C^2 d\ell$, we have
\[ C\sqrt{d\ell/n} \leq 1/2, \text{ and we obtain on the event in the previous bound} \]

\[ \mathbb{P}\left[ \left| -1 + \prod_{\ell'=1}^{\ell} \left\| g_{1}^{\ell'} \right\|_{2} \right| > \frac{1}{2} \right] \leq e^{-n/16} + C'\ell e^{-cd}. \]

In particular, on the complement of the event in the previous bound, the product lies in \([1/2, 3/2]\).

To conclude, we can linearize the square root near 1 to obtain an analogous bound for the product of the norms. Taylor expansion of the smooth function \(x \mapsto x^{1/2}\) about the point 1 gives

\[ \sqrt{x} - 1 = \frac{1}{2} (x - 1) - \frac{1}{8} k^{-3/2} (x - 1)^2, \]

where \(k\) lies between \(x\) and 1. In particular, if \(x \geq \frac{1}{2}\), we have

\[ \frac{1}{2} (x - 1) - \frac{1}{\sqrt{2}} (x - 1)^2 \leq \sqrt{x} - 1 \leq \frac{1}{2} (x - 1), \]

so that

\[ \left| (\sqrt{x} - 1) - \frac{1}{2} (x - 1) \right| \leq \frac{1}{\sqrt{2}} (x - 1)^2. \]

Thus, when \(x \geq \frac{1}{2}\) we have by the triangle inequality

\[ |\sqrt{x} - 1| \leq \frac{1}{\sqrt{2}} (x - 1)^2 + \frac{1}{2} |x - 1|. \]

from which we conclude based on a partition and our previous choices of large \(n\)

\[ \mathbb{P}\left[ \left| -1 + \prod_{\ell'=1}^{\ell} \left\| g_{1}^{\ell'} \right\|_{2} \right| > 2C\sqrt{\frac{d\ell}{n}} \right] \leq 2e^{-n/16} + 2C'\ell e^{-cd}, \]

which yields the claimed probability bound when \(n \geq 16d\). \(\square\)

**Lemma A.4.3.** There are absolute constants \(c, C, C_0 > 0\) and absolute constants \(K, K' > 0\) such
that for any \( L_{\text{max}} \in \mathbb{N} \) and any \( d \geq K \), if \( n \geq K' \max \{1, d^4 \log^4 n, d^3 L_{\text{max}} \log^3 n\} \), then one has

\[
\mathbb{P} \left[ \exists L \in [L_{\text{max}}] : \left| \hat{\nu}^L - \varphi^{(L)}(\nu^0) \right| > C_0 \sqrt{\frac{d^3 \log^3 n}{nL}} \right] \leq Cn^{-cd}. \tag{A.136}
\]

Proof. The proof uses a recursive construction involving \( L \in [L_{\text{max}}] \). Before beginning the main argument, we will define the key quantities that appear and enforce bounds on the parameters to obtain certain estimates. For each \( L \in [L_{\text{max}}] \), we define the event

\[
\mathcal{E}_L = \left\{ \left| \hat{\nu}^L - \varphi^{(L)}(\nu^0) \right| > C_0 \sqrt{\frac{d^3 \log^3 n}{nL}} \right\},
\]

where \( C_0 > 0 \) is an absolute constant whose value we will specify below, so that \( \mathcal{E}_L \in \mathcal{F}^L \), and our task is to produce an appropriate measure bound on \( \bigcup_{L \in [L_{\text{max}}]} \mathcal{E}_L \). For notational convenience, we also define \( \mathcal{E}_0 = \varnothing \). For each \( L \in [L_{\text{max}}] \) and each \( \ell \in [L] \), we define

\[
\Delta_L^\ell = \varphi^{(L-\ell)}(\hat{\nu}^{\ell}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}),
\]

so that for every \( L, \Delta_L^1, \ldots, \Delta_L^L \) is adapted to the sequence \( \mathcal{F}^1, \ldots, \mathcal{F}^L \), and we have the decomposition

\[
\hat{\nu}^L - \varphi^{(L)}(\nu^0) = \sum_{\ell=1}^L \Delta_L^\ell.
\]

In particular, we have

\[
\mathcal{E}_L = \left\{ \left| \sum_{\ell=1}^L \Delta_L^\ell \right| > C_0 \sqrt{\frac{d^3 \log^3 n}{nL}} \right\}.
\]

The sequences \( (\Delta_L^\ell)_{\ell \in L} \) are not quite martingale difference sequences, but we will show they are very nearly so: writing

\[
\Delta_L^\ell = \underbrace{\left( \Delta_L^\ell - \mathbb{E}[\Delta_L^\ell | \mathcal{F}^{\ell-1}] \right)}_{\widehat{\Delta}_L^\ell} + \mathbb{E}[\Delta_L^\ell | \mathcal{F}^{\ell-1}],
\]

213
we have that \((\bar{\Delta}_L^\ell)_{\ell \in L}\) is a martingale difference sequence, which can be controlled using truncation and martingale techniques, and the extra conditional expectation term can be controlled analytically. In particular, we have the following estimates: by Lemma A.4.14, we have if \(n \geq \max\{K_1 \log^4 n, K_2 L_{\max}\}\) that for every \(L \in [L_{\max}]\) and every \(\ell \in [L]\)

\[
|\mathbb{E}[\Delta_L^\ell \mid \mathcal{F}^{\ell-1}]| \leq C_1 \frac{\log n}{n} \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/64)(L - \ell)\hat{\nu}^{\ell-1}} (1 + \log L) + C_2 \frac{1}{n^2}; \tag{A.137}
\]

by the first result in Lemma A.4.15 we have for every \(d \geq \max\{K_3, 6/c_1\}\) that if \(n \geq K_4 d^{4} \log^{4} n\), then for every \(L \in [L_{\max}]\) and every \(\ell \in [L]\) (and after worsening constants)

\[
P\left[|\Delta_L^\ell| > 2C_3 \sqrt{\frac{d \log n}{n}} \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/64)(L - \ell)\hat{\nu}^{\ell-1}} + \frac{2C_2}{n^2} \bigg| \mathcal{F}^{\ell-1}\right] \leq C_5 n^{-c_1 d}; \tag{A.138}
\]

and by the second result in Lemma A.4.15, we have by our previous choices of \(n, d,\) and \(L_{\max}\) that for every \(L \in [L_{\max}]\) and every \(\ell \in [L]\) (after worsening constants)

\[
\mathbb{E}\left[\left(\Delta_L^\ell\right)^2 \bigg| \mathcal{F}^{\ell-1}\right] \leq 4C_3^2 \frac{d \log n}{n} \left(\frac{\hat{\nu}^{\ell-1}}{1 + (c_0/64)(L - \ell)\hat{\nu}^{\ell-1}}\right)^2 + \frac{C_4}{n^4}. \tag{A.139}
\]

The main line of the argument will consist of showing that a measure bound of the form (A.136) on \(\bigcup_{\ell \in [L-1]} \mathcal{E}_\ell\) implies one of the same form on \(\bigcup_{\ell \in [L]} \mathcal{E}_\ell\). For any \(L \in [L_{\max}]\), on the event \(\mathcal{E}_L^c\) we have

\[
\hat{\nu}^L \leq \varphi(L)(\hat{\nu}^0) + C_0 \sqrt{\frac{d^3 \log^3 n}{nL}} \leq \frac{2}{c_0 L} + C_0 \sqrt{\frac{d^3 \log^3 n}{nL}} \leq \frac{3}{c_0 L}, \tag{A.140}
\]

where the second inequality follows from Lemma A.3.9, and the third follows from the choice \(n \geq (C_0c_0)^2 d^3 L \log^3 n\). In particular, if we make the choice \(n \geq (C_0c_0)^2 d^3 L_{\max} \log^3 n\), we have (A.140) on \(\mathcal{E}_L^c\) for every \(L \in [L_{\max}]\). Accordingly, for every \(L \in [L_{\max}]\) and every \(\ell \in [L]\) we
define truncation events $G^\ell_L$ by

$$G^\ell_L = \left\{ |\Delta^\ell_L| \leq 2C_3 \sqrt{\frac{d \log n}{n}} \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/64)(L - \ell)\hat{\nu}^{\ell-1}} + \frac{2C_2}{n^2} \right\} \cap \mathcal{E}^{c}_{\ell-1}. \quad (A.141)$$

We have $G^\ell_L \in \mathcal{F}^\ell$, and a union bound and (A.138) imply

$$P\left[ \left( \bigcap_{\ell \in [L]} G^\ell_L \right)^c \bigg| \mathcal{F}^{L-1} \right] \leq C_5 Ln^{-c_1d} + P\left[ \bigcup_{\ell' \in [L-1]} \mathcal{E}_{\ell'} \bigg| \mathcal{F}^{L-1} \right]
= C_5 Ln^{-c_1d} + \mathbb{1}_{\bigcup_{\ell' \in [L-1]} \mathcal{E}_{\ell'}},$$

where the second line uses the fact that $\mathcal{E}_{\ell'} \in \mathcal{F}^{\ell'}$. In particular, taking expectations recovers

$$P\left[ \left( \bigcap_{\ell \in [L]} G^\ell_L \right)^c \right] \leq C_5 Ln^{-c_1d} + P\left[ \bigcup_{\ell' \in [L-1]} \mathcal{E}_{\ell'} \right]. \quad (A.142)$$

In addition, by (A.140) we have on $\mathcal{E}^{c}_{\ell-1}$

$$\frac{\hat{\nu}^{\ell-1}}{1 + (c_0/64)(L - \ell)\hat{\nu}^{\ell-1}} \leq \frac{3}{c_0 (\ell - 1) + (3/64)(L - \ell)} \leq \frac{3}{c_0 (3/64)L + (61/64)\ell - 1} \leq \frac{64}{c_0(L - 1)} \leq \frac{128}{c_0L},$$

where the final inequality requires $L \geq 2$. Thus, when $L \geq 2$, we have on $G^\ell_L$ that

$$|\Delta^\ell_L| \leq \frac{256C_3}{c_0L} \sqrt{\frac{d \log n}{n}} + \frac{2C_2}{n^2} \leq \frac{512C_3}{c_0(2\kappa_0)^{nL^2}}. \quad (A.143)$$
where the final inequality holds when \( d \geq 1 \) and \( n \geq (C_2 c_0/128 C_3)^{2/3} L^{2/3} \). Similarly, when \( L \geq 2 \), on \( \mathcal{E}_{\ell-1}^c \) we have by (A.139)

\[
\mathbb{E} \left[ (\Delta_L^\ell)^2 \mid \mathcal{F}^{\ell-1} \right] \leq \frac{2^{16} C_3^2}{c_0^2} \frac{d \log n}{n L^2} + \frac{C_4}{n^4} \leq \frac{2^{17} C_3^2}{c_0^2} \frac{d \log n}{n L^2} = 2 K_0^2 \frac{d \log n}{n L^2},
\]

where the second inequality holds when \( d \geq 1 \) and \( n \geq (C_4 c_0^2/2^{17} C_3^2)^{1/3} L^{2/3} \); and in the same setting we have by (A.137)

\[
|\mathbb{E}[\Delta_L^\ell \mid \mathcal{F}^{\ell-1}]| \leq \frac{128 C_1}{c_0} \frac{(1 + \log L) \log n}{n L} + \frac{C_2}{n^2} \leq \frac{256 C_1}{c_0} \frac{(1 + \log L) \log n}{n L},
\]

where the second inequality holds when \( n \geq (C_2 c_0/128 C_1) L \). In particular, if we enforce these conditions with \( L_{\text{max}} \) in place of \( L \), we have that (A.143), (A.144) and (A.145) hold for all \( 2 \leq L \leq L_{\text{max}} \) (with (A.144) and (A.145) holding on \( \mathcal{E}_{\ell-1}^c \)).

We begin the recursive construction. We will enforce \( C_0 = \max\{4\pi C_3, 6K_0\} \) for the absolute constant in the definition of \( \mathcal{E}_\ell \). The main tool is the elementary identity

\[
P \left[ \bigcup_{\ell \in [L]} \mathcal{E}_\ell \right] = P \left[ \bigcup_{\ell \in [L-1]} \mathcal{E}_\ell \right] + P \left[ \mathcal{E}_L \cap \bigcap_{\ell \in [L-1]} \mathcal{E}_\ell^c \right], \tag{A.146}
\]

which allows us to leverage an inductive argument provided we can control \( P[\mathcal{E}_L \cap \bigcap_{\ell \in [L-1]} \mathcal{E}_\ell^c] \), the probability that the \( L \)-th angle deviates above its nominal value subject to all prior angles being controlled. The case \( L = 1 \) can be addressed directly: (A.138) gives

\[
P \left[ |\Delta_1^\ell| > 2\pi C_3 \sqrt{\frac{d \log n}{n} + \frac{2C_2}{n^2}} \right] \leq C_5 n^{-c_1 d},
\]
and as long as \( d \geq 1 \) and \( n \geq (C_2/\pi C_3)^{2/3} \), this implies
\[
P \left[ \left| \Delta_1 \right| > 4\pi C_3 \sqrt{\frac{d \log n}{n}} \right] \leq C_5 n^{-c_1 d}. \tag{A.147}
\]
This gives a suitable measure bound on \( E_1 \), after choosing \( d \geq 1 \) and \( n \geq e \) so that \( d^3 \log^3 n \geq d \log n \). We now assume \( L \geq 2 \). By the triangle inequality, we have
\[
\left| \sum_{\ell=1}^L \Delta_L^\ell \right| \leq \left| \sum_{\ell=1}^L \bar{\Delta}_L^\ell \right| + \sum_{\ell=1}^L E[\Delta_L^\ell | \mathcal{F}^{\ell-1}], \tag{A.148}
\]
and we therefore have for any \( t > 0 \)
\[
P \left[ \left\{ \sum_{\ell=1}^L \Delta_L^\ell \right\} \cap \cap_{\ell \in [L-1]} \mathcal{E}_t^\ell \right] \\ \leq P \left[ \left\{ \sum_{\ell=1}^L \bar{\Delta}_L^\ell \right\} + \sum_{\ell=1}^L E[\Delta_L^\ell | \mathcal{F}^{\ell-1}] > t \right] \cap \cap_{\ell \in [L-1]} \mathcal{E}_t^\ell \tag{A.149}
\]
\[
= P \left[ \mathbb{1}_{\cap_{\ell \in [L-1]} \mathcal{E}_t^\ell} \sum_{\ell=1}^L \bar{\Delta}_L^\ell + \mathbb{1}_{\cap_{\ell \in [L-1]} \mathcal{E}_t^\ell} \sum_{\ell=1}^L E[\Delta_L^\ell | \mathcal{F}^{\ell-1}] > t \right]. \]
By (A.145), we have
\[
\mathbb{1}_{\cap_{\ell \in [L-1]} \mathcal{E}_t^\ell} \sum_{\ell=1}^L E[\Delta_L^\ell | \mathcal{F}^{\ell-1}] \leq 256 C_1 (1 + \log L) \log n \left/ c_0 \right. \tag{A.150}
\]
For the remaining term, we have by the triangle inequality
\[
\left| \sum_{\ell=1}^L \bar{\Delta}_L^\ell \right| \leq \left| \sum_{\ell=1}^L \Delta_L^\ell \mathbb{1} g_L^\ell \right| + \left| \sum_{\ell=1}^L \Delta_L^\ell \mathbb{1} g_L^\ell - E \left[ \Delta_L^\ell \mathbb{1} g_L^\ell | \mathcal{F}^{\ell-1} \right] \right| \\
+ \left| \sum_{\ell=1}^L E \left[ \Delta_L^\ell \mathbb{1} g_L^\ell | \mathcal{F}^{\ell-1} \right] - E \left[ \Delta_L^\ell | \mathcal{F}^{\ell-1} \right] \right|. \tag{A.151}
\]
By (A.141), an integration of (A.138), and a union bound, we have

\[
P \left[ \mathbb{1}_{\ell \in [L-1]} \mathbb{E}_{t}^{c} \left| \sum_{\ell=1}^{L} \Delta_{L,t}^{\ell} - \Delta_{L,t}^{\ell} \mathbb{1}_{G_{L,t}^{\ell}} \right| > 0 \right]
\leq P \left[ \bigcup_{\ell \in [L]} \left\{ |\Delta_{L,t}^{\ell}| > 2C_{3} \sqrt{\frac{d \log n}{n}} \frac{\varphi_{L,t}^{\ell-1}}{1 + (c_{0}/64)(L - \ell) \varphi_{L,t}^{\ell-1}} + \frac{2C_{2}}{n^{2}} \right\} \right]
\leq C_{5} L n^{-c_{1}d},
\]

and we have

\[
\left| \sum_{\ell=1}^{L} \mathbb{E} \left[ \Delta_{L,t}^{\ell} \mathbb{1}_{G_{L,t}^{\ell}} \big| \mathbb{F}^{t-1} \right] - \mathbb{E} \left[ \Delta_{L,t}^{\ell} \big| \mathbb{F}^{t-1} \right] \right|
\leq \sum_{\ell=1}^{L} \mathbb{E} \left[ |\Delta_{L,t}^{\ell}| \mathbb{1}_{(G_{L,t}^{\ell})^{c}} \big| \mathbb{F}^{t-1} \right]
\leq \pi \sum_{\ell=1}^{L} P \mathbb{E} \left[ (G_{L}^{\ell})^{c} \big| \mathbb{F}^{t-1} \right]
\leq \pi \sum_{\ell=1}^{L} \left\{ \left| \Delta_{L,t}^{\ell} \right| > 2C_{3} \sqrt{\frac{d \log n}{n}} \frac{\varphi_{L,t}^{\ell-1}}{1 + (c_{0}/64)(L - \ell) \varphi_{L,t}^{\ell-1}} + \frac{2C_{2}}{n^{2}} \right\} \cup \mathcal{E}_{t-1} \left| \mathbb{F}^{t-1} \right]
\leq \pi C_{5} L n^{-c_{1}d} + \pi \sum_{\ell=1}^{L-1} \mathbb{1}_{\mathcal{E}_{t-1}}
\]

where the first line uses linearity of the conditional expectation and the triangle inequality for sums and for the integral; the second line uses the worst-case bound of \( \pi \) on the magnitude of the increments \( \Delta_{L,t}^{\ell} \); the third line uses (A.141); and the fourth line uses a union bound, \( \mathcal{E}_{t-1} \in \mathbb{F}^{t-1} \), and (A.138). Multiplying both sides of the final bound by \( \mathbb{1}_{\ell \in [L-1]} \mathbb{E}_{t}^{c} \), we conclude

\[
\mathbb{1}_{\ell \in [L-1]} \mathbb{E}_{t}^{c} \left| \sum_{\ell=1}^{L} \mathbb{E} \left[ \Delta_{L,t}^{\ell} \mathbb{1}_{G_{L,t}^{\ell}} \big| \mathbb{F}^{t-1} \right] - \mathbb{E} \left[ \Delta_{L,t}^{\ell} \big| \mathbb{F}^{t-1} \right] \right|
\leq \mathbb{1}_{\ell \in [L-1]} \mathbb{E}_{t} \pi C_{5} L n^{-c_{1}d} \leq \pi C_{5} L n^{-c_{1}d}.
\]  

(A.153)
For the remaining term in (A.151), we first observe
\[
\mathbb{E} \left[ \left( \Delta^\ell_L 1_{G^\ell_L} - \mathbb{E} \left[ \Delta^\ell_L 1_{G^\ell_L} \mid \mathcal{F}^{\ell-1} \right] \right)^2 \mid \mathcal{F}^{\ell-1} \right] \leq \mathbb{E} \left[ \left( \Delta^\ell_L \right)^2 1_{G^\ell_L} \mid \mathcal{F}^{\ell-1} \right] \\
\leq \mathbb{E} \left[ \left( \Delta^\ell_L \right)^2 \mid \mathcal{F}^{\ell-1} \right],
\]
where the first line uses the centering property of the \(L^2\) norm, and the second line uses \((\Delta^\ell_L)^2 \geq 0\) to drop the indicator for \(G^\ell_L\). For notational simplicity, we define
\[
V^L = \sum_{\ell=1}^L \mathbb{E} \left[ \left( \Delta^\ell_L 1_{G^\ell_L} - \mathbb{E} \left[ \Delta^\ell_L 1_{G^\ell_L} \mid \mathcal{F}^{\ell-1} \right] \right)^2 \mid \mathcal{F}^{\ell-1} \right],
\]
so that our previous bound and (A.144) imply
\[
\bigcap_{\ell \in [L-1]} \mathcal{E}^c_\ell \subset \left\{ V^L \leq 2K_0^2 \frac{d \log n}{nL} \right\}.
\]
This implies that for any \(t > 0\)
\[
P \left[ \left\{ \bigcap_{\ell \in [L-1]} \mathcal{E}^c_\ell \right\} \cap \left\{ \sum_{\ell=1}^L \Delta^\ell_L 1_{G^\ell_L} - \mathbb{E} \left[ \Delta^\ell_L 1_{G^\ell_L} \mid \mathcal{F}^{\ell-1} \right] > t \right\} \right]
= P \left[ \left\{ \bigcap_{\ell \in [L-1]} \mathcal{E}^c_\ell \right\} \right. \\
\left. \cap \left\{ \left\{ \sum_{\ell=1}^L \Delta^\ell_L 1_{G^\ell_L} - \mathbb{E} \left[ \Delta^\ell_L 1_{G^\ell_L} \mid \mathcal{F}^{\ell-1} \right] > t \right\} \right\} \right]
\leq P \left[ V^L \leq 2K_0^2 \frac{d \log n}{nL} \right. \\
\left. \cap \left\{ \sum_{\ell=1}^L \Delta^\ell_L 1_{G^\ell_L} - \mathbb{E} \left[ \Delta^\ell_L 1_{G^\ell_L} \mid \mathcal{F}^{\ell-1} \right] > t \right\} \right].
\]
The previous term can be controlled using Lemma A.6.5 and (A.143):
\[
P \left[ \left\{ V^L \leq 2K_0^2 \frac{d \log n}{nL} \right\} \cap \left\{ \sum_{\ell=1}^L \Delta^\ell_L 1_{G^\ell_L} - \mathbb{E} \left[ \Delta^\ell_L 1_{G^\ell_L} \mid \mathcal{F}^{\ell-1} \right] > t \right\} \right]
\leq 2 \exp \left( - \frac{t^2/2}{2K_0^2 \frac{d \log n}{nL} + (2K_0/3) t \sqrt{\frac{d \log n}{nL^2}}} \right).
\]
Setting $t = 3K_0 \sqrt{d^3 \log^3 n/nL}$, we obtain

$$
P\left[ \bigcap_{\ell \in [L-1]} E_\ell^c \left| \sum_{\ell=1}^L \Delta_L^{\ell} \mathbf{1}_{g_L^\ell} - \mathbb{E} \left[ \Delta_L^{\ell} \mathbf{1}_{g_L^\ell} \right] \right| > 3K_0 \sqrt{\frac{d^3 \log^3 n}{nL}} \right] 
\leq 2 \exp \left( -\frac{9}{4} \frac{d^2 \log^2 n}{1 + \frac{d \log n}{\sqrt{L}}} \right) \leq 2n^{-(9/8)d},$$

(A.154)

where the last line uses the bounds $L \geq 1$ and $d \log n / (1 + d \log n) \geq \frac{1}{2}$ if $d \geq 1$ and $n \geq e$.

Combining (A.152), (A.153) and (A.154) in (A.151) via a union bound, we obtain

$$
P \left[ \bigcap_{\ell \in [L-1]} E_\ell^c \left| \sum_{\ell=1}^L \Delta_L^{\ell} \mathbf{1}_{g_L^\ell} - \mathbb{E} \left[ \Delta_L^{\ell} \mathbf{1}_{g_L^\ell} \right] \right| > 3K_0 \sqrt{\frac{d^3 \log^3 n}{nL}} + \pi C_5 L n^{-c_1 d} \right] \leq C_5 L n^{-c_1 d} + 2n^{-(9/8)d}.$$

Applying this result and (A.150) to (A.149) via a union bound, we obtain

$$
P \left[ \left\{ \left| \sum_{\ell=1}^L \Delta_L^{\ell} \right| > 3K_0 \sqrt{\frac{d^3 \log^3 n}{nL}} + \pi C_5 L n^{-c_1 d} + \frac{256C_1}{c_0} (1 + \log L) \log n \right\} \cap \bigcap_{\ell \in [L-1]} E_\ell \right] \leq C_5 L n^{-c_1 d} + 2n^{-(9/8)d}.$$

If $d \geq 2/c_1$ and $n \geq L_{\max}$, we have $C_5 L n^{-c_1 d} \leq C_5 n^{-c_1 d/2}$; under these condition on $d$ and $n$, we have $\pi C_5 L n^{-c_1 d} \leq \pi C_5 n^{-1}$, and so $\pi C_5 n^{-c_1 d/2} + (256C_1/c_0)(1 + \log L)(\log n)/n \leq C(1 + \log L)(\log n)/n$; and if $d \geq 1$ and $n \geq L_{\max}$, we have

$$
3K_0 \sqrt{\frac{d^3 \log^3 n}{nL}} \geq C \frac{(1 + \log L) \log n}{n}
$$

provided $n \geq C'(C/3K_0)^2 L_{\max} \log L_{\max}$. Under these conditions, our previous bound simplifies to

$$
P \left[ \left\{ \left| \sum_{\ell=1}^L \Delta_L^{\ell} \right| > 6K_0 \sqrt{\frac{d^3 \log^3 n}{nL}} \right\} \cap \bigcap_{\ell \in [L-1]} E_\ell \right] \leq (2 + C_5) n^{-\min(c_1/2,9/8)d}.$$
In particular, applying this bound to (A.146), we have shown that for any $L \geq 2$

$$
\mathbb{P}\left[\bigcup_{\ell \in [L]} \mathcal{E}_\ell\right] = \mathbb{P}\left[\bigcup_{\ell \in [L-1]} \mathcal{E}_\ell\right] + (2 + C_5)n^{-\min\{c_1/2,9/8\}d}.
$$

Unraveling the recursion with (A.147) (and worst-casing the constants there), we conclude

$$
\mathbb{P}\left[\bigcup_{\ell \in [L]} \mathcal{E}_\ell\right] \leq (2 + C_5)Ln^{-\min\{c_1/2,9/8\}d},
$$

which proves the claim, after possibly choosing $n$ to be larger than another absolute constant multiple of $L_{\max}$ to remove the leading $L$ factor. \qed

**Backward Feature Control**

Having established concentration of the feature norms and the angles between them, it remains to control the inner products of backward features that appear in $\Theta$. The core of the technical approach will once again be martingale concentration. We establish the following control on the backward feature inner products:

**Lemma A.4.4.** Fix $x, x' \in S^{n_0-1}$ and denote $\nu = \angle(x, x')$. If $n \geq \max\{KL \log n, K'Ld_b, K''\}$, $d_b \geq K''' \log L$ for suitably chosen $K, K', K''$, $K'''$ then

$$
\mathbb{P}\left[\bigcap_{\ell=0}^{L-1} \left\{\left\|\beta^\ell(x)\right\|_2^2 \leq Cn\right\}\right] \geq 1 - e^{-c\frac{n}{L}}.
$$

If additionally $n, L, d$ satisfy the requirements of lemmas A.4.3 and A.5.16, we have

$$
\mathbb{P}\left[\bigcap_{\ell=0}^{L-1} \left\{\left|\langle\beta^\ell(x), \beta^\ell(x')\rangle - \frac{n}{2} \prod_{i=\ell}^{L-1} \left(1 - \varphi^{(i)}(\nu)\right)\right| \leq \log^2(n)\sqrt{d^4Ln}\right\}\right] \geq 1 - e^{-cd}
$$

where $\varphi^{(i)}$ denotes $i$ applications of the angle evolution function defined in lemma A.5.2, and $c > 0, C$ are absolute constants.
A.4.3 Uniformization Estimates

Nets and Covering Numbers

We appeal to Lemma A.3.1 to obtain estimates for the covering number of $M$, which we will use throughout this section. In the remainder of this section, we will use the notation $N_\varepsilon$ to denote the $\varepsilon$-nets for $M$ constructed in Lemma A.3.1, and for any $\tilde{x} \in N_\varepsilon$, we will also use the notation $N_\varepsilon(\tilde{x}) = \mathbb{B}(\tilde{x}, \varepsilon) \cap M_\square$, where $\square \in \{+, -\}$ is the component of $\tilde{x}$, to denote the relevant connected neighborhood of the specific point in the net we are considering. Here we are implicitly assuming that $M_\pm$ are themselves connected, but this construction evidently generalizes to cases where $M_\pm$ themselves have a positive number of connected components, as treated in Lemma A.3.1. Focusing on this simpler case in the sequel will allow us to keep our notation concise.

Controlling Support Changes Uniformly

The quantities we have studied in Section A.4.2 are challenging to uniformize due to discontinuities in the support projections $P_\ell(\cdot)$. We will get around this difficulty by carefully tracking (with high probability) how much the supports can change by when we move away from the points in our net $N_\varepsilon$. It seems intuitively obvious that when $\varepsilon$ is exponentially small in all problem parameters, there should be almost no support changes when moving away from our net; the challenge is to show that this property also holds when $\varepsilon$ is not so relatively small.

Introduce the following notation for the network preactivations at level $\ell$, where $\ell \in [L]$:

$$\rho^\ell(x) = W^\ell \alpha^{\ell-1}(x),$$

so that $\alpha^\ell(x) = [\rho^\ell(x)]_+$. We also let $\mathcal{F}^\ell$ denote the $\sigma$-algebra generated by all weight matrices up to level $\ell$ in the network, and let $\mathcal{F}^0$ denote the trivial $\sigma$-algebra.
Definition A.4.1. Let $\varepsilon, \Delta > 0$, and let $\bar{x} \in \mathcal{N}_\varepsilon$. For $\ell \in [L]$, a feature $(\alpha^\ell(\bar{x}))_i$ is called $\Delta$-risky if $|(\rho^\ell(\bar{x}))_i| \leq \Delta$; otherwise, it is called $\Delta$-stable. If for all $x \in \mathcal{N}_\varepsilon(\bar{x})$ we have

$$\forall \ell' \in [\ell], \|\rho^\ell(x) - \rho^\ell(\bar{x})\|_\infty \leq \Delta,$$

we say that stable sign consistency holds up to layer $\ell$. We will abbreviate this condition as $\text{SSC}(\ell, \varepsilon, \Delta)$ at $\bar{x}$, with the dependence on $\bar{x}, \varepsilon,$ and $\Delta$ suppressed when it is clear from context.

If $\text{SSC}(\ell)$ holds at $\bar{x}$ and if $(\alpha^{\ell'}(\bar{x}))_i$ is stable, we can write for any $x \in \mathcal{N}_\varepsilon(\bar{x})$

$$\text{sign}((\rho^{\ell'}(x))_i) = \text{sign}((\rho^{\ell'}(\bar{x}))_i + (\rho^{\ell'}(x))_i - (\rho^{\ell'}(\bar{x}))_i) = \text{sign}((\rho^{\ell'}(\bar{x}))_i),$$

so that no stable feature supports change on $\mathcal{N}_\varepsilon(\bar{x})$, and we only need to consider changes due to the risky features. Moreover, observe that

$$\mathbb{P}[(\rho^\ell(\bar{x}))_i \in \{\pm \Delta\}] = \mathbb{E}[\mathbb{P}[\|\alpha^{\ell-1}(x)\|_2(e_i, g) \in \{\pm \Delta\} \mid \mathcal{F}^{\ell-1}]] = 0,$$

(A.155)

where $g \sim N(0, (2/n)I)$ is independent of everything else in the problem, since $\Delta > 0$. It follows that when considering the network features over any countable collection of points $\bar{x} \in \mathcal{M}$, we have almost surely that the risky features are witnessed in the interior of $[-\Delta, +\Delta]$.

Below, we will show that with appropriate choices of $\varepsilon$ and $\Delta$, with very high probability: (i) each point in the net $\bar{x}$ has very few risky features; and (ii) $\text{SSC}(L)$ holds uniformly over the net under reasonable conditions involving $n, L, d$. We write $R_\ell(\bar{x}, \Delta) \subset [n]$ for the random variable consisting of the set of indices of $\Delta$-risky features at level $\ell$ with input $\bar{x} \in \mathcal{N}_\varepsilon$.

Lemma A.4.5. There is an absolute constant $K > 0$ such that for any $\bar{x} \in \mathcal{M}$ and any $d > 0$, if $n \geq \max\{KdL, 4\}$ and $\Delta \leq d \log n/(6n^{3/2}L)$, then one has

$$\mathbb{P}\left[\sum_{\ell=1}^{L} |R_\ell(\bar{x}, \Delta)| > d \log n \right] \leq 2n^{-d} + L^2 e^{-cn/L}.$$
Proof. For any $\bar{x} \in N_\varepsilon$, Lemma A.4.2 (with a suitable choice of $d$ in that context) gives
\[
\mathbb{P}\left[\|\alpha^\ell(\bar{x})\|_2 - 1 > \frac{1}{2}\right] \leq C\ell e^{-c\frac{\varepsilon}{\ell}},
\]
so that if additionally $n \geq (2/c)\ell \log(C)$, one has
\[
\mathbb{P}\left[\|\alpha^\ell(\bar{x})\|_2 - 1 > \frac{1}{2}\right] \leq \ell e^{-c\frac{\varepsilon}{\ell}}. \tag{A.156}
\]
Let $\mathcal{G}_\ell = \{1/2 \leq \|\alpha^\ell(\bar{x})\|_2 \leq 2\}$, so that $\mathcal{G}_\ell$ is $\mathcal{F}_\ell$-measurable, and $\mathcal{G} = \cap_{\ell \in [L-1]} \mathcal{G}_\ell$; then by (A.156) and a union bound, we have $\mathbb{P} [\mathcal{G}] \geq 1 - L^2 e^{-cn/L}$. We also let $\mathcal{G}_0 = \emptyset$. For $i \in [n]$ and $\ell \in [L]$, consider the random variables $X_{i\ell} = |(\rho^\ell(\bar{x}))_i|$, and moreover define
\[
\tilde{X}_{i\ell} = \frac{X_{i\ell}}{\|\alpha^{\ell-1}(\bar{x})\|_2^{-1}}1_{\mathcal{G}_{\ell-1}}.
\]
We have $\sum_{i,\ell} 1_{X_{i\ell} \leq \Delta} = \sum_{\ell} |R_\ell(\bar{x})|$, which is the total number of $\Delta$-risky features at $\bar{x}$, and the corresponding sum with the random variables $\tilde{X}_{i\ell}$ is thus an upper bound on the number of risky features at $\bar{x}$. Notice that $X_{i\ell}$ and $\tilde{X}_{i\ell}$ are $\mathcal{F}_\ell$-measurable, and additionally notice that on $\mathcal{G}$, we have $X_{i\ell}/2 \leq \tilde{X}_{i\ell} \leq 2X_{i\ell}$. For any $K \in \{0, 1, \ldots, nL - 1, nL\}$, we have by disjointness of the events in the union and a partition
\[
\mathbb{P}\left[\sum_{i,\ell} 1_{X_{i\ell} \leq \Delta} > K\right] \leq L^2 e^{-cn/L} + \sum_{k=K+1}^{nL} \mathbb{P}\left[\mathcal{G} \cap \left\{\sum_{i,\ell} 1_{X_{i\ell} \leq \Delta} = k\right\}\right]\]
\[
\leq L^2 e^{-cn/L} + \sum_{k=K+1}^{nL} \mathbb{P}\left[\mathcal{G} \cap \left\{\sum_{i,\ell} 1_{\tilde{X}_{i\ell} \leq 2\Delta} = k\right\}\right],
\]
so it is essentially equivalent to consider the $\tilde{X}_{i\ell}$. By another partitioning we can write
\[
\mathbb{P}\left[\mathcal{G} \cap \left\{\sum_{i,\ell} 1_{\tilde{X}_{i\ell} \leq 2\Delta} = k\right\}\right] = \sum_{S \in \{0, 1\}^{n \times L}} \mathbb{E}\left[\prod_{\ell=1}^{L} \left\{1_{\mathcal{G}_{\ell-1}} \prod_{i=1}^{n} 1_{1_{\tilde{X}_{i\ell} \leq 2\Delta} = S_i}\right\}\right],
\]
where $\{0, 1\}^{n \times L}$ is the set of $n \times L$ matrices with entries in $\{0, 1\}$. Using the tower rule and $\mathcal{F}^{L-1}$-
measurability of all factors with $\ell < L$, we can then write

$$E \left[ \prod_{\ell=1}^{L} 1_{G_{\ell-1}} \prod_{i=1}^{n} 1_{X_{\ell i} \leq 2\Delta = S_{i\ell}} \right]$$

$$= E \left[ \left( \prod_{\ell=1}^{L} 1_{G_{\ell-1}} \prod_{i=1}^{n} 1_{X_{\ell i} \leq 2\Delta = S_{i\ell}} \right) \mid \mathcal{F}^{L-1} \right]$$

$$= E \left[ \left( \prod_{\ell=1}^{L-1} 1_{G_{\ell-1}} \prod_{i=1}^{n} 1_{X_{\ell i} \leq 2\Delta = S_{i\ell}} \right) 1_{G_{L-1}} E \left[ \prod_{i=1}^{n} 1_{X_{i L} \leq 2\Delta = S_{i L}} \mid \mathcal{F}^{L-1} \right] \right].$$

We study the inner conditional expectation as follows: because $\rho^{L}(\bar{x}) = W^{L} \alpha^{L-1}(\bar{x})$, we can apply rotational invariance in the conditional expectation to obtain

$$1_{G_{L-1}} E \left[ \prod_{i=1}^{n} 1_{X_{i L} \leq 2\Delta = S_{i L}} \mid \mathcal{F}^{L-1} \right] = 1_{G_{L-1}} E \left[ \prod_{i=1}^{n} 1_{(w_{L})_{i} \leq 2\Delta = S_{i L}} \mid \mathcal{F}^{L-1} \right]$$

$$= 1_{G_{L-1}} E \left[ \prod_{i=1}^{n} 1_{(w_{L})_{i} \leq 2\Delta = S_{i L}} \mid \mathcal{F}^{L-1} \right],$$

where $w_{L} \sim N(0, (2/n)I)$ is the first column of $W^{L}$, and the last equality takes advantage of the presence of the indicator for $1_{G_{L-1}}$ multiplying the conditional expectation. We then write using independence

$$E \left[ \prod_{i=1}^{n} 1_{(w_{L})_{i} \leq 2\Delta = S_{i L}} \mid \mathcal{F}^{L-1} \right] = P \left[ \bigcap_{i=1}^{n} \{ 1_{(w_{L})_{i} \leq 2\Delta = S_{i L}} \} \mid \mathcal{F}^{L-1} \right]$$

$$= \prod_{i=1}^{n} P \left[ 1_{(w_{L})_{i} \leq 2\Delta = S_{i L}} \mid \mathcal{F}^{L-1} \right],$$

and putting $p_{L} = P[| (w_{L})_{1} | \leq 2\Delta]$, we have by identically-distributedness

$$\prod_{i=1}^{n} P \left[ 1_{(w_{L})_{i} \leq 2\Delta = S_{i L}} \mid \mathcal{F}^{L-1} \right] = \prod_{i=1}^{n} p_{S_{i L}}^{S_{i L}} (1 - p_{L})^{1-S_{i L}}.$$

After removing the indicator for $G_{L-1}$ by nonnegativity of all factors in the expectation, this leaves
us with

\[
\mathbb{E}\left[\prod_{\ell=1}^{L} \mathbb{1}_{G_{\ell-1}} \prod_{i=1}^{n} \mathbb{1}_{\tilde{x}_{i\ell} \leq 2\Delta = S_{i\ell}}\right] \leq \left(\prod_{i=1}^{n} p_{L}^{S_{iL}} (1 - p_{L})^{1 - S_{iL}}\right) \mathbb{E}\left[\prod_{\ell=1}^{L-1} \mathbb{1}_{G_{\ell-1}} \prod_{i=1}^{n} \mathbb{1}_{\tilde{x}_{i\ell} \leq 2\Delta = S_{i\ell}}\right].
\]

This process can evidently be iterated \( L - 1 \) additional times with analogous definitions—we observe that the fact that all weight matrices \( \mathbf{W}_{\ell} \) have the same column distribution implies that \( p_{1} = \cdots = p_{L} \), so we write \( p = p_{1} \) henceforth—and by this we obtain

\[
\mathbb{E}\left[\prod_{\ell=1}^{L} \mathbb{1}_{G_{\ell-1}} \prod_{i=1}^{n} \mathbb{1}_{\tilde{x}_{i\ell} \leq 2\Delta = S_{i\ell}}\right] \leq \left(\prod_{i=1}^{n} p_{L}^{S_{iL}} (1 - p_{L})^{1 - S_{iL}}\right),
\]

and in particular

\[
\mathbb{P}\left[\mathcal{G} \cap \left\{ \sum_{i,\ell} \mathbb{1}_{\tilde{x}_{i\ell} \leq 2\Delta} = k \right\} \right] \leq \sum_{S \in \{0,1\}^{n \times L} : \|S\|_{F} = k} \prod_{\ell=1}^{L} \prod_{i=1}^{n} p_{L}^{S_{i\ell}} (1 - p_{L})^{1 - S_{i\ell}}.
\]

For \( i \in [n] \) and \( \ell \in [L] \), let \( Y_{i\ell} \) denote \( nL \) i.i.d. Bern\( (p) \) random variables; we recognize this last sum as the probability that \( \sum_{i,\ell} Y_{i\ell} = k \). In particular, using our previous work we can assert for any \( t > 0 \)

\[
\mathbb{P}\left[\sum_{i,\ell} \mathbb{1}_{X_{i\ell} \leq \Delta} > t \right] \leq \mathbb{P}\left[\sum_{i,\ell} Y_{i\ell} > t \right] + L^{2} e^{-cn/L},
\]

so to conclude it suffices to articulate some binomial tail probabilities and estimate \( p \). We have

\[
p = \mathbb{P}\left[\| (\mathbf{w}_{L})_{i} \| \leq 2\Delta \right] = \sqrt{n/2\pi} \int_{0}^{2\Delta} \exp\left(-\frac{nt^{2}}{4}\right) \, dt \leq \Delta \sqrt{\frac{2n}{\pi}}, \tag{A.157}
\]

and we can write with the triangle inequality and a union bound

\[
\mathbb{P}\left[\sum_{i,\ell} Y_{i\ell} > t \right] \leq \mathbb{P}\left[\left|\sum_{i,\ell} Y_{i\ell} - \mathbb{E}[Y_{i\ell}] \right| > t \right] + \mathbb{P}\left[\sum_{i,\ell} \mathbb{E}[Y_{i\ell}] > t \right].
\]
By (A.157), we have $\sum_{i,\ell} \mathbb{E}[Y_{i\ell}] \leq n^{3/2} L \Delta$. We calculate using independence
\[
\mathbb{E} \left[ \left( \sum_{i,\ell} Y_{i\ell} - \mathbb{E}[Y_{i\ell}] \right)^2 \right] \leq \sum_{i,\ell} \mathbb{E}[Y_{i\ell}] \leq n^{3/2} L \Delta,
\]
so an application of Lemma A.6.3 yields
\[
\mathbb{P} \left[ \left| \sum_{i,\ell} Y_{i\ell} - \mathbb{E}[Y_{i\ell}] \right| > t \right] \leq 2 \exp \left( -\frac{t^2/2}{n^{3/2} L \Delta + t/3} \right).
\]
For any $d > 0$, if we choose $t = d \log n$ and enforce $\Delta \leq d \log n / (6n^{3/2} L)$, we obtain
\[
\mathbb{P} \left[ \sum_{i,\ell} Y_{i\ell} > d \log n \right] \leq 2^{-d},
\]
from which we conclude as sought
\[
\mathbb{P} \left[ \sum_{i,\ell} 1_{X_{i\ell} \leq \Delta} > d \log n \right] \leq 2^{-d} + L^2 e^{-cn/L}.
\]

The next task is to study the stable sign condition at a point $\bar{x}$ as a function of $\varepsilon$ and $\Delta$, assuming $\Delta$ at least satisfies the hypotheses of Lemma A.4.5. In particular, we will be interested in conditions under which we can guarantee that $\text{SSC}(\ell - 1)$ holding implies that $\text{SSC}(\ell)$ holds. Let $S_{\ell}(\bar{x}, \Delta) = [n] \setminus R_{\ell}(\bar{x}, \Delta)$ denote the $\Delta$-stable features at level $\ell$ with input $\bar{x}$, and define for $0 \leq \ell' \leq \ell \leq L$

\[
T_{x}^{\ell,\ell'} = P_{S_{\ell}(\bar{x})} P_{l_{\ell}(x)} W^\ell P_{S_{\ell-1}(\bar{x})} P_{l_{\ell-1}(x)} W^{\ell-1} \cdots P_{S_{\ell'+1}(\bar{x})} P_{l_{\ell'+1}(x)} W^{\ell'+1};
\]
\[
\Phi_{x}^{\ell,\ell'} = W^{\ell} T_{x}^{\ell-1,\ell'},
\]

so that $\Phi_{x}^{\ell,\ell'} x$ carries an input $x \in N_{\varepsilon}(\bar{x})$ applied at the features at level $\ell'$ (in particular, $\ell' = 0$ corresponds to $a^{0}(x) = x$, the network input) to the preactivations at level $\ell$ in a network restricted
to only the stable features at $\bar{x}$. We can write

$$
\rho^\ell(x) = W^\ell P_{l_{\ell-1}(x)} W^{\ell-1} \cdots P_{l_1(x)} W^1 x;
$$

$$
\alpha^\ell(x) = P_{l_\ell(x)} W^\ell P_{l_{\ell-1}(x)} W^{\ell-1} \cdots P_{l_1(x)} W^1 x,
$$

which gives us a useful representation if we disregard all levels with no risky features: let $r = \sum_{\ell=1}^{L} 1_{|R_\ell(\bar{x},\Delta)| > 0}$ be the number of levels in the network with risky features, and let $\ell_1 < \ell_2 < \cdots < \ell_r$ denote the levels at which risky features occur. If no risky features occur at a level $\ell$, we of course have $P_{S_\ell(\bar{x})} = I$. Assume to begin that $\ell > \ell_r$, and start by writing

$$
\rho^\ell(x) = \Phi^\ell \left( P_{S_\ell(\bar{x})} + P_{R_\ell(\bar{x})} \right) P_{l_\ell(x)} \Phi^\ell_{\ell,\ell-1} \left( P_{S_{\ell-1}(\bar{x})} + P_{R_{\ell-1}(\bar{x})} \right) P_{l_{\ell-1}(x)} \cdots
$$

$$
\cdots \Phi^\ell_{\ell,1} \left( P_{S_{l_1}(\bar{x})} + P_{R_{l_1}(\bar{x})} \right) P_{l_{l_1}(x)} \Phi^\ell_{1,0}.
$$

Now we distribute from left to right, and recombine everything to right on the term corresponding to the projection onto the risky features at $\ell_r$; this gives

$$
\rho^\ell(x) = \Phi^\ell \left( P_{R_\ell(\bar{x})} \right) \alpha^\ell(x) + \Phi^\ell_{\ell,\ell-1} \left( P_{S_{\ell-1}(\bar{x})} + P_{R_{\ell-1}(\bar{x})} \right) P_{l_{\ell-1}(x)} \cdots
$$

$$
\cdots \Phi^\ell_{\ell,1} \left( P_{S_{l_1}(\bar{x})} + P_{R_{l_1}(\bar{x})} \right) P_{l_{l_1}(x)} \Phi^\ell_{1,0}.
$$

We can write

$$
\Phi^\ell_{x} P_{R_\ell(\bar{x})} \alpha^\ell(x) = \Phi^\ell_{x} \big|_{R_\ell(\bar{x})} \alpha^\ell(x) \big|_{R_\ell(\bar{x})},
$$

where the restriction notation emphasizes that we are considering a column submatrix of the transfer operator induced by the risky features. Iterating the previous argument, we obtain

$$
\rho^\ell(x) = \Phi^\ell_{x} x + \sum_{i=1}^{r} \Phi^\ell_{x} \big|_{R_{l_i}(\bar{x})} \alpha^\ell_{l_i}(x) \big|_{R_{l_i}(\bar{x})}.
$$

It is clear that an analogous argument can be used in the case where $\ell \leq \ell_r$ by adapting which risky
features can be visited: we can thus assert

\[ \rho^\ell(x) = \Phi_x^\ell,0 x + \sum_{i \in \mathcal{r} : \ell_i < \ell} \Phi_x^\ell,\ell_i \bigg|_{R_{\ell_i}(\bar{x})} \alpha^\ell_i(x) \bigg|_{R_{\ell_i}(\bar{x})}. \]  

(A.159)

Furthermore, we note that under \( \text{SSC}(\ell - 1) \), no stable feature supports change on \( \mathcal{N}_\varepsilon(\bar{x}) \), and so one has for every \( x \in \mathcal{N}_\varepsilon(\bar{x}) \)

\[ \Phi^{\ell,\ell'} = \Phi_x^{\ell,\ell'}, \]

so under \( \text{SSC}(\ell - 1) \) we have by (A.159)

\[ \rho^\ell(x) - \rho^\ell(\bar{x}) = \Phi_x^{\ell,0} (x - \bar{x}) + \sum_{i \in \mathcal{r} : \ell_i < \ell} \Phi_x^{\ell,\ell_i} \bigg|_{R_{\ell_i}(\bar{x})} \left( \alpha^\ell_i(x) \bigg|_{R_{\ell_i}(\bar{x})} - \alpha^\ell_i(\bar{x}) \bigg|_{R_{\ell_i}(\bar{x})} \right). \]

The ReLU \( [\cdot]_+ \) is 1-Lipschitz with respect to \( \| \cdot \|_\infty \), and by monotonicity of the max under restriction and \( \text{SSC}(\ell - 1) \) we have

\[ \left\| \alpha^\ell_i(x) \bigg|_{R_{\ell_i}(\bar{x})} - \alpha^\ell_i(\bar{x}) \bigg|_{R_{\ell_i}(\bar{x})} \right\|_\infty \leq \left\| \rho^\ell_i(x) \bigg|_{R_{\ell_i}(\bar{x})} - \rho^\ell_i(\bar{x}) \bigg|_{R_{\ell_i}(\bar{x})} \right\|_\infty \]

\[ \leq \left\| \rho^\ell_i(x) - \rho^\ell_i(\bar{x}) \right\|_\infty \leq \Delta. \]

Thus, by the triangle inequality, we have under \( \text{SSC}(\ell - 1) \) a bound

\[ \left\| \rho^\ell(x) - \rho^\ell(\bar{x}) \right\|_\infty \leq \varepsilon \left\| \Phi_x^{\ell,0} \right\|_{\ell_2 \rightarrow \ell_\infty} + \Delta \sum_{i \in \mathcal{r} : \ell_i < \ell} \left\| \Phi_x^{\ell,\ell_i} \right\|_{\ell_\infty \rightarrow \ell_\infty}. \]  

(A.160)

This suggests an inductive approach to establishing \( \text{SSC}(\ell) \) provided we have established it at previous layers—we just need to control the transfer coefficients in (A.160).

**Lemma A.4.6.** There are absolute constants \( c, c', C, C', C'', C''' > 0 \) and absolute constants \( K, K' > 0 \) such that for any \( 1 \leq \ell' < \ell \leq L \), any \( d \geq K \log n \) and any \( \bar{x} \in S^{n_0-1} \), if \( \Delta \leq cn^{-5/2} \) and
\( n \geq K' \max\{d^4L, 1\} \), then one has

\[
P\left[ \left\| \Phi_{\bar{x}}^{\ell,0} \right\|_{\ell^2 \rightarrow \ell^\infty} \leq C \left( 1 + \sqrt{\frac{n_0}{n}} \right) \right] \geq 1 - C'' e^{-cd},
\]

and for any fixed \( S \subset [n] \), one has

\[
P\left[ \left\| \Phi_{\bar{x}}^{\ell,\ell'} S \right\|_{\ell^\infty \rightarrow \ell^\infty} \leq C' \sqrt{\frac{|S|}{n}} \geq 1 - C''' e^{-c'd} \right].
\]

**Proof.** We will use [221, Lemmas D.14, D.23] to bound the transfer coefficients, so let us first verify the hypotheses of these lemmas. In our setting, the transfer matrices differ only from the ‘nominal’ transfer matrices by restriction to the stable features at \( \bar{x} \); we have \( S_\ell(\bar{x}) \cap I_\ell(\bar{x}) = [n] \setminus R_\ell(\bar{x}) \), which is an admissible support random variable for [221, Lemmas D.14, D.23], and in particular

\[
\left\| (P_{S_\ell(\bar{x})} P_{I_\ell(\bar{x})} - P_{I_\ell(\bar{x})}) \rho^{\ell}(\bar{x}) \right\|_2 = \| P_{R_\ell(\bar{x})} a^{\ell}(\bar{x}) \|_2 \leq \sqrt{n} \Delta
\]

by Lemma A.6.10 and the definition of \( R_\ell(\bar{x}) \). Additionally, using Lemma A.4.5, we have if \( d \geq 1, n \geq KdL \), and \( \Delta \leq cd/n^{5/2} \), there is an event \( \mathcal{E} \) with measure at least \( 1 - 2e^{-d} - L^2 e^{-cn/L} \) on which there are no more than \( d \) risky features at \( \bar{x} \). Worsening constants in the scalings of \( n \) if necessary and requiring moreover \( d \geq K' \log n \) and \( n \geq K'' d^4 \), it follows that we can invoke [221, Lemmas D.14, D.23] to bound the probability of events involving transfer coefficients multiplied by \( 1_{\mathcal{E}} \). Let us also check the residuals we will obtain when applying [221, Lemma D.23]: in the notation there, the vector \( d \) has as its \( \ell \)-th entry \( R_\ell(\bar{x}) \), and so we have bounds \( \|d\|_{1/2} \leq \|d\|_1^2 \) and \( \|d\|_1 = \sum_\ell R_\ell(\bar{x}) \), which means on the event \( \mathcal{E} \), the residual is dominated by the \( C\sqrt{d^4nL} \) term in the scalings we have assumed.

The \( \ell^2 \rightarrow \ell^\infty \) operator norm of a matrix is the maximum \( \ell^2 \) norm of a row of the matrix, and
the $\ell^\infty \to \ell^\infty$ operator norm is the maximum $\ell^1$ norm of a row. Thus

$$\left\| \Phi_{\tilde{x}}^{\ell,0} \right\|_{\ell^2 \to \ell^\infty} = \max_{i=1,\ldots,n} \left\| e_i^* \Phi_{\tilde{x}}^{\ell,0} \right\|_2 = \max_{i=1,\ldots,n} \left\| (W^\ell)_i^* T_{\tilde{x}}^{\ell-1,0} \right\|_2$$

where $(W^\ell)_i^*$ is the $i$-th row of $W^\ell$, which is $n_0$-dimensional when $\ell = 1$ and $n$-dimensional otherwise. In particular, we have

$$\left\| \Phi_{\tilde{x}}^{1,0} \right\|_{\ell^2 \to \ell^\infty} = \max_{i=1,\ldots,n} \left\| (W^1)_i \right\|_2,$$

and so taking a square root and applying Lemma A.6.2 and independence of the rows of $W^1$, we have

$$\mathbb{P} \left[ \left\| \Phi_{\tilde{x}}^{1,0} \right\|_{\ell^2 \to \ell^\infty} \leq 2 \left( 1 + \sqrt{\frac{n_0}{n}} \right) \right] \geq 1 - 2ne^{-cn},$$

for $c > 0$ an absolute constant. When $\ell > 1$, we can write

$$\max_{i=1,\ldots,n} \left\| (W^\ell)_i^* T_{\tilde{x}}^{\ell-1,0} \right\|_2 = \max_{i=1,\ldots,n} \left\| (W^\ell)_i^* T_{\tilde{x}}^{\ell-1,1} P_{l_1(\tilde{x})} W^1 \right\|_2 \leq \left\| W^1 \right\| \max_{i=1,\ldots,n} \left\| (W^\ell)_i^* T_{\tilde{x}}^{\ell-1,1} P_{l_1(\tilde{x})} \right\|_2,$$

where the second line applies Cauchy-Schwarz. Using, say, rotational invariance, Gauss-Lipschitz concentration, and Lemma A.5.48 (or [225, Theorem 4.4.5]), we have

$$\mathbb{P} \left[ \left\| W^1 \right\| > C \left( 1 + \sqrt{\frac{n_0}{n}} \right) \right] \leq 2e^{-cn}$$

for absolute constants $c, C > 0$. On the other hand, note that $\left\| (W^\ell)_i^* T_{\tilde{x}}^{\ell-1,1} P_{l_1(\tilde{x})} \right\|_2$ has the same distribution as the square root of one of the index-0 diagonal terms studied in [221, Lemma D.23] in a network truncated at level $\ell - 1$ instead of $L$ and scaled by $2/n$; and so applying this result together with a union bound and the choice $n \geq \max \{ K, K' d^4 L \}$ for absolute constants $K, K' > 0$ gives

$$\mathbb{P} \left[ \mathbb{1} \mathbb{E} \max_{i=1,\ldots,n} \left\| (W^\ell)_i^* T_{\tilde{x}}^{\ell-1,1} P_{l_1(\tilde{x})} \right\|_2 > C' \right] \leq C'' n e^{-c'd}$$

231
where \( C', c', C'' > 0 \) are absolute constants. We conclude by another union bound

\[
P \left[ \left\| \Phi_x^{\ell,0} \right\|_{\ell^2 \to \ell^\infty} \leq C(1 + \sqrt{\frac{n_0}{n}}) \right] \geq 1 - 2e^{-cn} - C'ne^{-c'd} - C''L^2e^{-c'n/L}.
\]

We can reduce the study of the partial risky propagation coefficients to a similar calculation. We have

\[
\left\| \Phi_x^{\ell',\ell'} \right\|_{\ell^0 \to \ell^\infty} = \max_{j=1,\ldots,n} \left\| (W_j^{\ell'} )^* \left( T_{\bar{x}}^{\ell-1,\ell'} \right)_S \right\|_1,
\]

where by construction we have that \( \ell > \ell' \). In the case \( \ell = \ell' + 1 \), the form is slightly different; we have

\[
\left\| \Phi_x^{\ell'+1,\ell'} \right\|_{\ell^0 \to \ell^\infty} = \max_{j=1,\ldots,n} \left\| (W_j^{\ell'+1} )^* \right\|_1 \leq |S| \max_{j=1,\ldots,n} \left\| (W_j^{\ell'+1} )^* \right\|_\infty,
\]

where the inequality uses Lemma A.6.10. The classical estimate for the gaussian tail gives

\[
P \left[ \left( W_j^{\ell'+1} \right)_j > \sqrt{\frac{2d}{n}} \right] \leq 2e^{-d}, \quad (A.161)
\]

for each \( k \in [n] \), so a union bound gives

\[
P \left[ \max_{j=1,\ldots,n} \left\| (W_j^{\ell'+1} )^* \right\|_\infty > \sqrt{\frac{2d}{n}} \right] \leq 2ne^{-d},
\]

and we conclude

\[
P \left[ \left\| \Phi_x^{\ell'+1,\ell'} \right\|_{\ell^0 \to \ell^\infty} \leq |S| \sqrt{\frac{2d}{n}} \right] \geq 1 - 2e^{-d/2},
\]

where the final bound holds if \( d \geq 2 \log n \). Next, we assume \( \ell > \ell' + 1 \). In this case, Lemma A.6.10 gives

\[
\left\| \Phi_x^{\ell',\ell'} \right\|_{\ell^0 \to \ell^\infty} = \max_{j=1,\ldots,n} \left\| (W_j^{\ell'} )^* \left( T_{\bar{x}}^{\ell-1,\ell'} \right)_S \right\|_1
\]

\[
\leq |S| \max_{j=1,\ldots,n} \left\| (W_j^{\ell'} )^* \left( \left( T_{\bar{x}}^{\ell-1,\ell'} \right)_S \right) \right\|_\infty.
\]
For the second term on the RHS of the inequality, we write

\[
\max_{j=1,\ldots,n} \left\| (W^\ell_j)^* T^{\ell-1,\ell'}_x \right\|_{\infty} = \max_{j=1,\ldots,n} \max_{k \in S} \left| (W^\ell_j)^* T^{\ell-1,\ell'}_x e_k \right|
\]

then apply rotational invariance of the distribution of \((W^\ell_j)\) and \(\mathcal{T}^{\ell-1}\)-measurability of \(T^{\ell-1,\ell'}_x\) to obtain

\[
\max_{j=1,\ldots,n} \max_{k \in S} \left| (W^\ell_j)^* T^{\ell-1,\ell'}_x e_k \right| = \max_{j=1,\ldots,n} \max_{k \in S: \left\| T^{\ell-1,\ell'}_x e_k \right\|_2 > 0} \left| (W^\ell_j)^* T^{\ell-1,\ell'}_x e_k \right|
\]

\[
\overset{d}{=} \max_{j=1,\ldots,n} \max_{k \in S} |g_j| \left\| T^{\ell-1,\ell'}_x e_k \right\|_2,
\]

where \(g \sim \mathcal{N}(0,(2/n)I)\) is independent of everything else in the problem. We have by [221, Lemma D.14] based on our previous choices of \(n\) and \(d\)

\[
P\left[ 1_{E} \left\| T^{\ell-1,\ell'}_x e_k \right\|_2 \leq C \right] \geq 1 - C'e^{-cn/L},
\]

and (A.161) applied to \(g\) controls the remaining term. Taking a union bound over \([n]\) in these two estimates and partitioning with \(E\), we conclude

\[
P\left[ \max_{j=1,\ldots,n} \left\| (W^\ell_j)^* \left( T^{\ell-1,\ell'}_x \right)_S \right\|_{\infty} > C \sqrt{\frac{d}{n}} \right] \leq 2ne^{-d} + C'ne^{-cn/L} + 2e^{-d} + L^2 e^{-c'n/L}
\]

and thus

\[
P\left[ \left\| \Phi^{\ell,\ell'}_x \right\|_{\ell^\infty_{\ell^\infty}} > C|S| \sqrt{\frac{d}{n}} \right] \leq C'e^{-d/2} + C''n^2 e^{-cn/L},
\]

where the last bound holds if \(d \geq 2 \log n\). Choosing \(n \geq KL \log n\) for a suitable absolute constant \(K > 0\) allows us to simplify the residual terms in the probability bounds to the forms we have claimed.

\[\square\]

**Lemma A.4.7.** There is an absolute constant \(c > 0\) and absolute constants \(k, k', K, K' > 0\) such that for any \(d \geq K \log n\), if \(n \geq K' \max \{d^4L, 1\}\), \(\Delta \leq kn^{-5/2}\), and \(\varepsilon \leq k'\Delta \left( 1 + \sqrt{\frac{m_0}{n}} \right)^{-1}\), then one
has for any \( \bar{x} \in N_e \\

\[ P \left[ \{ \text{SSC}(L) \text{ holds at } \bar{x} \} \cap \left\{ \sum_{\ell=1}^{L} |R_\ell(\bar{x}, \Delta)| \leq d \right\} \right] \geq 1 - e^{-cd}. \]

**Proof.** We start by constructing a high-probability event on which we have control of every possible propagation coefficient. For any \( d \geq K \log n \), choosing \( \Delta \leq cn^{-5/2} \) and \( n \geq K'd^4L \) and applying the first conclusion in Lemma A.4.6 and a union bound, we have

\[ P \left[ \exists \ell \in [L], \left\| \Phi_{\bar{x}}^{\ell,0} \right\|_{\ell^2 \rightarrow \ell^\infty} > C_1 (1 + \sqrt{\frac{n_0}{n}}) \right\| \leq Ce^{-cd} \]  

(A.162)

and under the same hypotheses, for any \( 1 \leq \ell' < \ell \leq L \) and any \( S \subset [n] \), the second conclusion in Lemma A.4.6 gives

\[ P \left[ \left\| \Phi_{\bar{x}}^{\ell,\ell'} \right\|_{S^\ell \rightarrow S^\ell^\infty} > C_2 |S| \sqrt{\frac{d}{n}} \right] \leq C' e^{-16d}. \]

Using Lemma A.4.5, we have if \( n \geq \max\{KdL, 4\} \) and \( \Delta \leq K'/n^{5/2} \)

\[ P \left[ \sum_{\ell=1}^{L} |R_\ell(\bar{x}, \Delta)| > d \right] \leq 2e^{-d} + L^2 e^{-cn/L}. \]  

(A.163)

Denote the complement of the event in the previous bound as \( \mathcal{E} \). On \( \mathcal{E} \), there are no more than \( d \) levels in the network with risky features. There are at most \( \sum_{k=0}^{[d]} \binom{n}{k} \) ways to choose a subset \( S \subset [n] \) with cardinality at most \([d]\); using \( n \geq e \) and \( d \geq 1 \), we have

\[ \sum_{k=0}^{[d]} \binom{n}{k} \leq 1 + [d]n^{[d]} \leq 4dn^{2d}. \]

In addition, there are at most \( L^2 \) ways to pick two indices \( 1 \leq \ell' < \ell \leq L \). Using \( n \geq L \), this yields

234
at most $4d n^{2+2d} \leq n^{8d}$ items to union bound over, i.e.,

$$\mathbb{P} \left[ \bigcup_{S \subseteq [n]} \bigcup_{1 \leq \ell, \ell' \leq L} \left\{ \left\| \Phi^{\ell, \ell'}_x \right\|_{S \to \ell' \to \ell} > C_2 |S| \sqrt{\frac{d}{n}} \right\} \right] \leq C e^{-8d} \quad (A.164)$$

if $d \geq K \log n$ and $n \geq \max\{K' d^4 L, n_0\}$. Denote the complement of the union of the events in the bounds (A.162), (A.163) and (A.164) and $\mathcal{E}$ as $\mathcal{G}$; taking additional union bounds and worst-casing absolute constants, we have shown

$$\mathbb{P}[\mathcal{G}] \geq 1 - C e^{-cd}.$$ 

If we enumerate the levels of the network that have risky features as $1 \leq \ell_1 < \cdots < \ell_r \leq L$, it follows from our previous counting argument that on $\mathcal{G}$, we have transfer coefficient bounds (for any $\ell \in [L]$ and any $\ell_i < \ell$)

$$\left\| \Phi^{\ell, 0}_x \right\|_{\ell^2 \to \ell^\infty} \leq C_1 (1 + \sqrt{\frac{n_0}{n}}), \quad \left\| \Phi^{\ell, \ell_i}_{\ell, \ell_i}(\bar{x}) \right\|_{\ell^\infty \to \ell^\infty} \leq C_2 |R_{\ell_i}(\bar{x})| \sqrt{\frac{d}{n}}.$$ 

Now we begin the induction. Let $x \in N_\varepsilon(\bar{x})$. For $\text{SSC}(1)$, we have from the definitions

$$\left\| \rho^1(x) - \rho^1(\bar{x}) \right\|_\infty \leq \varepsilon \left\| \Phi^{1,0}_x \right\|_{\ell^2 \to \ell^\infty} \leq C_1 (1 + \sqrt{\frac{n_0}{n}}) \varepsilon,$$

where the last inequality holds on $\mathcal{G}$. So we have $\text{SSC}(1)$ on $\mathcal{G}$ if $\varepsilon \leq \Delta (C_1 (1 + \sqrt{n_0/n}))^{-1}$. Continuing, we suppose that we have established $\text{SSC}(\ell - 1)$ on $\mathcal{G}$. We can therefore apply (A.160) together with our transfer coefficient bounds to get

$$\left\| \rho^\ell(x) - \rho^\ell(\bar{x}) \right\|_\infty \leq C_1 (1 + \sqrt{\frac{n_0}{n}}) \varepsilon + C_2 \Delta \sqrt{\frac{d}{n}} \sum_{i \in [r]: \ell_i < \ell} |R_{\ell_i}(\bar{x})|$$

$$\leq C_1 (1 + \sqrt{\frac{n_0}{n}}) \varepsilon + C_2 \Delta \sqrt{\frac{d^3}{n}}.$$ 

235
Notice that the last bound does not depend on $\ell$. Thus, if we choose $\varepsilon \leq \Delta (2C_1 (1 + \sqrt{\frac{n_0}{n}}))^{-1}$ and $n \geq 4C_2^2 d^3$, we obtain $\|\rho'(x) - \rho' (\bar{x})\|_\infty \leq \Delta$; by induction, we can conclude that $\text{SSC}(L)$ holds on $\mathcal{G}$, which implies the claim; we obtain the final simplified probability bound by worsening the constant in the exponent. $\square$

**Uniformizing Forward Features Under SSC**

Under the $\text{SSC}(L)$ condition, we can uniformize forward and backward features. A prerequisite of our approach, which we also used to establish $\text{SSC}(L)$ in the previous section, is control of certain residuals that appear when a small number of supports can change off the nominal forward and backward correlations.

In previous sections, most of our results (e.g. Lemma A.4.1) feature a lower bound of the type $n \geq K$ in their hypotheses. After uniformizing, we will discard this hypothesis using our extra assumption that $n_0 \geq 3$, which gives us a lower bound on the logarithmic terms of the form $\log(Cnn_0)$ that appear as lower bounds on $d$ after uniformizing, and the fact that our lower bounds on $n$ always involve a polynomial in $d$. Thus, by adjusting absolute constants, we can achieve the same effect as previously.

**Lemma A.4.8.** There are absolute constants $c, C > 0$ and an absolute constant $K, K' > 0$ such that for any $d \geq Kd_0 \log(nn_0C_M)$, if $n \geq K'd^4L$ then one has

$$
\mathbb{P} \left[ \bigcap_{\bar{x} \in \mathcal{N}} \left\{ \text{SSC}(L, n^{-3}n_0^{-1/2}, Cn^{-3}) \text{ holds at } \bar{x} \right\} \cap \left\{ \sum_{\ell=1}^{L} \left| R_\ell (\bar{x}, Cn^{-3}) \right| \leq d \right\} \right] 
\geq 1 - e^{-cd}.
$$

**Proof.** Following the discussion in Section A.4.3, if $0 < \varepsilon \leq 1$ then $|N_{\varepsilon}| \leq e^{d_0 \log(CM/\varepsilon)}$; to apply Lemma A.4.7 we at least need $\Delta \leq kn^{-5/2}$ and $\varepsilon \leq k'\Delta \left(1 + \sqrt{\frac{n_0}{n}}\right)^{-1}$, so it suffices to put $\Delta = Cn^{-3}$ when $n$ is chosen larger than an absolute constant, and require $\varepsilon \leq \min \left\{ 1, k'Cn^{-5/2} \left(1 + \sqrt{\frac{n_0}{n}}\right)^{-1} \right\}$. Fixing $\varepsilon = n^{-3}n_0^{-1/2}$, which again is admissible when $n$ is sufficiently large compared to an absolute
constant, for any \( d \geq Kd_0 \log(nn_0C_M) \), we choose \( n \geq K \max\{1, d^4L\} \) and take a union bound to obtain the claim (using here that \( C_M \geq 1 \)). □

**Lemma A.4.9.** There is an absolute constant \( c > 0 \) and absolute constants \( K, K', K'' > 0 \) such that for any \( d \geq Kd_0 \log(nn_0C_M) \), if \( n \geq Kd_L \), then one has

\[
P\left[ \bigcap_{x \in M} \left\{ \forall \ell \in [L], \|\alpha^\ell(x)\|_2 - 1 \leq \frac{1}{2} \right\} \right] \geq 1 - e^{-cd}.
\]

**Proof.** Let \( \bar{x} \in N_{n^{-3}n_0^{-1/2}} \). Lemma A.4.2 and a union bound give

\[
P\left[ \exists \ell \in [L] : \|\alpha^\ell(\bar{x})\|_2 - 1 > \frac{1}{4} \right] \leq C'L^2e^{-cn/L} \leq e^{-c'd}
\]

if \( n \geq KdL \) and \( d \geq K' \log n \). If additionally \( d \geq K'd_0 \log(nn_0C_M) \), we obtain by the discussion in Section A.4.3 and another union bound

\[
P\left[ \bigcup_{\bar{x} \in N_{n^{-3}n_0^{-1/2}}} \left\{ \exists \ell \in [L] : \|\alpha^\ell(\bar{x})\|_2 - 1 > \frac{1}{4} \right\} \right] \leq e^{-cd}.
\]

Let \( \mathcal{E} \) denote the event studied in Lemma A.4.8; choose \( d \geq Kd_0 \log(nn_0C_M) \) and \( n \) sufficiently large to make the measure bound applicable. A union bound gives

\[
P\left[ \mathcal{E}^c \cup \bigcup_{\bar{x} \in N_{n^{-3}n_0^{-1/2}}} \left\{ \exists \ell \in [L] : \|\alpha^\ell(\bar{x})\|_2 - 1 > \frac{1}{4} \right\} \right] \leq e^{-cd}.
\]

Let \( \mathcal{G} \) denote the complement of the event in the previous bound. For any \( x \in M \), we can find a point \( \bar{x} \in N_{n^{-3}n_0^{-1/2}} \cap N_{n^{-3}n_0^{-1/2}}(x) \). On \( \mathcal{G} \), SSC\((L, n^{-3}n_0^{-1/2}, Cn^{-3})\) holds at every point in the net \( N_{n^{-3}n_0^{-1/2}} \), which implies that on \( \mathcal{G} \)

\[
\forall \ell \in [L], \|\rho^\ell(x) - \rho^\ell(\bar{x})\|_\infty \leq \frac{C}{n^3}, \tag{A.165}
\]

237
and by the 1-Lipschitz property of $[ \cdot ]_+$ and Lemma A.6.10, this also implies that on $G$

$$\forall \ell \in [L], \left\| \alpha^{\ell}(x) - \alpha^{\ell}(\bar{x}) \right\|_2 \leq \frac{C}{n^{5/2}}.$$ 

Choosing $n \geq (4C)^{2/5}$, the RHS of this bound is no larger than $1/4$. We write using the triangle inequality

$$\left| \left\| \alpha^{\ell}(x) \right\|_2 - 1 \right| \leq \left| \left\| \alpha^{\ell}(x) \right\|_2 - \left\| \alpha^{\ell}(\bar{x}) \right\|_2 \right| + \left| \left\| \alpha^{\ell}(\bar{x}) \right\|_2 - 1 \right|.$$ 

Using the triangle inequality again, the first term on the RHS is no larger than $1/4$ for any $\ell$ on $G$. The second term is also no larger than $1/4$ on $G$ by control over the net. We conclude that on $G$

$$\forall \ell \in [L], \left| \left\| \alpha^{\ell}(x) \right\|_2 - 1 \right| \leq \frac{1}{2}.$$ 

This implies that the event $G$ is contained in the set

$$\bigcap_{x \in M} \left\{ \forall \ell \in [L], \left| \left\| \alpha^{\ell}(x) \right\|_2 - 1 \right| \leq \frac{1}{2} \right\},$$

which is closed, by continuity of $\| \cdot \|_2$ and of the features as a function of the parameters, and is therefore also an event. The claim follows. \hfill \Box

**Lemma A.4.10.** There are absolute constants $c, C > 0$ and absolute constants $K, K' > 0$ such that for any $d \geq Kd_0 \log(n n_0 C_M)$, if $n \geq K'd^4 L$, then one has

$$\mathbb{P} \left[ \bigcap_{(x,x') \in M \times M} \left\{ \forall \ell \in [L], \left| \langle \alpha^{\ell}(x), \alpha^{\ell}(x') \rangle - \cos \varphi^{(\ell)}(\angle(x,x')) \right| \leq C \sqrt{\frac{d^3 \ell}{n}} \right\} \right] \geq 1 - e^{-cd}.$$ 

**Proof.** Let $\bar{x}, \bar{x}' \in N_{n^{-3/2}}$. Lemma A.4.1 and a union bound give

$$\mathbb{P} \left[ \exists \ell \in [L] : \left| \langle \alpha^{\ell}(\bar{x}), \alpha^{\ell}(\bar{x}') \rangle - \cos \varphi^{(\ell)}(\angle(\bar{x}, \bar{x}')) \right| > C \sqrt{\frac{d^3 \ell}{n}} \right] \leq C' L e^{-cd} \leq e^{-c'd}$$

238
if $d \geq K \log n$ and $n \geq \max\{K'd^3L, K''d^4, K'''\}$. If additionally $d \geq Kd_0 \log(nn_0C_M)$, we obtain by the discussion in Section A.4.3 and another union bound

$$\Pr \left[ \bigcup_{(\bar{x}, \bar{x}') \in \mathcal{N}_{n^{3/2}}^{n^{3/2}} \cap \mathcal{N}_{n^{3/2}}^{n^{3/2}}(x)} \left\{ \exists \ell \in [L] : \left| \langle \alpha^\ell(\bar{x}), \alpha^\ell(\bar{x}') \rangle - \cos \varphi^{(\ell)}(\angle(\bar{x}, \bar{x}')) \right| > C \sqrt{\frac{d^3\ell}{n}} \right\} \right] \leq e^{-cd},$$

where with an abuse of notation we write $S \times S$ to denote $S \times S$ for a set $S$. Let $E_1$ denote the event studied in Lemma A.4.8, and let $E_2$ denote the event studied in Lemma A.4.9; choose $n$ sufficiently large to make the measure bounds applicable. A union bound gives

$$\Pr \left[ E_1^c \cup E_2^c \cup \bigcup_{(\bar{x}, \bar{x}') \in \mathcal{N}_{n^{3/2}}^{n^{3/2}} \cap \mathcal{N}_{n^{3/2}}^{n^{3/2}}(x)} \left\{ \exists \ell \in [L] : \left| \langle \alpha^\ell(\bar{x}), \alpha^\ell(\bar{x}') \rangle - \cos \varphi^{(\ell)}(\angle(\bar{x}, \bar{x}')) \right| > C \sqrt{\frac{d^3\ell}{n}} \right\} \right] \leq e^{-cd}$$

after adjusting constants. Let $\mathcal{G}$ denote the complement of the event in the previous bound. For any $(x, x') \in \mathcal{M} \times \mathcal{M}$, we can find a point $\bar{x} \in \mathcal{N}_{n^{-3/2}n_0^{-1/2}} \cap \mathcal{N}_{n^{-3/2}n_0^{-1/2}}(x)$ and a point $\bar{x}' \in \mathcal{N}_{n^{-3/2}n_0^{-1/2}} \cap \mathcal{N}_{n^{-3/2}n_0^{-1/2}}(x')$. On $\mathcal{G}$, SSC$(L, n^{-3}n_0^{-1/2}, Cn^{-3})$ holds at every point in the net $\mathcal{N}_{n^{-3}n_0^{-1/2}}$, which implies that on $\mathcal{G}$

$$\forall \ell \in [L], \| \rho^\ell(x) - \rho^\ell(\bar{x}) \|_\infty \leq \frac{C}{n^3}, \quad \text{and} \quad \forall \ell \in [L], \| \rho^\ell(x') - \rho^\ell(\bar{x}') \|_\infty \leq \frac{C}{n^3},$$

and by the 1-Lipschitz property of $[\cdot]_+$ and Lemma A.6.10, this also implies that on $\mathcal{G}$

$$\forall \ell \in [L], \| \alpha^\ell(x) - \alpha^\ell(\bar{x}) \|_2 \leq \frac{C}{n^{5/2}}, \quad \text{and} \quad \forall \ell \in [L], \| \alpha^\ell(x') - \alpha^\ell(\bar{x}') \|_2 \leq \frac{C}{n^{5/2}}. \quad \text{(A.166)}$$
For any $\ell \in [L]$, we write using the triangle inequality

\[
\left| \langle \alpha^\ell(x), \alpha^\ell(x') \rangle - \cos \varphi^{(\ell)}(\angle(x, x')) \right| \leq \left| \langle \alpha^\ell(x), \alpha^\ell(x') \rangle - \langle \alpha^\ell(\bar{x}), \alpha^\ell(x') \rangle \right|
+ \left| \langle \alpha^\ell(\bar{x}), \alpha^\ell(x') \rangle - \langle \alpha^\ell(\bar{x}), \alpha^\ell(\bar{x}') \rangle \right|
+ \left| \langle \alpha^\ell(\bar{x}), \alpha^\ell(\bar{x}') \rangle - \cos \varphi^{(\ell)}(\angle(\bar{x}, \bar{x}')) \right|
+ \left| \cos \varphi^{(\ell)}(\angle(\bar{x}, \bar{x}')) - \cos \varphi^{(\ell)}(\angle(x, x')) \right|.
\]

(A.167)

Using Cauchy-Schwarz, we have on $G$

\[
\left| \langle \alpha^\ell(x), \alpha^\ell(x') \rangle - \langle \alpha^\ell(\bar{x}), \alpha^\ell(x') \rangle \right| \leq \left\| \alpha^\ell(x) - \alpha^\ell(\bar{x}) \right\| \left\| \alpha^\ell(x') \right\| \leq \frac{2C}{n^{5/2}},
\]

with the same bound holding for the second term in (A.167) by an analogous argument. For the third term, we have on $G$

\[
\left| \langle \alpha^\ell(\bar{x}), \alpha^\ell(\bar{x}') \rangle - \cos \varphi^{(\ell)}(\angle(\bar{x}, \bar{x}')) \right| \leq C \sqrt{\frac{d^3 \ell}{n}}.
\]

For the last term, we use 1-Lipschitzness of $\cos$ and 1-Lipschitzness of the $\varphi^{(\ell)}$, which follows from Lemma A.5.5 and the chain rule, to obtain

\[
\left| \cos \varphi^{(\ell)}(\angle(\bar{x}, \bar{x}')) - \cos \varphi^{(\ell)}(\angle(x, x')) \right| \leq |\angle(\bar{x}, \bar{x}') - \angle(x, x')|.
\]

Using Lemma A.3.4 and several applications of the triangle inequality, we get

\[
|\angle(\bar{x}, \bar{x}') - \angle(x, x')| \leq \sqrt{2} \left( \left| \|x - x'\|_2 - \|\bar{x} - \bar{x}'\|_2 \right| \right.
\leq \sqrt{2} \left( \|x - x'\|_2 - (\|\bar{x} - \bar{x}'\|_2 \right)
\leq \sqrt{2} \|x - \bar{x}\|_2 + \sqrt{2} \|x' - \bar{x}'\|_2 \leq \frac{2\sqrt{2}}{n^3}.
\]
and so returning to (A.167), we have shown

$$\left| \langle \alpha^\ell(x), \alpha^\ell(x') \rangle - \cos \varphi^{(\ell)}(\angle(x, x')) \right| \leq C \sqrt{\frac{d^3 \ell}{n}} + \frac{C'}{n^{5/2}} \leq (C + C') \sqrt{\frac{d^3 \ell}{n}}$$

for every $\ell \in [L]$. This implies that the event $G$ is contained in the set

$$\left\{ \forall \ell \in [L], \left| \langle \alpha^\ell(x), \alpha^\ell(x') \rangle - \cos \varphi^{(\ell)}(\angle(x, x')) \right| \leq C \sqrt{\frac{d^3 \ell}{n}} \right\},$$

which is closed, by continuity of the inner product and of the features as a function of the parameters, and is therefore also an event. The claim follows. □

**Lemma A.4.11.** Assume $n, L, d$ satisfy the requirements of lemma A.4.10 and additionally $d \geq 1, n \geq K\sqrt{L}$ for some $K$. Then

$$\mathbb{P} \left[ \|f_{\theta_0}\|_{L^{\infty}} \leq \sqrt{d} \right] \geq 1 - e^{-cd},$$

$$\mathbb{P} \left[ \|\xi\|_{L^{\infty}} \leq \sqrt{d} \right] \geq 1 - e^{-cd}.$$

Define

$$\hat{\xi}(x) = -f_*(x) + \int_{M} f_{\theta_0}(x') d\mu^\infty(x').$$

Then under the same assumptions

$$\mathbb{P} \left[ \|\hat{\xi} - \zeta\|_{L^{\infty}} \leq \sqrt{\frac{d}{L^2} + d^{5/2} \sqrt{\frac{L}{n}}} \right] \geq 1 - e^{-cd}$$

for some numerical constant $c$.

**Proof.** See [221, Lemma D.11]. □

**Lemma A.4.12.** For some integer $d_0$, there exist absolute constants $K, K', K'', K''', K'''' > 0$ such that for any $d \geq \max\{Kd_0 \log(nn_0C_M), K' \log L\}$, if $n \geq \max\{K''d^4L, K'''L \log n, K''''\}$ then
1. If $d_0 = 1$ and $n \geq K'''' \max \left\{ d^2 L, \left( \frac{\kappa}{c_\tau} \right)^{1/3}, \kappa^{2/5} \right\}$ where $K''''$ is some absolute constant. $\kappa$ and $c_\tau$ are the extrinsic curvature and injectivity coefficient defined in Section A.1, then on an event of probability at least $1 - e^{-cd}$, one has

$$\left\| f_{\theta_0} \right\|_{\text{Lip}} \leq \sqrt{d}$$

for a numerical constant $c$, and where the Lipschitz seminorm is taken with respect to the Riemannian distance on $M_\pm$.

2. If $M = S^{n_0 - 1}$ so that $d_0 = n_0 - 1$, then on an event of probability at least $1 - e^{-cd}$, one has

$$\left\| f_{\theta_0} \right\|_{S^{n_0 - 1}} \leq \sqrt{d}$$

for a numerical constant $c$.

Proof. See [221, Lemma D.12].

**Lemma A.4.13.** There are absolute constants $c, C > 0$ and absolute constants $K_1, \ldots, K_4 > 0$ such that for any $d \geq K d_0 \log(n_0 C_M)$, if $n \geq K'd^4 L$, then there exists an event $E$ such that

1. On $E$, we have

$$\forall \ell \in [L], \left| \langle \beta^{\ell-1}(x), \beta^{\ell-1}(x') \rangle - \frac{n}{2} \prod_{\ell' = \ell-1}^{L-1} \left( 1 - \frac{\varphi(\ell') \angle(x, x'))}{\pi} \right) \right| \leq C \sqrt{d^4 n L}$$

simultaneously for every $(x, x') \in M \times M$;

2. $\mathbb{P}[E] \geq 1 - e^{-cd}$.

Proof. Let $E_1$ denote the event studied in Lemma A.4.8, with $C_0$ denoting the absolute constant appearing in the $\text{SSC}(L)$ condition there; choose $d \geq K d_0 \log(n_0 C_M)$ and $n$ sufficiently large to make the measure bound applicable. We will need to apply [221, Lemma D.23] together with a derandomization argument to prove the claim; we appeal to the same residual checks at the
beginning of the proof of Lemma A.4.6 to see that on $E_1$, the dominating residual in [221, Lemma D.23] under the scalings of $d$ and $n$ we enforce here is of size $C\sqrt{d^3 n L}$.

For any subset $S \subset [L] \times [n]$, we write $S_\ell = \{i \in [n] \mid (\ell, i) \in S\}$, and we define $S(S) = \{-1, +1\}^{|S|} \times \cdots \times \{-1, +1\}^{|S_L|}$ for the set of “lists” of sign patterns with sizes adapted to these projections of $S$, with the convention $\{-1, +1\}^0 = \{0\}$. If $\Sigma = \{\sigma_1, \ldots, \sigma_L\} \in S(S)$ is such a list of sign vectors and $\Delta \geq 0$, we define

$$\tilde{I}_\ell(x, S, \Sigma, \Delta) = \text{supp} \left( 1_{\mu^\ell(x) > \sum_{i \in S_\ell(\sigma_i, \Delta)} e_i} \right),$$

which is a sort of two-sided robust analogue of the support of $\alpha^\ell(x)$: notice that when $S = \emptyset$ we have $\tilde{I}_\ell(x, S, \Sigma, \Delta) = I_\ell(x)$. We also define for $\ell = 0, 1, \ldots, L - 1$

$$\tilde{\beta}_{S, \Sigma, \Delta}^\ell(x) = \left( W^{L+1} P_{I_L(x, S, \Sigma, \Delta)} W^L P_{I_{L-1}(x, S, \Sigma, \Delta)} \cdots W^{L+2} P_{I_2(x, S, \Sigma, \Delta)} \right)^*,$$

a generalized backward feature induced by these robust support patterns. Writing for concision

$$\mathcal{S}_{\bar{x}, \bar{x}', S, S', \Sigma, \Sigma}' = \begin{cases} \exists \ell \in [L] : & \left| \left( \tilde{\beta}_{S, \Sigma, C^{n-3}}^{\ell-1}(\bar{x}), \tilde{\beta}_{S', \Sigma', C^{n-3}}^{\ell-1}(\bar{x}') \right) \right| \\ & -\frac{n}{2} \prod_{\ell'=1}^{L-1} \left( 1 - \frac{\varphi^{\ell'}(\angle(\bar{x}, \bar{x}'))}{\pi} \right) \\ & > C_1 \sqrt{d^4 n L \log^4 n} \end{cases},$$

where $C_1 > 0$ is an absolute constant we will specify below to make the event hold with high probability, we then define the event\footnote{To see that this set is indeed an event, use that $\tilde{\beta}_{S, \Sigma, \Delta}^\ell(x)$ is a continuous function of the network weights except with respect to the support projections; but $x \mapsto 1_{x > 0}$ is increasing, hence Borel-measurable, and so the set consists of a finite union of Borel-measurable sets.} $\mathcal{E}_2$
There are no more than \( \sum_{k=0}^{d} \binom{nL}{k} \leq n^{4d} \) ways to choose the subset \( S \) in this union, and for a fixed \( S \) there are no more than \( 2^d \) ways to choose the sign pattern \( \Sigma \). Thus, there are no more than \( \exp(10d \log n + 12d_0 \log(nn_0C_M)) \) elements in the union, and under the condition on \( d \) this number is no larger than \( n^{11d} \). For concision, write

\[
\xi_{\ell}(\bar{x}, \bar{x}') = \frac{n}{2} \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{\varphi(\ell') (\angle(\bar{x}, \bar{x}'))}{\pi} \right).
\]

For any instantiation of these parameters, [221, Lemma D.23] and a union bound give

\[
\mathbb{P} \left[ \exists \ell \in [L] : \left| \left( \tilde{\mathcal{P}}_{S,\Sigma,\mathcal{S}C_{0n^{-3}}}^{\ell-1}(\bar{x}), \tilde{\mathcal{P}}_{S,\Sigma,\mathcal{S}C_{0n^{-3}}}^{\ell-1}(\bar{x}') \right) - \xi_{\ell-1}(\bar{x}, \bar{x}') \right| > C\sqrt{d^4nL} \right]
\]

\[
\leq \mathbb{P}[E_1^c] + \mathbb{P} \left[ \exists \ell \in [L] : \left| \mathbb{I}_{E_1} \left( \tilde{\mathcal{P}}_{S,\Sigma,\mathcal{S}C_{0n^{-3}}}^{\ell-1}(\bar{x}), \tilde{\mathcal{P}}_{S,\Sigma,\mathcal{S}C_{0n^{-3}}}^{\ell-1}(\bar{x}') \right) - \xi_{\ell-1}(\bar{x}, \bar{x}') \right| > C\sqrt{d^4nL} \right]
\]

\[
\leq e^{-cd}
\]

for any \( d \geq K \log n \) and \( n \geq K'd^4L \). Thus, if we set \( C_1 = C \) and enforce the conditions \( d \geq Kd_0 \log(nn_0C_M)/\log n \) and \( n \geq \max K'd^4L \log^4 n \), we have by a union bound

\[
\mathbb{P}[E_1 \cup E_2] \leq n^{-cd}.
\]

Let \( \mathcal{G} = E_1^c \cap E_2^c \). For any \((x, x') \in M \times M\), we can find a point \( \bar{x} \in N_{n^{-3}n_0^{-1/2}} \cap N_{n^{-3}n_0^{-1/2}}(x) \) and a point \( \bar{x}' \in N_{n^{-3}n_0^{-1/2}} \cap N_{n^{-3}n_0^{-1/2}}(x') \). On \( \mathcal{G} \), \( \text{SSC}(L, n^{-3}n_0^{-1/2}, Cn^{-3}) \) holds at every point in the net \( N_{n^{-3}n_0^{-1/2}} \), and there are no more than \( d \) \( Cn^{-3} \)-risky features at any point in the net \( N_{n^{-3}n_0^{-1/2}} \), and in addition, following (A.155), we have almost surely on \( \mathcal{G} \) that all risky features are realized for magnitudes in \((-\Delta, +\Delta)\). This implies that on \( \mathcal{G} \), the support sets \( \bigsqcup_{\ell \in [L]} I_{\ell}(x) \) at any point \( x \in N_{n^{-3}n_0^{-1/2}}(\bar{x}) \) differ by the support sets \( \bigsqcup_{\ell \in [L]} I_{\ell}(\bar{x}) \) at the base point in the net by no more than \( d \) entries, consisting only of a subset of the risky features at \( \bar{x} \); the analogous statement is of course true for \( x' \) and \( \bar{x}' \). At the same time, notice that on the event \( E_2^c \) we have constructed, we have control of every possible backward feature inner product obtained by modifying the supports at the base points \( \bar{x}, \bar{x}' \) at no more than \( d \) risky features (each), since, for example, if \((\rho_{\ell}(\bar{x}))_i\) is risky,
then \(1_{(\rho^f(\bar{x})), > \Delta}\) corresponds to “turning off” the feature, and \(1_{(\rho^f(\bar{x})), > -\Delta}\) corresponds to “turning on” the feature. Formally, we have established that on \(G\)

\[
\forall \ell \in [L], \quad \left| \beta^{\ell-1}(x), \beta^{\ell-1}(x') \right| - \frac{n}{2} \prod_{\ell' = \ell-1}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(\angle(\bar{x}, \bar{x}'))}{\pi} \right) \leq C_1 \sqrt{d^4nL \log^4 n}.
\]

We can use differentiability properties for the rest: following the proof of Lemma A.4.10, we have

\[
|\angle(\bar{x}, \bar{x}') - \angle(x, x')| \leq \sqrt{2}\|x - \bar{x}\|_2 + \sqrt{2}\|x' - \bar{x}'\|_2 \leq \frac{2\sqrt{2}}{n^2},
\]

so we just need a Lipschitz property for the function \(q(v) = (n/2) \prod_{\ell' = \ell}^{L-1} (1 - \pi^{-1} \varphi^{(\ell')}(v))\). For this we appeal to Lemma A.5.5, which shows that the function \(\varphi\) is smooth, increasing and concave; therefore by the chain rule, the functions \(\varphi^{(\ell')}\) are increasing and concave, and by the Leibniz rule, \(q\) is decreasing and convex. It therefore suffices to calculate \(q'(0)\); this is done in Lemma A.3.15, which gives \(q'(0) = -n(L - \ell)/(2\pi)\), and in particular \(|q'(0)| \leq c\ell n\). It follows

\[
\left| \frac{n}{2} \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(\angle(\bar{x}, \bar{x}'))}{\pi} \right) - \frac{n}{2} \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(\angle(x, x'))}{\pi} \right) \right| \leq \frac{c L}{n^2},
\]

so that by the triangle inequality

\[
\forall \ell \in [L], \quad \left| \beta^{\ell-1}(x), \beta^{\ell-1}(x') \right| - \frac{n}{2} \prod_{\ell' = \ell-1}^{L-1} \left( 1 - \frac{\varphi^{(\ell')}(\angle(x, x'))}{\pi} \right) \leq 2C_1 \sqrt{d^4nL \log^4 n},
\]

where the residual simplification is valid when \(n \geq KL\). We conclude that the set

\[
\bigcap_{(x, x') \in M \times M} \left\{ \forall \ell \in [L], \left| \beta^{\ell-1}(x), \beta^{\ell-1}(x') \right| - \xi_{\ell-1}(x, x') \leq 2C_1 \sqrt{d^4nL \log^4 n} \right\}
\]

contains the event \(G\), which satisfies the claimed properties and completes the proof (after rescaling \(d\) by \(1/\log n\), which updates the lower bound on \(d\)). \(\square\)
A.4.4 Auxiliary Results

Lemma A.4.14. There are absolute constants $c_1, C, C' > 0$ and absolute constants $K, K' > 0$ such that for any $L \in \mathbb{N}$, if $n \geq \max\{K \log^4 n, K'L\}$, then for every $\ell \in [L]$ one has

$$\left| \mathbb{E} \left[ \varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \right] \right| \leq C \frac{\log n}{n} \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/64)(L - \ell)\hat{\nu}^{\ell-1}} (1 + \log L) + \frac{C'}{n^2}. $$

The constant $c_1$ is the absolute constant appearing in Lemma A.5.1.

Proof. The case of $\ell = L$ follows immediately from Lemma A.5.1 with an appropriate choice of $d \geq K''$ for $K'' > 0$ some absolute constant. Henceforth we assume $\ell \in [L-1]$. We Taylor expand (with Lagrange remainder) the smooth function $\varphi^{(L-\ell)}$ about the point $\varphi(\hat{\nu}^{\ell-1})$, obtaining for any $t \in [0, \pi]$

$$\varphi^{(L-\ell)}(t) = \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) + \varphi^{(L-\ell)}(\varphi(\hat{\nu}^{\ell-1}))(t - \varphi(\hat{\nu}^{\ell-1})) + \frac{\varphi^{(L-\ell)}(\xi)}{2} (t - \varphi(\hat{\nu}^{\ell-1}))^2,$$

where $\xi$ is some point of $[0, \pi]$ lying in between $t$ and $\varphi(\hat{\nu}^{\ell-1})$. In particular, putting $t = \hat{\nu}^{\ell}$, we obtain

$$\varphi^{(L-\ell)}(\hat{\nu}^{\ell}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) = \varphi^{(L-\ell)}(\varphi(\hat{\nu}^{\ell-1}))(\hat{\nu}^{\ell} - \varphi(\hat{\nu}^{\ell-1})) + \frac{\varphi^{(L-\ell)}(\xi)}{2} (\hat{\nu}^{\ell} - \varphi(\hat{\nu}^{\ell-1}))^2,$$

where $\xi$ is some point of $[0, \pi]$ lying in between $\hat{\nu}^{\ell}$ and $\varphi(\hat{\nu}^{\ell-1})$. By (A.110) and (A.113), we have that $\varphi^{(L-\ell)}(\hat{\nu}^{\ell}) \leq 0$, whence

$$\varphi^{(L-\ell)}(\hat{\nu}^{\ell}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \leq \varphi^{(L-\ell)}(\varphi(\hat{\nu}^{\ell-1}))(\hat{\nu}^{\ell} - \varphi(\hat{\nu}^{\ell-1})). \quad (A.168)$$

Using Lemma A.5.5 and an induction, we have that $\varphi^{(L-\ell)}$ is decreasing, and moreover by the concavity property we have $\varphi(\hat{\nu}^{\ell-1}) \geq \hat{\nu}^{\ell-1}/2$. An application of Lemmas A.5.1 and A.3.12 then
yields
\[
\mathbb{E}\left[\varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \mid \mathcal{F}^{\ell-1}\right] \leq \left(C\hat{\nu}^{\ell-1}\frac{\log n}{n} + C' n^{-c_1}\right) \frac{1}{1 + (c_0/4)(L-\ell)\hat{\nu}^{\ell-1}} \leq C \frac{\log n}{n} \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/4)(L-\ell)\hat{\nu}^{\ell-1}} + C' n^{-c_1},
\]
as long as \( d \geq K \) and \( n \geq K'd^4 \log^4 n \). In particular, we can choose \( d = \max\{K, 2/c_1\} \) to obtained the claimed error for the upper bound. Next, for the lower bound, we make use of the estimate
\[
\varphi^{(L-\ell)}(\nu) \geq -C \frac{1}{1 + (c_0/8)(L-\ell)\nu} \left(1 + \frac{1}{(c_0/8)\nu} \log \left(1 + \frac{(c_0/8)(L-\ell-1)\nu}{1}\right)\right),
\]
which follows from Lemma A.3.13 and \( \varphi^{(L-\ell)} \leq 0 \); by that lemma, we have that \( f \) is increasing. By Lemma A.5.3, as long as \( n \geq K' \log^4 n \), there is an event \( \mathcal{E} \) on which \( |\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})| \leq C\hat{\nu}^{\ell-1}\sqrt{\log n/n} + C' n^{-3} \) and which satisfies \( \mathbb{P}[\mathcal{E} \mid \mathcal{F}^{\ell-1}] \geq 1 - C'' n^{-3} \). In particular, on the event \( \mathcal{E} \) we have \( \hat{\nu}^\ell \geq \hat{\nu}^{\ell-1}/4 - C'/n^3 \) provided \( n \geq 16C^2 \log n \), and so on the event \( \mathcal{E} \) we have \( \xi \geq \min\{\varphi(\hat{\nu}^{\ell-1}), \hat{\nu}^{\ell-1}/4 - C'/n^3\} \geq \hat{\nu}^{\ell-1}/4 - C'/n^3 \). We can thus write
\[
\varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1})
\geq \varphi^{(L-\ell)}(\varphi(\hat{\nu}^{\ell-1})) \left(\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})\right) + \frac{f(\xi)}{2} \left(\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})\right)^2
\geq \varphi^{(L-\ell)}(\varphi(\hat{\nu}^{\ell-1})) \left(\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})\right) + (1_\mathcal{E} + 1_{\mathcal{E}^c}) \frac{f(\xi)}{2} \left(\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})\right)^2
\geq \varphi^{(L-\ell)}(\varphi(\hat{\nu}^{\ell-1})) \left(\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})\right) + 1_\mathcal{E} \frac{f\left(\frac{\hat{\nu}^{\ell-1}}{4} - \frac{C_1}{n^2}\right)}{2} \left(\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})\right)^2 - (2C''\pi^2 L) 1_{\mathcal{E}^c}
\geq \varphi^{(L-\ell)}(\varphi(\hat{\nu}^{\ell-1})) \left(\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})\right) + \frac{f\left(\frac{\hat{\nu}^{\ell-1}}{4} - \frac{C_1}{n^2}\right)}{2} \left(\hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})\right)^2 - (2C''\pi^2 L) 1_{\mathcal{E}^c}
\]
where the inequality in the third line follows from boundedness of the angles and the magnitude estimate on \( f \) in Lemma A.3.13, together with our estimate on \( \xi \) on \( \mathcal{E} \), and the inequality in the final line is a consequence of \( f \leq 0 \), which allows us to drop the indicator for \( \mathcal{E} \) and obtain a lower bound. Taking conditional expectations using the previous lower bound and applying \( \mathcal{F}^{\ell-1}-\)}
measurability of \( \hat{\nu}^{\ell-1} \) and boundedness of the angles together with our conditional measure bound on \( \mathcal{E}^c \), we obtain

\[
\mathbb{E}\left[ \varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \mid \mathcal{F}^{\ell-1} \right] \geq -C \frac{\log n}{n} \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/4)(L-\ell)\hat{\nu}^{\ell-1}} - \frac{C'}{n^2} \frac{C_5 L}{n^3} \\
+ \frac{f(\frac{\hat{\nu}^{\ell-1}}{4} - \frac{C_4}{n^3})}{2} \mathbb{E}\left[ \varphi^{\ell} - \varphi(\hat{\nu}^{\ell-1}) \mid \mathcal{F}^{\ell-1} \right],
\]

where we also apply the complementary bound obtained by our previous work following (A.168). Since the CL estimate in Lemma A.3.13 applies also to \( f \), and since \( f \leq 0 \), an application of Lemma A.5.4 with an appropriate choice of \( d \) and the choice \( n \geq K' \log^4 n \) then yields (with a larger absolute constant \( C' \))

\[
\mathbb{E}\left[ \varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \mid \mathcal{F}^{\ell-1} \right] \geq -C \frac{\log n}{n} \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/4)(L-\ell)\hat{\nu}^{\ell-1}} - \frac{C'}{n^2} \frac{C_6 L}{n^3} \\
+ \frac{C_7 \log n}{n} \left( \frac{\hat{\nu}^{\ell-1}}{4} \right)^2 f \left( \frac{\hat{\nu}^{\ell-1}}{4} - \frac{C_4}{n^3} \right). \]

If we choose \( n \geq (C_6/C')L \), we can simplify this last estimate to

\[
\mathbb{E}\left[ \varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \mid \mathcal{F}^{\ell-1} \right] \geq -C \frac{\log n}{n} \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/4)(L-\ell)\hat{\nu}^{\ell-1}} - \frac{2C'}{n^2} \\
+ \frac{C_7 \log n}{n} \left( \frac{\hat{\nu}^{\ell-1}}{4} \right)^2 f \left( \frac{\hat{\nu}^{\ell-1}}{4} - \frac{C_4}{n^3} \right). \]

To conclude, we divide our analysis into two cases: when \( \hat{\nu}^{\ell-1} \geq 8C_4/n^3 \), we have \( \hat{\nu}^{\ell-1}/4 - C_4 n^{-3} \geq \hat{\nu}^{\ell-1}/8 \), and so

\[
\left( \frac{\hat{\nu}^{\ell-1}}{4} \right)^2 f \left( \frac{\hat{\nu}^{\ell-1}}{4} - \frac{C_4}{n^3} \right) \geq \left( \frac{\hat{\nu}^{\ell-1}}{8} \right)^2 f \left( \frac{\hat{\nu}^{\ell-1}}{8} \right) = 64 \left( \frac{\hat{\nu}^{\ell-1}}{8} \right)^2 f \left( \frac{\hat{\nu}^{\ell-1}}{8} \right) \geq -\frac{8C_4 \log(L-\ell)}{1 + (c_0/64)(L-\ell)\hat{\nu}^{\ell-1}} \left( 1 + \frac{8 \log \left( 1 + \frac{c_0 \pi}{\hat{\nu}^{\ell-1}} \right)}{c_0 \pi} \right),
\]

248
where the last inequality follows from Lemma A.3.13. On the other hand, when \( \hat{\nu}^{\ell-1} \leq 8C_4/n^3 \), the CL estimate in Lemma A.3.13 implies

\[
\left( \hat{\nu}^{\ell-1} \right)^2 \frac{f \left( \frac{\hat{\nu}^{\ell-1} - C_4}{4 \ n^3} \right)}{\frac{64CC_4^2 L}{n^6}} \geq - \frac{64CC_4^2 L}{n^3} \geq - \frac{2C'}{n^2},
\]

where the last estimate holds when \( n \geq (32CC_4^2/C')L \). Adding these two estimates together, we obtain one that is valid regardless of the value of \( \hat{\nu}^{\ell-1} \), and choosing \( n \geq C_7 \log n \) to combine the residuals, we obtain (after worst-case adjusting the constants)

\[
E \left[ \varphi^{(L-\ell)}(\hat{\nu}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \mid \mathcal{F}^{\ell-1} \right] \leq C \frac{\log n}{n} \frac{\hat{\nu}^{\ell-1} (1 + \log(L - \ell))}{1 + (c_0/64)(L - \ell)\hat{\nu}^{\ell-1}}.
\]

Combining with our previous work, we obtain

\[
\left| E \left[ \varphi^{(L-\ell)}(\hat{\nu}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \mid \mathcal{F}^{\ell-1} \right] \right| \leq C \frac{\log n}{n} \frac{\hat{\nu}^{\ell-1} (1 + \log L)}{1 + (c_0/64)(L - \ell)\hat{\nu}^{\ell-1}} + C' \frac{1}{n^2},
\]

after worst-casing constants.

**Lemma A.4.15.** There are absolute constants \( c_1, C, C', C'', C''' > 0 \) and \( K, K' > 0 \) such that for any \( d \geq K \), if \( n \geq K'd^4 \log^4 n \), then for every \( L \in \mathbb{N} \) and every \( \ell \in [L] \) one has

\[
E \left[ \varphi^{(L-\ell)}(\hat{\nu}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \mid \mathcal{F}^{\ell-1} \right] \leq \sqrt{\frac{d \log n}{n}} \frac{2C\hat{\nu}^{\ell-1}}{1 + (c_0/8)(L - \ell)\hat{\nu}^{\ell-1}} + 2C'n^{-c_1d/2} \right] \mid \mathcal{F}^{\ell-1} \right| \leq 1 - C''n^{-c_1d},
\]

and

\[
E \left[ \left( \varphi^{(L-\ell)}(\hat{\nu}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \right)^2 \mid \mathcal{F}^{\ell-1} \right] \leq 4C'^2 \frac{d \log n}{n} \left( \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/8)(L - \ell)\hat{\nu}^{\ell-1}} \right)^2 + C''n^{-c_1d/2}.
\]

The constant \( c_1 \) is the absolute constant appearing in Lemma A.5.1.
Proof. We will fix the meaning of the absolute constants $C, C', C'' > 0$ throughout the proof below.

By Lemma A.5.3, we have if $d \geq K$ and $n \geq K'd^4 \log^4 n$ that for every $\ell \in [L]$

$$
P \left[ \left| \hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1}) \right| \leq C\hat{\nu}^{\ell-1} \sqrt{\frac{d \log n}{n}} + C'n^{-c_1d} \right] \geq 1 - C'' n^{-c_1d}. \quad (A.169)$$

By Lemma A.3.12, we have the estimate

$$
\left| \varphi^{(\ell)}(t) \right| \leq \frac{1}{1 + (c_0/2)\ell t},
$$
valid for any $\ell \in \mathbb{N}_0$. Writing $\Xi^\ell = \hat{\nu}^\ell - \varphi(\hat{\nu}^{\ell-1})$ so that $\hat{\nu}^\ell = \varphi(\hat{\nu}^{\ell-1}) + \Xi^\ell$, we have that $(\Xi^\ell)$ is adapted to $(F^\ell)$, and by the fundamental theorem of calculus

$$
\left| \varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \right| = \left| \int_{\varphi(\hat{\nu}^{\ell-1}) + \Xi^\ell}^{\varphi(\hat{\nu}^{\ell-1}) + (c_0/2)(L-\ell)t} \frac{dr}{1 + (c_0/2)(L-\ell)t} \right|.
$$

The integrand is nonnegative, so by Jensen’s inequality we have

$$
\left( \varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \right)^2 \leq |\Xi^\ell| \int_{\varphi(\hat{\nu}^{\ell-1}) + \Xi^\ell}^{\varphi(\hat{\nu}^{\ell-1}) + (c_0/2)(L-\ell)t} \frac{dr}{1 + (c_0/2)(L-\ell)t} \frac{dr}{(1 + (c_0/2)(L-\ell)t)^2},
$$
and an integration then yields

$$
\left( \varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \right)^2 \leq \frac{|\Xi^\ell|^2}{\left[ 1 + (c_0/2)(L-\ell)\varphi(\hat{\nu}^{\ell-1}) \right] \left[ 1 + (c_0/2)(L-\ell)(\varphi(\hat{\nu}^{\ell-1}) + \Xi^\ell) \right]}.
$$

Choosing $d \geq 1/c_1$, we can guarantee that whenever $\nu \geq \frac{C'}{C} n^{-c_1d/2}$, one has

$$
CV \sqrt{\frac{d \log n}{n}} + C' n^{-c_1d} \leq 2CV \sqrt{\frac{d \log n}{n}}, \quad (A.171)
$$
and choosing \( n \geq 64C^2d \log n \), we can guarantee that

\[
\frac{\nu}{2} - 2C\sqrt{\frac{d \log n}{n}} \geq \frac{\nu}{4}.
\]

(A.172)

In particular, the last condition guarantees \( 2C\sqrt{d \log n/n} \leq 1/4 \). By concavity of \( \varphi \), shown in Lemma A.5.5, we have \( \varphi(\hat{\nu}^{\ell-1}) \geq \hat{\nu}^{\ell-1}/2 \), and using (A.169) to obtain

\[
P \left[ |\Xi| \leq C\hat{\nu}^{\ell-1} \sqrt{\frac{d \log n}{n}} + C'n^{-c_1d} \right| \mathcal{F}^{\ell-1} \right] \geq 1 - C''n^{-c_1d},
\]

we have by (A.171) and (A.172) as well as the concavity lower bound on \( \varphi \)

\[
P[\varphi(\hat{\nu}^{\ell-1}) + \Xi \geq \hat{\nu}^{\ell-1}/4 \ | \mathcal{F}^{\ell-1}] \geq 1 - C''n^{-c_1d}
\]

as long as \( \hat{\nu}^{\ell-1} \geq (C'/C)n^{-c_1d/2} \). In particular, plugging these bounds into (A.170) and taking square roots, we obtain by a union bound

\[
P \left[ \left| \varphi^{(L-\ell)}(\hat{\nu}^{\ell}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \right| \leq 2C\sqrt{\frac{d \log n}{n}} \hat{\nu}^{\ell-1} \right| \mathcal{F}^{\ell-1} \right] \geq 1 - 2C''n^{-cd}
\]

whenever \( \hat{\nu}^{\ell-1} \geq (C'/C)n^{-c_1d/2} \). Meanwhile, when \( \hat{\nu}^{\ell-1} \leq (C'/C)n^{-c_1d/2} \), if we choose \( n \geq d \log n \) we have

\[
C\hat{\nu}^{\ell-1} \sqrt{\frac{d \log n}{n}} + C'n^{-c_1d} \leq 2C'n^{-c_1d/2},
\]

and we can use the 1-Lipschitz property of \( \varphi^{(L-\ell)} \), which follows from Lemma A.5.5, to obtain using (A.169)

\[
P \left[ \left| \varphi^{(L-\ell)}(\hat{\nu}^{\ell}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \right| \leq 2C'n^{-c_1d/2} \right| \mathcal{F}^{\ell-1} \right] \geq P \left[ \varphi^{(L-\ell)}(\hat{\nu}^{\ell}) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \leq C\hat{\nu}^{\ell-1} \sqrt{\frac{d \log n}{n}} + C'n^{-c_1d} \right| \mathcal{F}^{\ell-1} \right]
\]

251
\[ \begin{align*}
&\geq P\left[ |\hat{\varphi}^\ell - \varphi(\hat{\varphi}^{\ell-1})| \leq C\hat{\varphi}^{\ell-1}\sqrt{\frac{d \log n}{n}} + C'n^{-c_1d} \right| \mathcal{F}^{\ell-1} \\
&\geq 1 - C''n^{-c_1d}.
\end{align*} \]

Because \( |\varphi(L^{\ell})\hat{\varphi}^\ell - \varphi(L^{\ell+1})\hat{\varphi}^{\ell-1}| \geq 0 \), we can then obtain using a union bound
\[
P\left[ |\varphi(L^{\ell})\hat{\varphi}^\ell - \varphi(L^{\ell+1})\hat{\varphi}^{\ell-1}| \leq \sqrt{\frac{d \log n}{n}} \frac{2C\hat{\varphi}^{\ell-1}}{1 + (c_0/8)(L - \ell)\hat{\varphi}^{\ell-1}} + \frac{2C'}{n^{c_1d/2}} \right| \mathcal{F}^{\ell-1} \\
\geq 1 - 3C''n^{-c_1d},
\]
which holds regardless of the value of \( \hat{\varphi}^{\ell-1} \). We can then obtain the second bound using this one, via a partition of the expectation: let
\[
\mathcal{E} = \left\{ |\varphi(L^{\ell})\hat{\varphi}^\ell - \varphi(L^{\ell+1})\hat{\varphi}^{\ell-1}| \leq \sqrt{\frac{d \log n}{n}} \frac{2C\hat{\varphi}^{\ell-1}}{1 + (c_0/8)(L - \ell)\hat{\varphi}^{\ell-1}} + \frac{2C'}{n^{c_1d/2}} \right\},
\]
so that \( \mathcal{E} \in \mathcal{F}^{\ell} \), and \( P[\mathcal{E} | \mathcal{F}^{\ell-1}] \geq 1 - 3C''n^{-c_1d} \) by our work above. Then we have
\[
\begin{align*}
&\mathbb{E}\left[ \left( \varphi(L^{\ell})\hat{\varphi}^\ell - \varphi(L^{\ell+1})\hat{\varphi}^{\ell-1} \right)^2 \right| \mathcal{F}^{\ell-1} \\
&\quad \leq \mathbb{E}\left[ \left( \varphi(L^{\ell})\hat{\varphi}^\ell - \varphi(L^{\ell+1})\hat{\varphi}^{\ell-1} \right)^2 \right| \mathcal{F}^{\ell-1} + \pi^2 \mathbb{E}\left[ 1_{\mathcal{E}} \right| \mathcal{F}^{\ell-1} \\
&\quad \leq \mathbb{E}\left[ \left( 2C\sqrt{\frac{d \log n}{n}} \frac{\hat{\varphi}^{\ell-1}}{1 + (c_0/8)(L - \ell)\hat{\varphi}^{\ell-1}} + \frac{2C'}{n^{c_1d/2}} \right)^2 \right| \mathcal{F}^{\ell-1} + C''n^{-c_1d} \\
&\quad \leq \left( 2C\sqrt{\frac{d \log n}{n}} \frac{\hat{\varphi}^{\ell-1}}{1 + (c_0/8)(L - \ell)\hat{\varphi}^{\ell-1}} + \frac{2C'}{n^{c_1d/2}} \right)^2 + C''n^{-c_1d}
\end{align*}
\]

where the first inequality uses the triangle inequality to obtain \( \left( \varphi(L^{\ell})\hat{\varphi}^\ell - \varphi(L^{\ell+1})\hat{\varphi}^{\ell-1} \right)^2 \leq \pi^2 \); the second inequality applies the definition of \( \mathcal{E} \), uses nonnegativity to drop the indicator in the first term, and applies the conditional measure bound on \( \mathcal{E}^c \); and the final inequality integrates. Using the fact that our previous choices of large \( n \) imply \( 2C\sqrt{d \log n/n} \leq 1/4 \), and that \( |\hat{\varphi}^{\ell-1}/1 + (c_0/8)(L - \ell)\hat{\varphi}^{\ell-1}| \leq \pi \), we can distribute the square in this final bound and worst-case constants to
obtain

\[
\mathbb{E} \left[ \left( \varphi^{(L-\ell)}(\hat{\nu}^\ell) - \varphi^{(L-\ell+1)}(\hat{\nu}^{\ell-1}) \right)^2 \mid \mathcal{F}^{\ell-1} \right]
\leq 4C^2 \frac{d \log n}{n} \left( \frac{\hat{\nu}^{\ell-1}}{1 + (c_0/8)(L-\ell)\hat{\nu}^{\ell-1}} \right)^2 + C''''n^{-c_1d/2},
\]

as claimed. \hfill \Box

**Lemma A.4.16.** Let \( X_1, \ldots, X_L \) be independent chi-squared random variables, having respectively \( d_1, \ldots, d_L \) degrees of freedom. Write \( d_{\min} = \min_{i \in L} d_i \) and let \( \xi_i = \frac{1}{d_i} X_i \). Then there are absolute constants \( c, C > 0 \) and an absolute constant \( 0 < K \leq \frac{1}{4} \) such that for any \( 0 < t \leq K \), one has

\[
P \left[ -1 + \prod_{i=1}^{L} \xi_i > t \right] \leq CLe^{-cd_{\min}t^2/L}.
\]

In particular, there are absolute constants \( C', C'' > 0 \) and an absolute constant \( K' > 0 \) such that for any \( d > 0 \), if \( d_{\min} \geq K'dL \) then one has

\[
P \left[ -1 + \prod_{i=1}^{L} \xi_i > C' \sqrt{\frac{dL}{d_{\min}}} \right] \leq C''Le^{-d}.
\]

**Proof.** For any \( t \geq 0 \), we have by the AM-GM inequality

\[
P \left[ \prod_{i=1}^{L} \xi_i > 1 + t \right] = P \left[ \left( \prod_{i=1}^{L} \xi_i \right)^{1/L} > (1 + t)^{1/L} \right] \leq P \left[ \frac{1}{L} \sum_{i=1}^{L} \xi_i > (1 + t)^{1/L} \right].
\]

By convexity of the exponential, we have \( (1 + t)^{1/L} \geq 1 + \frac{1}{L} \log(1 + t) \), and by concavity of the logarithm we have \( \log(1 + t) \geq t \log 2 \) if \( t \leq 1 \). This implies

\[
P \left[ \prod_{i=1}^{L} \xi_i > 1 + t \right] \leq P \left[ -L + \sum_{i=1}^{L} \xi_i > Kt \right],
\]

where \( K = \log(2) \). Decomposing each \( X_i \) into a sum of \( d_i \) i.i.d. squared gaussians and applying
Lemma A.6.2, we obtain

\[
\mathbb{P} \left[ \left| -L + \sum_{i=1}^{L} \xi_i \right| > t \right] \leq 2 \exp \left( -c \min \left\{ \frac{t^2}{\sum_{i=1}^{L} \frac{C}{d_i}}, \frac{t}{C \max_i \frac{1}{d_i}} \right\} \right)
\leq 2 \exp \left( -c' \min \{ t^2/CL, t \} \right)
\leq 2 \exp \left( -c'' \frac{d_{\min} t^2}{L} \right),
\]

where the last inequality holds provided \( t \leq CL \), where \( C > 0 \) is an absolute constant. Thus, as long as \( t \leq CL/K \), we have suitable control of the upper tail of the product \( \prod_i \xi_i \). For the lower tail, writing \( \log(0) = -\infty \), we have for any \( 0 \leq t < 1 \)

\[
\mathbb{P} \left[ \prod_{i=1}^{L} \xi_i < 1 - t \right] = \mathbb{P} \left[ \sum_{i=1}^{L} \log \xi_i < \log(1 - t) \right] \leq \mathbb{P} \left[ \sum_{i=1}^{L} \log \xi_i < -t \right],
\]

where the inequality uses concavity of \( t \mapsto \log(1 - t) \). By Lemma A.6.2, we have for each \( i \in [L] \) and every \( 0 \leq t \leq C \) (where \( C > 0 \) is an absolute constant)

\[
\mathbb{P}[|\xi_i - 1| < t] \leq 2e^{-cd_t^2},
\]

so that by a union bound and for \( t \leq C \sqrt{L} \), we have with probability at least \( 1 - 2Le^{-cd_{\min} t^2/L} \) that \( 1 - t/\sqrt{L} \leq \xi_i \leq 1 + t/\sqrt{L} \) for every \( i \in [L] \). Meanwhile, Taylor expansion of the smooth function \( x \mapsto \log x \) in a neighborhood of 1 gives

\[
\log x = (x - 1) - \frac{1}{2k^2} (x - 1)^2,
\]

where \( k \) is a number lying between 1 and \( x \). In particular, if \( x \geq \frac{1}{2} \) we have \( \log x \geq (x - 1) - 2(x - 1)^2 \),
whence for $t \leq \min\{C\sqrt{L}, \frac{1}{2}\}$

$$
\mathbb{P}\left[ \prod_{i=1}^{L} \xi_i < 1 - t \right] \leq 2L e^{-cd_{\text{min}} t^2 / L} + \mathbb{P}\left[ -L + \sum_{i=1}^{L} \xi_i < -t + 2t^2 \right] \\
\leq 2L e^{-cd_{\text{min}} t^2 / L} + \mathbb{P}\left[ -L + \sum_{i=1}^{L} \xi_i < -\frac{t}{2} \right],
$$

where the final inequality requires in addition $t \leq \frac{1}{4}$. An application of (A.173) then yields the claimed lower tail provided $t \leq CL$, which establishes the first claim. For the second claim, we consider the choice $t = \sqrt{dL/cd_{\text{min}}}$, for which we have $t \leq K$ whenever $d_{\text{min}} \geq dL/cK^2$, and $cd_{\text{min}} t^2 / L = d$. \hfill \Box

**Lemma A.4.17.** Let $X_1, \ldots, X_L$ be independent Binom$(n, \frac{1}{2})$ random variables, and write $\xi_i = \frac{2}{n} X_i$. Then for any $0 < t \leq \frac{1}{4}$, one has

$$
\mathbb{P}\left[ -1 + \prod_{i=1}^{L} \xi_i > t \right] \leq 4L e^{-nt^2/8L}.
$$

In particular, for any $d > 0$, if $n \geq 128dL$ then one has

$$
\mathbb{P}\left[ -1 + \prod_{i=1}^{L} \xi_i > 4\sqrt{\frac{dL}{n}} \right] \leq 4L e^{-d}.
$$

**Proof.** The proof is very similar to that of Lemma A.4.16. For any $t \geq 0$, we have by the AM-GM inequality

$$
\mathbb{P}\left[ \prod_{i=1}^{L} \xi_i > 1 + t \right] = \mathbb{P}\left[ \left( \prod_{i=1}^{L} \xi_i \right)^{1/L} > (1 + t)^{1/L} \right] \leq \mathbb{P}\left[ \frac{1}{L} \sum_{i=1}^{L} \xi_i > (1 + t)^{1/L} \right].
$$

By convexity of the exponential, we have $(1 + t)^{1/L} \geq 1 + \frac{1}{L} \log(1 + t)$, and by concavity of the
logarithm we have $\log(1 + t) \geq t \log 2$ if $t \leq 1$. This implies

$$
\mathbb{P}\left[ \prod_{i=1}^{L} \xi_i > 1 + t \right] \leq \mathbb{P}\left[ -L + \sum_{i=1}^{L} \xi_i > Kt, \right],
$$

where $K = \log(2)$. Decomposing each $X_i$ into a sum of $n$ i.i.d. $\text{Bern}(\frac{1}{2})$ random variables and applying Lemma A.6.1 twice, we obtain

$$
\mathbb{P}\left[ \left| -L + \sum_{i=1}^{L} \xi_i \right| > t \right] \leq 2e^{-nt^2/2L}. \quad (A.174)
$$

This gives suitable control of the upper tail of the product $\prod_i \xi_i$. For the lower tail, writing $\log(0) = -\infty$, we have for any $0 \leq t < 1$

$$
\mathbb{P}\left[ \prod_{i=1}^{L} \xi_i < 1 - t \right] = \mathbb{P}\left[ \sum_{i=1}^{L} \log \xi_i < \log(1 - t) \right] \leq \mathbb{P}\left[ \sum_{i=1}^{L} \log \xi_i < -t \right],
$$

where the inequality uses concavity of $t \mapsto \log(1 - t)$. By Lemma A.6.1, we have for each $i \in [L]$

$$
\mathbb{P}[|\xi_i - 1| < t] \leq 2e^{-nt^2/2},
$$

so that by a union bound, we have that $1 - t/\sqrt{L} \leq \xi_i \leq 1 + t/\sqrt{L}$ for every $i \in [L]$ with probability at least $1 - 2Le^{-nt^2/2L}$. Meanwhile, Taylor expansion of the smooth function $x \mapsto \log x$ in a neighborhood of 1 gives

$$
\log x = (x - 1) - \frac{1}{2k^2} (x - 1)^2,
$$

where $k$ is a number lying between 1 and $x$. In particular, if $x \geq \frac{1}{2}$ we have $\log x \geq (x - 1) - 2(x - 1)^2$, whence for $t \leq 1/2$

\[
\mathbb{P}\left[ \prod_{i=1}^{L} \xi_i < 1 - t \right] \leq 2Le^{-nt^2/2L} + \mathbb{P}\left[ -L + \sum_{i=1}^{L} \xi_i < -t + 2t^2 \right] \\
\leq 2Le^{-nt^2/2L} + \mathbb{P}\left[ -L + \sum_{i=1}^{L} \xi_i < -\frac{t}{2} \right],
\]
where the final inequality requires in addition $t \leq 1/4$. An application of (A.174) then yields the claimed lower tail, which establishes the first claim. For the second claim, we consider the choice $t = \sqrt{8dL/n}$, for which we have $t \leq 1/4$ whenever $n \geq 128dL$, and $nt^2/8L = d$. □

A.5 Sharp Bounds on the One-Step Angle Process

In this section, we characterize the process by which angles between features for different pairs of points evolve as they are propagated across one layer of the zero-time network. This section is self-contained, and as such it will occasionally overload notation used elsewhere in the document for different local purposes. In particular, we will use the notation $\sigma(x) = [x]_+$ for the ReLU in this section (and only in this section), and $\dot{\sigma}(g) = 1_{g>0}$ for its weak derivative.

A.5.1 Definitions and Preliminaries

Let $n \in \mathbb{N}$, with $n \geq 2$. Let $g_1$ and $g_2$ be i.i.d. $\mathcal{N}(0, (2/n)I)$ random vectors; we use $\mu$ to denote the joint law of these random variables. We write $G \in \mathbb{R}^{n \times 2}$ for the matrix with first column $g_1$ and second column $g_2$, and $g_1, \ldots, g_n$ for the $n$ rows of $G$. If $S \subset [n]$ is nonempty and $A \in \mathbb{R}^{n \times m}$, we write $A_S \in \mathbb{R}^{|S| \times m}$ to denote the submatrix of $A$ consisting of the rows indexed by $S$ in increasing index order. In such situations $S^c$ will always denote the complement relative to $[n]$.

For $0 \leq \nu \leq 2\pi$, define random variables

$$v_{\nu}(g_1, g_2) = \sigma(g_1 \cos \nu + g_2 \sin \nu),$$

and

$$\dot{v}_{\nu}(g_1, g_2) = \dot{\sigma}(g_1 \cos \nu + g_2 \sin \nu) \odot (g_2 \cos \nu - g_1 \sin \nu).$$

Because $\dot{v}_{\nu}$ separates over coordinates of its arguments and has each of its coordinates the product of a nondecreasing function and a continuous function, it is Borel measurable. A key property that we will use throughout this section is that the joint distribution of $(g_1, g_2)$ is rotationally invariant;
in particular, it is invariant to rotations of the type

$$G \mapsto G \begin{bmatrix} \cos \nu & \sin \nu \\ \sin \nu & -\cos \nu \end{bmatrix},$$

where $\nu \in [0, 2\pi]$. Since we can write

$$v_\nu = \sigma \left( G \begin{bmatrix} \cos \nu \\ \sin \nu \end{bmatrix} \right), \quad \dot{v}_\nu = \sigma \left( G \begin{bmatrix} \cos \nu \\ \sin \nu \end{bmatrix} \right) \odot \left( G \begin{bmatrix} -\sin \nu \\ \cos \nu \end{bmatrix} \right),$$

where all of the $\mathbb{R}^2$ vectors appearing above are elements of $\mathbb{S}^1$, it follows by applying rotational invariance and the specific rotation given above that

$$(v_\nu, \dot{v}_\nu) \overset{d}{=} (v_0, -\dot{v}_0).$$

This equivalence is useful for evaluating expectations and differentiating with respect to $\nu$.

If $0 < c \leq 0.5$ and $m \in \mathbb{N}_0$ with $m < n$, define an event

$$\mathcal{E}_{c,m} = \bigcap_{S \subset [n]} \bigcap_{\nu \in [0,2\pi]} \{ (g_1, g_2) \mid c \leq \|I_S v_\nu(g_1, g_2)\|_2 \leq c^{-1} \}.$$

For each $c, m$, the set $\mathcal{E}_{c,m}$ is closed, since $\|Av_\nu\|$ is a continuous function of $(g_1, g_2) \in \mathcal{E}_m$ for any linear map $A$. We further define

$$\mathcal{E}_{0,m} = \bigcup_{k \in \mathbb{N}} \mathcal{E}_{1/(2k),m},$$

so that

$$\mathcal{E}_{0,m} = \bigcap_{S \subset [n]} \bigcap_{\nu \in [0,2\pi]} \{ (g_1, g_2) \mid 0 < \|I_S v_\nu(g_1, g_2)\|_2 \},$$

and $\mathcal{E}_{0,m}$ is Borel measurable. If $c$ is omitted, we take the constant $c$ in the definition to be 0.5. On
we guarantee that \( \|v_\nu\|_0 \geq m \) uniformly on \([0, \pi]\). Define a function \( X_\nu \) by

\[
X_\nu = E_1 \left( \frac{v_0}{\|v_0\|_2}, \frac{v_\nu}{\|v_\nu\|_2} \right).
\]

On \( E_1 \), we guarantee that \( v_\nu \neq 0 \) for every \( \nu \), so \( X_\nu \) is well defined; because \( E_1 \) is Borel measurable, we have that \( X_\nu \) is Borel measurable, and moreover \( |X_\nu| \leq 1 \), so \( X_\nu \in L^p_\mu \) for every \( p \geq 1 \). Finally, define for \( 0 \leq \nu \leq \pi \)

\[
\bar{\varphi}(\nu) = \mathbb{E}_{g_1, g_2} \left[ \cos^{-1} X_\nu \right], \quad \varphi(\nu) = \cos^{-1} \mathbb{E}_{g_1, g_2} [\langle v_0, v_\nu \rangle].
\]

A.5.2 Main Results

**Lemma A.5.1.** There exist absolute constants \( c, C, C' > 0 \) and absolute constants \( K, K' > 0 \) such that if \( d \geq K \) and \( n \geq K' d^4 \log^4 n \), then one has

\[
|\bar{\varphi}(\nu) - \varphi(\nu)| \leq C \frac{\log n}{n} + C' n^{-cd}
\]

**Proof.** Using the triangle inequality, we can write

\[
|\bar{\varphi}(\nu) - \varphi(\nu)| \leq |\bar{\varphi}(\nu) - \cos^{-1} \mathbb{E}[X_\nu]| + \left| \cos^{-1} \mathbb{E}[X_\nu] - \varphi(\nu) \right|.
\]

Choose \( n \) sufficiently large to satisfy the hypotheses of Lemmas A.5.6 and A.5.7; applying these lemmas to bound the first and second terms, we conclude the claimed result (after choosing \( n \) larger than an absolute constant multiple of \( d \log n \) so that the \( n^{-cd} \) error dominates the \( e^{-c'n} \) error).

**Lemma A.5.2.** One has

\[
\varphi(\nu) = \cos^{-1} \left( \left( 1 - \frac{\nu}{\pi} \right) \cos \nu + \frac{\sin \nu}{\pi} \right).
\]

**Proof.** See [226].

**Lemma A.5.3.** There exist absolute constants \( c, C, C', C'' > 0 \) and absolute constants \( K, K' > 0 \)
such that if \( d \geq K \) and \( n \geq K'd^4 \log^4 n \), then one has with probability at least \( 1 - C'' n^{-cd} \)

\[
\left| \cos^{-1} X_v - \varphi(v) \right| \leq C \sqrt{n} \frac{d \log n}{n} + C'n^{-cd}.
\]

The constant \( c \) is the same as the constant appearing in Lemma A.5.1.

**Proof.** Under our hypothesis, the second result in Lemma A.5.6 together with Lemma A.5.1 and the triangle inequality imply the claimed result (after worst-case multiplicative constants). \( \square \)

**Lemma A.5.4.** There exist absolute constants \( c, C, C' > 0 \) and absolute constants \( K, K' > 0 \) such that if \( d \geq K \) and \( n \geq K'd^4 \log^4 n \), then one has

\[
\mathbb{E} \left[ \left( \cos^{-1} X_v - \varphi(v) \right)^2 \right] \leq C \nu^2 \frac{d \log n}{n} + C'n^{-cd}.
\]

The constant \( c \) is the same as the constant appearing in Lemma A.5.1.

**Proof.** Under our hypotheses, Lemma A.5.3 is applicable; we let \( \mathcal{E} \) denote the event corresponding to the bound in this lemma. By boundedness of \( \cos^{-1} \), nonnegativity of \( X_v \), and \( \varphi \leq \pi/2 \) from Lemma A.5.2, we have \( \| \cos^{-1} X_v - \varphi(v) \|_{L^2} \leq \pi \). Thus

\[
\begin{align*}
\mathbb{E} \left[ \left( \cos^{-1} X_v - \varphi(v) \right)^2 \right] &\leq \mathbb{E} \left[ 1_{\mathcal{E}} \left( \cos^{-1} X_v - \varphi(v) \right)^2 \right] + C'' \pi^2 n^{-cd} \\
&\leq \left( C \sqrt{n} \frac{d \log n}{n} + C'n^{-cd} \right)^2 + C'' \pi^2 n^{-cd} \\
&\leq C^2 \nu^2 \frac{d \log n}{n} + C'' n^{-cd},
\end{align*}
\]

as claimed. \( \square \)

**Lemma A.5.5.** One has

1. \( \varphi \in C^\infty(0, \pi) \), and \( \dot{\varphi} \) and \( \ddot{\varphi} \) extend to continuous functions on \([0, \pi]\);

2. \( \varphi(0) = 0 \) and \( \varphi(\pi) = \pi/2 \); \( \dot{\varphi}(0) = 1 \), \( \dot{\varphi}(0) = -2/(3\pi) \), and \( \ddot{\varphi}(0) = -1/(3\pi^2) \); and \( \dddot{\varphi}(\pi) = \dddot{\varphi}(\pi) = 0 \);
3. \( \varphi \) is concave and strictly increasing on \([0, \pi]\) (strictly concave in the interior);

4. \( \varphi < -c < 0 \) for an absolute constant \( c > 0 \) on \([0, \pi/2]\);

5. \( 0 < \varphi < 1 \) and \( 0 > \varphi \geq -C \) on \((0, \pi)\) for some absolute constant \( C > 0 \);

6. \( \nu(1 - C_1 \nu) \leq \varphi(\nu) \leq \nu(1 - c_1 \nu) \) on \([0, \pi]\) for some absolute constants \( C_1, c_1 > 0 \).

**Proof.** Deferred to Section A.5.4. \( \square \)

### A.5.3 Supporting Results

#### Core Supporting Results

**Lemma A.5.6.** There exist constants \( c, C, C', C'', C'''> 0 \) and an absolute constant \( K > 0 \) such that for any \( d \geq 1 \), if \( n \) and \( d \) satisfy the hypotheses of Lemmas A.5.9 and A.5.10 and moreover \( n \geq K d \log n \), then one has

\[
\left| \mathbb{E}_{g_1, g_2} \left[ \cos^{-1} X_{\nu} \right] - \cos^{-1} \mathbb{E} \left[ X_{\nu} \right] \right| \leq C \nu \frac{\log n}{n} + C' n^{-cd},
\]

and with probability at least \( 1 - C'' n^{-cd} \), one has

\[
\left| \cos^{-1} X_{\nu} - \mathbb{E} \left[ \cos^{-1} X_{\nu} \right] \right| \leq C''' \nu \sqrt{\frac{d \log n}{n}} + C''' n^{-cd}.
\]

**Proof.** Fix \( \nu \in [0, \pi] \). The function \( \cos^{-1} \) is smooth on \((-\delta, 1)\) if \( 0 < \delta < 1 \), and Taylor expansion with Lagrange remainder on this domain about the point \( \mathbb{E}[X_{\nu}] \) (by Lemma A.5.23, we have \( 0 \leq \mathbb{E}[X_{\nu}] < 1 \) if \( \nu > 0 \); we will handle \( \nu = 0 \) separately below) gives

\[
\cos^{-1}(x) = \cos^{-1} \mathbb{E}[X_{\nu}] - \frac{1}{\sqrt{1 - \mathbb{E}[X_{\nu}]}^2} (x - \mathbb{E}[X_{\nu}]) - \frac{\xi}{2 (1 - \xi^2)^{3/2}} (x - \mathbb{E}[X_{\nu}])^2,
\]

where \( \xi \) lies between \( x \) and \( \mathbb{E}[X_{\nu}] \). Using the fact that \( X_{\nu} \neq 1 \) almost surely if \( \nu > 0 \), which is
established in Lemma A.5.23, we plug in \( x = X_\nu \) to get

\[
\cos^{-1} \mathbb{E}[X_\nu] - \cos^{-1}(X_\nu) = \frac{1}{\sqrt{1 - \mathbb{E}[X_\nu]^2}} (x - \mathbb{E}[X_\nu]) + \frac{\xi(X_\nu)}{2 \sqrt{1 - \xi(X_\nu)^2}} (X_\nu - \mathbb{E}[X_\nu])^2,
\]

(A.175)

where we now express \( \xi \) as a function of \( X_\nu \). From Jensen’s inequality it is clear

\[
\mathbb{E}[\cos^{-1} X_\nu] \leq \cos^{-1} \mathbb{E}[X_\nu],
\]

(A.176)

so all that remains is to obtain a matching upper bound for the righthand side of (A.175). We will make use of the following facts, proved in subsequent sections: there are absolute constants \( C_i > 0, i \in [6] \), and \( c_i > 0, i \in [5] \), such that

1. \( \mathbb{E}[X_\nu] \leq 1 - c_5 \nu^2 + C_1 e^{-c_1 n}. \) (Lemma A.5.8)

2. For each \( \nu \), \( \text{Var}[X_\nu] \leq c_5 \nu^4 \log n/n + C_2 e^{-c_2 n}. \) (Lemma A.5.9)

3. With probability at least \( 1 - C_3 n^{-c_3 d} \), one has \( |X_\nu - \mathbb{E}[X_\nu]| \leq C_6 \nu^2 \sqrt{d \log n/n} + C_4 e^{-c_4 n}. \) (Lemma A.5.10)

Let \( \mathcal{E} \) denote the event on which property 3 holds. Combining properties 1 and 3, we obtain with probability at least \( 1 - C_3 n^{-c_3 d} \)

\[
X_\nu \leq 1 - (c_5/2) \nu^2 + C_1 e^{-c_1 n} + C_4 e^{-c_4 n},
\]

provided \( n \) is chosen larger than an absolute constant multiple of \( d \log n \). Thus, defining

\[
\nu_0 = \frac{4}{c_5} (C_1 e^{-c_1 n} + C_4 e^{-c_4 n}),
\]

we obtain for \( \nu \geq \nu_0 \)

\[
\mathbb{E}[X_\nu] \leq 1 - \frac{c_5}{4} \nu^2, \quad X_\nu \leq 1 - \frac{c_5}{4} \nu^2,
\]

(A.177)
with the second bound holding with probability at least \( 1 - C_3 n^{-c_3 d} \). Considering first \( 0 \leq \nu \leq \nu_0 \), we obtain using the triangle inequality, Lemma A.5.20 and property 3

\[
|\cos^{-1} E[X_\nu] - E[\cos^{-1}(X_\nu)]| \leq E[1_{E}|\cos^{-1} E[X_\nu] - \cos^{-1}(X_\nu)|] \\
+ E[1_{E}|\cos^{-1} E[X_\nu] - \cos^{-1}(X_\nu)|] \\
\leq E[1_{E}\sqrt{|X_\nu - E[X_\nu]|}] + E[1_{E} \pi/2] \\
\leq C e^{-\epsilon n} + C' n^{-c' d}, \tag{A.178}
\]

with the final inequality following from the triangle inequality for the \( \ell^2 \) norm and the fact that \( \nu \leq \nu_0 \). Meanwhile, if \( \nu \geq \nu_0 \), we have by (A.177)

\[
0 \leq \xi(X_\nu) \leq \max\{X_\nu, E[X_\nu]\} \leq 1 - \frac{c_5}{4} \nu^2
\]

with probability at least \( 1 - C_3 n^{-c_3 d} \). Using \( 1 - \nu^2 = (1 + \nu)(1 - \nu) \) and \( E[X_\nu] \geq 0, \xi(X_\nu) \geq 0 \), we have under this condition on \( \nu \)

\[
\frac{1}{\sqrt{1 - E[X_\nu]^2}} \leq \frac{1}{\sqrt{1 - E[X_\nu]}} \leq \frac{2}{c_5 \nu} \tag{A.179}
\]

and similarly

\[
\frac{\xi(X_\nu)}{2 (1 - \xi(X_\nu))^2} \leq \frac{4}{c_5^3 \nu^3} \frac{1}{1 - \xi(X_\nu)} + \frac{\pi}{2} \frac{1}{1}_{E \xi}. \tag{A.180}
\]

Applying (A.180) and taking expectations in (A.175), we obtain by property 2

\[
\cos^{-1} E[X_\nu] - E[\cos^{-1} X_\nu] \leq C \nu \frac{\log n}{n} + C' e^{-\epsilon n} + C'' n^{-c_3 d}. \tag{A.181}
\]

Together, (A.176), (A.178) and (A.181) establish the first claim provided \( n \) is chosen larger than an absolute constant multiple of \( d \log n \).
For the second claim, we begin by using the triangle inequality to write

\[ \left| \cos^{-1} X_\nu - \mathbb{E}[\cos^{-1} X_\nu] \right| \leq \left| \cos^{-1} X_\nu - \cos^{-1} \mathbb{E}[X_\nu] \right| + \left| \cos^{-1} \mathbb{E}[X_\nu] - \mathbb{E}[\cos^{-1} X_\nu] \right|, \]

and then observe that our proof of the first claim implies suitable control of the second term. For the first term, if \( \nu \leq \nu_0 \) we use Lemma A.5.20 to immediately obtain with probability at least \( 1 - C_3 n^{-c_3 d} \) that this term is at most \( C e^{-c n} \). For \( \nu \geq \nu_0 \), we apply property 3 and the bounds (A.179) and (A.180) in the expression (A.175) to obtain with probability at least \( 1 - C_3 n^{-c_3 d} \)

\[ \left| \cos^{-1} X_\nu - \cos^{-1} \mathbb{E}[X_\nu] \right| \leq C \nu \frac{\sqrt{d \log n}}{n} + C' \frac{\nu d \log n}{n}, \]

which is of the claimed order when \( n \) is chosen larger than an absolute constant multiple of \( d \log n \).

\[ \square \]

**Lemma A.5.7.** There exist absolute constants \( c, C, C', C'' > 0 \) such that if \( n \geq C \log n \), one has

\[ \left| \varphi(\nu) - \cos^{-1} \mathbb{E}_{g_1,g_2}[X_\nu] \right| \leq C' e^{-c n} + C'' \frac{\nu}{n}. \]

**Proof.** Write \( f(\nu) = \cos \varphi(\nu) \), and

\[ h(\nu) = \mathbb{E}[X_\nu] - f(\nu), \]

so that \( h \) is the residual between the two terms whose images we are trying to tie together. We will make use of the following results:

1. The function \( \cos^{-1} \) is \( \frac{1}{2} \)-Hölder continuous on \([0, 1]\), so that \( |\cos^{-1} x - \cos^{-1} y| \leq \sqrt{|x - y|} \) if \( x, y \geq 0 \). (Lemma A.5.20)

2. For \( \nu \in [0, \pi] \), we have \( 1 - \frac{1}{2} \nu^2 \leq f(\nu) \leq 1 - c_2 \nu^2 \). (Lemma A.5.14)

3. For all \( 0 \leq \nu \leq \pi \), \( |h(\nu)| \leq C_1 e^{-c_1 n} + C_2 \nu^2 / n \). (Lemma A.5.15)
We choose \( n \) large enough that the hypotheses of Lemma A.5.15 are satisfied. Define \( \nu_0 = 2\sqrt{C_1/c_2}e^{-c_1n/2} \). We split the analysis into two sub-intervals: \( I_1 := [0, \nu_0] \), and \( I_2 := [\nu_0, \pi] \).

Choosing \( n \) larger than an absolute constant multiple of \( \log n \), we guarantee that \( I_1 \) and \( I_2 \) both have positive measure.

On \( I_1 \), we proceed as follows:

\[
|\cos^{-1} f - \cos^{-1}(f + h)| \leq \sqrt{|h|}
\]
\[
\leq \sqrt{C_1e^{-c_1n} + C_2\nu^2/n}
\]
\[
\leq \sqrt{C_1e^{-c_1n} + 4C_1C_2\nu^2e^{-c_1n}}
\]
\[
\leq Ce^{-\frac{1}{2}c_1n}.
\]

The first inequality uses Hölder continuity of \( \cos^{-1} \), the second uses our bound on the residual, the third uses the definition of \( I_1 \), and the fourth worst-cases the constants.

On \( I_2 \), we calculate

\[
|f + h| \leq |f| + |h| \leq C_1e^{-c_1n} + C_2\frac{\nu^2}{n} + 1 - c_2\nu^2,
\]

using the triangle inequality and our bounds on \( |h| \) and \( f \). Using the conditions \( \nu \geq \nu_0 \) and choosing \( n \geq 4C_2/c_2 \), we can rearrange to get

\[
C_1e^{-c_1n} + C_2\frac{\nu^2}{n} \leq \frac{c_2\nu^2}{2},
\]

which implies \( |f + h| \leq 1 - c_2\nu^2/2 \). By the control \( f(\nu) \leq 1 - c_2\nu^2 \), valid on \( I_2 \), we get that both \( f \) and \( f + h \) are bounded above by \( 1 - c_2\nu^2/2 \) on \( I_2 \); moreover, because \( f \geq 0 \) and \( f + h \geq 0 \), we can apply local Lipschitz properties of \( \cos^{-1} \) on \( I_2 \). This yields

\[
|\cos^{-1} f - \cos^{-1}(f + h)| \leq \frac{|h|}{\sqrt{1 - (\sup_{I_2} \max\{f, f + h\})^2}}
\]

265
\[ \leq \frac{C_1 e^{-c_1 n} + C_2 y^2 / n}{\sqrt{1 - (1 - (c_2 / 2)y^2)^2}} \]
\[ = \frac{C_1 e^{-c_1 n}}{\sqrt{\frac{1}{2} c_2 y^2 (2 - \frac{1}{2} c_2 y^2)}} + \frac{C_2 y^2 / n}{\sqrt{\frac{1}{2} c_2 y^2 (2 - \frac{1}{2} c_2 y^2)}} \]
\[ \leq C y^{-1} e^{-c_1 n} + C' v / n \]
\[ \leq C e^{-\frac{1}{2} c_1 n} + C' v / n. \]

Above, the first inequality is the instantiation of the local Lipschitz property; the second applies our upper and lower bounds on \( f \) and \( f + h \) derived above, and our bound on the residual \( |h| \); the fourth applies the bound \( 0 \leq f(\nu) \leq 1 - \frac{1}{2} c_2 \nu^2 \) to conclude \( 2 - \frac{1}{2} c_2 \nu^2 \geq 1 \) on \( I_2 \), and cancels a factor of \( \nu \) in the second term; and in the last line, we apply \( \nu \in I_2 \) to get \( \nu \geq 2 \sqrt{C_1 / c_2 e^{-c_1 n / 2}} \), which allows us to cancel the \( \nu^{-1} \) factor in the first term of the previous line.

To wrap up, we can choose the largest of the constants appearing in the bounds derived for \( I_1 \) and \( I_2 \) above and then conclude, since \( I_1 \cup I_2 = [0, \pi] \) under our condition on \( n \). \( \square \)

**Proving Lemma A.5.6**

**Lemma A.5.8.** There exist absolute constants \( c, c', C, C' > 0 \) such that if \( n \geq C \) and if \( n \) is sufficiently large to satisfy the hypotheses of Lemma A.5.15, one has

\[ 1 - C'' \nu^2 - C' e^{-c' n} \leq \mathbb{E}_{g_1,g_2} [X_\nu] \leq 1 - c \nu^2 + C' e^{-c' n}. \]

**Proof.** By the triangle inequality, we have

\[ |\cos \varphi(\nu)| - |\mathbb{E}[X_\nu] - \cos \varphi(\nu)| \leq \mathbb{E}[X_\nu] \leq |\cos \varphi(\nu)| + |\mathbb{E}[X_\nu] - \cos \varphi(\nu)|. \]

Applying Lemmas A.5.14 and A.5.15 with \( m = 0 \), we get

\[ 1 - C'' \nu^2 - C e^{-c' n} - C' \nu^2 / n \leq \mathbb{E}[X_\nu] \leq 1 - c \nu^2 + C e^{-c' n} + C' \nu^2 / n, \]

266
which proves the claim if we choose \( n \geq 2C' / c \). □

**Lemma A.5.9.** There exist absolute constants \( c, C, C' > 0 \) such that if \( n \) satisfies the hypotheses of Lemmas A.5.11 and A.5.12, then one has for each \( \nu \in [0, \pi] \)

\[
\text{Var}[X_\nu] \leq \frac{C \nu^4 \log n}{n} + C' e^{-cn}.
\]

**Proof.** We use the following elementary fact for a random variable with finite first and second moments, easily proved using \( \text{Var}[X_\nu] = \mathbb{E}[X_\nu^2] - \mathbb{E}[X_\nu]^2 \) and Fubini’s theorem: in this setting one has

\[
\text{Var}[X_\nu] = \mathbb{E}_g[\text{Var}[X_\nu(g_1, \cdot)]] + \mathbb{E}_g[\text{Var}[X_\nu(\cdot, g_2)]].
\]

By Lemma A.5.11, there is an event \( E \) of probability at least \( 1 - C e^{-cn} \) on which \( \text{Var}[X_\nu(g_1, \cdot)] \leq C' \nu^4 / n + C'' e^{-c'n} \). Invoking as well Lemma A.5.12, we obtain

\[
\text{Var}[X_\nu] \leq \frac{C \nu^4 \log n}{n} + C' e^{-cn} + \mathbb{P}[E^c] \mathbb{E}_{g_1}[\text{Var}[X_\nu(g_1, \cdot)]^2]^{1/2}
\]

as claimed, where in the second line we applied nonnegativity of the variance and the Schwarz inequality, and in the third line we used the fact that \( \|X\|_{L^2} \leq \|X\|_{L^\infty} \) for any random variable \( X \) in \( L^\infty \). □

**Lemma A.5.10.** There exist absolute constants \( c, c', C, C', C'' > 0 \) and absolute constants \( K, K' > 0 \) such that for any \( d \geq 1 \) such that \( n \) and \( d \) satisfy the hypotheses of Lemmas A.5.11 and A.5.13 and \( n \geq \max\{K \log n, K'd\} \), for any \( \nu \in [0, \pi] \), one has

\[
\mathbb{P}\left[ \left| X_\nu - \mathbb{E}_{g_1, g_2} [X_\nu] \right| \leq C'' \nu^2 \sqrt{\frac{d \log n}{n}} + C e^{-cn} \right] \geq 1 - C' n^{-c'd}.
\]
Proof. By Lemma A.5.11, we have

\[ P \left[ \left| X_\nu - \mathbb{E}[X_\nu] \right| \leq C'' \nu^2 \sqrt{\frac{d}{n}} + Ce^{-cn} \right] \geq 1 - C' e^{-c'd}. \]

Let \( \psi = \psi_{0.25} \) denote the cutoff function defined in Lemma A.5.31, and write

\[ Y_\nu(g_1, g_2) = \left\{ \frac{v_0(g_1, g_2)}{\psi(\|v_0(g_1, g_2)\|_2)}, \frac{v_\nu(g_1, g_2)}{\psi(\|v_\nu(g_1, g_2)\|_2)} \right\}. \]

By Lemma A.5.13, we have

\[ P \left[ \left| \mathbb{E}[Y_\nu] - \mathbb{E}[Y_\nu] \right| \leq C'' \nu^2 \sqrt{\frac{d \log n}{n}} + Cne^{-cn} \right] \geq 1 - C'n^{-c'd}. \]

We have \( X_\nu = Y_\nu \) on the event \( E_1 \), by Lemma A.5.16, and we thus calculate using the triangle inequality

\[ \left| \mathbb{E}[Y_\nu] - \mathbb{E}[X_\nu] \right| \leq \mathbb{E}_{g_1, g_2} \left[ |X_\nu - Y_\nu| \right] = \mathbb{E}_{g_1, g_2} \left[ 1_{E_1} \mathbb{E}^c Y_\nu \right] \leq Cne^{-cn}, \]

where the last inequality uses Hölder’s inequality and the measure bound in Lemma A.5.16. Again using the triangle inequality, we have

\[ \left| \mathbb{E}[X_\nu] - \mathbb{E}[Y_\nu] \right| \leq \mathbb{E}_{g_2} \left[ |X_\nu - Y_\nu| \right], \]

and so using our previous calculation and Markov’s inequality, we can assert

\[ P \left[ \left| \mathbb{E}[Y_\nu] - \mathbb{E}[X_\nu] \right| \leq Cne^{-cn/2} \right] \geq 1 - e^{-cn/2}. \]

The claim then follows from the triangle inequality, a union bound, and a choice of \( n \) larger than an absolute constant multiple of \( \log n \) and an absolute constant multiple of \( d \). \( \square \)

Lemma A.5.11. There exist absolute constants \( c, c', c'', c''' \), \( C, C', C'', C''' \), \( C_4, C_5 > 0 \), and absolute constants \( K, K' > 0 \) such that for any \( d \geq 1 \), if \( n \geq \max\{Kd, K' \log n\} \), then for every
\( \nu \in [0, \pi] \) one has with probability at least \( 1 - Ce^{-cn} \)

\[
\text{Var}[X_\nu(g_1, \cdot)] \leq \frac{C_4 \nu^4}{n} + C'e^{-c'n},
\]

and with \((g_1, g_2)\) probability at least \(1 - C''e^{-c''d}\) one has

\[
\left| X_\nu - \mathbb{E}_{g_2}[X_\nu] \right| \leq C_5 \nu^2 \sqrt{\frac{d}{n}} + C'''e^{-c'''n}.
\]

**Proof.** Fix \( \nu \in [0, \pi] \). Let \( E_1 = E_{0,5,1} \) denote the event in Lemma A.5.16 which is in the definition of \( X_\nu \). We start by treating the case of \( \nu = 0 \) or \( \nu = \pi \). We have \( X_\pi = 0 \) deterministically, so the variance is zero and it equals its partial expectation over \( g_2 \) with probability one. For the other case, one has \( X_0 = \mathbb{1}_{E_1} \); we have

\[
\text{Var}[X_0(g_1, \cdot)] = \mathbb{E}_{g_2}[\mathbb{1}_{E_1}] - \mathbb{E}_{g_2}[\mathbb{1}_{E_1}]^2 \leq \left( 1 - \mathbb{E}_{g_2}[\mathbb{1}_{E_1}] \right),
\]

and since \( \mathbb{E}[\mathbb{1}_{E_1}] = 1 - Cne^{-cn} \) by Lemma A.5.16, we obtain by Markov’s inequality

\[
P\left[ \text{Var}[X_0(g_1, \cdot)] \geq Cne^{-cn/2} \right] \leq e^{-cn/2}.
\]

This gives a suitable bound on the variance with suitable probability. For deviations, we note that

\[
\mathbb{E}_{g_2}\left[ X_0 - \mathbb{E}_{g_2}[X_0] \right] = 0,
\]

and following our previous variance inequality but taking expectations over both \( g_1 \) and \( g_2 \) gives \( \text{Var}[X_0] \leq Cne^{-cn} \), which implies by Chebyshev’s inequality

\[
P\left[ \left| X_0 - \mathbb{E}_{g_2}[X_0] \right| \geq \sqrt{Cne^{-cn/2}} \right] \leq e^{-cn/2}
\]

which is a suitable deviations bound that we can union bound with the event constructed below,
which controls deviations uniformly for the remaining values of $\nu$. We therefore assume below that $0 < \nu < \pi$.

Let $\psi(x) = \max\{x, \frac{1}{8}\}$, which is continuous and differentiable except at $x = \frac{1}{8}$, with derivative $\psi'(x) = 1_{x > 1/8}$. We note in addition that $x \leq \psi(x)$, and since $\psi \geq \frac{1}{8}$ we have for $x \geq 0$ the bound $x/\psi(x) \leq 1$. Define

$$Y_\nu(g_1, g_2) = \frac{\langle v_0(g_1, g_2), v_\nu(g_1, g_2) \rangle}{\psi(\|v_0(g_1, g_2)\|_2) \psi(\|v_\nu(g_1, g_2)\|_2)}.$$

We first show that it is enough to prove the claims for $Y_\nu$, which will be preferable for technical reasons. On $\mathcal{E}_1$, we have $Y_\nu = X_\nu$. We have $|Y_\nu| \leq 1$, and we calculate

$$\mathbb{E}_{g_2}[(Y_\nu - X_\nu)^2] = \mathbb{E}_{g_2}[1_{\mathcal{E}_1^c}(Y_\nu - X_\nu)^2] \leq C \mathbb{E}_{g_2}[1_{\mathcal{E}_1^c}]^{1/2} \mathbb{E}_{g_2}[(Y_\nu - X_\nu)^4]^{1/2} \leq C \mathbb{E}_{g_2}[1_{\mathcal{E}_1^c}]^{1/2},$$

where the first inequality uses the Schwarz inequality, and the last inequality uses that $|X_\nu| \leq 1$ and the triangle inequality, and where $C > 0$ is an absolute constant. We have by Tonelli’s theorem and Lemma A.5.16

$$\mathbb{E}_{g_1} \left[ \mathbb{E}_{g_2}[1_{\mathcal{E}_1^c}]^{1/2} \right] \leq C n e^{-c n},$$

so Markov’s inequality implies

$$\mathbb{P} \left[ \mathbb{E}_{g_2}[1_{\mathcal{E}_1^c}]^{1/2} \geq C n e^{-c n/2} \right] \leq e^{-c n/2}.$$

Thus, with probability at least $1 - e^{-c n/2}$, we have

$$\mathbb{E}_{g_2}[(Y_\nu - X_\nu)^2] \leq C' n e^{-c n/2},$$

so that an application of Lemma A.5.32 yields that with probability at least $1 - e^{-c n/2}$

$$\text{Var}[X_\nu(g_1, \cdot)] \leq \text{Var}[Y_\nu(g_1, \cdot)] + C'' n e^{-c' n},$$
where we have worst-cased constants and the exponent on $n$. For deviations, we write using the triangle inequality

$$\left| X_\nu - \mathbb{E}[X_\nu] \right| \leq |X_\nu - Y_\nu| + \left| Y_\nu - \mathbb{E}[Y_\nu] \right| + \left| \mathbb{E}[Y_\nu] - \mathbb{E}[X_\nu] \right|,$$

and then note that the first term is identically zero on the event $E_1$, which has probability at least $1 - Ce^{-cn}$, whereas for the third term, we have

$$\left| \mathbb{E}[Y_\nu] - \mathbb{E}[X_\nu] \right| \leq \mathbb{E}[(Y_\nu - X_\nu)^2]^{1/2} \leq C'ne^{-cn/2},$$

where the first inequality uses the triangle inequality and the Lyapunov inequality, and the second inequality holds with probability at least $1 - e^{-cn/2}$, and leverages the argument we used to control the difference in variances. Ultimately taking union bounds, we can conclude that it sufficient to prove the claimed properties for $Y_\nu$.

With $0 < \nu < \pi$ fixed, we introduce the notation

$$u_{g_1} = v_0(g_1); \quad v_{g_1, g_2} = v_\nu(g_1, g_2),$$

so that

$$Y_\nu = \left( \frac{u_{g_1}}{\psi (\|u_{g_1}\|_2)}, \frac{v_{g_1, g_2}}{\psi (\|v_{g_1, g_2}\|_2)} \right).$$

For fixed $g_1$, we will write $Y_\nu(g_2) = Y_\nu(g_1, g_2)$ with an abuse of notation. For $\tilde{g} \in \mathbb{R}^n$ arbitrary and $g_2$ fixed, we consider the function $f(t) = Y_\nu(g_2 + t\tilde{g})$ for $t \in [0, 1]$. Writing $f'$ for the derivative of $f$ where it exists, at any point of differentiability, we calculate by the chain rule

$$f'(t) = \langle \nabla_{g_2} Y_\nu(g_2 + t\tilde{g}), \tilde{g} \rangle.$$
where
\[
\nabla_{g_2} Y_{\nu}(g_2) = \frac{\sin \nu}{\psi(||u_{g_1}||_2)\psi(||v_{g_1,g_2}||_2)} \left( I - \frac{\psi'(||v_{g_1,g_2}||_2)v_{g_1,g_2}v_{g_1,g_2}^*}{\psi(||v_{g_1,g_2}||_2)||v_{g_1,g_2}||_2} \right) \left( \mathbb{1}_{v_{g_1,g_2} > 0} \circ u_{g_1} \right).
\]

Using the fact that
\[
\mathbb{1}_{v_{g_1,g_2} > 0} \circ u_{g_1} = P_{\{v_{g_1,g_2} > 0\}}u_{g_1},
\]
where \(P_{\{v_{g_1,g_2} > 0\}}\) is the orthogonal projection onto the coordinates where \(v_{g_1,g_2}\) is positive, together with the fact that
\[
v_{g_1,g_2}v_{g_1,g_2}^*P_{\{v_{g_1,g_2} > 0\}} = P_{\{v_{g_1,g_2} > 0\}}v_{g_1,g_2}v_{g_1,g_2}^* = v_{g_1,g_2}v_{g_1,g_2}^*,
\]
we can also write
\[
\nabla_{g_2} Y_{\nu}(g_2) = \frac{\sin \nu}{\psi(||u_{g_1}||_2)\psi(||v_{g_1,g_2}||_2)} P_{\{v_{g_1,g_2} > 0\}} \left( I - \frac{\psi'(||v_{g_1,g_2}||_2)v_{g_1,g_2}v_{g_1,g_2}^*}{\psi(||v_{g_1,g_2}||_2)||v_{g_1,g_2}||_2} \right) u_{g_1}.
\]

We next argue that \(f\) does not fail to be differentiable at too many points of \([0, 1]\). Because \(\psi > 0\), it will suffice to show that (i) \(t \mapsto v_{g_1,g_2+tg}\) and (ii) \(t \mapsto \psi(||v_{g_1,g_2+tg}||_2)\) are differentiable at all but a set of isolated points in \([0, 1]\). For the latter function, we note that at any point where \(||v_{g_1,g_2+tg}||_2 < \frac{1}{8}\), by continuity we have that \(t \mapsto \psi(||v_{g_1,g_2+tg}||_2)\) is locally constant, and therefore differentiable at such points. At other points, by Lemma A.5.21 it suffices to characterize \(t \mapsto ||v_{g_1,g_2+tg}||_2\) as differentiable at all but isolated points, and because \(||v_{g_1,g_2+tg}||_2 \geq \frac{1}{8}\) by assumption, the norm is differentiable and by the chain rule it suffices to characterize differentiability of each coordinate of \(t \mapsto v_{g_1,g_2+tg}\), which settles the question of all-but-isolated differentiability of (i) as well. We have \(v_{g_1,g_2+tg} = \sigma(g_1 \cos \nu + g_2 \sin \nu + t\tilde{g} \sin \nu)\), so again by Lemma A.5.21, we conclude from differentiability of \(t \mapsto g_1 \cos \nu + g_2 \sin \nu + t\tilde{g} \sin \nu\) that \(t \mapsto v_{g_1,g_2+tg}\) is differentiable at all but isolated points, and consequently so is \(f\). In particular, \(f\) is differentiable at all but countably many points of \([0, 1]\). Next, we show that \(f'\) has suitable integrability properties. Indeed, we calculate
using (A.182)

\[ \| \nabla_{g_2} Y_v(g_2) \|_2 \leq 8\nu \left\| \left( I - \frac{\psi'(\|v_{g_1,g_2}\|_2)v_{g_1,g_2}v_{g_1,g_2}^*}{\psi(\|v_{g_1,g_2}\|_2)\|v_{g_1,g_2}\|_2} \right) \frac{u_{g_1}}{\psi(\|u_{g_1}\|_2)} \right\|_2 \]

\[ = 8\nu \sqrt{1 - \psi'(\|v_{g_1,g_2}\|_2)} Y_v(g_2)^2, \quad \text{(A.183)} \]

where we used Cauchy-Schwarz and \( \psi \geq \frac{1}{8} \) in the first inequality and distributed and applied \( (\psi')^2 = \psi' \) and the estimate \( x/\psi(x) \leq 1 \) in the second inequality. In particular, this implies that

\[ |f'(t)| \leq C\|\tilde{g}\|_2, \] which is a \( t \)-integrable upper bound for every \( \tilde{g} \). Because \( Y_v(g_1, \cdot) \) is continuous by continuity of \( \sigma, \psi \), and the fact that \( \psi \) becomes constant whenever \( \|v_{g_1,g_2}\|_2 < \frac{1}{8} \), we can apply [219, Theorem 6.3.11] to get

\[ Y_v(g_2 + \tilde{g}) = Y_v(g_2) + \int_0^1 \langle \nabla_{g_2} Y_v(g_2 + t\tilde{g}), \tilde{g} \rangle \, dt, \]

and since \( \tilde{g} \) was arbitrary, for any \( g'_2 \in \mathbb{R}^n \) we can put \( \tilde{g} = g'_2 - g_2 \) to get

\[ Y_v(g'_2) = Y_v(g_2) + \int_0^1 \langle \nabla_{g_2} Y_v(tg'_2 + (1 - t)g_2), g'_2 - g_2 \rangle \, dt. \]

Performing the expansion with \( g_2 \) and \( g'_2 \) reversed and applying the triangle inequality and Cauchy-Schwarz then implies the estimate

\[ |Y_v(g'_2) - Y_v(g_2)| \leq \|g'_2 - g_2\|_2 \int_0^1 \|\nabla_{g_2} Y_v(tg'_2 + (1 - t)g_2)\|_2 \, dt. \quad \text{(A.184)} \]

This relation is enough to conclude the result for angles satisfying \( \nu \geq c_0 \), where \( 0 < c_0 \leq \pi/4 \) is an absolute constant. Indeed, (A.183) and (A.184) imply that \( Y_v \) is \( C \)-Lipschitz, where \( C > 0 \) is an absolute constant; so the Gaussian Poincaré inequality implies

\[ \mathbb{E}_{g_2} \left[ \left( Y_v - \mathbb{E}_{g_2}[Y_v] \right) \right] \leq \frac{C'}{n}, \]
and Gauss-Lipschitz concentration implies for any \( d \geq 0 \)
\[
\mathbb{P} \left[ \left| Y_v - \mathbb{E}[Y_v] \right| \geq C'' \sqrt{\frac{d}{n}} \right] \leq 2e^{-d}.
\]

Because \( v \geq c_0 \), we can adjust these bounds to involve \( v^4 \) and \( v^2 \) (respectively) while only paying increases in the constant factors. We proceed assuming \( 0 < v \leq c_0 \).

Let \( 0 \leq \tau_{g_1} \leq 1 \) denote a median of \( Y_v(g_1, \cdot) \), i.e., a number satisfying \( \mathbb{P}_{g_2}[Y_v \geq \tau_{g_1}] \geq \frac{1}{2} \) and \( \mathbb{P}_{g_2}[Y_v \leq \tau_{g_1}] \geq \frac{1}{2} \), and for each \( 0 \leq s < \tau_{g_1} \) define
\[
w_s(g_2) = \max\{Y_v(g_2), \tau_{g_1} - s\}.
\]

For any \( 0 \leq s < \tau_{g_1} \), notice that \( w_s \geq Y_v \), which implies that \( \mathbb{P}[w_s \geq \tau_{g_1}] \geq \mathbb{P}[Y_v \geq \tau_{g_1}] \geq \frac{1}{2} \), because \( \tau_{g_1} \) is a median of \( Y_v \); and similarly \( \mathbb{P}[w_s \leq \tau_{g_1}] \geq \mathbb{P}[Y_v \leq \tau_{g_1}] \geq \frac{1}{2} \), so that \( \tau_{g_1} \) is also a median of \( w_s \). The fact that \( w_s \geq Y_v \) implies for any \( t > 0 \) that \( \mathbb{P}[Y_v - \tau_{g_1} > t] \leq \mathbb{P}[w_s - \tau_{g_1} > t] \), and additionally if \( Y_v \leq \tau_{g_1} - s \) we have \( w_s = \tau_{g_1} - s \), so that \( \mathbb{P}[Y_v - \tau_{g_1} \leq -s] \leq \mathbb{P}[w_s - \tau_{g_1} \leq -s] \).

In particular, the tails of \( Y_v \) can be controlled in terms of those of \( w_s \) for appropriate choices of \( s \). Additionally, by Lemma A.5.21, we have that for each \( s, t \mapsto w_s(g_2 + t \bar{g}) \) is differentiable at all but countably many points of \([0, 1]\), and has derivative there equal to \( t \mapsto \langle \bar{g}, \nabla w_s(g_2) \rangle \), where
\[
\nabla_{g_2} w_s(g_2) = \mathbb{1}_{w_s(g_2) > \tau_{g_1} - s} \nabla_{g_2} Y_v(g_2),
\]

which, following from (A.183), satisfies a strengthened gradient norm estimate
\[
\| \nabla_{g_2} w_s(g_2) \|_2 \leq 8v \mathbb{1}_{w_s(g_2) > \tau_{g_1} - s} \sqrt{1 - \psi'(\|v_{g_1,g_2}\|_2)Y_v(g_2)^2} \]
\[
\leq 8v \sqrt{1 - \psi'(\|v_{g_1,g_2}\|_2)(\tau_{g_1} - s)^2}.
\]

In particular, we obtain a nearly-Lipschitz estimate of the form (A.184):
\[
|w_s'(g_2') - w_s'(g_2)| \leq \|g_2' - g_2\|_2 \int_0^1 8v \sqrt{1 - \psi'(\|v_{g_1,g_2'+(1-t)g_2}\|_2)(\tau_{g_1} - s)^2} \, dt.
\]

(274)
For each $g_1$, we define a set $S_{g_1} = \{ g_2 \mid \|v_{g_1,g_2}\|_2 \geq \frac{1}{4}\}$. Noting that the function $g_2 \mapsto \|\sigma(g_1\cos \nu + g_2\sin \nu]\|_2$ is a convex 1-Lipschitz function (given that $|\sin \nu| \leq 1$), we have by Gauss-Lipschitz concentration

$$
P_{g_2}[\|v_{g_1,g_2}\|_2 \leq \mathbb{E}_{g_2}[\|v_{g_1,g_2}\|_2] - t] \leq e^{-cnt^2},$$

and by Jensen’s inequality

$$
\mathbb{E}_{g_2}[\|v_{g_1,g_2}\|_2] \geq |\cos \nu|\|u_{g_1}\|_2 \geq \frac{\|u_{g_1}\|_2}{\sqrt{2}},
$$

where the last line holds because $\nu \leq \pi/4$. By Lemma A.5.16, there is a $g_1$ event $\mathcal{E}$ having probability at least $1 - Ce^{-cn}$ on which $\|u_{g_1}\|_2 \geq \frac{1}{2}$, so that for any $g_1 \in \mathcal{E}$, we have by a suitable choice of $t$ in our Gauss-Lipschitz bound $P_{g_2}[S_{g_1}] \geq 1 - e^{-cn}$. Thus, using the first line of (A.185), the Gaussian Poincaré inequality and the Lipschitz property of $w_s$ (which follows from (A.186) after bounding by an absolute constant) and Rademacher’s theorem on a.e. differentiability of Lipschitz functions, we have whenever $g_1 \in \mathcal{E}$

$$\Var[w_s] \leq \frac{2}{n} \mathbb{E}_{g_2}[\|\nabla_{g_2} w_s(g_2)\|_2^2] \leq \frac{128\nu^2}{n} \mathbb{E}_{g_2}[\left(1 - \psi'(\|v_{g_1,g_2}\|_2)Y_\nu(g_2)^2\right)]$$

$$\leq \frac{128\nu^2}{n} \mathbb{E}_{g_2}[\mathbb{1}_{\mathcal{E}} \left(1 - \psi'(\|v_{g_1,g_2}\|_2)Y_\nu(g_2)^2\right)]$$

$$+ \mathbb{E}_{g_2}[\mathbb{1}_{\mathcal{E}^c} \left(1 - \psi'(\|v_{g_1,g_2}\|_2)Y_\nu(g_2)^2\right)]$$

$$\leq \frac{256\nu^2}{n} \mathbb{E}_{g_2}[1 - Y_\nu(g_2)] + Ce^{-cn}, \quad (A.187)$$

where we also make use of the fact that $0 \leq Y_\nu \leq 1$. Now, we calculate for $g_1 \in \mathcal{E}$ and $g_2 \in S_{g_1}$

$$Y_\nu = 1 - \frac{1}{2} \frac{\|u_{g_1} - v_{g_1,g_2}\|_2^2}{\|u_{g_1}\|_2^2}$$

$$\geq 1 - 2 \frac{\|u_{g_1} - v_{g_1,g_2}\|_2^2}{\|u_{g_1}\|_2^2}$$

275
\[
\geq 1 - 8\|u_{g_1} - v_{g_1, g_2}\|_2^2 \\
\geq 1 - 8\|g_1 - (g_1 \cos \nu + g_2 \sin \nu)\|_2^2, \tag{A.188}
\]

where the second inequality uses \(g_1 \in \mathcal{E}\), the third uses nonexpansiveness of \(\sigma\), and the first requires a proof; we will show that for any nonzero vectors \(x, y \in \mathbb{R}^n\), one has

\[
\left\| \frac{x}{\|x\|_2} - \frac{y}{\|y\|_2} \right\| \leq 2\|x - y\|_2 \|y\|_2. \tag{A.189}
\]

To see this, write \(\theta\) for the angle between \(x\) and \(y\), and distribute to obtain equivalently

\[
-\frac{1}{2} \|y\|_2^2 (1 + \cos \theta) \leq \|x\|_2^2 - 2\|x\|_2 \|y\|_2 \cos \theta.
\]

Divide through by \(\|x\|_2^2\), write \(K = \|y\|_2 \|x\|_2^{-1}\), and rearrange to obtain the equivalent expression

\[
K^2 (1 + \cos \theta) - 4K \cos \theta + 2 \geq 0.
\]

It suffices to minimize the LHS of the previous inequality with respect to \(K\) subject to the constraint \(K > 0\) and then study the resulting function of \(\theta\) to ascertain the validity of the bound. Given that \(1 + \cos \theta \geq 0\), the LHS is a convex function of \(K\), with minimizer \(K = 2 \cos \theta (1 + \cos \theta)^{-1}\), and therefore for any \(\theta \geq \pi/2\), the LHS subject to the constraint \(K > 0\) is minimized at \(K = 0\), where the inequality is easily seen to be true. If \(\theta < \pi/2\), we have that the minimizer is positive, and we verify that after substituting the bound becomes

\[
1 + \cos \theta \geq 2 \cos^2 \theta,
\]

which is also seen to be true for \(\theta < \pi/2\), for example by showing that the polynomial \(x \mapsto \ldots\).
−2x^2 + x + 1 is nonnegative on [0, 1]. This proves the inequality, so returning to (A.188), we have

\[ Y_\nu \geq 1 - 8((1 - \cos \nu)^2 \|g_1\|_2^2 + \sin^2 \nu \|g_2\|_2^2 - 2(\sin \nu)(1 - \cos \nu)\langle g_1, g_2 \rangle) \]

\[ \geq 1 - 8((1 - \cos \nu)^2 \|g_1\|_2^2 + \sin^2 \nu \|g_2\|_2^2 - 2(\sin \nu)(1 - \cos \nu)\|g_1\|_2 \|g_2\|_2) \]

using Cauchy-Schwarz in the second inequality. By Gauss-Lipschitz concentration (e.g. following the proof of the third assertion in Lemma A.5.17), there is a \( g_1 \) event \( \mathcal{E}' \) and a \( g_2 \) event \( \mathcal{E}'' \), each with probability at least \( 1 - Ce^{-cn} \), on which we have (respectively) \( \|g_i\|_2 \leq 2 \) for \( i = 1, 2 \). Then using \( (\sin \nu)(1 - \cos \nu) \geq 0 \), we obtain that when \( g_1 \in \mathcal{E} \cap \mathcal{E}' \) and when \( g_2 \in S_{g_1} \cap \mathcal{E}'' \)

\[ Y_\nu \geq 1 - 32((1 - \cos \nu)^2 + \sin^2 \nu) = 1 - 64(1 - \cos \nu) \geq 1 - 32\nu^2, \]

where the final inequality uses the standard estimate \( \cos \nu \geq 1 - 0.5\nu^2 \), which can be proved via Taylor expansion. By a union bound, we can assert that with \( g_1 \)-probability at least \( 1 - Ce^{-cn} \), with conditional (in \( g_2 \)) probability at least \( 1 - C'e^{-c'n} \) we have \( Y_\nu \geq 1 - 32\nu^2 \), so that in particular, by nonnegativity of \( Y_\nu \), and choosing \( n \) larger than an absolute constant, we guarantee with \( g_1 \)-probability at least \( 1 - Ce^{-cn} \)

\[ \mathbb{E}_{g_2}[Y_\nu] \geq 1 - 32\nu^2 - C'e^{-cn}, \quad \tau_{g_1} \geq 1 - 32\nu^2. \quad (A.190) \]

Plugging the mean estimate into (A.187), we conclude with probability at least \( 1 - C''e^{-c'n} \)

\[ \text{Var}[w_s] \leq \frac{C\nu^4}{n} + C'e^{-cn}. \quad (A.191) \]

We could have just as well applied this exact argument to \( Y_\nu \) instead of \( w_s \), so we conclude the claimed variance bound from this expression. We have stated the result in terms of the truncations \( w_s \) so that it can be applied towards deviations control in the sequel. As an immediate application, we use the fact that any median is a minimizer of the quantity \( c \mapsto \mathbb{E}[|X - c|] \) for any integrable
X and $c \in \mathbb{R}$ to get with probability at least $1 - C''e^{-c'n}$

$$\left| \mathbb{E}[w_s] - \tau_{g_1} \right| \leq \mathbb{E}\left| w_s - \tau_{g_1} \right| \leq \mathbb{E}\left[ \left| w_s - \mathbb{E}[w_s] \right| \right] \leq \sqrt{\text{Var}[w_s]} \leq \frac{Cy^2}{\sqrt{n}} + C' e^{-cn}, \quad (A.192)$$

where we also applied Jensen’s inequality for the first inequality and the Lyapunov inequality for the third. In particular, the same argument yields

$$\left| \mathbb{E}[Y_v] - \tau_{g_1} \right| \leq \frac{Cy^2}{\sqrt{n}} + C' e^{-cn}. \quad (A.193)$$

We turn to removing the $t$ dependence in (A.186) without sacrificing the dependence on $\tau_{g_1}$. To obtain a Lipschitz estimate on the subset $S_{g_1}$, we need to control the norm of $\nabla g_2 w_s$ on the line segment between $g_2, g'_2 \in S_{g_1}$. For this, write $\sigma_y(x) = \max\{x - y, 0\}$ for any $y \in \mathbb{R}$, and make the following observations:

1. $v_{g_1, g_2} = (\sin \nu)\sigma_{-g_1 \cot \nu}(g_2)$, so that

$$Y_v(g_2) = \left\{ \frac{u_{g_1}}{\psi\left(\|u_{g_1}\|_2\right)}, \frac{(\sin \nu)\sigma_{-g_1 \cot \nu}(g_2)}{\psi\left(\|\sin \nu\sigma_{-g_1 \cot \nu}(g_2)\|_2\)} \right\};$$

2. for any $x, y$, $\sigma_y(x) = \max\{x, y\} - y$; $\mathcal{P} : \mathbb{R} \to \mathcal{P} = \max\{x, y\}$ is the projection onto the convex set $\{x \mid x_i \geq y, \forall i\}$, so in particular $\mathcal{P} : x \mapsto \sigma_y(x)$ is nonexpansive, has convex range, and satisfies $\sigma_y(\sigma_y(x) + y) = \sigma_y(x)$; and thus

3. for any $g_2, Y_v(g_2) = Y_v(\sigma_{-g_1 \cot \nu}(g_2) - g_1 \cot \nu)$.

We write $\bar{g}_2 = \sigma_{-g_1 \cot \nu}(g_2) - g_1 \cot \nu, \bar{g}'_2 = \sigma_{-g_1 \cot \nu}(g'_2) - g_1 \cot \nu$, so that (A.186) becomes

$$\left| w_s(g'_2) - w_s(g_2) \right| = \left| w_s(\bar{g}'_2) - w_s(\bar{g}_2) \right|$$

$$\leq \|\bar{g}'_2 - \bar{g}_2\|_2 \int_0^1 8\nu \sqrt{1 - \psi'(\|v_{g_1, \tau\bar{g}'_2 + (1-\tau)\bar{g}_2\|_2)}(\tau_{g_1} - s)^2 \, dt$$

$$\leq \|g'_2 - g_2\|_2 \int_0^1 8\nu \sqrt{1 - \psi'(\|v_{g_1, \tau\bar{g}'_2 + (1-\tau)\bar{g}_2\|_2)}(\tau_{g_1} - s)^2 \, dr,$$
where the second line follows from nonexpansiveness and translation invariance of the distance.

Having reduced to the study of points along the segment between $\tilde{g}_2$ and $\tilde{g}_2'$, we now observe

$$\sigma_{-g_1 \cot \nu} \left( t \tilde{g}_2' + (1 - t) \tilde{g}_2 \right) = \sigma \left( t \sigma_{-g_1 \cot \nu} (g_2) + (1 - t) \sigma_{-g_1 \cot \nu} (g_2) \right)$$

$$= t \sigma_{-g_1 \cot \nu} (g_2) + (1 - t) \sigma_{-g_1 \cot \nu} (g_2),$$

because $\sigma_{-g_1 \cot \nu}$ has image included in the nonnegative orthant, which is convex. It then follows from (1) above that

$$\| v_{g_1, t \tilde{g}_2' + (1 - t) \tilde{g}_2} \|_2 = (\sin \nu) \| t \sigma_{-g_1 \cot \nu} (g_2) + (1 - t) \sigma_{-g_1 \cot \nu} (g_2) \|_2$$

$$= \| v_{g_1, g_2} + (1 - t) v_{g_1, g_2'} \|_2,$$

and in particular

$$\| v_{g_1, g_2} + (1 - t) v_{g_1, g_2'} \|_2^2 = t^2 \| v_{g_1, g_2} \|_2^2 + 2t(1 - t) \langle v_{g_1, g_2}, v_{g_1, g_2'} \rangle + (1 - t)^2 \| v_{g_1, g_2'} \|_2^2$$

$$\geq \frac{1}{16} \left( t^2 + (1 - t)^2 \right) \geq \frac{1}{32},$$

where the first inequality uses that $\sigma \geq 0$ and $g_2, g_2' \in S_{g_1}$, and the second minimizes the convex function of $t$ in the previous bound. We conclude that $g_2, g_2' \in S_{g_1}$ implies that $\| v_{g_1, t \tilde{g}_2' + (1 - t) \tilde{g}_2} \|_2 > \frac{1}{8}$ for every $t \in [0, 1]$, and consequently (A.186) becomes (after an additional simplification of the quantity under the square root using $\tau_{g_1} \leq 1$)

$$| w_s(g_2') - w_s(g_2) | \leq 16\nu \sqrt{1 - (\tau_{g_1} - s)} \| g_2' - g_2 \|_2,$$  \hspace{1cm} (A.194)

so that $w_s$ is $16\nu \sqrt{1 - (\tau_{g_1} - s)}$-Lipschitz on $S_{g_1}$. Then by an application of the median bound in (A.190), if $0 \leq s < 1 - 32\nu^2$, with $g_1$ probability at least $1 - Ce^{-cn}$ we have that $w_s$ is $16\nu \sqrt{32\nu^2 + s}$-Lipschitz on $S_{g_1}$. For the previous assertion to be nonvacuous, we need to take $\nu$ small; in particular, we have $1 - 32\nu^2 > \frac{1}{2}$ if $\nu < 1/8$, which we can take to be the value
of the absolute constant $c_0$ we left unspecified previously. Then for each such $s$, define $L_s = 16\nu \sqrt{32\nu^2} + s$, and define

$$
\hat{w}_s(g_2) = \sup_{g'_2 \in S_{g_1}} \{ w_s(g'_2) - L_s \| g'_2 - g_2 \|_2 \}.
$$

Then $\hat{w}_s$ is $L_s$-Lipschitz on $\mathbb{R}^n$, and satisfies $\hat{w}_s = w_s$ on $S_{g_1}$ [227, §3.1.1 Theorem 1]. By the Gaussian Poincaré inequality, we obtain immediately $\text{Var}[\hat{w}_s] \leq L_s$, and using $\hat{w}_s = w_s$ on $S_{g_1}$, we compute

$$
\left| \mathbb{E}_{g_2} [w_s - \hat{w}_s] \right| 
\leq \mathbb{E}_{g_2} \left[ L_{g_1}^{S_{g_1}} \| w_s - \hat{w}_s \| \right]
\leq \mathbb{P}_{g_2} \left[ S_{g_1}^c \right]^{1/2} \| w_s - \hat{w}_s \| L_2
\leq Ce^{-cn} (\| w_s \|_{L^2} + \| \hat{w}_s \|_{L^2}) \leq C'e^{-cn},
$$

where the second inequality follows from the Schwarz inequality, the third holds given that $g_1 \in \mathcal{E}$ and by the Minkowski inequality, and the final uses that $w_s$ and $\hat{w}_s$ are both Lipschitz with Lipschitz constants bounded above by absolute constants together with the Gaussian Poincaré inequality.

Meanwhile, by Gauss-Lipschitz concentration, we obtain a Bernstein-type lower tail

$$
\mathbb{P}_{g_2} \left[ \hat{w}_s \leq \mathbb{E}_{g_2} [\hat{w}_s] - s \right] \leq \exp \left( -\frac{cns^2}{\nu^2 (32\nu^2 + s)} \right),
$$

and for the upper tail, it will be sufficient to consider $\hat{w}_0$, which satisfies a subgaussian tail (for any $t \geq 0$)

$$
\mathbb{P}_{g_2} \left[ \hat{w}_0 \leq \mathbb{E}_{g_2} [\hat{w}_0] - t \right] \leq \exp \left( -\frac{c'nt^2}{\nu^4} \right).
$$

Using the results (A.192), (A.193), (A.195), and the fact that $w_s = \hat{w}_s$ on $S_{g_1}$, we get

$$
\mathbb{P}_{g_2} \left[ Y_y - \mathbb{E}_{g_2} [Y_y] \leq -s \right] \leq \mathbb{P}_{g_2} \left[ \hat{w}_s - \mathbb{E}_{g_2} [\hat{w}_s] \leq C \frac{\nu^2}{\sqrt{n}} + C'e^{-cn} - s \right] + \mathbb{P}_{g_2} \left[ S_{g_1}^c \right].
$$
Using \( d \geq 1 \), we put \( s = 2C\nu^2\sqrt{d/n} + C'e^{-cn} \) in this bound; using that \( \nu < 1/8 \), and in particular \( 1 - 32\nu^2 > \frac{1}{2} \), we can choose \( n \) larger than an absolute constant multiple of \( d \) to guarantee that for all \( 0 \leq \nu < 1/8 \), this choice of \( s \) is less than \( 1 - 32\nu^2 \), and that \( C\nu^2\sqrt{d/n} \leq 32\nu^2 \). Together with the lower tail bound (A.196), these facts imply

\[
\mathbb{P}_{g_2}
\left[
Y_\nu - \mathbb{E}[Y_\nu] \leq -2C\nu^2\sqrt{\frac{d}{n}} - C'e^{-cn}
\right] \leq \mathbb{P}_{g_2}
\left[
\hat{w}_s - \mathbb{E}[\hat{w}_s] \leq -C\nu^2\sqrt{\frac{d}{n}} + C''e^{-c'n}
\right] \leq e^{-c''d} + C''e^{-c'n}.
\]

Meanwhile, for the upper tail, we have for any \( t \geq 0 \)

\[
\mathbb{P}_{g_2}
\left[
Y_\nu - \mathbb{E}[Y_\nu] \geq t
\right] \leq \mathbb{P}_{g_2}
\left[
\hat{w}_0 - \mathbb{E}[\hat{w}_0] \geq t - C\frac{\nu^2}{\sqrt{n}} - C'e^{-cn}
\right] + \mathbb{P}_{g_2}
\left[
\mathcal{S}^c_{g_1}
\right], \quad (A.199)
\]

and if we put \( t = 2C\nu^2\sqrt{d/n} + C'e^{cn} \), our previous requirements on \( n \) and the upper tail bound (A.197) yield

\[
\mathbb{P}_{g_2}
\left[
Y_\nu - \mathbb{E}[Y_\nu] \geq 2C\nu^2\sqrt{\frac{d}{n}} + C'e^{-cn}
\right] \leq e^{-c'''d} + C''e^{-c'n}.
\]

Combining these two bounds gives control of absolute deviations about the mean. By independence, we conclude

\[
\mathbb{P}_{g_1, g_2}
\left[
Y_\nu - \mathbb{E}[Y_\nu] \leq 2C\nu^2\sqrt{\frac{d}{n}} + C'e^{-c'''n}
\right] \geq (1 - 2e^{-cd} - C'e^{-c'n})(1 - C'e^{-c'''n})
\]

\[
\geq 1 - 2e^{-cd} - C'e^{-c'n} - C'e^{-c'''n}.
\]

To conclude, we have shown that for every \( \nu \in [0, \pi] \) one has with probability at least \( 1 - Ce^{-cn} \)

\[
\text{Var}[X_\nu(g_1, \cdot)] \leq \frac{C'\nu^4}{n} + C''ne^{-c'n},
\]

281
and with \((g_1, g_2)\) probability at least \(1 - 2e^{-c''d} + C'''n e^{-c''n}\) one has

\[
\left| X_v - \mathbb{E}_{g_2} [X_v] \right| \leq C'''v^2 \sqrt{\frac{d}{n}} + C'''n e^{-c''n}.
\]

To simplify these bounds, we may in addition choose \(n\) larger than an absolute constant multiple of \(\log n\), and \(n\) larger than an absolute constant multiple of \(d\), to obtain that with probability at least \(1 - Ce^{-cn}\)

\[
\text{Var}[X_v(g_1, \cdot)] \leq \frac{C_4 v^4}{n} + C'e^{-c'n},
\]

and with \((g_1, g_2)\) probability at least \(1 - C''e^{-c''d}\) one has

\[
\left| X_v - \mathbb{E}_{g_2} [X_v] \right| \leq C_5 v^2 \sqrt{\frac{d}{n}} + C'''e^{-c''n},
\]

which was to be shown. \(\square\)

**Lemma A.5.12.** There exist absolute constants \(c, C, C' > 0\) and an absolute constant \(K > 0\) such that if \(n \geq K \log^d n\), then for every \(\nu \in [0, \pi]\) one has

\[
\text{Var} \left[ \mathbb{E}_{g_2} [X_v(\cdot, g_2)] \right] \leq \frac{C v^4 \log n}{n} + C'e^{-cn}.
\]

**Proof.** Define

\[
Y_v(g_1, g_2) = \frac{\langle \nu_0(g_1, g_2), \nu_v(g_1, g_2) \rangle}{\psi(\|\nu_0(g_1, g_2)\|_2) \psi(\|\nu_v(g_1, g_2)\|_2)},
\]

where \(\psi = \psi_{0.25}\) is as in Lemma A.5.31. Then by Cauchy-Schwarz and the second property in Lemma A.5.31 (the case where either \(\|\nu_0\|_2 = 0\) or \(\|\nu_v\|_2 = 0\) is treated separately, since in this case \(Y_v = 0\), we obtain \(|Y_v| \leq 4\), and

\[
\mathbb{E}_{g_1, g_2} \left[ \left( \mathbb{E}_{g_2} [X_v] - \mathbb{E}_{g_2} [Y_v] \right)^2 \right] \leq \mathbb{E}_{g_1, g_2} \left[ \left( \mathbb{E}_{g_2} \left[ \mathbb{E}_{g_2} \left[ \frac{\langle \nu_0, \nu_v \rangle}{\psi(\|\nu_0\|_2) \psi(\|\nu_v\|_2)} \right] \right] \right)^2 \right]
\]

282
\[
\leq 16\mu(E_m^c) \leq C n e^{-cn},
\]

where we use the fact that if \((g_1, g_2) \in E_m\) then \(\|v_\nu\|_2 \geq \frac{1}{2}\) for every \(0 \leq \nu \leq \pi\) and hence \(\psi(\|v_\nu\|_2) = \|v_\nu\|_2\) in the first line, apply Jensen’s inequality in the second line, and combine our bound on \(Y_\nu\) with Hölder’s inequality and the measure bound in Lemma A.5.16 in the third line. An application of Lemma A.5.32 then yields

\[
\text{Var}_{g_2} \left[ E_{g_2} [X_\nu(\cdot, g_2)] \right] \leq \text{Var}_{g_2} \left[ E_{g_2} [Y_\nu(\cdot, g_2)] \right] + C n e^{-cn} \leq \text{Var}_{g_2} \left[ E_{g_2} [Y_\nu(\cdot, g_2)] \right] + C e^{-cn/2},
\]

where the last inequality holds when \(n\) is chosen to be larger than an absolute constant multiple of \(\log n\). It thus suffices to control the variance of \(Y_\nu\). Applying Lemma A.5.26, we get for almost all \(g_1 \in \mathbb{R}^n\)

\[
\mathbb{E}_{g_2} [Y_\nu(g_1, g_2)] = \frac{\|v_0\|_2^2}{\psi(\|v_0\|_2^2)} + \int_0^\nu \int_0^t \mathbb{E}_{g_2} \left[ (\Xi_1 + \Xi_2 + \Xi_3 + \Xi_4 + \Xi_5 + \Xi_6)(s, g_1, g_2) \right] ds \, dt,
\]

where we follow the notation defined in Lemma A.5.13. We start by removing the term outside of the integral from consideration. We have as above \(|Y_\nu| \leq 4\), so that \(|\mathbb{E}_{g_2}[Y_\nu]| \leq 4\). Moreover, following the proof of the measure bound in Lemma A.5.16, but using only the pointwise concentration result, we assert that if \(n \geq C\) an absolute constant there is an event \(E\) on which \(0.5 \leq \|v_0\|_2 \leq 2\) with probability at least \(1 - 2e^{-cn}\) with \(c > 0\) an absolute constant. This implies that if \(g_1 \in E\) we have

\[
\frac{\|v_0\|_2^2}{\psi(\|v_0\|_2^2)} = 1,
\]

and since

\[
\left| \frac{\|v_0\|_2^2}{\psi(\|v_0\|_2^2)} \right| \leq 4,
\]

by the same argument used for \(Y_\nu\), we can calculate

\[
\left\| \frac{\|v_0\|_2^2}{\psi(\|v_0\|_2^2)} - 1 \right\|_{L^2} \leq \left\| \left( \frac{\|v_0\|_2^2}{\psi(\|v_0\|_2^2)} - 1 \right) 1_{E^c} \right\|_{L^2} \leq 5 \|1_{E^c}\|_{L^2} \leq C e^{-cn},
\]
by the Minkowski inequality and the triangle inequality. An application of Lemma A.5.32 implies that it is therefore sufficient to control the variance of the quantity

\[ f(\nu, g_1) = 1 + \int_0^\nu \int_0^t \mathbb{E} \left[ (\Xi_1 + \Xi_2 + \Xi_3 + \Xi_4 + \Xi_5 + \Xi_6)(s, g_1, g_2) \right] \, ds \, dt. \]

By Lemma A.5.37, the Lyapunov inequality, and Fubini’s theorem, we have

\[ (f(\nu, g_1) - \mathbb{E}[f(\nu, g_1)])^2 = \left( \int_0^\nu \int_0^t \left( \sum_{i=1}^6 \mathbb{E} \left[ \Xi_i(s, g_1, g_2) \right] - \mathbb{E} \left[ \Xi_i(s, g_1, g_2) \right] \right) \, ds \, dt \right)^2. \]

Using the elementary inequality

\[ \left( \int_0^\nu \int_0^t g(s) \, ds \, dt \right)^2 \leq \nu \int_0^\nu t \, dt \int_0^t g^2(s) \, ds, \]

valid for any square integrable \( g : [0, \pi] \to \mathbb{R} \) and proved with two applications of Jensen’s inequality, and Lemma A.5.37, we obtain

\[ (f(\nu, g_1) - \mathbb{E}[f(\nu, g_1)])^2 \leq \nu \int_0^\nu t \, dt \int_0^t \left( \sum_{i=1}^6 \mathbb{E} \left[ \Xi_i(s, g_1, g_2) \right] - \mathbb{E} \left[ \Xi_i(s, g_1, g_2) \right] \right)^2 \, ds \, dt. \]

Thus, again by Lemma A.5.37, the Lyapunov inequality, Fubini’s theorem, and compactness of \([0, \pi]\), we have

\[ \text{Var}[f(\nu, \cdot)] \leq \nu \int_0^\nu t \, dt \int_0^t \text{Var} \left[ \sum_{i=1}^6 \mathbb{E} \left[ \Xi_i(s, g_1, g_2) \right] \right] \, ds \, dt. \quad (A.200) \]

We can control the variance under the integral using a combination of Lemmas A.5.35 and A.5.37, together with the deviations control given by Lemmas A.5.39, A.5.41, A.5.42, A.5.43, A.5.44 and A.5.46, since we have chosen \( n \) according to the hypotheses of Lemma A.5.13. In particular,
these lemmas furnish deviation bounds of size at most

\[ C_i \left( \sqrt{\frac{d \log n}{n} + n^{-c_i d}} \right) + C'_i n e^{-c'_i n} \]

that hold with probabilities at least \( 1 - C''_i n^{-c''_i d} - C'''_i n e^{-c'''_i n} \), for any \( d \geq 1 \) larger than an absolute constant and suitable absolute constants specified above. We can simplify these bounds as follows: first, choose \( n \) such that \( n \geq (2/c'''_i) \log n \) for each \( i \), which guarantees that the bounds hold with probability at least \( 1 - C''_i n^{-c''_i d} - C'''_i n e^{-c'''_i n/2} \). Next, choose \( n \geq (2c'_i / c''_i) d \log n \) for all \( i \), which implies that the bounds hold with probability at least \( 1 - 2 \max\{C''_i, C'''_i\} n^{-c''_i d} \). Similarly, we also choose \( n \) such that \( n \geq (2/c'_i) \log n \) for each \( i \), which guarantees that the error terms that are exponential in \( n \) in the bounds are upper bounded by \( C'_i e^{-c'_i n/2} \), and, choose \( n \geq (2c_i / c'_i) d \log n \) for all \( i \), which implies that for all \( i \)

\[ C_i \left( \sqrt{\frac{d \log n}{n} + n^{-c_i d}} \right) + C'_i n e^{-c'_i n} \leq C_i \sqrt{\frac{d \log n}{n}} + 2 \max\{C_i, C'_i\} n^{-c_i d}. \]

Finally, we make the particular choice \( d = 4 / \min_i \{c_i, c''_i\} \), or the minimum required value of \( d \), whichever is larger, so that there are absolute constants \( C, C', C'' > 0 \) such that with probability at least \( 1 - C'''' n^{-4} \) we have for all \( i \)

\[ \left| \mathbb{E}_{g_2} [\Xi_i (\nu, g_1, g_2)] - \mathbb{E}_{g_1, g_2} [\Xi_i (\nu, g_1, g_2)] \right| \leq C \sqrt{\frac{\log n}{n}} + C' n^{-4} \leq 2 C \sqrt{\frac{\log n}{n}}, \]

where the last inequality holds when \( n \) is larger than an absolute constant. With these bounds, we can now invoke Lemma A.5.35 with Lemma A.5.37 to get

\[ \text{Var} \left[ \sum_{i=1}^{6} \mathbb{E}_{g_2} [\Xi_i (s, g_1, g_2)] \right] \leq C \frac{\log n}{n} + C' \frac{n}{n^2} \leq C'' \frac{\log n}{n}, \]

for different absolute constants \( C, C', C'' > 0 \), and where the last inequality again holds \( n \) is larger
than an absolute constant. Plugging back into (A.200) and evaluating the integrals, we get

$$\text{Var}[f(\nu, \cdot)] \leq C\nu^4 \frac{\log n}{n},$$

which is enough to conclude. □

**Lemma A.5.13.** Write

$$Y_\nu(g_1, g_2) = \langle v_0(g_1, g_2), v_\nu(g_1, g_2) \rangle \frac{\psi(\|v_0(g_1, g_2)\|_2)\psi(\|v_\nu(g_1, g_2)\|_2)}{\psi(\|v_0(g_1, g_2)\|_2)\psi(\|v_\nu(g_1, g_2)\|_2)},$$

where $\psi = \psi_{0.25}$ is as in Lemma A.5.31. There exist absolute constants $c, c', C, C' > 0$ and absolute constants $K, K' > 0$ such that for any $d \geq 1$, if $n \geq Kd^4 \log^4 n$ and if $d \geq K'$, then there is an event $E$ such that

1. One has

$$\forall \nu \in [0, \pi], \quad \left| \mathbb{E}_{g_2}[Y_\nu] - \mathbb{E}_{g_1, g_2}(Y_\nu) \right| \leq C'' \nu^2 \sqrt{\frac{d \log n}{n}} + C e^{-cn}$$

if $g_1 \in E$;

2. One has

$$\mathbb{P}[E] \geq 1 - C'n^{-c'd}.$$

**Proof.** Fix $d > 0$, and write

$$f(\nu, g_1) = \mathbb{E}_{g_2}[Y_\nu(g_1, g_2)].$$

Applying Lemma A.5.26, we get for almost all $g_1 \in \mathbb{R}^n$

$$f(\nu, g_1) = \frac{\|v_0\|_2^2}{\psi(\|v_0\|_2)^2} + \int_0^\nu \int_0^t \mathbb{E}_{g_2}[\Xi_1 + \Xi_2 + \Xi_3 + \Xi_4 + \Xi_5 + \Xi_6)(s, g_1, g_2)] ds dt, \quad (A.201)$$

where

$$\Xi_1(s, g_1, g_2) = \sum_{i=1}^n \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot s)}{\psi(\|v_0\|_2)\psi(\|v_i'\|_2)\sin^3 s}$$
\[
\Xi_2(s, g_1, g_2) = \frac{\langle v_0, v_s \rangle \psi'(||v_s||_2)||v_s||_2}{\psi(||v_0||_2)\psi(||v_s||_2)^2} - \frac{\langle v_0, v_s \rangle}{\psi(||v_0||_2)\psi(||v_s||_2)}
\]
\[
\Xi_3(s, g_1, g_2) = -\frac{\langle v_0, v_s \rangle \langle v_s, v_s \rangle \psi''(||v_s||_2)}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2^2}
\]
\[
\Xi_4(s, g_1, g_2) = -2\frac{\langle v_0, \hat{v}_s \rangle \langle v_s, \hat{v}_s \rangle \psi'||v_s||_2}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2}
\]
\[
\Xi_5(s, g_1, g_2) = -\frac{\langle v_0, v_s \rangle \|\hat{v}_s\|^2 \psi'||v_s||_2}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2}
\]
\[
\Xi_6(s, g_1, g_2) = 2\frac{\langle v_0, v_s \rangle \langle v_s, \hat{v}_s \rangle \psi'||v_s||_2}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2} + \frac{\langle v_0, v_s \rangle \langle v_s, \hat{v}_s \rangle \psi'||v_s||_2}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2^3}
\]

Here we put \( \Xi_1(0, g_1, g_2) = \Xi_1(\pi, g_1, g_2) = 0 \), which does not affect the integral and which is equal to the limits \( \lim_{\nu \to 0} \Xi_1(\nu, g_1, g_2) = \lim_{\nu \to \pi} \Xi_1(\nu, g_1, g_2) \) for every \((g_1, g_2)\).

Momentarily ignoring measurability issues, it is of interest to construct \( g_1 \) events \( \mathcal{E}_i \) of suitable probability on which we have

\[
\sup_{\nu \in [0,\pi]} \left| \mathbb{E}_{g_2}[\Xi_i(\nu, g_1, g_2)] - \mathbb{E}_{g_1, g_2}[\Xi_i(\nu, g_1, g_2)] \right| \leq C_i \left( \sqrt{\frac{d\log n}{n}} + n^{-c_i d} \right) + C' n e^{-c_i n} \quad (A.202)
\]

for each \( i = 1, \ldots, 6 \), and a \( g_1 \) event \( \mathcal{E}_7 \) on which we have

\[
\left| \frac{||v_0||_2^2}{\psi(||v_0||_2)^2} - \mathbb{E}_{g_1}[\frac{||v_0||_2^2}{\psi(||v_0||_2)^2}] \right| \leq C' e^{-c_i n}.
\]

We can then consider the event \( \mathcal{E} = \bigcap_{i=1}^7 \mathcal{E}_i \), possibly minus a negligible set on which (A.201) fails to hold, which has high probability via a union bound and on which we have simultaneously for all \( \nu \in [0, \pi] \)

\[
\left| f(\nu, g_1) - \mathbb{E}_{g_1}[f(\nu, g_1)] \right| \leq \left| \frac{||v_0||_2^2}{\psi(||v_0||_2)^2} - \mathbb{E}_{g_1}[\frac{||v_0||_2^2}{\psi(||v_0||_2)^2}] \right|
\]
\[
+ \sum_{i=1}^6 \int_0^\nu \int_0^\nu \left| \mathbb{E}_{g_2}[\Xi_i(s, g_1, g_2)] - \mathbb{E}_{g_1, g_2}[\Xi_i(s, g_1, g_2)] \right| ds \, dt
\]
\[
\leq C \nu^2 \left( \sqrt{\frac{d\log n}{n}} + n^{-c d} \right) + C' n e^{-c n},
\]

287
by Fubini’s theorem and Lemma A.5.37, the triangle inequality (for $|\cdot|$ and for the integral), (A.202), and using $v^2 \leq \pi^2$ and worst-casing the remaining constants.

To establish the bounds (A.202), we will employ lemma Lemma A.5.48, which shows that it is sufficient to obtain pointwise control and show a suitable $s$-Lipschitz property for each $i \in [6]$; following the lemma, these properties also imply Lebesgue measurability of the suprema immediately.

**Reduction to product space events.** Fix $v$. By the triangle inequality, we have for each $i = 1, \ldots, 6$

$$
\left| \mathbb{E}_{g_2} \left[ \Xi_i(v, g_1, g_2) \right] - \mathbb{E}_{g_1, g_2} \left[ \Xi_i(v, g_1, g_2) \right] \right| \leq \mathbb{E}_{g_2} \left[ \Xi_i(v, g_1, g_2) \right] - \mathbb{E}_{g_1, g_2} \left[ \Xi_i(v, g_1, g_2) \right]. \tag{A.203}
$$

Suppose we can construct $(g_1, g_2)$ events $E_i'$ such that

1. If $(g_1, g_2) \in E_i'$, then

$$
\left| \Xi_i(v, g_1, g_2) - \mathbb{E}_{g_1, g_2} \left[ \Xi_i(v, g_1, g_2) \right] \right| \leq C_i \left( \frac{d \log n}{n} + n^{-c_{id}} \right) + C'_i ne^{-c'_i n};
$$

2. One has $\mathbb{P}[E_i'] \geq 1 - C''_i n^{-c''_id} - C'''_i ne^{-c'''_i n}$.

Then for each such $i$, we can write

$$
\mathbb{E}_{g_2} \left[ \Xi_i(v, g_1, g_2) - \mathbb{E}_{g_1, g_2} \left[ \Xi_i(v, g_1, g_2) \right] \right] = \mathbb{E}_{g_2} \left[ \mathbb{1}_{E_i'} + \mathbb{1}_{(E_i')^c} \right] \left| \Xi_i(v, g_1, g_2) - \mathbb{E}_{g_1, g_2} \left[ \Xi_i(v, g_1, g_2) \right] \right| \\
\leq C_i \left( \frac{d \log n}{n} + n^{-c_{id}} \right) + C'_i ne^{-c'_i n} \\
+ \mathbb{E}_{g_2} \left[ \mathbb{1}_{(E_i')^c} \right] \left| \Xi_i(v, g_1, g_2) - \mathbb{E}_{g_1, g_2} \left[ \Xi_i(v, g_1, g_2) \right] \right| \tag{A.204}
$$
using nonnegativity of the integrand and boundedness of the indicator for $E'_i$ in the second line. The random variable remaining in the second line is nonnegative, and by Fubini’s theorem (with Lemma A.5.37 for joint integrability) and the Schwarz inequality we have

$$
\mathbb{E}_{g_1, g_2} \left[ \mathbb{E} \left[ 1_{(E'_i)^c} \right] \Xi_i(v, g_1, g_2) - \mathbb{E} \left[ \Xi_i(v, g_1, g_2) \right] \right] \\
\leq \mathbb{E}_{g_1, g_2} \left[ 1_{(E'_i)^c} \right]^{1/2} \mathbb{E}_{g_1, g_2} \left[ \Xi_i(v, g_1, g_2) - \mathbb{E} \left[ \Xi_i(v, g_1, g_2) \right] \right]^{1/2} \\
\leq C \left( C'' n^{-c''_i d} + C''' n^{-c'''_i n} \right)^{1/2},
$$

where the second line applies Lemma A.5.37 and the Lyapunov inequality. We can replace this last inequality with one equivalent to the measure bound on $\left( (E'_i)^c \right)$ using subadditivity of the square root and reducing the constants $c'_i$ and $c''_i$ by a factor of 2. Using this last inequality, Markov’s inequality implies for any $t \geq 0$

$$
\mathbb{P} \left[ \mathbb{E} \left[ 1_{(E'_i)^c} \right] \Xi_i(v, g_1, g_2) - \mathbb{E} \left[ \Xi_i(v, g_1, g_2) \right] \right] \geq C n^{-\frac{1}{2} c''_i d} + C' n^{1/2} e^{-\frac{1}{2} c'''_i n} \\
\leq C n^{-\frac{1}{2} c''_i d} + C' n^{1/2} e^{-\frac{1}{2} c'''_i n},
$$

which, together with (A.203) and after worst-casing some exponents and constants, implies that there is a $g_1$ event $E_i$ that satisfies (the constants $C$ and $C'$ are scoped across properties 1 and 2)

1. If $g_1 \in E_i$, then

$$
\left| \mathbb{E}_{g_2} \left[ \Xi_i(v, g_1, g_2) \right] - \mathbb{E}_{g_1, g_2} \left[ \Xi_i(v, g_1, g_2) \right] \right| \leq C_i \sqrt{\frac{d \log n}{n}} + (C_i + C) n^{-\frac{1}{2} c_i d} \\
+ (C'_i + C') n e^{-\frac{1}{2} \min(c'_i, c''_i) n};
$$

2. One has $\mathbb{P}[E_i] \geq 1 - C n^{-\frac{1}{2} c'_i d} - C' n e^{-\frac{1}{2} c'''_i n}$.

Thus, we can pass from $(g_1, g_2)$ events to $g_1$ events with only a worsening of constants, and it

289
suffices to construct the events $E'_i$.

Additionally, we can leverage this same framework to pass $\nu$-uniform control from the product space to $g_1$-space. Suppose we can construct $(g_1, g_2)$ events $E'_i$ such that

1. If $(g_1, g_2) \in E'_i$, then

   \[
   \forall \nu \in [0, \pi], \left| \Xi_t(\nu, g_1, g_2) - \mathbb{E}_{g_1, g_2} [\Xi_t(\nu, g_1, g_2)] \right| \leq C_i \left( \sqrt{\frac{d \log n}{n}} + n^{-c_i d} \right) + C_i' n e^{-c'_i n},
   \]

2. One has $\mathbb{P}[E'_i] \geq 1 - C''_i n^{-c''_i d} - C''_i n e^{-c''_i n}$.

Then following (A.204), we can assert

\[
\forall \nu \in [0, \pi], \mathbb{E}_{g_2} \left[ \left| \Xi_t(\nu, g_1, g_2) - \mathbb{E}_{g_1, g_2} [\Xi_t(\nu, g_1, g_2)] \right| \right] \leq C_i \left( \sqrt{\frac{d \log n}{n}} + n^{-c_i d} \right) + C_i' n e^{-c'_i n} + \mathbb{E}_{g_1, g_2} [\mathbf{1}(E'_i) \mathbb{E}_{g_2} [\Xi_t(\nu, g_1, g_2) - \mathbb{E}_{g_1, g_2} [\Xi_t(\nu, g_1, g_2)]]].
\]

To get uniform control of this last random variable, we can use Lemma A.5.37, which tells us that we have a bound

\[
\forall \nu \in [0, \pi], \mathbb{E}_{g_2} \left[ \left| \Xi_t(\nu, g_1, g_2) - \mathbb{E}_{g_1, g_2} [\Xi_t(\nu, g_1, g_2)] \right| \right] \leq C_i \left( \sqrt{\frac{d \log n}{n}} + n^{-c_i d} \right) + C_i' n e^{-c'_i n} + \mathbb{E}_{g_1, g_2} [\mathbf{1}(E'_i) \mathbb{E}_{g_2} \mathbf{1}(E'_i) f_i(g_1, g_2)],
\]

where $f_i$ is in $L^4(\mathbb{R}^n \times \mathbb{R}^n)$, and has $L^4$ norm bounded by an absolute constant $C_i > 0$. Then Fubini’s theorem and the Schwarz inequality allow us to assert

\[
\mathbb{E}_{g_1, g_2} [\mathbf{1}(E'_i) f_i(g_1, g_2)] \leq C_i \mathbb{E}_{g_1, g_2} [\mathbf{1}(E'_i)]^{1/2},
\]

which can be controlled exactly as in the pointwise control argument. In particular, an application
of Markov’s inequality gives

\[
P \left[ E \mathbf{1}(E') e f_i(g_1, g_2) \right] \geq C n^{-\frac{1}{2} c_i'' d} + C' n^{1/2} e^{-\frac{1}{2} c_i'''} n \]

so that, returning to (A.205), we have uniform control of the quantity

\[
| E [\Xi_i(v, g_1, g_2)] - E [\Xi_i(v, g_1, g_2)] |
\]

on an event of appropriately high probability. In particular, we have incurred only losses in the constants compared to the pointwise case.

**Approach to Lipschitz estimates.** We will use this framework for controlling the $\Xi_1$ and $\Xi_5$ terms only. Accordingly, the sections for those terms below will produce results of the following type, for absolute constants $c_i, c'_i, c''_i, C_i, C'_i, C''_i$ and parameters $d \geq 1, \delta > 0$ such that $d$ and $\delta$ are larger than (separate) absolute constants and $n$ satisfies certain conditions involving $d$:

1. For each $\nu \in [0, \pi]$ fixed, with probability at least $1 - C_1'' n^{-c_1'' d} - C_1''' n e^{-c_1''' n}$, we have that

\[
| E [\Xi_i(v, g_1, g_2)] - E [\Xi_i(v, g_1, g_2)] | \leq C_1 \sqrt{d \log n} n + C_1' n^{-c_1 d} + C_1''' n e^{-c_1 n};
\]

2. With probability at least $1 - C_2' n^{-c_2' n} - C_2'' n^{-\delta}$, we have that

\[
| E [\Xi_i(v, g_1, g_2)] - E [\Xi_i(v, g_1, g_2)] |
\]

is $(C_2 + C_2'n^{1+\delta})$-Lipschitz.

We show here that we can use these properties to obtain uniform concentration of the relevant quantities. Write $M = C_1 \sqrt{d \log n} n + C_1' n^{-c_1 d} + C_1''' n e^{-c_1 n}$; we are interested in showing that uniform bounds of sizes close to $M$ hold with probability not much smaller than that of the pointwise
bounds. By Lemma A.5.48, it follows from the assumed properties that for any \(0 < \varepsilon < 1\) one has

\[
P \left[ \sup_{\nu \in [0,\pi]} \left| \mathbb{E}[\Xi_i(\nu, g_1, g_2)] - \mathbb{E}[\Xi_i(\nu, g_1, g_2)] \right| \leq M + \varepsilon \left( C_2 + C'_2 n^{1+\delta} \right) \right] \leq 1 - \left( C''_1 n^{-c''_1 d} + C''_1 n e^{-c''_1 n} \right) K e^{-1} - \left( C''_2 e^{-c''_2 n} + C''_2 n^{-\delta} \right),
\]

where \(K > 0\) is an absolute constant. To make the RHS of the bound on the supremum of size comparable to \(M\), it suffices to choose \(\varepsilon = C_1 \sqrt{d \log n} / (C_2 + C'_2 n^{1+\delta})\). We have \(C_2 + C'_2 n^{1+\delta} \leq K'n^{1+\delta}\) for \(K' > 0\) an absolute constant, and so we have \(\varepsilon^{-1} \leq K'n^{3/2+\delta}\) for \(K' > 0\) another absolute constant. This gives

\[
\left( C''_1 n^{-c''_1 d} + C''_1 n e^{-c''_1 n} \right) \varepsilon^{-1} \leq K'n^{3/2+\delta} \left( C''_1 e^{-c''_1 d \log n} + C''_1 e^{-c''_1 n / 2n} \right)
\]

\[
\leq K'n^{3/2+\delta} e^{-c''_1 d \log n}
\]

\[
\leq K'n^{-c''_1 d / 2},
\]

where \(K' > 0\) is an absolute constant whose value changes from line to line; and where the first inequality assumes that \(n \geq (2/c''_1) \log n\), the second inequality assumes that \(n \geq (2c''_1 / c''_1) d \log n\), and the third assumes that \(\delta \leq c''_1 d / 2 - 3/2\). Choosing \(d\) so that the value \(c''_1 d / 2 - 3/2\) is larger than the minimum value for \(\delta\) (i.e., larger than an absolute constant), then choosing \(\delta = c''_1 d / 2 - 3/2\), and finally choosing \(d \geq 6/c''_1\), we obtain

\[
P \left[ \sup_{\nu \in [0,\pi]} \left| \mathbb{E}[\Xi_i(\nu, g_1, g_2)] - \mathbb{E}[\Xi_i(\nu, g_1, g_2)] \right| \leq 2M \right] \geq 1 - Kn^{-c''_1 d / 2} - C''_2 e^{-c''_2 n} - C''_2 n^{-c''_1 d / 4},
\]

where \(K > 0\) is an absolute constant, which is an acceptable level of uniformization.

**Completing the proof.** To obtain the desired control, we apply the uniform framework for the terms \(\Xi_i, i = 2, 3, 4, 6\); and the pointwise with Lipschitz control framework for the terms \(\Xi_i, i = 1, 5\). We also establish high probability control of the zero-order term in Lemma A.5.38. The events we need for the pointwise framework terms are constructed in Lemmas A.5.39, A.5.40, A.5.44
and A.5.45. The events we need for the uniform framework are constructed in Lemmas A.5.41, A.5.42, A.5.43 and A.5.46. Because \( n \) and \( d \) are chosen appropriately by our hypotheses here, we can invoke each of these lemmas to construct the necessary sub-events and obtain an event \( E \) which satisfies

1. One has
\[
\forall \nu \in [0, \pi], \quad \left| \mathbb{E}_{g_2}[Y_\nu] - \mathbb{E}_{g_1, g_2}[Y_\nu] \right| \leq C\nu^2 \left( \sqrt{\frac{d \log n}{n}} + n^{-cd} \right) + C'ne^{-c'n}
\]
if \( g_1 \in E \);

2. One has
\[
P[E] \geq 1 - C''n^{-c''d} - C'''ne^{-c'''n}.
\]

We can adjust \( d \) and \( n \) slightly to obtain an event with the properties claimed in the statement of the lemma. Indeed, choosing \( n \) to be larger than an absolute constant multiple of \( \log n \), we can obtain \( C'ne^{-c'n} \leq C'e^{-c'n/2} \) and \( C'''ne^{-c'''n} \leq C''''e^{-c''''n/2} \); choosing \( n \) to be larger than an absolute constant multiple of \( d \log n \), we can obtain \( C''n^{-c''d} + C'''e^{-c'''n/2} \leq 2C''n^{-c''d} \); and choosing \( d \) to be larger than an absolute constant, we can assert \( \sqrt{d \log n/n} + n^{-cd} \leq 2\sqrt{d \log n/n} \). This turns the guarantees of \( E \) into the guarantees claimed in the statement of the lemma, and completes the proof.

\[\square\]

**Proving Lemma A.5.7**

**Lemma A.5.14.** One has bounds

\[
1 - \frac{\nu^2}{2} \leq \cos \varphi(\nu) \leq 1 - c\nu^2, \quad \nu \in [0, \pi].
\]

**Proof.** Write \( f(\nu) = \cos \varphi(\nu) = \cos \nu + \pi^{-1}(\sin \nu - \nu \cos \nu) \), where the last equality follows from Lemma A.5.2. We start by obtaining quadratic bounds on \( f(\nu) \) for \( \nu \in [0, 0.1] \). In particular, we
will show

\[ 1 - \frac{1}{2} \nu^2 \leq f(\nu) \leq 1 - \frac{1}{4} \nu^2, \quad \nu \in [0, 0.1]. \tag{A.206} \]

We readily calculate

\[
\begin{align*}
  f'(\nu) &= -\sin \nu + \pi^{-1} \nu \sin \nu, \\
  f''(\nu) &= -\cos \nu + \pi^{-1}(\nu \cos \nu + \sin \nu).
\end{align*}
\]

Taylor expanding at \( \nu = 0 \) gives

\[
1 + \frac{\inf_{t \in [0,0.1]} f''(t)}{2} \nu^2 \leq f(\nu) \leq 1 + \frac{\sup_{t \in [0,0.1]} f''(t)}{2} \nu^2.
\]

We have \( f''(0) = -1 \), and \( \sin \nu \leq \sin 0.1 \) on our interval of interest by monotonicity. The derivative of \( \nu \cos \nu \) is \( \cos \nu - \nu \sin \nu \); \( \nu \sin \nu \) is increasing as the product of two increasing functions (given \( \nu \leq 0.1 \)), and one checks that \( \cos(0.1) - 0.1 \sin(0.1) > 0 \); therefore \( \nu \cos \nu \leq 0.1 \cos(0.1) \) on our domain of interest. One checks numerically

\[
-\cos(0.1) + \pi^{-1}(0.1 \cos(0.1) + \sin(0.1)) < -\frac{1}{2} < 0,
\]

and this establishes \( f(\nu) \leq 1 - \frac{1}{4} \nu^2 \) on \( [0,0.1] \). If \( \nu \leq \pi/2 \), we have \( \cos \geq 0 \) and \( \sin \geq 0 \), so that \( \nu \cos \nu + \sin \nu \geq 0 \) on this domain. This implies \( f''(\nu) \geq -\cos \nu \geq -1 \) for \( 0 \leq \nu \leq \pi/2 \), which proves \( \inf_{t \in [0,\pi/2]} f''(t) = -1 \), and establishes the lower bound on \( [0, \pi/2] \).

To obtain (possibly) looser bounds on \([0, \pi]\), we use a bootstrapping approach. The lower bound is more straightforward; to assert the lower bound on \([0, \pi]\), we evaluate constants numerically to find that the lower bound’s value at \( \pi/2 \) is \( 1 - \pi^2/8 < 0 \), and given that \( f \geq 0 \) by Lemma A.5.5 and the concave quadratic bound is maximized at \( \nu = 0 \), it follows that the bound holds on the entire interval.
For bootstrapping the upper bound, we note that the equation

\[ f'(v) = -\sin v + \pi^{-1} v \sin v = \sin v \left( \frac{v}{\pi} - 1 \right) \]

shows immediately that \( f \) is a strictly decreasing function of \( v \) on \((0, \pi)\). Therefore \( f(v) \leq f(0.1) \) on \([0.1, \pi]\), and so the quadratic function \( v \mapsto 1 - \pi^{-2}(1 - f(0.1))v^2 \), which is lower bounded by \( 1 - v^2/4 \) on \([0, \pi]\) by the fact that both concave quadratic functions are maximized at 0 and the verification \( 1 - \pi^2/4 < 0 \leq f(0.1) \), is an upper bound for \( f \) on all of \([0, \pi]\); so the claim holds with \( c = \pi^{-2}(1 - f(0.1)) \).

\[ \square \]

**Lemma A.5.15.** There exist absolute constants \( c, C, C', C'' > 0 \) such that if \( n \geq C \log n \), then one has

\[ \left| \mathbb{E}_{g_1, g_2}[X_v] - \cos \varphi(v) \right| \leq C' e^{-cn} + C'' v^2/n. \]

**Proof.** Write \( h(v) = \cos \varphi(v) - \mathbb{E}[X_v] \). By Lemmas A.5.24 and A.5.25, we have a second-order Taylor formula

\[ h(v) = h(0) + \int_0^v \left( h'(0) + \int_0^t h''(s) \, ds \right) \, dt. \]

We calculate \( h'(0) = 0 \), since \( \mathbb{E}[\langle v_0, \dot{v}_0 \rangle] = \mathbb{E}[\langle \sigma(g_1), g_2 \rangle] = 0 \), and \( P_{v_0}^1 v_0 = 0 \). We also have \( h(0) = \mathbb{E}[\|v_0\|_2^2] - \mathbb{E}[1_{E_m}] = \mu(E_m^c) \) (writing \( m = 1 \)), so this formula yields

\[ |h(v)| \leq \mu(E_m^c) + \frac{v^2}{2} \esssup_{v' \in [0, \pi]} |h''(v')|, \]

and we see that it suffices to bound \( h'' \). We will use the (Lebesgue-a.e.) expression

\[ |h''(v)| = \left| \mathbb{E}[\langle \dot{v}, \dot{v}_0 \rangle] - \mathbb{E} \left[ 1_{E_m} \left( \frac{1}{\|v\|_2} \left( I - \frac{v_0 v_0^*}{\|v\|_2^2} \right) \dot{v}_v, \frac{1}{\|v\|_2} \left( I - \frac{v_0^* v_0}{\|v_0\|_2^2} \right) \dot{v}_0 \right) \right] \right|. \]

Distributing over the inner product and applying rotational invariance to combine the two cross
terms, then using the triangle inequality, we obtain the bound

\[
|h''(\nu)| \leq \mathbb{E}[\langle \hat{v}_v, \hat{v}_0 \rangle] - \mathbb{E}\left[1_{E_m}\frac{\langle \hat{v}_0, \hat{v}_v \rangle}{\|v_0\|_2\|v_v\|_2}\right]_{\Xi_1(\nu)} + 2\mathbb{E}\left[1_{E_m}\frac{\langle \hat{v}_0, v_v \rangle\langle v_v, \hat{v}_v \rangle}{\|v_0\|_2\|v_v\|_2}\right]_{\Xi_2(\nu)} + \mathbb{E}\left[1_{E_m}\frac{\langle \hat{v}_0, v_0 \rangle\langle v_0, v_v \rangle\langle v_v, \hat{v}_v \rangle}{\|v_0\|_2\|v_v\|_2}\right]_{\Xi_3(\nu)}.
\]

We proceed by giving magnitude bounds for \(\Xi_i(\nu), i = 1, 2, 3\). Because we are working with expectations, it suffices to fix one value \(\nu \in [0, \pi]\) and prove pointwise \(\nu\)-independent bounds; we will exploit this in the sequel to easily define extra good events without having to uniformize, and we will generally suppress the notational dependence of \(\Xi_i\) on \(\nu\) as a result. We will also repeatedly use the fact that we have \(\mu(E_1^c) \leq Cne^{-cn}\) for some absolute constants \(c, C > 0\) by Lemma A.5.16.

We will accrue a large number of additive \(C/n\) and \(C'n^{pm}e^{-cn}\) errors as we bound the \(\Xi_i\) terms; at the end of the proof we will worst-case the constants in each additive error and assert a bound of the form claimed.

**\(\Xi_1\) control.** Let \(E = \{\|v_v\|_2 \leq 2\} \cap \{\|v_0\|_2 \leq 2\}\). By Lemma A.5.17 and a union bound, we have \(\mu(E^c) \leq Cne^{-cn}\). Define an event \(E_1 = E_m \cap E\). The first step is to pass to the control of

\[
\Xi_1 := \mathbb{E}\left[1_{E_1}\langle \hat{v}_v, \hat{v}_0 \rangle\left(1 - \frac{1}{\|v_0\|_2\|v_v\|_2}\right)\right].
\]

The triangle inequality gives

\[
|\Xi_1 - \Xi| \leq \mathbb{E}\left[1_{E_1^c}\langle \hat{v}_v, \hat{v}_0 \rangle\right] + \mathbb{E}\left[1_{E_m\setminus E}\frac{\langle \hat{v}_v, v_v \rangle}{\|v_v\|_2\|v_v\|_2}\right].
\]

The first term is readily controlled from two applications of the Schwarz inequality, a union bound,
and rotational invariance together with Lemma A.5.29:

\[
\mathbb{E} \left[ \mathbf{1}_{\mathcal{E}_1} \langle \hat{v}, \hat{v}_0 \rangle \right] \leq \mathbb{E} \left[ \mathbf{1}_{\mathcal{E}_1} \right]^{1/2} \mathbb{E} \left[ \| \hat{v} \|_2^4 \right]^{1/4} \mathbb{E} \left[ \| \hat{v}_0 \|_2^4 \right]^{1/4}
\]

\[
\leq (\mu(\mathcal{E}_m^c) + Ce^{-cn})^{1/2} \mathbb{E} \left[ \| \hat{v}_0 \|_2^4 \right]^{1/2}
\]

\[
\leq \left( Cne^{-cn} + C'e^{-cn} \right)^{1/2} \left( 1 + \frac{Cn}{n} \right)^{1/2}
\]

\[
\leq Cn^{1/2} e^{-cn},
\]

where in the last line we require \(n\) to be at least the value of a large absolute constant. The calculation is similar for the normalized term, except we also apply the definition of \(\mathcal{E}_m\) to get some extra cancellation:

\[
\mathbb{E} \left[ \mathbf{1}_{\mathcal{E}_m} \mathcal{E} \left( \frac{\langle \hat{v}, \hat{v}_0 \rangle}{\| v \|_2 \| v_0 \|_2} \right) \right] \leq \mathbb{E} \left[ \mathbf{1}_{\mathcal{E}_m} \mathcal{E} \left( \frac{\langle \hat{v}, \hat{v}_0 \rangle}{\| v \|_2 \| v_0 \|_2} \right) \right] \leq 4 \mathbb{E} \left[ \mathbf{1}_{\mathcal{E}_m} \mathcal{E} \left( \frac{\langle \hat{v}, \hat{v}_0 \rangle}{\| v \|_2 \| v_0 \|_2} \right) \right]
\]

\[
\leq 4 \mathbb{E} \left[ \mathbf{1}_{\mathcal{E}} \langle \hat{v}, \hat{v}_0 \rangle \right]
\]

\[
\leq 4 \mathbb{E} \left[ \mathbf{1}_{\mathcal{E}} \right]^{1/2} \mathbb{E} \left[ \| \hat{v} \|_2^4 \right]^{1/4} \mathbb{E} \left[ \| \hat{v}_0 \|_2^4 \right]^{1/4}
\]

\[
\leq C e^{-cn} \left( 1 + \frac{C}{n} \right)^{1/2} \leq Ce^{-cn},
\]

where in the last line we apply our bounds from the first term and use \(n \geq 1\) to obtain the final inequality. Next, Taylor expansion of the smooth convex function \(x \mapsto x^{-1/2}\) on the domain \(x > 0\) about the point \(x = 1\) gives

\[
x^{-1/2} = 1 - \frac{1}{2} (x - 1) + \frac{3}{4} \int_1^x (x - t)t^{-5/2} \, dt. \quad (A.207)
\]

Given that \(\mathcal{E}_m\) guarantees \(\| v \|_2 \geq \frac{1}{2}\), we can apply this to get a bound

\[
\mathbf{1}_{\mathcal{E}_1} \left( 1 - \frac{1}{\| v_0 \|_2 \| v \|_2} \right)
\]

\[
= \mathbf{1}_{\mathcal{E}_1} \left( \frac{1}{2} (\| v_0 \|_2^2 \| v \|_2^2 - 1) - \frac{3}{4} \int_1^\infty (\| v_0 \|_2^2 \| v \|_2^2 - t) \, dt^{-5/2} \right).
\]

297
On $\mathcal{E}'$, we also have $\|v_0\|_2^{-2} \|v_r\|_2^{-2} \leq 2^4$, so we can control the integral residual as

$$0 \leq \frac{3}{4} E_1 \int_1^{\|v_0\|_2^{2}} \left( \|v_0\|_2^2 \|v_r\|_2^2 - t \right) t^{-5/2} dt \leq \frac{1}{E_1} \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right)^2,$$

where we replace the tighter bound that we get in the case $\|v_0\|_2^2 \|v_r\|_2^2 \geq 1$ with the worst-case bound from the other case. This gives bounds

$$1 E_1 \left( \frac{1}{2} \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right) - 384 \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right)^2 \right) \leq 1 E_1 \left( 1 - \frac{1}{\|v_0\|_2 \|v_r\|_2} \right) \leq \frac{1}{E_1} \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right).$$

Given that $\|\psi_r\|_2 \leq 2$ on $\mathcal{E}''$, it follows $|\langle \psi_0, \psi_r \rangle| \leq 4$ on $E_1$, so that $\langle \psi_0, \psi_r \rangle + 4 \geq 0$ here. Writing

$$1 E_1 \langle \psi_0, \psi_r \rangle \left( 1 - \frac{1}{\|v_0\|_2 \|v_r\|_2} \right) = 1 E_1 \left( \langle \psi_0, \psi_r \rangle + 4 \right) \left( 1 - \frac{1}{\|v_0\|_2 \|v_r\|_2} \right) - 4 \left( 1 - \frac{1}{\|v_0\|_2 \|v_r\|_2} \right),$$

we can apply nonnegativity to obtain upper and lower bounds

$$\overline{E}_1 \leq E \left[ 1 E_1 \langle \psi_0, \psi_r \rangle \left( \frac{1}{2} \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right) + 4 C \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right)^2 \right) \right],$$

$$\underline{E}_1 \geq E \left[ 1 E_1 \langle \psi_0, \psi_r \rangle \left( \frac{1}{2} \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right) - 5 C \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right)^2 \right) \right],$$

where $C = 384$.

We continue with bounding the quadratic term arising in the previous equation. We have

$$E \left[ 1 E_1 \langle \psi_0, \psi_r \rangle \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right)^2 \right] \leq 4 E \left[ \left( \|v_0\|_2^2 \|v_r\|_2^2 - 1 \right)^2 \right]$$

$$= 4 E \left[ \|v_0\|_2^4 \|v_r\|_2^4 - 2 \|v_0\|_2^2 \|v_r\|_2^2 + 1 \right]$$

$$\leq 4 \left( 1 - 2 E \left[ \|v_0\|_2^2 \|v_r\|_2^2 \right] + E \left[ \|v_0\|_2^8 \right] \right)$$

$$\leq 4 \left( 1 - 2 (1 - (C n^{-1} + C' e^{-cn})^2) \right) \left( 1 + \frac{C''}{n} \right)$$

298
\[ \leq Cn^{-1}e^{-cn} + C' e^{-c'n} + \frac{C''}{n}. \]

The first inequality applies the triangle inequality for the integral, the definition of \( \mathcal{E}_1 \) and Cauchy-Schwarz, then drops the indicator for \( \mathcal{E}_1 \) because the remaining terms are nonnegative; the second line is just distributing; the third line rearranges and applies the Schwarz inequality; and the fourth inequality applies Jensen’s inequality and Lemma A.5.18 to control the second term (to apply this lemma, we need to choose \( n \) larger than an absolute constant; we assume this is done), and Lemma A.5.29 to control the third term. Since \( n \geq 1 \), this gives a \( C/n + C' e^{-cn} \) bound on the quadratic term.

Next is the linear term; our first step will be to get rid of the indicator. By the triangle inequality, it suffices to get control of the corresponding term with the indicator for \( \mathcal{E}_1^c \) instead; we control it as follows:

\[
\left| \mathbb{E} \left[ \mathbb{1}_{\mathcal{E}_1^c} \langle \hat{\psi}_0, \hat{\psi}_\nu \rangle \left( \| \hat{\psi}_0 \|_2^2 \| \psi_\nu \|_2^2 - 1 \right) \right] \right| \\
\leq \mathbb{E} \left[ \mathbb{1}_{\mathcal{E}_1^c} \right]^{1/2} \mathbb{E} \left[ \langle \hat{\psi}_0, \hat{\psi}_\nu \rangle^2 \left( \| \hat{\psi}_0 \|_2^2 \| \psi_\nu \|_2^2 - 1 \right)^2 \right]^{1/2} \\
\leq \left( Cn e^{-cn} + C' e^{-c'n} \right)^{1/2} \mathbb{E} \left[ \| \hat{\psi}_0 \|_2^2 \| \psi_\nu \|_2^2 \left( \| \hat{\psi}_0 \|_2^2 \| \psi_\nu \|_2^2 - 1 \right)^2 \right]^{1/2} \\
\leq \left( Cn e^{-cn} + C' e^{-c'n} \right)^{1/2} \mathbb{E} \left[ \| \hat{\psi}_0 \|_2^8 \| \psi_\nu \|_2^4 \right]^{1/4} \mathbb{E} \left[ \| \hat{\psi}_0 \|_2^6 \| \psi_\nu \|_2^4 \right]^{1/4} + \mathbb{E} \left[ \| \hat{\psi}_0 \|_2^4 \| \psi_\nu \|_2^4 \right]^{1/2} \\
\leq \left( Cn^{-cn} + C' e^{-c'n} \right)^{1/2} \left( \left( 1 + \frac{C_1}{n} \right)^{1/4} \left( 1 + \frac{C_2}{n} \right) + \left( 1 + \frac{C_3}{n} \right)^{1/2} \right) \\
\leq Cn^{1/2} e^{-cn} + C' e^{-c'n}.
\]

The first line is the Schwarz inequality; the second line is the good event measure bound and Cauchy-Schwarz; the third line distributes and drops the cross term, given that all factors are nonnegative; the fourth line applies subadditivity of the square root function, then the Schwarz inequality to the resulting separate terms; the fifth line applies Lemma A.5.29; and the last line
again uses square root subadditivity and treats the remaining terms as multiplicative constants, since \( n \geq 1 \). Therefore passing to the linear term without the indicator incurs only an additional exponential factor. Proceeding, we drop the indicator and distribute to get for the linear term

\[
E \langle \hat{v}_0, \hat{v}_\nu \rangle \left( \|v_0\|_2^2 \|v_\nu\|_2^2 - 1 \right) = E \left[ \langle \hat{v}_0, \hat{v}_\nu \rangle \|v_0\|_2^2 \|v_\nu\|_2^2 \right] - E \left[ \langle \hat{v}_0, \hat{v}_\nu \rangle \right];
\]

it is of interest to apply Lemma A.5.30 to these two terms to get the proper cancellation, and for this we just need to check that the coordinates of each factor in the product have subexponential moment growth with the proper rate. For even powers of \( \ell^2 \) norms of \( v_\nu \), this follows immediately from Lemma A.6.11 after scaling by \( \sqrt{2/n} \); for the inner product term, the coordinate functions are \( \hat{\sigma}(g_{1i})g_{2i}\hat{\sigma}(g_{1i}\cos \nu + g_{2i}\sin \nu)(g_{2i}\cos \nu - g_{1i}\sin \nu) \), and we have from the Schwarz inequality and rotational invariance

\[
E \left[ |\hat{\sigma}(g_{1i})g_{2i}\hat{\sigma}(g_{1i}\cos \nu + g_{2i}\sin \nu)(g_{2i}\cos \nu - g_{1i}\sin \nu)|^k \right] \leq E \left[ \hat{\sigma}(g_{1i})g_{2i}^{2k} \right],
\]

which has subexponential moment growth with rate \( Cn^{-1} \) by Lemma A.5.17 and Lemma A.6.11 after rescaling by \( \sqrt{2/n} \). These formulas also show that when \( k = 1 \), we have a bound of precisely \( n^{-1} \). This makes Lemma A.5.30 applicable, so we can assert bounds

\[
\left| E \left[ \langle \hat{v}_0, \hat{v}_\nu \rangle \left( \|v_0\|_2^2 \|v_\nu\|_2^2 - 1 \right) \right] - \left( n^3 E[ (\hat{v}_0)_1(\hat{v}_\nu)_1] E[ (\sigma(g_{11}))^2 \right] - n E[ (\hat{v}_0)_1(\hat{v}_\nu)_1] \right| \leq \frac{C}{n}
\]

Because \( E[ (\sigma(g_{11}))^2 \] = \( n^{-2} \), this is enough to conclude a \( C/n \) bound on the magnitude of the linear term. Thus, in total, we have shown

\[
|\Xi_1| \leq \frac{C}{n} + C' e^{-\epsilon n} + C'' n^{1/2} e^{-\epsilon' n},
\]

where we combine the different constant that appear in the various exponential additive errors throughout our work by choosing the largest magnitude scaling factor and the smallest magnitude.
constant in the exponent to assert the previous expression.

**Ξ₂ control.** The approach is similar to what we have used to control Ξ₁. We start with exactly the same $E_1$ event definition, and as previously define

$$\Xi_2 = \mathbb{E} \left[ I_{E_1} \frac{\langle \dot{v}_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \|v_0\|_2^2}{\|v_0\|_2^3 \|v_\nu\|_2^3} \right],$$

and then calculating

$$|\Xi_2 - \Xi_2| = \mathbb{E} \left[ I_{E_1 \setminus E} \frac{\langle \dot{v}_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \|v_0\|_2^2}{\|v_0\|_2^3 \|v_\nu\|_2^3} \right]$$

$$\leq 2^6 \mathbb{E} \left[ I_{E_1 \setminus E} \|\langle \dot{v}_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \|v_0\|_2^2 \right]$$

$$\leq 2^6 \mathbb{E} \left[ I_{E_1} \langle \dot{v}_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \|v_0\|_2^2 \right]$$

$$\leq 2^6 \mathbb{E} \left[ I_{E_1} \langle \dot{v}_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \|v_0\|_2^2 \right]$$

$$\leq 2^6 \mathbb{E} \left[ I_{E_1} \right]^{1/2} \mathbb{E} \left[ \langle \dot{v}_0, v_\nu \rangle^4 \right]^{1/4} \mathbb{E} \left[ \langle v_\nu, \dot{v}_\nu \rangle^8 \right]^{1/8} \mathbb{E} \left[ \|v_0\|_2^8 \right]^{1/8}$$

$$\leq 2^6 \mathbb{E} \left[ I_{E_1} \right]^{1/2} \mathbb{E} \left[ \|\dot{v}_0\|_2^8 \right]^{1/8} \mathbb{E} \left[ \|v_\nu\|_2^8 \right]^{1/8} \mathbb{E} \left[ \|\dot{v}_\nu\|_2^{16} \right]^{1/16} \mathbb{E} \left[ \|v_\nu\|_2^{16} \right]^{1/16} \mathbb{E} \left[ \|v_0\|_2^{16} \right]^{1/16}$$

$$\leq Ce^{-cn} + C'n^{1/2}e^{-c'n},$$

using the same ideas as in the previous section, plus several applications of the Schwarz inequality and a final application of Lemma A.5.29. We can therefore pass to $\Xi_2$ with a small additive error. Next, we Taylor expand in the same way as previously, except that larger powers in the denominator force the constant in our residual bound to be $3 \cdot 2^{27}$, and the event $E_1$ now gives us a bound $|\langle \dot{v}_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \|v_0\|_2^2| \leq 2^6$ on the numerator, which we add and subtract as before to exploit nonnegativity. We get

$$\Xi_2 \leq \mathbb{E} \left[ I_{E_1} \langle \dot{v}_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \|v_0\|_2^2 \left( \frac{1}{2} \left( 3 - \|v_0\|_2^6 \|v_\nu\|_2^6 \right) \right) \right]$$

$$\Xi_2 \geq \mathbb{E} \left[ I_{E_1} \langle \dot{v}_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \|v_0\|_2^2 \left( \frac{1}{2} \left( 3 - \|v_0\|_2^6 \|v_\nu\|_2^6 \right) \right) \right].$$
with \( C = 3 \cdot 2^{27} \). Proceeding to control the quadratic term, we have

\[
\left| \mathbb{E} \left[ 1_{\mathcal{E}_i} \langle \dot{v}_0, v_r \rangle \langle v_r, \dot{v}_r \rangle \|v_0\|_2 \left( \|v_0\|_2^6 \|v_r\|_2^6 - 1 \right)^2 \right] \right|
\leq 4^3 \mathbb{E} \left( \|v_0\|_2^6 \|v_r\|_2^6 - 1 \right)^2
\leq 2^6 \mathbb{E} \left[ \|v_0\|_2^6 \|v_r\|_2^6 - 2 \|v_0\|_2^6 \|v_r\|_2^6 + 1 \right]
\leq 2^6 \left( 1 - 2 \mathbb{E} \left[ \|v_0\|_2^6 \|v_r\|_2^6 \right] + \mathbb{E} \left[ \|v_0\|_2^{24} \right] \right)
\leq 2^6 \left( 1 - 2(1 - (Cn^{-1} + C'e^{-cn}))^6 + (1 + C''n^{-1}) \right)
\leq Cn^{-1} + \sum_{k=1}^{3} \left( 6 \left( 2k - 1 \right) \right) \left( C'n^{-1} + C'' e^{-cn} \right)^{2k-1}
\leq Cn^{-1} + C' \sum_{k=1}^{3} \sum_{j=0}^{2k-1} n^{-(2k-1-j)} e^{-cnj}
\leq Cn^{-1} + C'e^{-cn}.
\]

The justifications for the first four lines are identical to those of the previous section. In the last three lines, we use the binomial theorem twice to expand the sixth power term, and we assert the final line by the fact that \( k > 0 \), so that each term in the sum corresponding to a \( j = 0 \) has a positive inverse power of \( n \) attached, and when \( j = 2k - 1 \) we pick up an exponential factor. Moving on to the linear term, as in the previous section we start by dropping the indicator. We control the residual as follows:

\[
\left| \mathbb{E} \left[ 1_{\mathcal{E}_i} \langle \dot{v}_0, v_r \rangle \langle v_r, \dot{v}_r \rangle \|v_0\|_2 \left( \|v_0\|_2^6 \|v_r\|_2^6 - 3 \right) \right] \right|
\leq \mathbb{E} \left[ 1_{\mathcal{E}_i} \right]^{1/2} \mathbb{E} \left[ \|\dot{v}_0\|_2^6 \|\dot{v}_r\|_2^6 \|v_0\|_2^4 \left( \|v_0\|_2^6 \|v_r\|_2^6 - 3 \right)^2 \right]^{1/2}
\leq \mathbb{E} \left[ 1_{\mathcal{E}_i} \right]^{1/2} \mathbb{E} \left[ \|\dot{v}_0\|_2^2 \|\dot{v}_r\|_2^2 \|v_0\|_2^4 \left( \|v_0\|_2^6 \|v_r\|_2^6 + 3 \|\dot{v}_0\|_2^2 \|v_0\|_2^2 \|v_0\|_2^2 \|v_r\|_2^4 \right) \right]^{1/2}
\leq \mathbb{E} \left[ 1_{\mathcal{E}_i} \right]^{1/2} \left( \mathbb{E} \left[ \|\dot{v}_0\|_2^2 \|\dot{v}_r\|_2^2 \|v_0\|_2^4 \|v_0\|_2^2 \|v_0\|_2^2 \|v_r\|_2^4 \right]^{1/2} + 3 \mathbb{E} \left[ \|\dot{v}_0\|_2^2 \|\dot{v}_r\|_2^2 \|v_0\|_2^2 \|v_0\|_2^2 \|v_r\|_2^4 \right]^{1/2} \right)
\leq C e^{-cn} + C'n^{1/2} e^{-cn}.
\]
The justifications are almost the same as the previous section, although we have compressed some steps into fewer lines here and we have omitted the final simplifications which follow from applying the Schwarz inequality to each of the two expectations in the second-to-last line 3 times and then applying Lemma A.5.29. Dropping the indicator and distributing now gives:

$$\mathbb{E}\left[\langle \hat{v}_0, v_r \rangle \langle v_r, \hat{v}_r \rangle \|v_0\|_2^2 \left(\|v_0\|_2^6 \|v_r\|_2^6 - 3\right)\right] = \mathbb{E}\left[\langle \hat{v}_0, v_r \rangle \langle v_r, \hat{v}_r \rangle \|v_0\|_2^2 \|v_0\|_2^6 \|v_r\|_2^6\right] - 3 \mathbb{E}\left[\langle \hat{v}_0, v_r \rangle \langle v_r, \hat{v}_r \rangle \|v_0\|_2^2\right];$$

to apply Lemma A.5.30, we check the two new coordinate functions that appear in this linear term: for \(\langle \hat{v}_0, v_r \rangle\), we have

$$\mathbb{E}\left[|\hat{\sigma}(g_{1i})g_{2i}\sigma(g_{1i} \cos \nu + g_{2i} \sin \nu)|^k\right] \leq \mathbb{E}\left[|\hat{\sigma}(g_{1i})g_{2i}^{2k}|^{1/2}\right] \mathbb{E}\left[|\sigma(g_{1i})^{2k}|^{1/2}\right], \quad (A.208)$$

and for \(\langle \hat{v}_r, v_r \rangle\), we have likewise

$$\mathbb{E}\left[|\sigma(g_{1i} \cos \nu + g_{2i} \sin \nu)(g_{2i} \cos \nu - g_{1i} \sin \nu)|^k\right] \leq \mathbb{E}\left[|\hat{\sigma}(g_{1i})g_{2i}^{2k}|^{1/2}\right] \mathbb{E}\left[|\sigma(g_{1i})^{2k}|^{1/2}\right], \quad (A.209)$$

both by the Schwarz inequality and rotational invariance. As before, an appeal to Lemmas A.6.11 and A.5.17 implies that these two coordinate functions satisfy the hypotheses of Lemma A.5.30, so we have a bound

$$\left|\mathbb{E}\left[\langle \hat{v}_0, v_r \rangle \langle v_r, \hat{v}_r \rangle \|v_0\|_2^2 \left(\|v_0\|_2^6 \|v_r\|_2^6 - 3\right)\right] - n^9 \mathbb{E}\left[\langle \hat{v}_0 \rangle_1 \langle v_r \rangle_1\right] \mathbb{E}\left[\langle v_r \rangle_1 \langle \hat{v}_r \rangle_1\right] \mathbb{E}\left[\sigma(w_{11})^2\right]^7 + 3n^3 \mathbb{E}\left[\langle \hat{v}_0 \rangle_1 \langle v_r \rangle_1\right] \mathbb{E}\left[\langle v_r \rangle_1 \langle \hat{v}_r \rangle_1\right] \mathbb{E}\left[\sigma(w_{11})^2\right]\right| \leq \frac{C}{n}.$$

Noticing that

$$\mathbb{E}[\langle v_r, \hat{v}_r \rangle] = -\mathbb{E}[\langle v_0, \hat{v}_0 \rangle] = -\mathbb{E}[\langle \sigma(g_1), g_2 \rangle] = 0,$$

by rotational invariance and independence, we conclude by identically-distributedness of the coor-
dinates of $v_\nu$ and $\dot{v}_\nu$

$$n^9 \mathbb{E}[(\dot{v}_0)_1(v_\nu)_1] \mathbb{E}[(\dot{v}_\nu)_1(v_\nu)_1] \mathbb{E}[\sigma(g_{11})^2] \left(1 - 3n^3 \mathbb{E}[(\dot{v}_0)_1(\dot{v}_\nu)_1] \mathbb{E}[(\dot{v}_\nu)_1(v_\nu)_1] \mathbb{E}[\sigma(g_{11})^2] \right) = 0,$$

which establishes the desired control on $\Xi_2$. Thus, in total, we have shown

$$|\Xi_2| \leq C_n + C' e^{-cn} + C'' n^{1/2} e^{-c'n},$$

where we combine the different constant that appear in the various exponential additive errors throughout our work by choosing the largest magnitude scaling factor and the smallest magnitude constant in the exponent to assert the previous expression.

$\Xi_3$ control. The argument for control of this term is very similar to the previous section, since the degrees of the denominators now match. We start by defining

$$\Xi_3 = \mathbb{E} \left[ \mathbb{I}_{E_1} \frac{\langle \dot{v}_0, v_0 \rangle \langle v_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle}{\|v_0\|^3 \|v_\nu\|^3} \right],$$

with the same $E_1$ event as previously, and then calculating

$$|\Xi_3 - \Xi_3| = |\mathbb{E} \left[ \mathbb{I}_{E_{m1} \setminus E} \frac{\langle \dot{v}_0, v_0 \rangle \langle v_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle}{\|v_0\|^3 \|v_\nu\|^3} \right]| \leq 2^6 \mathbb{E} \left[ \mathbb{I}_{E_{m1} \setminus E} \frac{\langle \dot{v}_0, v_0 \rangle \langle v_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle}{\|v_0\|^3 \|v_\nu\|^3} \right]$$

$$\leq 2^6 \mathbb{E} \left[ \mathbb{I}_{E_{m1} \setminus E} \langle \dot{v}_0, v_0 \rangle \langle v_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \right]$$

$$\leq 2^6 \mathbb{E} \left[ \mathbb{I}_{E_{m1} \setminus E} \langle \dot{v}_0, v_0 \rangle \langle v_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle \right]$$

$$\leq 2^6 \mathbb{E} \left[ \mathbb{I}_{E_{m1} \setminus E} \langle \dot{v}_0, v_0 \rangle^4 \right]^{1/4} \mathbb{E} \left[ \langle v_0, v_\nu \rangle^8 \right]^{1/8} \mathbb{E} \left[ \langle v_\nu, \dot{v}_\nu \rangle^8 \right]^{1/8}$$

$$\leq 2^6 \mathbb{E} \left[ \mathbb{I}_{E_{m1} \setminus E} \langle \dot{v}_0, v_0 \rangle^4 \right]^{1/4} \mathbb{E} \left[ \langle v_0, v_\nu \rangle^8 \right]^{1/8} \mathbb{E} \left[ \langle v_\nu, \dot{v}_\nu \rangle^8 \right]^{1/8}$$

$$\leq C n^{1/2} e^{-cn} + C e^{-c'n},$$

using the same ideas as in the previous section. We can therefore pass to $\Xi_3$ with an exponentially
small error. Next, we Taylor expand in the same way as previously, obtaining

\[
\tilde{\Xi}_3 \leq E \left[ 1_{\mathcal{E}_1} \langle \dot{v}_0, v_0, \dot{v}_v, v_v \rangle \left( \frac{1}{2} \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 \right) + (4^3 + 1)C \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 - 1 \right) \right) \right];
\]

\[
\tilde{\Xi}_3 \geq E \left[ 1_{\mathcal{E}_1} \langle \dot{v}_0, v_0, \dot{v}_v, v_v \rangle \left( \frac{1}{2} \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 \right) - 4^3C \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 - 1 \right) \right) \right],
\]

with \( C = 3 \cdot 2^{27} \). Proceeding to control the quadratic term, we notice

\[
\left| E \left[ 1_{\mathcal{E}_1} \langle \dot{v}_0, v_0, \dot{v}_v, v_v \rangle \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 - 1 \right) \right] \right| \leq 4^3E \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 - 1 \right) \]

\[
= Cn^{-1} + C'e^{-cn},
\]

since the final term was controlled in the previous section. Moving on to the linear term, as in the previous section we start by dropping the indicator. We control the residual as follows:

\[
\left| E \left[ 1_{\mathcal{E}_1^c} \langle \dot{v}_0, v_0, \dot{v}_v, v_v \rangle \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 - 3 \right) \right] \right|
\]

\[
\leq E \left[ 1_{\mathcal{E}_1^c} \right]^{1/2} E \left[ \left\| \dot{v}_0 \right\|_2^2 \left\| \dot{v}_v \right\|_2^2 \left\| v_0 \right\|_2^4 \left\| v_v \right\|_2^4 \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 - 3 \right) \right]^{1/2}
\]

\[
\leq E \left[ 1_{\mathcal{E}_1^c} \right]^{1/2} E \left[ \left\| \dot{v}_0 \right\|_2^2 \left\| \dot{v}_v \right\|_2^2 \left\| v_0 \right\|_2^{16} \left\| v_v \right\|_2^{16} + 3 \left\| \dot{v}_0 \right\|_2^2 \left\| \dot{v}_v \right\|_2^2 \left\| v_0 \right\|_2^4 \left\| v_v \right\|_2^4 \right]^{1/2}
\]

\[
\leq E \left[ 1_{\mathcal{E}_1^c} \right]^{1/2} \left( E \left[ \left\| \dot{v}_0 \right\|_2^2 \left\| \dot{v}_v \right\|_2^2 \left\| v_0 \right\|_2^{16} \left\| v_v \right\|_2^{16} \right]^{1/2} + 3E \left[ \left\| \dot{v}_0 \right\|_2^2 \left\| \dot{v}_v \right\|_2^2 \left\| v_0 \right\|_2^4 \left\| v_v \right\|_2^4 \right]^{1/2} \right)
\]

\[
\leq C'e^{-cn} + C'n^{1/2}e^{-c'n},
\]

by the same argument as in the previous section. Dropping the indicator and distributing now gives:

\[
E \left[ \langle \dot{v}_0, v_0 \rangle \langle \dot{v}_v, v_v \rangle \left( \left\| v_0 \right\|_2^6 \left\| v_v \right\|_2^6 - 3 \right) \right] = E \left[ \langle \dot{v}_0, v_0 \rangle \langle \dot{v}_v, v_v \rangle \left| v_0 \right|_2^6 \left| v_v \right|_2^6 \right] - 3E \left[ \langle \dot{v}_0, v_0 \rangle \langle \dot{v}_v, v_v \rangle \langle v_0 \rangle \langle v_v \rangle \right];
\]

to apply Lemma A.5.30, we check the one new coordinate function that appears in this linear term:

305
for \( \langle \nu_0, \nu \rangle \), we have

\[
E \left[ |\sigma(g_{1i}) \sigma(g_{1i} \cos \nu + g_{2i} \sin \nu)|^k \right] \leq E \left[ |\sigma(g_{1i})|^{2k} \right].
\] (A.210)

by the Schwarz inequality and rotational invariance. As before, an appeal to Lemmas A.6.11 and A.5.17 implies that this coordinate function satisfies the hypotheses of Lemma A.5.30, so we have a bound

\[
\begin{align*}
&\left| E \left[ \langle \hat{\nu}_0, \nu_0 \rangle \langle \hat{\nu}_0, \nu \rangle \langle \nu_0, \nu \rangle \left( \|\nu_0\|^6_2 \|\nu\|^6_2 - 3 \right) \right] \\
&\quad - n^9 E[\langle \hat{\nu}_0 \rangle_1 (\nu_0)_1] E[\langle \hat{\nu}_0 \rangle_1 (\nu_0)_1] E[\langle \nu_0 \rangle_1 (\nu_0)_1] E\left[ \sigma(g_{1i})^2 \right]^6 \right| \\
&\quad + 3n^3 E[\langle \hat{\nu}_0 \rangle_1 (\nu_0)_1] E[\langle \hat{\nu}_0 \rangle_1 (\nu_0)_1] E[\langle \nu_0 \rangle_1 (\nu_0)_1] \\
&\quad \leq n^{-1}.
\end{align*}
\]

As in the previous section, using that \( E[\langle \nu, \hat{\nu} \rangle] = 0 \) then allows us to conclude the desired control on \( \Xi_3 \). Thus, in total, we have shown

\[
|\Xi_3| \leq C \frac{n}{n} + C' e^{-c/n} + C'' n^{1/2} e^{-c'n},
\]

where we combine the different constant that appear in the various exponential additive errors throughout our work by choosing the largest magnitude scaling factor and the smallest magnitude constant in the exponent to assert the previous expression.

To wrap up, we take the largest of the scaling constants in the estimates we have derived, and the smallest of the constants-in-the-exponent that we have derived, in order to assert

\[
|h''(\nu)| \leq \frac{C}{n} + C' n^{1/2} e^{-c/n}.
\]

Matching constants in the exponent and choosing \( n \) larger than an absolute constant multiple of \( \log n \), it follows

\[
|h(\nu)| \leq C e^{-c/n} + C' \frac{\nu^2}{n},
\]
which was to be proved. □

General Properties

Lemma A.5.16. Consider the event

\[
\mathcal{E}_{c,m} = \bigcap_{S \subseteq [n]} \bigcap_{|S| = m} \{ (g_1, g_2) \mid c \leq \| I_{S^c} v (g_1, g_2) \|_2 \leq c^{-1} \}.
\]

Suppose \( n \geq \max\{2m, m + 20\} \). Then we have the following properties:

1. \( \mu(\mathcal{E}_{c,m}^c) \leq Cn^m e^{-c'n} \),

2. We have \( \mathcal{E}_{c,m} = \mathcal{E}_{c,m} Q \) for every \( Q \in O(2) \), so that in particular \( 1_{\mathcal{E}_{c,m}}(GQ) = 1_{\mathcal{E}_{c,m}}(G) \).

Above, \( O(n) \) denotes the set of \( n \times n \) orthogonal matrices.

Proof. We will show the second property first. For each \( c > 0 \), if \( Q \in O(2) \), notice that

\[
\mathcal{E}_{c,m} Q = \bigcap_{S \subseteq [n]} \bigcap_{|S| = m} \left\{ GQ \mid c \leq \left\| I_{S^c} \sigma \left( G \begin{bmatrix} \cos \nu \\ \sin \nu \end{bmatrix} \right) \right\|_2 < c^{-1} \right\}
\]

\[
= \bigcap_{S \subseteq [n]} \bigcap_{|S| = m} \left\{ G \mid c \leq \left\| I_{S^c} \sigma (GQ^* u) \right\|_2 < c^{-1} \right\}
\]

\[
= \mathcal{E}_{c,m},
\]

since the vector \([\cos \nu, \sin \nu]^* \in S^1\), and \( O(2) \) acts transitively on \( S^1 \). This proves the second property when \( c > 0 \); the result for \( c = 0 \) is obtained by applying the preceding argument to each set in the infinite union defining the \( c = 0, m \) event.

For the measure bound, we observe that \( \mathcal{E}_{c,m} \subset \mathcal{E}_{c',m} \) if \( c \geq c' \), so it suffices to bound the measure of the complement for the particular choice \( c = \frac{1}{2} \). We start by controlling pointwise the
measure of the complement of the event

\[ \mathcal{E}_{0.6,m,u} = \bigcap_{S \subseteq [n], |S| = m} \{ G \mid 0.6 < \| I_{S^c} \sigma (Gu) \|_2 < 5/3 \} \]

for each \( u \in S^1 \), then uniformize over the one-dimensional manifold \( S^1 \); we need to begin with \( c = 0.6 \) instead of \( c = \frac{1}{2} \) to survive some loosening of the bounds when we uniformize. We have

\[ \mathcal{E}_{0.6,m,u}^c = \bigcup_{S \subseteq [n], |S| = m} \{ G \mid \| I_{S^c} \sigma (Gu) \|_2 \leq 0.6 \} \cup \{ G \mid \| I_{S^c} \sigma (Gu) \|_2 \geq 5/3 \} \]

so that a union bound implies

\[
\mu \left( \mathcal{E}_{0.6,m,u}^c \right) \leq \sum_{S \subseteq [n], |S| = m} \mathbb{P} \left[ \| I_{S^c} \sigma (Gu) \|_2 \leq 0.6 \right] + \mathbb{P} \left[ \| I_{S^c} \sigma (Gu) \|_2 \geq 5/3 \right] \\
\leq \left( \frac{n}{m} \right) \left( \mathbb{P} \left[ \| I_{[m]^c} \sigma (g_1) \|_2 \leq 0.6 \right] + \mathbb{P} \left[ \| I_{[m]^c} \sigma (g_1) \|_2 \geq 5/3 \right] \right),
\]

(A.211)

where the final inequality follows from right-rotational invariance of \( \mu \) and the fact that the coordinates of \( g_1 \) are identically distributed. Let \( \tilde{g} \in \mathbb{R}^{n-m} \) be distributed as \( \mathcal{N}(0, (2/n)I) \), so that \( \sigma(\tilde{g}) \) has the same distribution as \( I_{[m]^c} \sigma (g_1) \) By Gauss-Lipschitz concentration [228, Theorem 5.6], we have

\[
\mathbb{P}[\| \sigma(\tilde{g}) \|_2 \geq \mathbb{E}[\| \sigma(\tilde{g}) \|_2] + t] \leq e^{-ct^2}, \quad \mathbb{P}[\| \sigma(\tilde{g}) \|_2 \leq \mathbb{E}[\| \sigma(\tilde{g}) \|_2] - t] \leq e^{-ct^2},
\]

since \( \sigma \) is 1-Lipschitz and nonnegative homogeneous. After rescaling, we apply Lemma A.5.19 to get

\[
\sqrt{1 - \frac{m}{n}} - \frac{2}{\sqrt{n} \sqrt{n-m}} \leq \mathbb{E}[\| \sigma(\tilde{g}) \|_2] \leq \sqrt{1 - \frac{m}{n}} \leq 1.
\]

308
Plugging these estimates into the Gauss-Lipschitz bounds gives

\[ \mathbb{P}[\|\sigma(\tilde{g})\|_2 \geq 1 + t] \leq e^{-cn^2}, \quad \mathbb{P}\left[\|\sigma(\tilde{g})\|_2 \leq \sqrt{1 - \frac{m}{n} - \frac{2}{n\sqrt{n-m}}} - t \right] \leq e^{-cn^2}. \]

Putting \( t = 2/3 \) in the upper tail bound gives the control we need for one half of (A.211). For the lower tail, we note that the assumption \( n \geq \max\{2m, m + 20\} \) yields the estimates

\[ \sqrt{1 - \frac{m}{n}} \geq \frac{1}{\sqrt{2}}, \quad \frac{2}{n\sqrt{n-m}} \leq \frac{2}{n-m} \leq \frac{1}{10}, \]

so that

\[ \sqrt{1 - \frac{m}{n} - \frac{2}{n\sqrt{n-m}}} - t \geq \frac{1}{\sqrt{2}} - \frac{1}{10} - t, \]

and one checks numerically that \( 2^{-1/2} - (1/10) > 0.6 \). Putting therefore \( t = 2^{-1/2} - (1/10) - 0.6 \) in the lower tail bound yields

\[ \mathbb{P}[\|\sigma(\tilde{g})\|_2 \leq 0.6] \leq e^{-cn}. \]

Plugging these results into (A.211) gives the pointwise measure bound

\[ \mu \left( \mathcal{E}_{0.6,m,u}^c \right) \leq 2 \binom{n}{m} e^{-cn} \]

for some constant \( c > 0 \).

For uniformization, fix \( S \subset [n] \) with \( |S| = m \) and consider the function \( f_S : \mathbb{R}^2 \to \mathbb{R} \) defined by

\[ f_S(u) = \|I_S \sigma(Gu)\|_2. \]

By Gauss-Lipschitz concentration, we have

\[ \mathbb{P}[\|G\| > \mathbb{E}[\|G\|] + t] \leq e^{-cn^2}. \]
and by [229, Theorem 2.6], we have

\[ \mathbb{E}[\|G\|] \leq \sqrt{2} + \frac{2}{\sqrt{n}} \leq 4. \]

Let \( E = \{\|G\| \leq 5\} \); then it follows that \( \mu(E) \geq 1 - e^{-cn} \). On \( E \), for every \( S \), we have that \( f_S \) is a 5-Lipschitz function of \( u \). Let \( T_\varepsilon \subset S^1 \) be a family of sets with the property that \( u \in S^1 \) implies that there is \( u' \in T_\varepsilon \) such that \( \|u' - u\|_2 \leq \varepsilon \) for each \( \varepsilon > 0 \); by standard results [225, Corollary 4.2.13], \( T_\varepsilon \) exists and we have \( |T_\varepsilon| \leq (1 + 2\varepsilon^{-1})^2 \). Define

\[ E_{0.6,m,\varepsilon} = \bigcap_{u \in T_\varepsilon} E_{0.6,m,u}. \]

Then a union bound together with our pointwise concentration result gives

\[ \mu \left( E_{0.6,m,\varepsilon} \right) \leq 2 \left( \frac{n}{m} \right) \left( 1 + \frac{2}{\varepsilon} \right)^2 e^{-cn}. \]

On \( E \cap E_{0.6,m,\varepsilon} \), for any \( u \in S^1 \) and any \( S \), there is \( u' \in T_\varepsilon \) such that \( |f_S(u) - f_S(u')| \leq 5\varepsilon \). But since on this event \( 0.6 \leq f_S(u') \leq 5/3 \), we conclude \( 0.6 - 5\varepsilon \leq f_S(u) \leq 5/3 + 5\varepsilon \), and therefore the choice \( \varepsilon = 1/50 \) gives \( 0.5 \leq f_S(u) \leq 2 \). This implies

\[ E \cap E_{0.6,m,1/50} \subset E_{0.5,m}. \]

Thus, by a union bound and our previous results, we have

\[ \mu \left( E_{0.5,m}^c \right) \leq \mu \left( E_{0.5,m}^c \cup E_{0.6,m,1/50}^c \right) \]
\[ \leq \mu \left( E_{0.6,m,1/50}^c \right) + e^{-cn} \]
\[ \leq 2 \cdot 150^2 \left( \frac{n}{m} \right) e^{-c'n} + e^{-cn}, \]

which is the desired measure bound. \( \square \)
Lemma A.5.17. We have for each fixed \( \nu \in [0, \pi] \) that:

1. The coordinates of \( \hat{v}_\nu \) have subgaussian moment growth

\[
\mathbb{E}[(v_\nu)_i^p] \leq \frac{1}{2} \left( \frac{2p}{n} \right)^{p/2};
\]

2. The event \( \{\|\hat{v}_\nu\|_2 \leq 2\} \) has probability at least \( 1 - e^{-cn} \);

3. The event \( \{\forall \nu \in [0, \pi] \|\hat{v}_\nu\|_2 \leq 4\} \) has probability at least \( 1 - e^{-c'\nu} \).

Proof. We have that the coordinates of \( \hat{v}_\nu \) are i.i.d., and

\[
(\hat{v}_\nu)_i \overset{d}{=} \hat{\sigma}(g_{1i})g_{2i},
\]

by rotational invariance. By independence of \( g_1 \) and \( g_2 \), we compute

\[
\mathbb{E}[(\hat{v}_\nu)_i^p] = \mathbb{E}[\hat{\sigma}(g_{1i})g_{2i}^p] = \frac{1}{2} \mathbb{E}[g_{2i}^p] \leq \frac{2^{p/2}}{2n^{p/2}2^{p/2}},
\]

for each \( p \geq 1 \); the last inequality follows from Lemma A.6.11. This shows that the coordinates of \( \hat{v}_\nu \) are independent subgaussian random variables with scale parameters at most \( C\sqrt{2/n} \), so we have a tail bound [225, Theorem 3.1.1]

\[
P[\|\hat{v}_\nu\|_2 \geq 1 + t] \leq e^{-cn t^2},
\]

also taking into account that \( \mathbb{E}[(\hat{v}_\nu)_i^2] = 1/n \). This shows that the event \( E'' = \{\|\hat{v}_\nu\|_2 \leq 2\} \) has probability at least \( 1 - e^{-cn} \).

For the third assertion, we use the triangle inequality to get \( \|\hat{v}_\nu\|_2 \leq \|g_2\|_2 + \|g_2\|_2 \), which has RHS independent of \( \nu \); then applying Gauss-Lipschitz concentration gives for \( t \geq 0 \)

\[
P[\|g_i\|_2 \geq \sqrt{2} + t] \leq e^{-cn t^2}.
\]
using that $\mathbb{E}[\|g_i\|_2] \leq \sqrt{\mathbb{E}[\|g_i\|_2^2]}$. Putting $t = 0.5$ in this bound and applying a union bound, we conclude that there is an event of probability at least $1 - e^{-cn}$ on which $\|\hat{v}_v\|_2 \leq 4$ uniformly in $v$.

\begin{lemma}
There exists an absolute constant $C > 0$ such that if $n \geq C$, one has

$$1 - \frac{C'}{n} - C''e^{-cn} \leq \mathbb{E}_{g_1, g_2}[\|v_0\|_2\|v_v\|_2] \leq 1,$$

where $c, C', C'' > 0$ are absolute constants.

\end{lemma}

\begin{proof}
For the upper bound, we apply the Schwarz inequality to get

$$\mathbb{E}[\|v_0\|_2\|v_v\|_2] \leq \mathbb{E}[\|v_0\|_2^2]^{1/2}\mathbb{E}[\|v_v\|_2^2]^{1/2} \leq 1,$$

by rotational invariance and Lemma A.6.11. For the lower bound, we will truncate and linearize the product using logarithms. Let $E = E_{0.5,0}$; by Lemma A.5.16, as long as $n \geq 20$ we have $\mu(E^c) \leq Ce^{-cn}$. Define $X = \|v_0\|_2\|v_v\|_21_E + 1_{E^c}$, so that

$$X(G) = \begin{cases} 
\|v_0(G)\|_2\|v_v(G)\|_2 & G \in E, \\
1 & \text{otherwise}. 
\end{cases}$$

We calculate

$$|\mathbb{E}[\|v_0\|_2\|v_v\|_2] - \mathbb{E}[X]| \leq \mu(E^c) + \mathbb{E}[1_E]^{1/2}\mathbb{E}[\|v_0\|_2^4]^{1/2} \leq Ce^{-cn} + C'e^{-c'n}(1 + C'/n)^{1/2}$$

using the triangle inequality, the Schwarz inequality, rotational invariance, and Lemmas A.5.16 and A.5.29. It follows

$$\mathbb{E}[\|v_0\|_2\|v_v\|_2] \geq \mathbb{E}[X] - C'e^{-cn},$$

312
so it suffices to prove the lower bound for $X$ instead. Factoring as $X = (\|v_0\|_2 \mathbb{1}_E + \mathbb{1}_{\mathcal{E}^c})(\|v_\nu\|_2 \mathbb{1}_E + \mathbb{1}_{\mathcal{E}^c})$, we apply concavity of $x \mapsto \log x$, Jensen’s inequality, and convexity of $x \mapsto e^x$ to get

$$
\mathbb{E}[X] \geq \exp(\mathbb{E}[\log(\|v_0\|_2 \mathbb{1}_E + \mathbb{1}_{\mathcal{E}^c})] + \mathbb{E}[\log(\|v_\nu\|_2 \mathbb{1}_E + \mathbb{1}_{\mathcal{E}^c})])
$$

$$
\geq 1 + \mathbb{E}[\log(\|v_0\|_2 \mathbb{1}_E + \mathbb{1}_{\mathcal{E}^c})] + \mathbb{E}[\log(\|v_\nu\|_2 \mathbb{1}_E + \mathbb{1}_{\mathcal{E}^c})]
$$

$$
\geq 1 + 2\mathbb{E}[\log(\|v_0\|_2 \mathbb{1}_E + \mathbb{1}_{\mathcal{E}^c})]
$$

where the last equality is due to rotational invariance. Now write $Y = \|v_0\|_2 \mathbb{1}_E + \mathbb{1}_{\mathcal{E}^c}$, so that by the definition of $\mathcal{E}$ we have $Y \geq \frac{1}{2}$. Taylor expansion with Lagrange remainder of the logarithm about $\mathbb{E}[Y] \geq \frac{1}{2}$ gives

$$
\log(Y) = \log(\mathbb{E}[Y]) - \frac{1}{\mathbb{E}[Y]} (Y - \mathbb{E}[Y]) - \frac{1}{2\xi(Y)^2} (Y - \mathbb{E}[Y])^2
$$

for some $\xi(Y)$ between $\mathbb{E}[Y]$ and $Y$. Using $Y \geq \frac{1}{2}$ and taking expectations on both sides, we get

$$
\mathbb{E}[\log Y] \geq \log \mathbb{E}[Y] - 2\text{Var}[Y].
$$

Moreover, we have

$$
|\mathbb{E}[Y] - \mathbb{E}[\|v_0\|_2]| \leq Ce^{-cn} + \mathbb{E}[\mathbb{1}_{\mathcal{E}^c}\|v_0\|_2] \leq Ce^{-cn} + C' e^{-c'n},
$$

by the Schwarz inequality, and this extra exponential error can be rolled into the exponential error accrued via our use of $X$. In particular, we have

$$
1 - \frac{2}{n} - Ce^{-cn} \leq \mathbb{E}[Y] \leq 1 + Ce^{-cn},
$$

by Lemma A.5.19. Since $n \geq 20$, if we also enforce $n \geq C_1 := c^{-1}\log(5C/2)$ we have $2/n +
\( C e^{-cn} \leq \frac{1}{2} \); it follows by concavity of \( x \mapsto \log(1 - x) \) that we have a bound

\[
\log \left( 1 - \frac{2}{n} - C e^{-cn} \right) \geq -2 \log(2) \left( \frac{2}{n} + C e^{-cn} \right),
\]

which has the form claimed. It remains to upper bound \( \text{Var}[Y] \); using that \( Y^2 = \|v_0\|^2_2 1_{E} + 1_{E^c} \), we have

\[
\text{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \leq 1 + C e^{-cn} - \left( 1 - \frac{2}{n} - C e^{-cn} \right)^2 = C e^{-cn} + 2 \left( \frac{2}{n} + C e^{-cn} \right) - \left( \frac{2}{n} + C e^{-cn} \right)^2 \leq \frac{4}{n} + 3 C e^{-cn},
\]

which is sufficient to conclude. \(\square\)

**Lemma A.5.19.** One has

\[
1 - \frac{2}{n} \leq \mathbb{E}_{g_1,g_2} [\|v_r\|_2] \leq 1.
\]

**Proof.** By rotational invariance, it is equivalent to characterize the expectation of \( \|\sigma(g_1)\|_2 \). By the Schwarz inequality, we have

\[
\mathbb{E}[\|v_0\|_2] \leq \mathbb{E}[\|v_0\|^2_2]^{1/2} = 1,
\]

by Lemma A.6.11. For the lower bound, we apply the Gaussian Poincaré inequality [228, Theorem 3.20] and the 1-Lipschitz property of \( g \mapsto \|\sigma(g)\|_2 \) to get

\[
\frac{n}{2} \mathbb{E}[(\|v_0\|_2 - \mathbb{E}[\|v_0\|_2])^2] \leq 1,
\]

so that after distributing and applying \( \mathbb{E}[\|v_0\|^2_2] = 1 \), we see that

\[
1 - \frac{2}{n} \leq \mathbb{E}[\|v_0\|_2]^2.
\]
Because $n \geq 2$, it follows
\[ \mathbb{E}[\|v_0\|_2] \geq \sqrt{1 - \frac{2}{n}} \geq 1 - \frac{2}{n}, \]
where the last bound holds because $1 - 2n^{-1} \leq 1$. □

**Lemma A.5.20.** If $0 \leq x, y \leq 1$, we have
\[ |\cos^{-1} x - \cos^{-1} y| \leq \sqrt{|x - y|}. \]

**Proof.** Let $0 \leq x, y \leq 1$, and assume to begin that $x \leq y$. We apply the fundamental theorem of calculus and knowledge of the derivative of $\cos^{-1}$ to get
\[ \cos^{-1} x - \cos^{-1} y = \int_x^y \frac{1}{\sqrt{1 - t^2}} \, dt \]
The integrand is nonnegative, so $\cos^{-1} x - \cos^{-1} y \geq 0$. Writing $\sqrt{1 - t^2} = \sqrt{1 - t}\sqrt{1 + t}$ and using $x \geq 0$, we get
\[ \cos^{-1} x - \cos^{-1} y \leq \int_x^y \frac{1}{\sqrt{1 - t}} \, dt = \sqrt{1 - x} - \sqrt{1 - y}. \]
This shows that $|\cos^{-1} x - \cos^{-1} y| \leq |\sqrt{1 - x} - \sqrt{1 - y}|$ when $x \leq y$. An almost-identical argument establishes the same when $y \leq x$, via the inequalities $0 \geq \cos^{-1} x - \cos^{-1} y \geq -(\sqrt{1 - x} - \sqrt{1 - y})$.
So we have shown
\[ |\cos^{-1} x - \cos^{-1} y| \leq |\sqrt{1 - x} - \sqrt{1 - y}| \]
for arbitrary $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Now notice
\[ |\sqrt{1 - x} - \sqrt{1 - y}|^2 \leq |\sqrt{1 - x} - \sqrt{1 - y}| |\sqrt{1 - x} + \sqrt{1 - y}| \leq |(1 - x) - (1 - y)| = |x - y|, \]

315
which establishes $|\cos^{-1} x - \cos^{-1} y| \leq \sqrt{|x - y|}$. □

**Differentiation Results**

**Lemma A.5.21.** For $a < b$, let $f : [a, b] \rightarrow \mathbb{R}$ be a continuous function that is differentiable on $(a, b)$ except at a set of isolated points in $(a, b)$, and let $c \in \mathbb{R}$. Then $\max\{f, c\}$ is differentiable except at a set of isolated points in $(a, b)$.

**Proof.** Let $A \subset (a, b)$ denote the set of points of differentiability of $f$, and let $B \subset (a, b)$ denote the set of points of nondifferentiability of $\max\{f, c\}$. Because finite unions of isolated sets of points in $(a, b)$ are isolated in $(a, b)$, it suffices to consider only points $x \in A$.

Fix $x \in A$, and consider the case $f(x) \neq c$. Then because $f$ is continuous, there is a neighborhood of $x$ on which $f \neq c$. If $f > c$ on this neighborhood, then we have $\max\{f, c\} = f$ on this neighborhood; if $f < c$, then we have $\max\{f, c\} = c$. In either case, this implies that $\max\{f, c\}$ is differentiable at $x$, and thus $x$ is not in $B$.

Next, consider the case where $f(x) = c$. First, suppose $f'(x) > 0$; then by Rolle’s theorem, we can find a neighborhood of $x$ on which $f(x') > c$ if $x' > x$ and $f(x') < c$ if $x' < x$. Possibly shrinking this neighborhood, we can assume every point of the neighborhood is a point of differentiability of $f$. Thus, for $x' < x$ in this neighborhood, we have $\max\{f(x'), c\} = f(x')$, and for $x' > x$, we have $\max\{f(x'), c\} = c$. We conclude that $\max\{f, c\}$ is differentiable at all points of this neighborhood except $x$, and in particular $x$ is an isolated point in $B$. A symmetric argument treats the case where $f'(x) < 0$, with the same conclusion.

On the other hand, if $f'(x) = 0$, we can write $f(x') = c + o(|x' - x|)$ for $x'$ in a neighborhood of $x$, which implies $\max\{f(x'), c\} = \max\{c, c + o(|x' - x|)\} = c + o(|x' - x|)$. In particular, $|\max\{f(x'), c\} - \max\{f(x), c\}| = o(|x' - x|)$, which shows that $\max\{f, c\}$ is differentiable at $x$, and thus $x$ is not in $B$. This shows that every point of $A \cap B$ is isolated in $A \cap B$, and we can therefore conclude that $\max\{f, c\}$ is differentiable except at isolated points of $(a, b)$. □
Lemma A.5.22. For $0 \leq \nu \leq \pi$, consider the function

$$
\tilde{\varphi}(\nu) = \mathbb{E}_{g_1, g_2 \sim \text{i.i.d.} \mathcal{N}(0, (2/n)I)} \left[ I_{E_1} \phi(\nu, g_1, g_2) \right],
$$

where

$$
\phi(\nu, g_1, g_2) = \cos^{-1} \left( \frac{\langle v_0, v_\nu \rangle}{\|v_0\|_2 \|v_\nu\|_2} \right).
$$

Then $\tilde{\varphi}$ is absolutely continuous on $[0, \pi]$, and satisfies the first-order Taylor expansion

$$
\tilde{\varphi}(\nu) = \tilde{\varphi}(0) - \int_0^\nu \mathbb{E}_{g_1, g_2} \left[ I_{E_1} \frac{v_0}{\|v_0\|_2 \|v_\nu\|_2^2} \left( I - \frac{v_\nu v_\nu^*}{\|v_\nu\|_2^2} \right) \right] \frac{d}{\sqrt{1 - \left( \frac{v_\nu}{\|v_0\|_2 \|v_\nu\|_2^2} \right)^2}} \, d\nu,
$$

and moreover $\tilde{\varphi}$ is 1-Lipschitz.

Proof. At points of $(0, \pi)$ where each of the functions composed in $\phi$ is differentiable, the chain rule gives for the derivative of the integrand as a function of $\nu$

$$
\phi'(\nu, g_1, g_2) = -\frac{\langle v_0, v_\nu \rangle}{\|v_0\|_2 \|v_\nu\|_2^2} \left( I - \frac{v_\nu v_\nu^*}{\|v_\nu\|_2^2} \right) \frac{\frac{v_\nu}{\|v_0\|_2 \|v_\nu\|_2^2}}{\sqrt{1 - \left( \frac{v_\nu}{\|v_0\|_2 \|v_\nu\|_2^2} \right)^2}},
$$

where we have used the result

$$
d \left( \frac{\cdot}{\|\cdot\|} \right)_x = \frac{1}{\|x\|_2} \left( I - \frac{xx^*}{\|x\|_2^2} \right),
$$

valid for any $x \neq 0$. Because $E_1$ guarantees that $v_\nu \neq 0$ for all $\nu \in [0, \pi]$, we see that the integrand $\phi$ is continuous. Similarly, given that $\|v_\nu\|_2 \geq \frac{1}{2}$ on $E_1$, we note that there are just two obstructions to differentiability:

1. The inverse cosine is not differentiable at $\{\pm 1\}$;
2. The activation $\sigma$ is not differentiable at $0$. 

317
First we characterize the issue of nondifferentiability with regards to the inverse cosine. We note that $\cos \phi(\nu, g_1, g_2) = 1$ if and only if the Cauchy-Schwarz inequality is tight, which is equivalent to $v_0$ and $v_\nu$ being linearly dependent. Suppose we have $(g_1, g_2) \in \mathcal{E}_1$ and $\nu_0 \in (0, \pi)$ such that $v_0(g_1, g_2)$ and $v_{\nu_0}(g_1, g_2)$ are linearly dependent. Because two vectors $u_1, u_2 \in \mathbb{R}^n$ have $\sigma(u_1)$ and $\sigma(u_2)$ linearly dependent if and only if $\sigma(u_1)$ and $\sigma(u_2)$ have the same support and are linearly dependent on the support, and given that $\|v_\nu\|_0 > 1$ for each $\nu$, we have that there is a $2 \times 2$ submatrix of $GM_{\nu_0}$ having positive entries and rank 1 (since the rank is zero if and only if the submatrix is zero), where

$$M_\nu = \begin{bmatrix} 1 & \cos \nu \\ 0 & \sin \nu \end{bmatrix}.$$

Write the corresponding $2 \times 2$ submatrix of $G$ as $X$. Because rank $M_{\nu_0} = 2$ by $\nu_0 \in (0, \pi)$, we have rank $X = 1$. On the other hand, if $G \sim_{\text{i.i.d.}} \mathcal{N}(0, 2/n)$, we have

$$\mathbb{P}[G \text{ has a singular } 2 \times 2 \text{ minor}] \leq \sum_{1 \leq i < j \leq n} \mathbb{P} \left[ \text{rank} \left( \begin{bmatrix} G_{1i} & G_{2i} \\ G_{1j} & G_{2j} \end{bmatrix} < 2 \right) \right]$$

$$= 0,$$

where the first line is a union bound, and the second line uses the fact that $2 \times 2$ submatrices of $G$ are i.i.d. $\mathcal{N}(0, 2/n)$, and that the complement of the set of full-rank $2 \times 2$ matrices is a positive-codimensional closed embedded submanifold of $\mathbb{R}^{2 \times 2}$. It follows that the subset of $\mathcal{E}_1$ of matrices having no singular $2 \times 2$ minor has full measure in $\mathcal{E}_1$, and we conclude that for almost all $(g_1, g_2)$, we have $\cos \phi(\nu, g_1, g_2) < 1$ for every $\nu \in (0, \pi)$. Next, we characterize nondifferentiability due to the activation $\sigma$; by the chain rule, it suffices to consider nondifferentiability of $v_\nu$ as a function of $\nu$, and then Lemma A.5.21 implies that for every $(g_1, g_2)$, $v_\nu$ is differentiable at all but at most countably many points of $[0, \pi]$. Next, we observe that whenever $v_\nu$ is nonvanishing, one has

$$\left\| \frac{v_0}{\|v_0\|_2} - \left( I - \frac{v_\nu v_\nu^*}{\|v_\nu\|_2^2} \right) \hat{v}_\nu \right\|_2 \leq \frac{\|P_{v_\nu} \hat{v}_\nu\|_2}{\|v_\nu\|_2} \left\| \left( I - \frac{v_\nu v_\nu^*}{\|v_\nu\|_2^2} \right) v_0 \right\|_2.$$
\[
\frac{\|P_{v_{\nu}}^{\perp} \dot{v}_{\nu}\|_2}{\|v_{\nu}\|_2} \sqrt{1 - \left( \frac{v_0}{\|v_0\|_2}, \frac{v_{\nu}}{\|v_{\nu}\|_2} \right)^2},
\]

where the first inequality is due squaring the orthogonal projection and Cauchy-Schwarz, and the second equality follows from distributing to evaluate the squared norm, cancelling, and taking square roots. Using the fact that orthogonal projections have operator norm 1, we thus conclude

\[
|\phi'(\nu, g_1, g_2)| \leq \frac{\|P_{v_{\nu}}^{\perp} \dot{v}_{\nu}\|_2}{\|v_{\nu}\|_2} \leq C\|\dot{v}_{\nu}\|_2,
\]

where the last inequality is valid whenever \((g_1, g_2) \in E_1\). Since

\[
\|\dot{v}_{\nu}\|_2 = \|\dot{\sigma}(g_1 \cos \nu + g_2 \sin \nu) \odot (g_2 \cos \nu - g_1 \sin \nu)\|_2
\]

\[
\leq \|g_2 \cos \nu - g_1 \sin \nu\|_2
\]

\[
\leq \|g_2\| + \|g_1\|_2,
\]

and this upper bound is jointly integrable in \(\nu\) and \((g_1, g_2)\) over \([0, \pi] \times \mathbb{R}^n \times \mathbb{R}^n\), we can apply [219, Theorem 6.3.11] to obtain that whenever \((g_1, g_2) \in E_1\) minus a negligible set, we have for every \(\nu \in [0, \pi]\)

\[
\phi(\nu, g_1, g_2) = \phi(0, g_1, g_2) + \int_0^\nu \phi'(t, g_1, g_2) \, dt.
\]

In particular, multiplying by the indicator for \(E_1\), taking expectations over \((g_1, g_2)\), and applying the previous joint integrability assertion for \(\phi'\) together with Fubini’s theorem yields

\[
\hat{\phi}(\nu) = \hat{\phi}(0) + \int_0^\nu \mathbb{E}_{g_1, g_2}[\phi'(t, g_1, g_2)] \, dt,
\]

so to conclude the Lipschitz estimate, it suffices to estimate \(\mathbb{E}_{g_1, g_2}[\phi'(\nu, g_1, g_2)]\). In light of (A.213) we calculate more precisely

\[
\mathbb{E}\left[ \mathbb{1}_{E_1} \frac{\|P_{v_{\nu}}^{\perp} \dot{v}_{\nu}\|_2}{\|v_{\nu}\|_2} \right] = \mathbb{E}\left[ \mathbb{1}_{E_1} \frac{\|P_{v_0}^{\perp} \dot{v}_0\|_2}{\|v_0\|_2} \right]
\]

319
\[
\begin{align*}
&= \mathbb{E} \left[ 1_{\mathcal{E}_1} \left\| \left( I - \frac{\sigma(g_1) \sigma(g_1)^*}{\|\sigma(g_1)\|^2} \right) (\hat{\sigma}(g_1) \odot g_2) \right\|_2 \right] \\
&\leq \mathbb{E} \left[ 1_{\|\sigma(g_1)\| > 1} \left\| \left( I - \frac{\sigma(g_1) \sigma(g_1)^*}{\|\sigma(g_1)\|^2} \right) (\hat{\sigma}(g_1) \odot g_2) \right\|_2 \right] \\
&= \sum_{k=2}^{n} 2^{-n} \binom{n}{k} \mathbb{E} \left[ \left\| \left( I - \frac{\sigma(g_1) \sigma(g_1)^*}{\|\sigma(g_1)\|^2} \right) (\hat{\sigma}(g_1) \odot g_2) \right\|_2 \right] \left\| \sigma(g_1) \right\|_0 = k \\
&= \sum_{k=2}^{n} 2^{-n} \binom{n}{k} \left( X \sim \chi(k-1) \frac{1}{Y} \right) \mathbb{E} \left[ X \right] \mathbb{E} \left[ Y \right].
\end{align*}
\]

In the first line, we apply rotational invariance and unpack notation; in the second line, we use nonnegativity of the integrand to pass to the containing event where \(v_0\) is at least 2-sparse; and in the third line, we condition on the size of the support of \(g_1\). In the fourth line, we use several facts; first, we note that \(P_{v_0}^\perp (\hat{\sigma}(g_1) \odot g_2) = P_{v_0}^\perp P_{\|\sigma(g_1)\| > 0} g_2\) for any \(g_2 \in \mathbb{R}^n\), and that the commutation relation \(P_{v_0}^\perp P_{\|\sigma(g_1)\| > 0} P_{v_0}^\perp = P_{\|\sigma(g_1)\| > 0} P_{v_0}^\perp\) implies that the operator \(P_{v_0}^\perp P_{\|\sigma(g_1)\| > 0}\) is itself an orthogonal projection, with range equal to the \((\|v_0\|_0 - 1)\)-dimensional subspace consisting of vectors with support \(\text{supp}(v_0)\) orthogonal to \(v_0\). In particular, \(\sigma(g_1)\) and \(P_{v_0}^\perp P_{\|\sigma(g_1)\| > 0} g_2\) are independent gaussian vectors, and conditioned on the size of the support of \(\sigma(g_1)\) the quantities \(\|\sigma(g_1)\|_2\) and \(\|P_{v_0}^\perp P_{\|\sigma(g_1)\| > 0} g_2\|_2\) are distributed as independent chi random variables with (respectively) \(k\) and \(k - 1\) degrees of freedom. An application of Lemma A.6.9 then gives

\[
\mathbb{E} \left[ 1_{\mathcal{E}_1} \frac{\|P_{v_0}^\perp \hat{\nu}_Y\|_2}{\|\nu_Y\|_2} \right] \leq 1,
\]  

(A.214) which is sufficient to conclude.

**Lemma A.5.23.** The random variable \(X_\nu\) satisfies the following regularity properties:

1. If \(0 < \nu \leq \pi\), we have \(X_\nu < 1\) almost surely.
2. If \((g_1, g_2) \in E_1\), then \(X_v\) is absolutely continuous on \([0, \pi]\), with a.e. derivative

\[
\dot{X}_v = \left( \frac{v_0}{\|v_0\|_2}, \frac{1}{\|v_v\|_2} \left( I - \frac{v_v v_v^*}{\|v_v\|_2^2} \right) \dot{v}_v \right),
\]

and moreover we have \(\mathbb{E}_{g_1, g_2}[|\dot{X}_v|] \leq 1\), so the analogous differentiation result applies to \(\mathbb{E}_{g_1, g_2}[X_v]\).

Proof. The first claim is a corollary of the proof of differentiability of the inverse cosine part of \(\tilde{\varphi}\) in Lemma A.5.22 and the observation that \(X_\pi = 0\). The second claim is also a direct consequence of the proof of Lemma A.5.22 and Fubini’s theorem. \(\square\)

Lemma A.5.24. Consider the function

\[
f(v) = \mathbb{E}_{g_1, g_2} [X_v] = \mathbb{E}_{g_1, g_2} \left[ \mathbb{1}_{E_1} \left( \frac{v_0}{\|v_0\|_2}, \frac{v_v}{\|v_v\|_2} \right) \right].
\]

Then \(f\) is continuously differentiable, with derivative

\[
f'(v) = \mathbb{E}_{g_1, g_2} \left[ \mathbb{1}_{E_1} \left( \frac{v_0}{\|v_0\|_2}, \frac{1}{\|v_v\|_2} \left( I - \frac{v_v v_v^*}{\|v_v\|_2^2} \right) \dot{v}_v \right) \right].
\]

Moreover, \(f'\) is absolutely continuous, with Lebesgue-a.e. derivative

\[
f''(v) = -\mathbb{E}_{g_1, g_2} \left[ \mathbb{1}_{E_1} \left( \frac{1}{\|v_v\|_2} \left( I - \frac{v_v v_v^*}{\|v_v\|_2^2} \right) \dot{v}_v, \frac{1}{\|v_0\|_2} \left( I - \frac{v_0 v_0^*}{\|v_0\|_2^2} \right) \dot{v}_0 \right) \right].
\]

Proof. The expression for \(f'\) is a direct consequence of Lemma A.5.23. To see that \(f'\) is actually continuous, apply rotational invariance of the Gaussian measure and of \(\mathbb{1}_{E_1}\) by Lemma A.5.16 to get

\[
f'(v) = -\mathbb{E}_{g_1, g_2} \left[ \mathbb{1}_{E_1} \left( \frac{v_v}{\|v_v\|_2}, \frac{1}{\|v_0\|_2} \left( I - \frac{v_0 v_0^*}{\|v_0\|_2^2} \right) \dot{v}_0 \right) \right],
\]

then notice that this expression is an integral of a continuous function of \(v\), which is therefore continuous. Moreover, the \(v\) dependence in this expression for \(f'\) mirrors exactly that of \(f\); in
particular, the integrand
\[
-\left( \frac{v_v}{\|v_v\|_2} \cdot \frac{1}{\|v_0\|_2} \left( I - \frac{v_0v_0^*}{\|v_0\|_2^2} \right) \dot{v}_0 \right)
\]
is absolutely continuous whenever \((g_1, g_2) \in E_1\) by Lemma A.5.23, with a.e. derivative
\[
-\left( \frac{1}{\|v_v\|_2} \left( I - \frac{v_vv_v^*}{\|v_v\|_2^2} \right) \dot{v}_v, \frac{1}{\|v_0\|_2} \left( I - \frac{v_0v_0^*}{\|v_0\|_2^2} \right) \dot{v}_0 \right).
\]

We can therefore conclude the claimed expression for \(f''\) provided we can show absolute integrability over \(E_1\) of this last expression, using Fubini’s theorem in a way analogous to the argument in Lemma A.5.22. But
\[
\mathbb{E}_{g_1, g_2} \left[ \mathbb{I}_{E_1} \left( \frac{\dot{v}_v}{\|v_v\|_2} \left( I - \frac{v_vv_v^*}{\|v_v\|_2^2} \right), \frac{\dot{v}_0}{\|v_0\|_2} \left( I - \frac{v_0v_0^*}{\|v_0\|_2^2} \right) \right) \right] \leq 4 \mathbb{E}_{g_1, g_2} \left[ \mathbb{I}_{E_1} \left[ \|P_{v_v} \dot{v}_v\|_2 \|P_{v_0} \dot{v}_0\|_2 \right] \right]
\leq 4 \mathbb{E} \left[ \|\dot{v}_v\|_2^2 \right] = 4,
\]
using, in sequence, Cauchy-Schwarz and the lower bound in the definition of \(E_1\); the operator norm of orthogonal projections being 1, the Schwarz inequality, nonnegativity of the integrand, and rotational invariance; and Lemma A.5.17. We can therefore conclude the claimed expression for \(f''\) and complete the proof. \(\square\)

**Lemma A.5.25.** For the heuristic cosine angle evolution function
\[
\cos \varphi(v) = \mathbb{E}_{g_1, g_2} \left[ \langle v_0, v_v \rangle \right],
\]
we have the following integral representations for its continuous derivatives:
\[
(\cos \circ \varphi)'(v) = \mathbb{E}_{g_1, g_2} \left[ \langle v_0, \dot{v}_v \rangle \right]
\]
\[
(\cos \circ \varphi)''(v) = -\mathbb{E}_{g_1, g_2} \left[ \langle \dot{v}_0, \dot{v}_v \rangle \right].
\]

**Proof.** The proof follows exactly the arguments of Lemma A.5.24, but with a simpler integrand
and different integrability checks; the continuity assertion relies on Lemma A.5.5. Indeed, this approach gives that $\langle v_0, v_\nu \rangle$ is absolutely continuous, with Lebesgue-a.e. derivative $\langle v_0, \dot{v}_\nu \rangle$; we check

$$E_{g_1, g_2} [|\langle v_0, \dot{v}_\nu \rangle|] \leq E_{g_1, g_2} \left[ \|v_0\|^2 \right]^{1/2} E_{g_1, g_2} \left[ \|\dot{v}_0\|^2 \right]^{1/2} \leq 1$$

by Cauchy-Schwarz, the Schwarz inequality, rotational invariance, and Lemma A.5.17. This verifies the claimed expression for $(\cos \circ \varphi)'$. For the second derivative, we apply rotational invariance to get

$$(\cos \circ \varphi)''(\nu) = -E_{g_1, g_2} [\langle v_\nu, \dot{v}_0 \rangle],$$

which has an absolutely continuous integrand, with Lebesgue-a.e. derivative

$$-\langle \dot{v}_0, \dot{v}_\nu \rangle.$$ 

Checking absolute integrability, we have as before

$$E_{g_1, g_2} [|\langle \dot{v}_0, \dot{v}_\nu \rangle|] \leq E_{g_1, g_2} \left[ \|\dot{v}_0\|^2 \right] \leq 1$$

by Cauchy-Schwarz, the Schwarz inequality, rotational invariance, and Lemma A.5.17. This establishes the claimed expression for $(\cos \circ \varphi)''$.

□

Lemma A.5.26. Let $\psi : \mathbb{R} \to \mathbb{R}$ be defined by $\psi(x) = \psi_{0.25}(x)$, where $\psi_{0.25}$ is the function constructed in Lemma A.5.31. Then the function

$$f(\nu, g_1) = E_{g_2} \left[ \frac{\langle v_0, v_\nu \rangle}{\psi(\|v_0\|_2)\psi(\|v_\nu\|_2)} \right]$$

satisfies for all $\nu \in [0, \pi]$ and Lebesgue-a.e. $g_1$ the second-order Taylor expansion

$$f(\nu, g_1) = \frac{\|v_0\|^2}{\psi(\|v_0\|_2)^2} + \int_0^\nu \int_0^t \left( \sum_{i=1}^n \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot s)}{\psi(\|v_0\|_2)\psi(\|v'_i\|_2)\sin^3 s} \right)$$

323
where previously-unspecified notation in this expression is introduced in (A.218).

Proof. Take \( g_1 \in \mathbb{R}^n \) such that \( f(\nu, \cdot) \) exists and is \( g_1 \)-integrable; by Fubini’s theorem such \( g_1 \) have full measure in \( \mathbb{R}^n \). Because \( \psi > 0 \) and \( \psi(||\nu||) \) is locally (as a function of \( \nu \)) constant whenever \( ||\nu|| < \frac{1}{4} \), we need only consider nondifferentiability of \( \sigma \) when assessing differentiability of \( f(\cdot, g_1) \). By Lemma A.5.21, we conclude that \( f(\cdot, g_1) \) is differentiable at all but at most countably many points of \((0, \pi)\); since \( \psi > 0 \) and \( \psi \) is smooth, \( f \) is continuous, and we can therefore apply Lebesgue differentiation theorems [219, Theorem 6.3.11] to \( f \) provided we satisfy the standard derivative product integrability checks. Writing

\[
\phi(\nu, g_1, g_2) = \frac{\langle v_0, v_\nu \rangle}{\psi(||v_0||^2)\psi(||v_\nu||^2)},
\]

the chain rule gives (at points of differentiability)

\[
\phi'(\nu, g_1, g_2) = \left( \frac{v_0}{\psi(||v_0||^2)\psi(||v_\nu||^2)} \hat{v}_\nu \right) - \left( \frac{\langle v_0, v_\nu \rangle \psi'(||v_\nu||^2)v_\nu}{\psi(||v_0||^2)\psi(||v_\nu||^2)^2||v_\nu||^2} \hat{v}_\nu \right).
\]

In this expression, we follow the convention \( 0/0 = 0 \) to account for the possibility that \( ||v_\nu||^2 = 0 \) (in this case, the \( \psi' \) term handles the denominator). For product integrability, we Lemma A.5.31 to get \( |\psi'| \leq C \) for some absolute constant \( C > 0 \) together with Cauchy-Schwarz and the triangle inequality to get

\[
|\phi'(\nu, g_1, g_2)| \leq 16||v_0||^2||\hat{v}_\nu||^2 + 64C||v_0||^2||v_\nu||^2||\hat{v}_\nu||^2.
\]
and applying the Schwarz inequality, rotational invariance (to eliminate $\nu$ dependence in the resulting expectations) and Lemma A.5.17, we conclude that $\phi'$ is jointly absolutely integrable over $[0, \pi] \times (\mathbb{R}^n \times 2, \mu \otimes \mu)$. We have therefore a first-order Taylor expansion

$$f(\nu, g_1) = f(0, g_1)$$

$$+ \int_0^\nu \left( \mathbb{E}_{g_2} \left[ \frac{v_0}{\psi(||v_0||_2)} \psi(||v_1||_2) \dot{v}_1 \right] - \mathbb{E}_{g_2} \left[ \frac{\langle v_0, v_t \rangle \psi'(||v_t||_2)v_t}{\psi(||v_0||_2)\psi(||v_1||_2)^2||v_t||_2^2} \dot{v}_t \right] \right) \, dt.$$

We have

$$f(0, g_1) = \mathbb{E}_{g_2} \left[ \frac{||v_0||_2^2}{\psi(||v_0||_2)^2} \right] = \frac{||v_0||_2^2}{\psi(||v_0||_2)^2},$$

since $v_0$ depends only on $g_1$. Next, we show $t$-differentiability of the inner expectation. Our aim is to apply Lemma A.5.27 to differentiate $\Xi_1$ and $\Xi_2$. We first focus on $\Xi_1$; distributing and applying linearity, we have

$$\Xi_1(\nu) = \sum_{i=1}^n \mathbb{E}_{g_2} \left[ \frac{\sigma(g_{1i})(g_{2i} \cos \nu - g_{1i} \sin \nu) \psi(||v_0||_2) \psi(||v_1||_2)}{\psi(||v_0||_2^2)} \dot{\sigma}(g_{1i} \cos \nu + g_{2i} \sin \nu) \right].$$

We have shown absolute integrability of the quantity inside the expectation above; we can therefore apply Fubini’s theorem and the previous definition to write

$$\Xi_1(\nu) = \sum_{i=1}^n \mathbb{E}_{g_2} \left[ \frac{\sigma(g_{1i})(g_{2i} \cos \nu - g_{1i} \sin \nu) \psi(||v_0||_2) \psi(||v_1||_2)}{\psi(||v_0||_2^2)} \dot{\sigma}(g_{1i} \cos \nu + g_{2i} \sin \nu) \right]. \quad (A.215)$$

For each $i \in [n]$, write $\pi_i : \mathbb{R}^n \to \mathbb{R}^{n-1}$ for the linear map that deletes the $i$-th coordinate from its input, and let $\hat{\pi}_i : \mathbb{R} \times \mathbb{R}^{n-1} \to \mathbb{R}^n$ be the linear map such that $\hat{\pi}_i(g_i, \pi_i(g)) = g$. With $g_2$ fixed (in the context of (A.215)), if we define

$$f_1(\nu, g) = \frac{\sigma(g_{1i})(g \cos \nu - g_{1i} \sin \nu)}{\psi(||v_0(g_1, \pi_i(g))(g_2)||_2)\psi(||v_1(g_1, \pi_i(g))(g_2)||_2)},$$

325
then we can write

\[ \Xi_1(\nu) = \sum_{i=1}^{n} \mathbb{E}_{(g_{2i})} \left[ \mathbb{E}_{g_{2i}} \left[ f_1(\nu, g_{2i}) \sigma(g_{1i} \cos \nu + g_{2i} \sin \nu) \right] \right]. \]

Thus, to differentiate \( \Xi_1 \), it suffices to check the regularity of \( f_1(\nu, g) \) and apply Lemma A.5.27. As before, \( \psi > 0 \) and \( \psi \) smooth implies that \( f_1 \) is continuous on \([0, \pi] \times \mathbb{R}\). For integrability of \( f \), we appeal to the Fubini’s theorem justification that we applied previously. For absolute continuity, we apply Lemma A.5.21 to get that the derivative of \( f \) with respect to \( \nu \) is, by the chain rule,

\[ f_1'(\nu, g) = -\sigma(g_{1i}) \left( \frac{g_{1i} \cos \nu + g \sin \nu}{\psi(||v_0||_2)\psi(||v_v||_2)} + \frac{(g \cos \nu - g_{1i} \sin \nu)\psi'(||v_v||_2)\langle v_v, v_v \rangle}{\psi(||v_0||_2)\psi(||v_v||_2)^2||v_v||_2} \right) \]

at all but at most countably many values of \( \nu \); and the triangle inequality, Cauchy-Schwarz, and Lemma A.5.31 yield

\[ |f_1'(\nu, g)| \leq \sigma(g_{1i}) \left( 16(|g_{1i}| + |g|) + 64C(|g| + |g_{1i}|)||v_v||_2) \right. \]

\[ \leq \sigma(g_{1i}) (|g| + |g_{1i}|) (16 + 64C(||g||_2 + ||g||_2)) \]

\[ \leq \sigma(g_{1i}) (|g| + |g_{1i}|) (16 + 64C(||g||_2 + ||g||_2) + |g|), \]  \hspace{1cm} \text{(A.216)}

(we apply square root subadditivity in the last line) which is jointly integrable over \([0, \pi] \times \mathbb{R}\), and moreover over \([0, \pi] \times \mathbb{R}^n\). We conclude absolute continuity of \( f_1(\cdot, g) \) and the integrability property of \( f_1' \). Finally, for the growth estimate, we obtain an estimate for \( f_1 \) similar to the one we just obtained for \( f_1' \) as follows:

\[ |f_1(\nu, g)| \leq 16|g_{1i}|(|g| + |g_{1i}|); \]  \hspace{1cm} \text{(A.217)}

the RHS of the final inequality above is a linear function of \( |g| \), and when \( |g| \geq 1 \) we can therefore obtain \( |f_1(\nu, g)| \leq 16(|g_{1i}| + |g_{1i}|^2)|g| \), which is a suitable growth estimate with \( p = 1 \). Then as long as \( g_{1i} \neq 0 \) for all \( i \) (such \( g_1 \) form a set of measure zero, which we can neglect), we can apply
Lemma A.5.27 to get

\[ \Xi_1(\nu) = \sum_{i=1}^{n} \mathbb{E}_{g_2} \left[ f_1(0, g_{2i}) \sigma(g_{1i}) \right] + \int_{0}^{\nu} \left( \mathbb{E}_{g_2} \left[ f_1'(t, g_{2i}) \sigma(g_{1i} \cos t + g_{2i} \sin t) \right] - g_{1i} f_1(t, -g_{1i} \cot t) \frac{\rho(-g_{1i} \cot t)}{\sin^2 t} \right) \, dt. \]

The estimates (A.216) and (A.217) show, respectively, that \( f_1' \) and \( f_1 \) are absolutely integrable functions of \((\nu, g_2)\). We have

\[ f_1(t, -g_{1i} \cot t) = -\sigma(g_{1i})^2 \frac{\sigma(g_{1i})}{\psi(\|v_0(g_1, g_2)\|_2)\psi(\|v_i(g_1, \hat{\pi}_i(-g_{1i} \cot t, \pi_i(g_2))\|_2)} \sin t, \]

so that Lemma A.5.31 and nonnegativity give

\[ \left| g_{1i} f_1(t, -g_{1i} \cot t) \rho(-g_{1i} \cot t) \right| \leq 16 \sigma(g_{1i})^3 \frac{\rho(-g_{1i} \cot t)}{\sin^3 t}. \]

As in the proof of Lemma A.5.37, in particular using the estimates (A.226) (A.227) to control the magnitude of the RHS for all values of \( t \), we can conclude that the Dirac term is absolutely integrable over \([0, \pi] \times \mathbb{R}^n\). An application of Fubini’s theorem then allows us to re-combine the split integrals in the previous expression:
We notice that

$$v_i(g_1, \tilde{\pi}_i(-g_{1i} \cot t, \pi_i(g_2))) = \begin{bmatrix} \sigma(g_{11} \cos \nu + g_{21} \sin \nu) \\ \vdots \\ \sigma(g_{1(i-1)} \cos \nu + g_{2(i-1)} \sin \nu) \\ 0 \\ \sigma(g_{1(i+1)} \cos \nu + g_{2(i+1)} \sin \nu) \\ \vdots \\ \sigma(g_{1n} \cos \nu + g_{2n} \sin \nu) \end{bmatrix},$$

and thus motivated introduce the notation

$$\tilde{g}_i(t, g_1, g_2) = \tilde{\pi}_i(-g_{1i} \cot t, \pi_i(g_2));$$

$$v_i'(g_1, g_2) = v_i(g_1, \tilde{g}_i(t, g_1, g_2)).$$

We can then write

$$-g_{1i} f_i(t, -g_{1i} \cot t) \rho(-g_{1i} \cot t) \sin^2 t = \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot t)}{\psi(\|v_0\|_2)\psi(\|v_i'\|_2) \sin^3 t}.$$ 

Finally, we apply linearity of the integral to move the summation over $i$ back inside the integrals, obtaining

$$\Xi_1(\nu) = \mathbb{E}_{g_2} \left[ \frac{\langle v_0, \hat{v}_0 \rangle}{\psi(\|v_0\|_2)^2} \right]$$

$$+ \int_0^\nu \left[ \mathbb{E}_{g_2} \left[ \sum_{i=1}^n \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot t)}{\psi(\|v_0\|_2)\psi(\|v_i'\|_2) \sin^3 t} \right] - \mathbb{E}_{g_2} \left[ \frac{\langle v_0, \hat{v}_r \rangle}{\psi(\|v_0\|_2)\psi(\|v_r\|_2)} \right] \right] dr.$$

Noting that, in the zero-order term, the only $g_2$ dependence is in $\hat{v}_0 = \sigma(g_1) \odot g_2$, we apply
independence of \( \mathbf{g}_1 \) and \( \mathbf{g}_2 \) to obtain finally

\[
\Xi_1(\nu) = \int_0^\nu \mathbb{E}_{\mathbf{g}_2} \left[ \sum_{i=1}^n \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot t)}{\psi(||v_0||_2) \psi(||v_i||_2) \sin^3 t} \right] \, dt
- \int_0^\nu \mathbb{E}_{\mathbf{g}_2} \left[ \frac{\langle v_0, v_i \rangle}{\psi(||v_0||_2) \psi(||v_i||_2)} + \frac{\langle v_0, \dot{v}_i \rangle \langle v_i, \dot{v}_i \rangle \psi'(||v_i||_2)}{\psi(||v_0||_2) \psi(||v_i||_2)^2 ||v_i||_2} \right] \, dt
\]

We run the same type of argument on \( \Xi_2 \) next. Distributing and applying linearity, we have

\[
\Xi_2(\nu) = \sum_{i=1}^n \mathbb{E}_{\mathbf{g}_2} \left[ I_m \sigma(g_{1i} \cos \nu + g_{2i} \sin \nu) A \right],
\]

where in the previous expression

\[
A = \frac{\langle v_0, v_i \rangle \sigma(g_{1i} \cos \nu + g_{2i} \sin \nu)(g_{2i} \cos \nu - g_{1i} \sin \nu) \psi'(||v_i||_2)}{\psi(||v_0||_2) \psi(||v_i||_2)^2 ||v_i||_2}.
\]

By the preceding (product) absolute integrability check when taking first derivatives, we can apply Fubini’s theorem to split the integral as we did with \( \Xi_1 \). We define, with \( \mathbf{g}_2 \) fixed, the function

\[
f_2(\nu, g) = B \frac{\langle v_0(\mathbf{g}_1), v_i(\mathbf{g}_1, \dot{\mathbf{g}}_i(g, \pi_i(\mathbf{g}_2))) \rangle}{\psi(||v_0(\mathbf{g}_1)||_2) \psi(||v_i(\mathbf{g}_1, \dot{\mathbf{g}}_i(g, \pi_i(\mathbf{g}_2)))||_2)^2 ||v_i(\mathbf{g}_1, \dot{\mathbf{g}}_i(g, \pi_i(\mathbf{g}_2)))||_2}
\]

where in the previous expression

\[
B = \sigma(g_{1i} \cos \nu + g \sin \nu)(g \cos \nu - g_{1i} \sin \nu) \psi'(||v_i(\mathbf{g}_1, \dot{\mathbf{g}}_i(g, \pi_i(\mathbf{g}_2)))||_2),
\]

so that

\[
\Xi_2(\nu) = \sum_{i=1}^n \mathbb{E}_{(\mathbf{g}_{2i}); j \neq i} \mathbb{E}_{\mathbf{g}_{2i}} \left[ f_2(\nu, g_{2i}) \sigma(g_{1i} \cos \nu + g_{2i} \sin \nu) \right].
\]

Now we check that the hypotheses of Lemma A.5.27 are satisfied for \( f_2 \). The continuity argument is identical to that employed for \( f_1 \), as is the joint absolute integrability property of \( f_2 \). For absolute continuity, we again use \( \psi > 0, \psi \) smooth, and Lemma A.5.21 to obtain the derivative at all but
continuity of infinitely many points of $[0, \pi]$ (by the chain rule and the Leibniz rule) as

$$f_2' (v, g) = \frac{\langle v_0, \hat{v}_{\gamma} \rangle \sigma (g_{11} \cos v + g \sin v) (g \cos v - g_{11} \sin v) \psi' (\|v_r\|_2)}{\psi (\|v_0\|_2) \psi (\|v_r\|_2)^2 \|v_r\|_2} + \frac{\langle v_0, v_r \rangle \sigma (g_{11} \cos v + g \sin v) (g \cos v - g_{11} \sin v)^2 \psi' (\|v_r\|_2)}{\psi (\|v_0\|_2) \psi (\|v_r\|_2)^2 \|v_r\|_2} - \frac{\langle v_0, v_r \rangle \sigma (g_{11} \cos v + g \sin v) (g \cos v - g_{11} \sin v)^2 \psi' (\|v_r\|_2) \langle v_r, \hat{v}_{\gamma} \rangle}{\psi (\|v_0\|_2) \psi (\|v_r\|_2)^3 \|v_r\|_2^3} - 2 \frac{\langle v_0, v_r \rangle \sigma (g_{11} \cos v + g \sin v) (g \cos v - g_{11} \sin v)^2 \psi' (\|v_r\|_2) \langle v_r, \hat{v}_{\gamma} \rangle}{\psi (\|v_0\|_2) \psi (\|v_r\|_2)^2 \|v_r\|_2^3} + \frac{\langle v_0, v_r \rangle \sigma (g_{11} \cos v + g \sin v) (g \cos v - g_{11} \sin v)^2 \psi' (\|v_r\|_2) \langle v_r, \hat{v}_{\gamma} \rangle}{\psi (\|v_0\|_2) \psi (\|v_r\|_2)^2 \|v_r\|_2^2}.$$

Because $\psi'$ or $\psi''$ and our convention handle cancellation in the case where $\|v_r\|_2 = 0$, we can proceed when necessary with the convenient estimate

$$\left| \frac{\sigma (g_{11} \cos v + g \sin v)}{\|v_r (g_1, \hat{v}_{\gamma} (g, \pi_i (g_2)))\|_2} \right| \leq 1,$$

which follows from the fact that $\|u\|_\infty \leq \|u\|_2$ for any $u \in \mathbb{R}^n$. As with $\Xi_1$, we then estimate the magnitude of $f_2'$ using Lemma A.5.31, Cauchy-Schwarz, the triangle inequality, and square-root subadditivity (skipping some steps that we wrote out in the $\Xi_1$ estimate):

$$|f_2' (v, g)| \leq 64 C (\|g\| + |g_{11}|) \|v_0\|_2 \left( \|\hat{v}_{\gamma}\|_2 \left( 2 + \frac{C'}{C} \right) + (\|g\| + |g_{11}|) (2 + 8 \|\hat{v}_{\gamma}\|_2) \right) \leq 64 C (\|g\| + |g_{11}|) \|g_1\|_2 \left( (\|g_1\|_2 + \|\pi_i (g_2)\|_2 + \|g\|) \left( 2 + \frac{C'}{C} \right) + (\|g\| + |g_{11}|) (2 + 8 (\|g_1\|_2 + \|\pi_i (g_2)\|_2 + \|g\|)) \right), \quad (A.219)$$

which is jointly integrable over $[0, \pi] \times \mathbb{R}$, and moreover over $[0, \pi] \times \mathbb{R}^n$. We conclude absolute continuity of $f_2 (\cdot, g)$ and the integrability property of $f_2'$. For the growth estimate, we argue...
similarly to our bound on $f'_2$ to get

$$|f_2(\nu, g)| \leq 64C\|v_0\|_2(|g| + |g_{1i}|)^2$$

$$\leq 64C\|g_1\|_2 \left(|g|^2 + 2|g_{1i}||g| + |g_{1i}|^2\right); \quad (A.220)$$

the RHS in the final inequality is a quadratic function of $|g|$, and we therefore obtain a suitable growth estimate with $p = 2$ and $C' = 64C\|g_1\|(1 + 2|g_{1i}| + |g_{1i}|^2)$ as soon as $|g| \geq 1$. We can therefore apply Lemma A.5.27 to get that for all but a negligible set of $g_1$ that

$$\Xi_2(\nu) = \sum_{i=1}^n \mathbb{E}_{g_2}(f_2(0, g_{2i})\sigma(g_{1i})) + \int_0^\nu \left( \mathbb{E}_{g_{2i}}[f'_2(t, g_{2i})\sigma(g_{1i} \cos t + g_{2i} \sin t)] \right) dt. \quad (A.219)$$

The estimates (A.219) and (A.220) show, respectively, that $f'_2$ and $f_2$ are absolutely integrable functions of $(\nu, g_2)$. Because $\sigma(g_{1i} \cos \nu - g_{1i} \cot \nu \sin \nu) = 0$, we have (fortuitously)

$$f_2(t, -g_{1i} \cot t) = 0,$$

so that there is no Dirac term in the derivative expression for $\Xi_2$. An application of Fubini’s theorem then allows us to re-combine the split integrals in the previous expression:

$$\Xi_2(\nu) = \sum_{i=1}^n \mathbb{E}_{g_2}[f_2(0, g_{2i})\sigma(g_{1i})] + \int_0^\nu \left( \mathbb{E}_{g_2}[f'_2(t, g_{2i})\sigma(g_{1i} \cos t + g_{2i} \sin t)] \right) dt.$$

We have by linearity of the integral

$$\sum_{i=1}^n \mathbb{E}_{g_2}[f_2(0, g_{2i})\sigma(g_{1i})] = \mathbb{E}_{g_1}\left[ \frac{\langle v_0, \dot{v}_0 \rangle\|v_0\|\psi'(\|v_0\|)}{\psi(\|v_0\|)^3} \right] = 0,$$

where the last equality applied independence of $g_1$ and $g_2$, as in the zero-order term of $\Xi_1$. Finally, we apply linearity of the integral to move the summation over $i$ back inside the remaining integrals,
Then \( q \) is density. Let \( \rho \) denote the distribution of a \( \mathcal{N}(0, 2/n) \) random variable, and let \( \rho \) denote its density. Let \( u \in \mathbb{R} \) and \( u \neq 0 \), and let \( f : [0, \pi] \times \mathbb{R} \to \mathbb{R} \) satisfy:

1. \( f \) is continuous in its second argument with its first argument fixed;

2. \( f \) is absolutely continuous in its first argument with its second argument fixed, with a.e. derivative \( f' \);

3. \( f \) and \( f' \) are absolutely integrable with respect to the product of Lebesgue measure and \( \rho \) over \([0, \pi] \times \mathbb{R}\);

4. There exist \( p \geq 1 \) and \( C > 0 \) constants independent of \( x \) such that \( |f(v, x)| \leq C|x|^p \) whenever \( |x| \geq 1 \).

Consider the function

\[
q(v) = \int_{\mathbb{R}} f(v, x) \rho(u \cos v + x \sin v) \, d\mu(x).
\]

Then \( q \) is absolutely continuous, and the following first-order Taylor expansion holds:

\[
q(v) = q(0) + \int_{0}^{v} \left( -u f'(t, -u \cot t) \rho(-u \cot t) \frac{\rho(-u \cot t)}{\sin^2 t} + \int_{\mathbb{R}} f''(t, x) \rho(u \cos t + x \sin t) \, d\mu(x) \right) \, dt.
\]
Proof. For $m \in \mathbb{N}$, define

$$
\hat{\sigma}_m(x) = \begin{cases} 
0 & x \leq 0 \\
mx & 0 \leq x \leq m^{-1} \\
1 & x \geq m^{-1}.
\end{cases}
$$

Then $0 \leq \hat{\sigma}_m \leq 1$; $\hat{\sigma}_m$ is continuous, hence Borel measurable; $\hat{\sigma}_m \to \hat{\sigma}$ pointwise as $m \to \infty$; and $\hat{\sigma}_m$ is differentiable on $\mathbb{R}$ except at $x \in \{0, \frac{m}{m-1}\}$, with derivative $\hat{\sigma}_m = m \mathbb{1}_{0 \leq x \leq m^{-1}}$. Moreover, we have

$$
\int_{\mathbb{R}} m \mathbb{1}_{0 \leq x \leq m^{-1}} \, dx = 1,
$$

and the first-order Taylor expansion

$$
\hat{\sigma}_m(x) = \int_{0}^{x} m \mathbb{1}_{0 \leq x' \leq m^{-1}} \, dx'.
$$

Define

$$
q_m(\nu) = \int_{\mathbb{R}} f(\nu, x) \hat{\sigma}_m(u \cos \nu + x \sin \nu) \, d\mu(x).
$$

Then at every $\nu \in [0, \pi]$, we have by assumption

$$
\int_{\mathbb{R}} |f(\nu, x) \hat{\sigma}_m(u \cos \nu + x \sin \nu)| \, d\mu(x) \leq \int_{\mathbb{R}} |f(\nu, x)| \, d\mu(x) < +\infty,
$$

so that the dominated convergence theorem implies

$$
\lim_{m \to \infty} q_m(\nu) = q(\nu).
$$

By the chain rule, the expression $\hat{\sigma}_m(x) = -\max\{-m \max\{x, 0\}, -1\}$, and Lemma A.5.21, $\nu \mapsto \hat{\sigma}_m(u \cos \nu + x \sin \nu)$ is an absolutely continuous function of $\nu \in [0, \pi]$, and we therefore have by the product rule for AC functions on an interval [219, Corollary 6.3.9]

$$
q_m(\nu) = q_m(0).
$$
\[ + \int_{\mathbb{R}} d\mu(x) \int_0^\nu dt \left( f'(t, x) \hat{\sigma}_m(u \cos t + x \sin t) \\
+ mf(t, x)(x \cos t - u \sin t) 1_{0 \leq u \cos \nu + x \sin \nu \leq m^{-1}} \right). \]

We have

\[ \int_{\mathbb{R}} \int_0^\pi |f'(t, x)\hat{\sigma}_m(u \cos t + x \sin t)| \, dt \, d\mu(x) \leq \int_{\mathbb{R}} \int_0^\pi |f'(t, x)| \, dt \, d\mu(x) < +\infty, \]

dominated by assumption, and

\[ \int_{\mathbb{R}} |f(t, x)(x \cos t - u \sin t) \mathbf{1}_{0 \leq u \cos t + x \sin t \leq m^{-1}}| \, d\mu(x) \]
\[ \leq \left( \int_{\mathbb{R}} f(t, x)^2 \, d\mu(x) \right)^{1/2} \left( \int_{\mathbb{R}} (x \cos t - u \sin t)^2 \, d\mu(x) \right)^{1/2} \]
\[ \leq C_f \left( |u| + \left( \int_{\mathbb{R}} x^2 \, d\mu(x) \right)^{1/2} \right) < +\infty, \]

by the growth assumption on \( f \) and the Schwarz inequality. Applying compactness of \([0, \pi]\) and the lack of \( \nu \) dependence in the final inequality above, an application of Fubini’s theorem therefore yields

\[ q_m(\nu) = q_m(0) \]
\[ + \int_0^\nu \int_{\mathbb{R}} d\mu(x) \, dt \left( f'(t, x) \hat{\sigma}_m(u \cos t + x \sin t) \\
+ mf(t, x)(x \cos t - u \sin t) 1_{0 \leq u \cos \nu + x \sin \nu \leq m^{-1}} \right). \]

By dominated convergence and the first of the preceding two product integrability checks, it is clear

\[ \lim_{m \to \infty} \int_0^\nu \int_{\mathbb{R}} f'(t, x)\hat{\sigma}_m(u \cos t + x \sin t) \, d\mu(x) \, dt = \int_0^\nu \int_{\mathbb{R}} f'(t, x)\hat{\sigma}(u \cos t + x \sin t) \, d\mu(x) \, dt. \]
For the second term, we need to proceed more carefully. For $k \in \mathbb{N}$ sufficiently large for the integral to be over a nonempty interval, we consider

$$q_{m,k}(\nu) := \int_{k-1}^{\nu-k^{-1}} \int_{\mathbb{R}} m f(t, x) (x \cos t - u \sin t) \frac{1}{\sqrt{2\pi c^2}} e^{-\frac{x^2}{2c^2}} 1_{0 \leq u \cos t + x \sin t \leq m^{-1}} \, dx,$$

which is a truncated version of the integral constituting the second term in $q_m$, with a change of variables applied to explicitly show the density corresponding to $\mu$, and where we write $c^2 = 2/n$. In particular, by the calculation used to apply Fubini’s theorem in this context previously, we have by dominated convergence

$$\lim_{k \to \infty} q_{m,k}(\nu) = \int_{0}^{\nu} \int_{\mathbb{R}} m f(t, x) (x \cos t - u \sin t) 1_{0 \leq u \cos t + x \sin t \leq m^{-1}} \, d\mu(x) \, dt.$$

By the product integrability assumption on $f$ and Fubini’s theorem, we can consider the inner $\mathbb{R}$-integral for fixed $t$, and due to our truncation we have $0 < t < \pi$; we therefore change variables $x \mapsto x \sin^{-1} t$ in the inner integral to get

$$q_{m,k}(\nu) = \int_{k-1}^{\nu-k^{-1}} dt \int_{\mathbb{R}} m f(t, x) \left(\frac{x \cos t}{\sin^2 t} - u \right) \frac{1}{\sqrt{2\pi c^2}} e^{-\frac{x^2}{2c^2 \sin^2 t}} 1_{0 \leq u \cos t + x \sin t \leq m^{-1}} \, dx.$$

If $0 < t < \pi$ and $x \in \mathbb{R}$, define

$$g(t, x) = f\left(t, \frac{x}{\sin t}\right) \left(\frac{x \cos t}{\sin^2 t} - u \right) \frac{1}{\sqrt{2\pi c^2}} e^{-\frac{x^2}{2c^2 \sin^2 t}},$$

so that, after an additional change of variables $x \mapsto x - u \cos t$, we obtain

$$q_{m,k}(\nu) = m \int_{k-1}^{\nu-k^{-1}} dt \int_{\mathbb{R}} g(t, x - u \cos t) 1_{0 \leq x \leq m^{-1}} \, dx.$$

Using the growth estimate for $f$, we have

$$|g(t, x - u \cos t)| \leq C \frac{|x - u \cos t|^p |x \cos t - u|}{\sin^{p+2} t} \exp\left(-\frac{(x - u \cos t)^2}{2c^2 \sin^2 t}\right),$$

335
where $C > 0$ depends only on $c$. We are going to bound this quantity under the assumption that $|x| \leq |u|/2$, where we use the assumption $|u| > 0$. First, note that when $\pi/4 \leq t \leq 3\pi/4$, we have $\sin t \geq 1/\sqrt{2}$, and we always have $\sin t \leq 1$ for $0 \leq t \leq \pi$; so in this regime

$$|g(t, x - u \cos t)| \leq C 2^{p/2+1} |x - u \cos t|^p |x \cos t - u| \exp \left( -\frac{(x - u \cos t)^2}{2c^2} \right),$$

which is a continuous function of $(t, x)$, and is therefore bounded by a constant depending only on $c, f, u$ over the compact set $[\pi/4, 3\pi/4] \times [-u/2, u/2]$. Next, we consider the case $0 < t \leq \pi/4$; by the symmetry $\sin(\pi - t) = \sin t$, controlling $|g(t, x - u \cos t)|$ in this regime implies control of it in the regime $3\pi/4 \leq t < \pi$. Here, we note that by our assumption on $t$ and the triangle inequality

$$|x - u \cos t| \geq |u||\cos t| - |x| \geq |u|(|\cos t| - \frac{1}{2}) \geq K|u|,$$

where we can take $K = 2^{-1/2} - 2^{-1} > 0$. Applying the triangle inequality and the condition on $|x|$ gives

$$|g(t, x - u \cos t)| \leq C (3/2)^{p+1} \frac{|u|^{p+1}}{\sin^{p+2} t} \exp \left( -\frac{K^2 u^2}{2c^2 \sin^2 t} \right),$$

which only depends on $t$. For any constants $c', C' > 0$, the continuous map $y \mapsto C|y|^{p+2} e^{-c'y^2}$ is a bounded function of $y \in \mathbb{R}$ by L'Hôpital’s rule applied to determine $\lim_{y \to \pm \infty} |y|^p e^{-y^2} = 0$ for any $p > 0$. It follows that there is a constant $M \geq 0$ depending only on $c, u, p$ such that $|g(t, x - u \cos t)| \leq M$ whenever $0 \leq t \pi/4$; we obtain the result for $t = 0$ by the previous limit calculation. Applying symmetry and taking the sum of our two bounds then yields $|g(t, x - u \cos t)| \leq M'$ for $M' \geq 0$ not depending on $k, m$ whenever $(t, x) \in [0, \pi] \times [-u/2, u/2]$.

Now, we have after one additional change of variables $x \mapsto xm^{-1}$

$$q_{m,k}(v) = \int_{k^{-1}}^{v-k^{-1}} \int_{\mathbb{R}} g(t, xm^{-1} - u \cos t) 1_{0 \leq x \leq 1} \, dx.$$

We can invoke our $M'$ bound when $xm^{-1} \leq |u|/2$, and the indicator enforces $|x| \leq 1$; thus, taking
\( m \geq 2/|u| \) (here we use \(|u| > 0 \) critically) implies
\[
\int_{k^{-1}}^{y-k^{-1}} dt \int_{\mathbb{R}} g(t, xm^{-1} - u \cos t) \mathbb{1}_{0 \leq x \leq 1} dx \leq M' \int_{k^{-1}}^{y-k^{-1}} dt < +\infty,
\]
so that by dominated convergence, we have
\[
\lim_{k \to \infty} q_{m,k}(\nu) = \int_0^y dt \int_{\mathbb{R}} g(t, xm^{-1} - u \cos t) \mathbb{1}_{0 \leq x \leq 1} dx.
\]
By the same estimate together with second-argument continuity of \( f \), hence of \( g \), we have by the dominated convergence theorem
\[
\lim_{m \to \infty} \lim_{k \to \infty} q_{m,k}(\nu) = \int_0^y g(t, -u \cos t) dt = -u \int_0^y f(t, -u \cot t) \frac{1}{\sin^2 t} \frac{e^{-u^2 \cot^2 t}}{2c^2} dt.
\]
Combining with our results on \( q_m \) and the first term, we conclude
\[
q(\nu) = q(0) + \int_0^y dt \left( -u \frac{f(t, -u \cot t)}{\sin^2 t} \frac{1}{\sqrt{2\pi c^2}} e^{-u^2 \cot^2 t} + \int_{\mathbb{R}} f'(t, x) \sigma(\nu \cos t + x \sin t) d\mu(x) \right),
\]
as claimed. \( \Box \)

**Miscellaneous Analytical Results**

**Lemma A.5.28.** If \( m > 0 \), then \( \tilde{\phi} \) is 1-Lipschitz.

**Proof.** We recall
\[
\tilde{\phi}(\nu) = \mathbb{E}_{g_1, g_2 \sim \text{i.i.d.} \mathcal{N}(0, (2/n)I)} \left[ \cos^{-1} X_\nu \right].
\]
Considering instead the related function \( \check{\phi} \) defined by
\[
\check{\phi}(\nu) = \mathbb{E}_{g_1, g_2 \sim \text{i.i.d.} \mathcal{N}(0, (2/n)I)} \left[ \mathbb{1}_{E_1} \phi(\nu, g_1, g_2) \right],
\]

337
where
\[ \phi(\nu, g_1, g_2) = \cos^{-1}\left( \frac{\langle v_0, \nu \rangle}{\|v_0\|_2 \|\nu\|_2} \right), \]
we notice
\[ \tilde{\phi}(\nu) = \tilde{\phi}(\nu) + (\pi/2)\mu(E_c^c). \]

It is therefore equivalent to show that \( \tilde{\phi} \) is 1-Lipschitz; but this follows from Lemma A.5.22. \( \square \)

**Lemma A.5.29** (Even Moments). If \( k \in \mathbb{N} \) and \( k \leq n \), one has
\[
| \mathbb{E}[\|v_\nu\|_2^{2k}] - 1 | \leq C_k n^{-1}, \quad | \mathbb{E}[\|\hat{v}_\nu\|_2^{2k}] - 1 | \leq C_k n^{-1},
\]
where \( C_k \leq (k-1)^2 4^{k-1} (2k-1)!! \).

**Proof.** First notice that the claim is immediate if \( k = 1 \), since \( \mathbb{E}[\|v_\nu\|_2^2] = 1 \). We therefore proceed assuming \( k > 1 \). Also notice that Lemmas A.6.11 and A.5.17 show that \( \hat{v}_\nu \) and \( v_\nu \) have matching even moments, so it suffices to prove the claim for \( v_\nu \). By rotational invariance, we can write
\[
\mathbb{E}[\|v_\nu\|_2^{2k}] = \frac{2^k}{n^k} \mathbb{E}_{g_i \sim \mathcal{N}(0,1)} \left[ \sum_{i=1}^{n} \sigma(g_i)^2 \right]^k
\]
\[
= \frac{2^k}{n^k} \sum_{1 \leq i_1, \ldots, i_k \leq n} \mathbb{E} \left[ \prod_{j=1}^{k} \sigma(g_{i_j})^2 \right],
\]
where the last sum is taken over all elements of \([n]^k\). We split this sum into a sum over terms whose expectations contain no repeated indices, and a sum over all other terms. There are exactly \( k! \binom{n}{k} \) ways to choose a k-multi-index from an alphabet of size n without repetitions—select the k distinct indices, then arrange them in every possible way—and multi-indices without repetitions correspond to terms in the sum where the expectation factors completely, by independence, so we
can write
\[
\mathbb{E}[\|v_v\|_{2}^{2k}] = \frac{2^{k}}{n^{k}} k! \binom{n}{k} \mathbb{E}[\sigma(g_1)^2]^k + \sum_{1 \leq i_1, \ldots, i_k \leq n} \mathbb{E} \left[ \prod_{j=1}^{k} \sigma(g_{ij})^2 \right].
\]

We will prove the elementary estimate
\[
\left| n^k - k! \binom{n}{k} \right| \leq (k - 1)^2 n^{k-1} 2^{k-2}.
\]  
(A.221)

Assuming it for the time being, we use that \( \mathbb{E}[\sigma(g_1)^2]^k = 2^{-k} \) to conclude
\[
\left| (2/n)^k k! \binom{n}{k} \mathbb{E}[\sigma(g_1)^2]^k - 1 \right| \leq (k - 1)^2 2^{k-2} n^{-1}.
\]

Next we study the expectation-of-products arising in the sum. The expectation factors over distinct indices; we can classify repeated indices in a multi-index by partitions \( j_1 + \ldots + j_m = k \), where each \( j_i \) is a positive integer. Formally, for each multi-index \((i_1, \ldots, i_k)\), there is a partition \( j_1 + \ldots + j_m = k \) such that
\[
\mathbb{E} \left[ \prod_{j=1}^{k} \sigma(g_{ij})^2 \right] = \prod_{l=1}^{m} \mathbb{E} \left[ \sigma(g_{ip(l)})^{2j_i} \right],
\]
where \( p : [m] \rightarrow [k] \) is injective. We can evaluate these expectations using \( \mathbb{E}[\sigma(g_1)^{2k}] = \frac{1}{2} (2k - 1)!! \), because the coordinates of \( g \) are i.i.d.:
\[
\prod_{l=1}^{m} \mathbb{E}[\sigma(g_{ip(l)})^{2j_i}] = \frac{1}{2m} \prod_{i=1}^{m} (2j_i - 1)!!.
\]

We claim that
\[
\frac{1}{2m} \prod_{i=1}^{m} (2j_i - 1)!! \leq \frac{1}{2} (2k - 1)!! \tag{A.222}
\]
which is the expectation obtained from a term with all indices equal, whence

\[
\frac{2^k}{n^k} \sum_{1 \leq i_1, \ldots, i_k \leq n} \text{only repeated indices} \mathbb{E} \left[ \prod_{j=1}^{k} \sigma(g_{i_j})^2 \right] \leq \frac{2^k}{n^k} n^{k-1} (k - 1)^2 2^{k-2} \mathbb{E} [\sigma_1^{2k}] \\
= ((k - 1)^2 2^{2k-3} (2k - 1)!!) n^{-1}
\]

by (A.221), which gives a bound on the number of terms in the sum. Noticing that this constant is larger than \((k - 1)^2 2^{k-2}\), we can conclude the claimed estimate on \(C_k\) provided we can justify (A.222). For this, it suffices to show

\[
1 \leq 2^{m-1} \frac{(2k - 1)!! \prod_{i=1}^{m} (2j_i - 1)!!}{\prod_{i=1}^{m} (2j_i - 1)!!}.
\]

Observe that \(m \geq 1\) for any partition, so \(2^{m-1} \geq 1\) and we need only study the second term on the righthand side. We write this term as

\[
\frac{(2k - 1)!! \prod_{i=1}^{m} (2j_i - 1)!!}{\prod_{i=1}^{m} (2j_i - 1)!!} = \frac{\prod_{i=1}^{k} (2i - 1)}{\prod_{i=1}^{m} \prod_{l=1}^{j_i} (2l - 1)}.
\]

The fact that \(j_1 + \cdots + j_m = k\) implies that there are \(k\) factors in the denominator, so we can put the factors in the numerator and denominator into one-to-one correspondence. Consider the ordering of the factors in the denominator \((\prod_{l=1}^{j_1} (2l - 1)) \cdots (\prod_{l=1}^{j_m} (2l - 1))\). Then

\[
\frac{\prod_{i=1}^{k} (2i - 1)}{\prod_{i=1}^{j_1} (2i - 1)} = \prod_{i=j_1+1}^{k} (2i - 1).
\]

If \(j_1 = k\), then this product is empty and \(m = 1\), so the claim is established. If not, then we proceed to the next group of factors in the denominator: we get

\[
\frac{\prod_{i=j_1+1}^{k} (2i - 1)}{\prod_{i=1}^{j_2} (2i - 1)} \geq 1,
\]

340
because \( j_i > 0 \) implies that every term in the numerator (ordered in ascending order) is larger than the corresponding term in the denominator. This gives the claim in the case \( m = 2 \); for \( m > 2 \), we conclude the claim by induction.

To close the loop, we prove (A.221). Using simple algebra, we observe

\[
n^k - k! \binom{n}{k} = n^k - n(n-1) \ldots (n-k+1) = n^k \left(1 - \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right)\right),
\]

and we note bounds

\[
\left(1 - \frac{k-1}{n}\right)^{k-1} \leq \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right) \leq 1.
\]

Working on the upper bound first, we obtain with the help of the binomial theorem

\[
1 - \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right) \leq 1 - \left(1 - \frac{k-1}{n}\right)^{k-1} = \sum_{j=1}^{k-1} \binom{k-1}{j} (-1)^{j+1} \left(\frac{k-1}{n}\right)^j \leq \frac{k-1}{n} \sum_{j=0}^{k-2} \binom{k-1}{j+1} \left(\frac{k-1}{n}\right)^j,
\]

where the last expression removes cancellation by making each term in the sum nonnegative, then applies a change of index. With the identity \( \binom{k-1}{j+1} = (k-1)/(j+1) \binom{k-2}{j} \), we proceed as

\[
\frac{k-1}{n} \sum_{j=0}^{k-2} \binom{k-1}{j+1} \left(\frac{k-1}{n}\right)^j = \frac{(k-1)^2}{n} \sum_{j=0}^{k-2} \binom{k-2}{j} \frac{1}{j+1} \left(\frac{k-1}{n}\right)^j \leq \frac{(k-1)^2}{n} \sum_{j=0}^{k-2} \binom{k-2}{j} \left(\frac{k-1}{n}\right)^j = \frac{(k-1)^2}{n} \left(1 + \frac{k-1}{n}\right)^{k-2}.
\]
given that $1/(j+1) \leq 1$. Since $n \geq k$, this gives

$$n^k - k! \binom{n}{k} \leq (k - 1)^2 n^{k-1} 2^{k-2}.$$ 

The upper bound on the product gives immediately $n^k - k! \binom{n}{k} \geq 0$, which completes the proof.

□

Lemma A.5.30 (Mixed Moments). Let $g^1, \ldots, g^n$ denote the $n$ (i.i.d. according to $N(0, (2/n)I_2)$) rows of the matrix $G$. Let $k \in [n]$, and for each $1 \leq j \leq k$ let $f_j : \mathbb{R}^2 \to \mathbb{R}$ be a function such that

1. $\mathbb{E}[|f_j(g^1)|^p]^{1/p} \leq Cn^{-1}p$, with $C > 0$ an absolute constant and $p \geq 1$;

2. $\mathbb{E}[|f_j(g^1)|] \leq n^{-1}$.

Consider the quantities

$$A = \mathbb{E} \left[ \prod_{j=1}^{k} \left( \sum_{i=1}^{n} f_j(g_i^j) \right) \right]; \quad B = n^k \prod_{j=1}^{k} \mathbb{E}[f_j(g^1)].$$

Then one has $|A - B| \leq Cn^{-1}$, with the constant depending only on $k$.

Proof. Start by writing

$$A = \sum_{1 \leq i_1, \ldots, i_k \leq n} \mathbb{E} \left[ \prod_{j=1}^{k} f_j(g_i^j) \right]$$

$$= k! \binom{n}{k} \mathbb{E}[f_j(g^1)] + \sum_{1 \leq i_1, \ldots, i_k \leq n} \mathbb{E} \left[ \prod_{j=1}^{k} f_j(g_i^j) \right]$$

$$= n^{-k} k! \binom{n}{k} B + \sum_{1 \leq i_1, \ldots, i_k \leq n} \mathbb{E} \left[ \prod_{j=1}^{k} f_j(g_i^j) \right]$$

as in Lemma A.5.29. Applying the triangle inequality and the first moment assumption on the
functions $f_j$, we get

$$\left| n^{-k} k! \binom{n}{k} B - B \right| = |B| k! \binom{n}{k} n^{-k} - 1 \leq (k - 1)^2 2^{k-2} n^{-1},$$

with the last inequality following from the estimate (A.221). For the remaining term, we have by the triangle inequality

$$\sum_{1 \leq i_1, \ldots, i_k \leq n} \mathbb{E} \left[ \prod_{j=1}^{k} f_j (g^{i_j}) \right] \leq \left| n^k - k! \binom{n}{k} \right| \sup_{(i_1, \ldots, i_k) \subset [n]^k} \mathbb{E} \left[ \prod_{j=1}^{k} f_j (g^{i_j}) \right] \leq (k - 1)^2 n^{k-1} 2^{k-2} \sup_{(i_1, \ldots, i_k) \subset [n]^k} \mathbb{E} \left[ \prod_{j=1}^{k} f_j (g^{i_j}) \right],$$

using again (A.221) to control the number of terms in the sum. To control the supremum, we apply the Schwarz inequality $k - 1$ times to get

$$\mathbb{E} \left[ \prod_{j=1}^{k} f_j (g^{i_j}) \right] \leq \mathbb{E} \left[ f_1 (g^{i_1})^2 \right]^{1/2} \mathbb{E} \left[ \prod_{j=2}^{k} f_j (g^{i_j})^2 \right]^{1/2} \leq \ldots \leq \left( \prod_{j=1}^{k-1} \mathbb{E} \left[ f_j (g^{i_j})^2 \right]^{2^{-j}} \right) \mathbb{E} \left[ f_k (g^{i_k})^{2^{k-1}} \right]^{2^{-k-1}}.$$

By the subexponential assumption on the functions $f_j$, we have moment growth control, and we therefore have a bound

$$\mathbb{E} \left[ \prod_{j=1}^{k} f_j (g^{i_j}) \right] \leq \left( \prod_{j=1}^{k-1} C_1 n^{-1} 2^{j} \right) C_1 n^{-1} 2^{k-1} = C_1^k n^{-k} 2^{(k-1)+\sum_{j=1}^{k-1} j} = C_1^k n^{-k} 2^{\frac{1}{2} (k-1)(k+2)},$$

343
and consequently

\[
\sum_{1 \leq i_1, \ldots, i_k \leq n \text{ only repeated indices}} \mathbb{E} \left[ \prod_{j=1}^{k} f_j(g^{i_j}) \right] \leq C_1^k (k-1)^2 2^{\frac{1}{2}k(k+3)} n^{-1},
\]

which proves the claim. \qed

**Lemma A.5.31.** For any \( 0 < c \leq \frac{1}{2} \), there exists a smooth function \( \psi_c : \mathbb{R} \to \mathbb{R} \) satisfying

1. \( \psi_c(x) = x \) if \( x \geq 2c \) and \( \psi_c(x) = c \) if \( x \leq c \), and \( \psi_c \) is between \( c \) and \( 2c \) if \( c \leq x \leq 2c \);

2. \( \psi_c(x) \geq \frac{1}{2}x \);

3. There are constants \( M_1, M_2 > 0 \) depending only on \( c \) such that \( |\psi'_c| \leq M_1 \) and \( |\psi''_c| \leq M_2 \).

**Proof.** The function \( f(x) = 1_{x>0} e^{-\frac{1}{x}} \) is smooth on \( \mathbb{R} \), and satisfies \( 0 \leq f \leq 1 \) and \( f = 0 \) if \( x \leq 0 \). The function

\[
\phi_c(x) = \frac{f(x)}{f(x) + f(c-x)}
\]

is therefore smooth, satisfies \( 0 \leq \phi_c \leq 1 \), and satisfies \( \phi_c(x) = 0 \) if \( x \leq 0 \) and \( \phi_c(x) = 1 \) if \( x \geq c \). Simplifying using the definitions, we can write

\[
\phi_c(x) = \begin{cases} 
0 & x \leq 0 \\
\frac{1}{1+\exp \left( \frac{c-2x}{x(c-x)} \right)} & 0 < x < c \\
1 & x \geq c.
\end{cases}
\]

It follows that \( x \mapsto x \phi_c(x) \) is zero when \( x \leq 0 \), \( x \) when \( x \geq c \), and in between otherwise. Thus, the function \( \psi_c(x) = c + (x-c)\phi_c(x-c) \) satisfies property 1.

For property 2, we note that \( \psi_c(x) = c + (x-c)\phi_c(x-c) \) implies that \( \psi_c \geq c \), since \( \phi_c(x-c) = 0 \) whenever \( x \leq c \) and \( \phi_c \geq 0 \). Since \( \psi_c(x) = x \) when \( x \geq 2c \), we can then conclude \( \psi_c(x) \geq \frac{1}{2}x \), since \( \frac{1}{2}x \leq c \) when \( x \leq 2c \) and \( \frac{1}{2}x \leq x \) when \( x \geq 2c \).
For property 3, we note that by property 1, \( \psi'_c(x) = 1 \) if \( x \geq 2c \) and \( \psi'_c(x) = 0 \) if \( x \leq 0 \); consequently \( \psi''_c(x) = 0 \) if \( x \not\in [0, 2c] \), and it suffices to control \( \psi'_c \) and \( \psi''_c \) in this region. By translation equivariance of the derivative, it then suffices to control the derivatives of \( h(x) = x\phi_c(x) \) for \( 0 < x < c \). We calculate

\[
h'(x) = x\phi'_c(x) + \phi_c(x),
\]

\[
h''(x) = x\phi''_c(x) + 2\phi_c(x), \tag{A.223}
\]

and

\[
\phi'_c(x) = \frac{f(c-x)f'(x) - f(x)f'(c-x)}{(f(x) + f(c-x))^2}, \tag{A.224}
\]

\[
\phi''_c(x) = \frac{(f(x) + f(c-x))(f(c-x)f''(x) + f(x)f''(c-x) - 2f'(x)f'(c-x))}{(f(x) + f(c-x))^3} - 2 \frac{(f'(x) - f'(c-x))(f(c-x)f'(x) - f(x)f'(c-x))}{(f(x) + f(c-x))^3}. \tag{A.225}
\]

Completely ignoring possible cancellation, we see that it suffices to get a lower bound on \( f(x) + f(c-x) \) and upper bounds on \( f' \) and \( f'' \) to bound \( |h'| \) and \( |h''| \). We calculate

\[
f'(x) + f'(c-x) = \frac{1}{x^2}e^{-\frac{1}{x}} 1_{x > 0} - \frac{1}{(c-x)^2}e^{-\frac{1}{c-x}} 1_{x < c},
\]

and since \( f(x) > 0 \) if \( x > 0 \) and \( c > 0 \), we see that any solution of \( f'(x) - f'(c-x) = 0 \) must occur for \( x \in (0, c) \), which implies as well \( c - x \in (0, c) \). Writing \( g(x) = x^2e^{-x} \) and using \( c^{-1} < x^{-1} < \infty \) for \( x \in (0, c) \), we note from our previous work that \( f'(x) - f'(c-x) = 0 \iff g(x^{-1}) = g((c-x)^{-1}) \). We calculate \( g'(x) = xe^{-x}(2-x) \), so that if \( x > 2 \) then \( g'(x) < 0 \), which implies that \( g \) is injective on \( (2, \infty) \). By assumption, we have \( c^{-1} > 2 \); consequently there is at most one solution to \( f'(x) - f'(c-x) = 0 \) in \( 0 < x < c \), and given that \( x = \frac{1}{2}c \) is a solution, there is exactly one solution. We check

\[
2f(c/2) < f(0) + f(c) \iff \log 2 < 1/c,
\]
where the first RHS is the value of \( f(x) + f(c - x) \) at both \( x = 0 \) and \( x = c \), and since \( 1/c \geq 2 \), we conclude that \( f(x) + f(c - x) \geq 2f(c/2) > 0 \). Next, we use

\[
f'(x) = \frac{1}{x^2} e^{-1/x} \mathbb{1}_{x > 0},
\]

\[
f''(x) = \left( \frac{1}{x^3} e^{-1/x} - \frac{2}{x^3} e^{-1/x} \right) \mathbb{1}_{x > 0},
\]

together with the bound \( x^p e^{-x} \leq p^p e^{-p} \) for \( p > 0 \), which is proved by differentiating \( x \mapsto x^p e^{-x} \), equating to zero, and comparing the values of the function at \( x = 0, x = p \), and \( x \to \infty \), to obtain with the triangle inequality

\[
|f'(x)| \leq 4/e^2, \quad |f''(x)| \leq 4^4 e^{-4} + 2 \cdot 3^3 e^{-3}.
\]

Combining these bounds with our lower bound on \( f(x) + f(c - x) \) and repeatedly applying the triangle inequality and modulus bounds in (A.224) and (A.225), then subsequently in (A.223) (using also \(|x| \leq c\)), we conclude the claimed bounds on \( |\phi'_c| \) and \( |\phi''_c| \).

\[\square\]

**Lemma A.5.32.** Let \( Z, \bar{Z} \in L^2 \) be square-integrable random variables. Suppose that \( \bar{Z} \leq C \) a.s. and \( \|Z - \bar{Z}\|_{L^2} \leq M \). Then

\[
\text{Var}[Z] \leq \text{Var}[\bar{Z}] + CM + M^2.
\]

**Proof.** This is a simple consequence of the triangle inequality and the centering inequality for the \( L^2 \) norm. We have

\[
\|Z - \mathbb{E}[Z]\|_{L^2} \leq \|Z - \bar{Z} - \mathbb{E}[Z - \bar{Z}]\|_{L^2} + \|\bar{Z} - \mathbb{E}[\bar{Z}]\|_{L^2},
\]

and additionally

\[
\|Z - \bar{Z} - \mathbb{E}[Z - \bar{Z}]\|_{L^2} \leq \|Z - \bar{Z}\|_{L^2} \leq M,
\]
so that, after squaring, we get

\[
\text{Var}[Z] \leq \text{Var}[\tilde{Z}] + M\|\tilde{Z} - \mathbb{E}[\tilde{Z}]\|_{L^2} + M^2 \\
\leq \text{Var}[\tilde{Z}] + M\|\tilde{Z}\|_{L^2} + M^2 \\
\leq \text{Var}[\tilde{Z}] + CM + M^2,
\]

by centering and the a.s. boundedness assumption. \qed

**Lemma A.5.33.** Let \(X, Y\) be square-integrable random variables, and let \(d > 0\). Suppose \(|X| \leq M_1\) a.s., and suppose \(\mathbb{P}[|Y - 1| \geq C\sqrt{d/n}] \leq C'e^{-cd}\) and \(\|Y - 1\|_{L^2} \leq M_2\). Then one has with probability at least \(1 - C'e^{-cd}\)

\[
|XY - \mathbb{E}[XY]| \leq |X - \mathbb{E}[X]| + 2CM_1\sqrt{\frac{d}{n}} + \sqrt{C'M_1M_2e^{-cd/2}}.
\]

**Proof.** We apply the triangle inequality:

\[
|XY - \mathbb{E}[XY]| \leq |XY - X| + |X - \mathbb{E}[X]| + |\mathbb{E}[X] - \mathbb{E}[XY]|
\]

\[
\leq M_1|Y - 1| + M_1\mathbb{E}[|Y - 1|] + |X - \mathbb{E}[X]|,
\]

where the second inequality also applies Jensen’s inequality. We have

\[
\mathbb{E}[|Y - 1|] = \mathbb{E}\left[\left(\mathbb{1}_{|Y-1| \geq C\sqrt{d/n}} + \mathbb{1}_{|Y-1| < C\sqrt{d/n}}\right)|Y - 1|\right]
\]

\[
\leq C\sqrt{\frac{d}{n}} + \mathbb{E}\left[\mathbb{1}_{|Y-1| \geq C\sqrt{d/n}}|Y - 1|\right]
\]

\[
\leq C\sqrt{\frac{d}{n}} + \mathbb{E}\left[\mathbb{1}_{|Y-1| \geq C\sqrt{d/n}}\right]^{1/2}\mathbb{E}[|Y - 1|]^{1/2}
\]

\[
\leq C\sqrt{\frac{d}{n}} + \sqrt{C'e^{-cd/2}}M_2,
\]

where we apply the Schwarz inequality in the third line. Consequently, with probability at least
\[ 1 - C' e^{-cd}, \] we have
\[
|XY - \mathbb{E}[XY]| \leq |X - \mathbb{E}[X]| + 2CM_1 \sqrt{\frac{d}{n}} + \sqrt{C' M_1 M_2 e^{-cd/2}},
\]
as claimed.

\[ \square \]

**Lemma A.5.34.** For \( i = 1, \ldots, n \), let \( X_i, Y_i \) be random variables in \( L^4 \), and let \( d > 0 \) and \( \delta > 0 \). Suppose \( X_i \geq 0 \) for each \( i \) and \( \sum_{i=1}^n \|X_i\|_{L^2} \leq M_3 \), and suppose \( \mathbb{P}[\forall i \in [n], |Y_i - 1| \geq C\sqrt{d/n}] \leq \delta \) and for each \( i \), \( \|Y_i - 1\|_{L^4} \leq M_2 \). Moreover, suppose that \( C\sqrt{d/n} \leq 1 \). Then one has with probability at least \( 1 - \delta \)
\[
\left| \sum_{i=1}^n X_i Y_i - \mathbb{E}[X_i Y_i] \right| \leq 2 \left| \sum_{i=1}^n X_i - \mathbb{E}[X_i] \right| + 2CM_3 \sqrt{\frac{d}{n}} + \delta^{1/4} M_2 M_3.
\]

**Proof.** The proof is a minor elaboration on Lemma A.5.33. We apply the triangle inequality:
\[
\left| \sum_{i=1}^n X_i Y_i - \mathbb{E}[X_i Y_i] \right| \leq \left| \sum_{i=1}^n X_i Y_i - X_i \right| + \left| \sum_{i=1}^n X_i - \mathbb{E}[X_i] \right| + \left| \sum_{i=1}^n \mathbb{E}[X_i] - \mathbb{E}[X_i Y_i] \right| \leq C \left( \sum_{i=1}^n |X_i| \right) \sqrt{\frac{d}{n}} + \sum_{i=1}^n \mathbb{E}[|X_i||Y_i - 1|] + \sum_{i=1}^n X_i - \mathbb{E}[X_i],
\]
where the second line holds with probability at least \( 1 - \delta \). Another application of the triangle inequality together with nonnegativity of the \( X_i \) gives
\[
\sum_{i=1}^n |X_i| = \sum_{i=1}^n X_i \leq \sum_{i=1}^n X_i - \mathbb{E}[X_i] + \sum_{i=1}^n \mathbb{E}[X_i] \leq M_3 + \sum_{i=1}^n X_i - \mathbb{E}[X_i],
\]
where the second line applies the Lyapunov inequality. By the Schwarz inequality and the Ly-
punov inequality, we have
\[
\sum_{i=1}^{n} \mathbb{E}[|X_i| | Y_i - 1|] \leq C \sqrt{\frac{d}{n}} \sum_{i=1}^{n} \mathbb{E}[|X_i|] + \sum_{i=1}^{n} \mathbb{E}[\mathbb{I}_{|Y_i - 1| \geq C \sqrt{d/n}} |X_i| | Y_i - 1|] \leq CM_3 \sqrt{\frac{d}{n}} + \delta^{1/4} M_2 M_3.
\]
Consequently, with probability at least \(1 - \delta\), we have
\[
\left| \sum_{i=1}^{n} X_i Y_i - \mathbb{E}[X_i Y_i] \right| \leq 2 \left| \sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \right| + 2CM_3 \sqrt{\frac{d}{n}} + \delta^{1/4} M_2 M_3
\]
as claimed, where we use that \(C \sqrt{d/n} \leq 1\) here.

\[\square\]

**Lemma A.5.35.** Let \(k \in \mathbb{N}\), and let \(X_1, \ldots, X_k\) be integrable random variables satisfying \(\|X_i - \mathbb{E}[X_i]\|_{L^4} \leq M_i\) for some constants \(M_i > 0\). Suppose moreover that with probability at least \(1 - \delta_i\), one has \(|X_i - \mathbb{E}[X_i]| \leq N_i\) for some constants \(N_i > 0\). Then one has
\[
\text{Var} \left[ \sum_{i=1}^{k} X_i \right] \leq \sum_{i,j=1}^{k} N_i N_j + \sqrt{\delta_i + \delta_j} M_i M_j.
\]

**Proof.** We start from the formula
\[
\text{Var} \left[ \sum_{i=1}^{k} X_i \right] = \sum_{i=1}^{n} \text{Var}[X_i] + 2 \sum_{i<j} \text{cov}[X_i, X_j],
\]
where \(\text{cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \mathbb{E}[X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]\); one establishes this formula by distributing in the definition of the variance. By assumption, there are events \(\mathcal{E}_i\) on which \(|X_i - \mathbb{E}[X_i]| \leq N_i\) and such that \(\mathbb{P}[\mathcal{E}_i] \geq 1 - \delta_i\). Partitioning the expectation, we therefore have
\[
\text{Var}[X_i] = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \leq N_i^2 + \mathbb{E}[\mathbb{1}_{\mathcal{E}_i^c} (X_i - \mathbb{E}[X_i])^2]
\]

349
\[
\leq N_i^2 + \mathbb{E}[\mathbf{1}_{E^c_i}]^{1/2}\mathbb{E}[(X_i - \mathbb{E}[X_i])^4]^{1/2} \\
\leq N_i^2 + \sqrt{\delta_i M_i^2},
\]

where the first line uses nonnegativity of the integrand to discard the indicator after applying the deviations bound, the second line applies the Schwarz inequality, and the third line uses fourth moment control. For the covariance terms, we apply Jensen’s inequality to obtain

\[
|\text{cov}[X_i, X_j]| = |\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]| \leq \mathbb{E}[(|X_i - \mathbb{E}[X_i]| |X_j - \mathbb{E}[X_j]|),
\]

so that, again partitioning the outermost expectation and applying our assumptions, we get

\[
|\text{cov}[X_i, X_j]| \leq N_i^2 + \mathbb{E}[\mathbf{1}_{E^c_i \cup E^c_j}] |X_i - \mathbb{E}[X_i]| |X_j - \mathbb{E}[X_j]| \\
\leq N_i^2 + \mathbb{E}[\mathbf{1}_{E^c_i \cup E^c_j}]^{1/2}\mathbb{E}[(X_i - \mathbb{E}[X_i])^4]^{1/4}\mathbb{E}[(X_j - \mathbb{E}[X_j])^4]^{1/4} \\
\leq N_i N_j + \sqrt{\delta_i + \delta_j M_i M_j},
\]

where in the first line we again use nonnegativity of the integrand to discard the indicator after applying the deviations bound, in the second line we apply the Schwarz inequality twice, and in the third line we use a union bound to control the indicator. Since \(\delta_i \leq 2\delta_i\), we conclude the claimed expression. \(\Box\)

**Lemma A.5.36.** If \(C > 0\) and \(p > 0\), the function \(g(t) = t^p e^{-Ct^2}\) for \(t \geq 0\) satisfies the bound \(g(t) \leq (p/(2Ce))^{p/2}\).

**Proof.** The function \(g\) is smooth and has derivatives

\[
g'(t) = t^{p-1} e^{-Ct^2} (p - 2 Ct^2)
\]

and

\[
g''(t) = t^{p-2} e^{-Ct^2} \left(p(p - 1) - 2(4p - 1)Ct^2 + 4C^2t^4\right).
\]
It therefore has at most two critical points, one possibly at \( t = 0 \) and one at \( t = \sqrt{p/(2C)} \), and these points are distinct when \( p > 0 \) and \( C > 0 \). We check the sign of \( g'' \) at the second critical point; since \( \sqrt{p/(2C)} > 0 \) we need only check the value of \( (p(p - 1) - 2(4p - 1)C^2 + 4C^2t^4) \) evaluated at \( t = \sqrt{p/(2C)} \), which is \(-2p^2 < 0\). Then since \( \lim_{t \to \pm \infty} g(t) = 0 \) and \( g(0) = 0 \), we conclude that \( g(t) \leq g(\sqrt{p/(2C)}) \), which gives the claimed bound. \( \square \)

**Lemma A.5.37.** Following Lemma A.5.26, consider the random variables

\[
\Xi_1(s, g_1, g_2) = \sum_{i=1}^{n} \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot s)}{\psi(||v_0||_2)\psi(||v_s^i||_2) \sin^3 s} \\
\Xi_2(v, g_1, g_2) = \frac{(v_0, v_s)\psi'(||v_s||_2)||v_s||_2}{\psi(||v_0||_2)\psi(||v_s||_2)^2} - \frac{(v_0, v_s)}{\psi(||v_0||_2)\psi(||v_s||_2)} \\
\Xi_3(v, g_1, g_2) = -\frac{(v_0, v_s)\langle v_s, v_s \rangle\psi''(||v_s||_2)}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2^2} \\
\Xi_4(v, g_1, g_2) = -\frac{(v_0, v_s)\langle v_s, v_s \rangle\psi'(||v_s||_2)}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2^2} \\
\Xi_5(v, g_1, g_2) = -\frac{(v_0, v_s)\langle v_s, v_s \rangle\psi'(||v_s||_2)}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2^2} \\
\Xi_6(v, g_1, g_2) = \frac{(v_0, v_s)\langle v_s, v_s \rangle\psi'(||v_s||_2)}{\psi(||v_0||_2)\psi(||v_s||_2)^3||v_s||_2^3} + \frac{(v_0, v_s)\langle v_s, v_s \rangle\psi'(||v_s||_2)}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2^\frac{3}{2}} + \frac{(v_0, v_s)\langle v_s, v_s \rangle\psi'(||v_s||_2)}{\psi(||v_0||_2)\psi(||v_s||_2)^2||v_s||_2^\frac{3}{2}}
\]

where \( \Xi_i(\cdot, g_1, g_2) \) is defined at \( \{0, \pi\} \) by continuity (following the proof of Lemma A.5.27, it is 0 here). Then for each \( i = 1, \ldots, 6 \), one has:

1. For each \( i \), there is a \( \mu \otimes \mu \)-integrable function \( f_i : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \) such that
   \[ |\Xi_i(\cdot, g_1, g_2) - \mathbb{E}[\Xi_i(\cdot, g_1, g_2)]| \leq f_i(g_1, g_2); \]

2. There is an absolute constant \( C_i > 0 \) such that for every \( 0 \leq \nu \leq \pi \), one has \( ||f_i||_{L^2} \leq C_i \), so that in particular \( ||\Xi_i(\nu, \cdot, \cdot) - \mathbb{E}[\Xi_i(\nu, \cdot, \cdot)]||_{L^2} \leq C_i \).

**Proof.** First we reduce to noncentered fourth moment calculations. If \( X \) is a random variable with finite fourth moment, we have by Minkowski’s inequality

\[ ||X - \mathbb{E}[X]||_{L^4} \leq ||X||_{L^4} + ||\mathbb{E}[X]||. \]
so that the triangle inequality for the expectation and the Lyapunov inequality imply

\[ \| X - \mathbb{E}[X] \|_{L^4} \leq 2 \| X \|_{L^4}. \]

We can therefore control the noncentered fourth moments of the random variables \( \Xi_i \) and pay only an extra factor of 2 in controlling the centered moments. For the proofs of property 1, we have similarly \( |X - \mathbb{E}[X]| \leq |X| + \mathbb{E}[|X|] \) from the triangle inequality, so that it again suffices to prove property 1 for the noncentered random variables \( |\Xi_i| \).

**\( \Xi_1 \) control.** If \( \nu = 0 \) or \( \nu = \pi \), the integrand is identically zero; we proceed assuming \( 0 < \nu < \pi \).

Using \( \psi \geq \frac{1}{4} \), we have

\[
0 \leq \Xi_1(\nu, g_1, g_2) \leq 16 \sum_{i=1}^{n} \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot \nu)}{\sin^3 \nu}.
\]

For property 1, by elementary properties of \( \cos \) we have for \( 0 \leq \nu \leq \pi/4 \) and \( 3\pi/4 \leq \nu \leq \pi \) that \( \cos^2 \nu \geq \frac{1}{2} \), so

\[
\rho(-g_{1i} \cot \nu) \leq \sqrt{\frac{n}{4\pi}} e^{-\frac{ng_{1i}^2}{8\sin^3 \nu}}.
\]

This gives

\[
\frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot \nu)}{\sin^3 \nu} \leq |g_{1i}|^3 \rho(-g_{1i} \cot \nu) = \sqrt{\frac{2}{\pi}} K^{1/2} |g_{1i}| \frac{g_{1i}}{\sin \nu} \left| e^{-\frac{g_{1i}^2}{8\sin^3 \nu}} \right|^2,
\]

where we define \( K = n/8 \). By Lemma A.5.36, we have that \( g \leq g(\sqrt{3/2K}) = CK^{-3/2} \), where \( C > 0 \) is an absolute constant. We conclude

\[
\frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot \nu)}{\sin^3 \nu} \leq C/n,
\]

(A.226)

provided \( \nu \) is not in \([\pi/4, 3\pi/4]\). On the other hand, if \( \pi/4 \leq \nu \leq 3\pi/4 \), we have \( \sin \nu \geq 1/\sqrt{2} \), so that

\[
\frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot \nu)}{\sin^3 \nu} \leq C \sqrt{n} \sigma(g_{1i})^3,
\]

(A.227)
where $C > 0$ is an absolute constant. Since these $\nu$ constraints cover $[0, \pi]$, we have for all $\nu$ and all $g_1$ (by the triangle inequality)

$$|\Xi_1(\nu, g_1, g_2)| \leq C + C' n^{3/2} \sigma(g_1)^3,$$

where $C, C' > 0$ are absolute constants, and by Lemma A.6.11, we have

$$\mathbb{E}[C + C' n^{3/2} \sigma(g_1)^3] = C + C'' ,$$

where $C'' > 0$ is an absolute constant. This proves property 1 with $f_1(g_1, g_2) = C + C' n^{3/2} \sigma(g_1)^3$, with different absolute constants, and property 2 follows from Lemma A.6.11 after applying the Minkowski inequality and calculating the integral, which has the necessary cancellation of the $n^{3/2}$ factor.

**$\Xi_2$ control.** By Lemma A.5.31, we have $|\psi'| \leq C$ for an absolute constant $C > 0$ and $x/\psi(x) \leq 2$. Cauchy-Schwarz then implies

$$\left| \frac{\langle v_0, v_s \rangle \psi'(\|v_s\|_2)\|v_s\|_2}{\psi(\|v_0\|_2)\psi(\|v_s\|_2)^2} \right| \leq 8C .$$

In an exactly analogous manner, we have

$$\left| \frac{\langle v_0, v_s \rangle}{\psi(\|v_0\|_2)\psi(\|v_s\|_2)} \right| \leq 4 .$$

Both bounds satisfy the requirements of property 1, with $f_2(g_1, g_2) = 16C + 8$. The triangle inequality and Minkowski’s inequality then implies $\|\Xi_1(\nu, \cdot, \cdot)\|_{L^1} \leq C'$.

**$\Xi_3$ control.** By Lemma A.5.31, we have $|\psi''| \leq C$ for an absolute constant $C > 0$, $\psi \geq \frac{1}{4}$, and $x/\psi(x) \leq 2$. Cauchy-Schwarz then implies

$$\left| \frac{\langle v_0, v_s \rangle \langle v_s, v_s \rangle^2 \psi''(\|v_s\|_2)}{\psi(\|v_0\|_2)\psi(\|v_s\|_2)^2\|v_s\|_2^2} \right| \leq 16C\|v_s\|_2^2 .$$
and the triangle inequality gives $\|\hat{v}_s\|_2^2 \leq \|g_1\|_2^2 + \|g_2\|_2^2 + 2\|g_1\|_2\|g_2\|_2$, whose expectation is bounded by 4, by the Schwarz inequality and Lemma A.6.11. We can therefore take $f_3(g_1, g_2) = C + C'\|g_1\|_2 + \|g_2\|_2^2$, and we have

$$\| (\|g_1\|_2 + \|g_2\|_2)^2 \|_{L^4} = \|g_1\|_2 + \|g_2\|_2\|_L^2 \leq (\|g_1\|_2 + \|g_2\|_2\|_L)^2 \leq C,$$

where $C > 0$ is a (new) absolute constant, by the Minkowski inequality and lemmas Lemmas A.6.10 and A.6.11. This establishes property 2.

$\Xi_4$ control. By Lemma A.5.31, we have $|\psi'| \leq C$ for an absolute constant $C > 0$, $\psi \geq \frac{1}{4}$, and $x/\psi(x) \leq 2$; Cauchy-Schwarz then implies

$$\left| \frac{2\langle v_0, v_s \rangle \langle v_s, \hat{v}_s \rangle \psi' (\|v_s\|_2)}{\psi (\|v_0\|_2) \psi (\|v_s\|_2)} \right| \leq 64C\|\hat{v}_s\|_2^2.$$

Following the argument for $\Xi_3$ exactly, we conclude property 1 and 2 from this bound with a suitable modification of the constant.

$\Xi_5$ control. We have

$$\left| \frac{\langle v_0, v_s \rangle \|\hat{v}_s\|_2^2 \psi' (\|v_s\|_2)}{\psi (\|v_0\|_2) \psi (\|v_s\|_2)} \right| \leq 32C\|\hat{v}_s\|_2^2,$$

following exactly the setup and instantiations in the argument for $\Xi_4$. Following the argument for $\Xi_3$ exactly, we conclude property 1 and 2 from this bound with a suitable modification of the constant.

$\Xi_6$ control. The triangle inequality gives

$$|\Xi_6(s, g_1, g_2)| \leq 2\left| \frac{\langle v_0, v_s \rangle \langle v_s, \hat{v}_s \rangle^2 \psi' (\|v_s\|_2)}{\psi (\|v_0\|_2) \psi (\|v_s\|_2)^2} \right| + \left| \frac{\langle v_0, v_s \rangle \langle v_s, \hat{v}_s \rangle^2 \psi' (\|v_s\|_2)}{\psi (\|v_0\|_2) \psi (\|v_s\|_2)^2} \right|,$$

and following the setup of $\Xi_4$ and $\Xi_5$ control gives $|\Xi_6(v, g_1, g_2)| \leq 128C\|\hat{v}_s\|_2^2 + 32C\|\hat{v}_s\|_2^2$. Following the argument for $\Xi_3$ exactly, we conclude property 1 and 2 from this bound with a
Lemma A.5.38. In the notation of Lemma A.5.13, there are absolute constants $c, c', C > 0$ and an absolute constant $K > 0$ such that if $n \geq K$, there is an event with probability at least $1 - 2e^{-cn}$ on which one has

$$\left| \frac{\|v_0\|^2}{\psi(||v_0||_2)^2} - \mathbb{E}\left[ \frac{\|v_0\|^2}{\psi(||v_0||_2)^2} \right] \right| \leq Ce^{-c'n}.$$ 

Proof. There is no $\nu$ dependence in this term, so we need only prove a single bound. Following the proof of the measure bound in Lemma A.5.16, but using only the pointwise concentration result, we assert that if $n \geq C$ an absolute constant there is an event $E$ on which $0.5 \leq \|v_0\|_2 \leq 2$ with probability at least $1 - 2e^{-cn}$ with $c > 0$ an absolute constant. This implies that if $g_1 \in E$ we have

$$\frac{\|v_0\|^2}{\psi(||v_0||_2)^2} = 1,$$

which we can use together with nonnegativity of the integrand to calculate

$$\mathbb{E}\left[ \frac{\|v_0\|^2}{\psi(||v_0||_2)^2} \right] = \mathbb{E}[1_E] + \mathbb{E}\left[ 1_E \frac{\|v_0\|^2}{\psi(||v_0||_2)^2} \right] \geq \mathbb{E}[1_E] \geq 1 - 2e^{-cn},$$

whence

$$\frac{\|v_0\|^2}{\psi(||v_0||_2)^2} - \mathbb{E}\left[ \frac{\|v_0\|^2}{\psi(||v_0||_2)^2} \right] \leq 1$$

whenever $g_1 \in E$. Similarly, we calculate

$$\mathbb{E}\left[ \frac{\|v_0\|^2}{\psi(||v_0||_2)^2} \right] = \mathbb{E}[1_E] + \mathbb{E}\left[ 1_E \frac{\|v_0\|^2}{\psi(||v_0||_2)^2} \right] \leq 1 + \mathbb{E}[1_E]^{1/2} \mathbb{E}\left[ \left( \frac{\|v_0\|^2}{\psi(||v_0||_2)} \right)^4 \right]^{1/2} \leq 1 + 16Ce^{-cn},$$

suitable modification of the constant. □

355
applying the Schwarz inequality, property 2 in Lemma A.5.31, and the measure bound on $E$, with $c', C' > 0$ absolute constants, whence

$$E \left[ \|v_0\|_2^2 \right] - \frac{\|v_0\|_2^2}{\psi(\|v_0\|_2)^2} \leq 16C' e^{-c'n}$$

whenever $g_1 \in E$. Worst-casing constants, we conclude

$$\left| \frac{\|v_0\|_2^2}{\psi(\|v_0\|_2)^2} - E \left[ \frac{\|v_0\|_2^2}{\psi(\|v_0\|_2)^2} \right] \right| \leq C e^{-c'n}$$

when $g_1 \in E$, which is sufficient for our purposes.

□

Lemma A.5.39. In the notation of Lemma A.5.13, if $d \geq 1$, there are absolute constants $c, c', c'', C, C', C'', C''' > 0$

and absolute constants $K, K' > 0$ such that if $n \geq K d^4 \log^4 n$ and $d \geq K'$, there is an event with probability at least $1 - C''' n^{-c'd/2} - C''' n e^{-c'n}$ on which one has

$$|\Xi_1(\nu, g_1, g_2) - E[\Xi_1(\nu, g_1, g_2)]| \leq C \sqrt{\frac{d \log n}{n}} + C' n^{-c\bar{d}} + C'' n e^{-c'n}.$$ 

Proof. If $\nu \in \{0, \pi\}$, then $\Xi_1(\nu, g_1, g_2) = 0$ for every $(g_1, g_2)$; we therefore assume $0 < \nu < \pi$ below.

We will apply Lemma A.5.34 to begin, with the instantiations

$$X_i = \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot \nu)}{\sin^3 \nu}, \quad Y_i = \frac{1}{\psi(\|v_0\|_2)\psi(\|v'_i\|_2)},$$
since then $\Xi_1(\nu, g_1, g_2) = \sum_i X_i Y_i$. We have $X_i \geq 0$; writing $k^2 = 2/n$, we calculate

$$
\mathbb{E}[X_i] = \frac{1}{\sqrt{8\pi k^2}} \frac{1}{\sqrt{2\pi k^2}} \int_{\mathbb{R}} \frac{g^3}{\sin^3 \nu} \exp \left( -\frac{1}{2k^2} \frac{g^2}{\sin^2 \nu} \right) \, dg
= \frac{2}{\pi n} \sin \nu
$$  \quad (A.228)

where the second line uses the change of variables $g \mapsto g \sin \nu$ and Lemma A.6.11. Additionally, we have

$$
\mathbb{E}[X_i^2] = \frac{k^4}{4\pi} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{g^6}{\sin^6 \nu} \exp \left( -\frac{1}{2} g^2 (1 + 2 \cot^2 \nu) \right) \, dg
= \frac{k^4 \sin \nu}{4\pi} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g^6 \exp \left( -\frac{1}{2} g^2 (1 + \cos^2 \nu) \right) \, dg
= \frac{k^4 \sin \nu}{4\pi (1 + \cos^2 \nu)^{7/2}} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g^6 e^{-g^2/2} \, dg
= \frac{15 \sin \nu}{\pi n^2 (1 + \cos^2 \nu)^{7/2}},
$$  \quad (A.229)

where in the second line we change variables $g \mapsto g \sin \nu$, in the third line we change variables $g \mapsto g / \sqrt{1 + \cos^2 \nu}$, and in the fourth line we use Lemma A.6.11. We can calculate the derivative of the map $g(\nu) = (1 + \cos^2 \nu)^{-7/2} \sin \nu$ as $g'(\nu) = \cos(\nu)(1 + \cos^2 \nu)^{-7} \left[ (1 + \cos^2 \nu)^{7/2} + 7 \sin^2(\nu)(1 + \cos^2 \nu)^{5/2} \right]$, which evidently has the same sign as $\cos(\nu)$; so $g$ is strictly increasing below $\pi/2$ and strictly decreasing above it, and is therefore maximized at $g(\pi/2)$. We conclude the bound

$$
\mathbb{E}[X_i^2] \leq \frac{15}{\pi n^2},
$$  \quad (A.230)

which shows that $\sum_i \|X_i\|_{L^2} = O(1)$. Next, we have $Y_i \leq 16$ by Lemma A.5.31, so by the Minkowski inequality $\|Y_i - 1\|_{L^4} \leq 17$ for each $i$, and it remains to control deviations. We consider the event $E = E_{0.5,1}$ in the notation of Lemma A.5.16, which has probability at least $1 - Cne^{-cn}$ and on which we have $\frac{1}{2} \leq \|v_i\|_2 \leq 2$ for all $i \in [n]$ and in particular $\frac{1}{2} \leq \|v_0\|_2$, and thus by
Lemma A.5.31

\[ Y_i = \frac{1}{\|v_0\|_2 \|v_i\|_2} \]

for all \( i \in [n] \). By Taylor expansion with Lagrange remainder of the smooth function \( x \mapsto x^{-1} \) on the domain \( x > 0 \) about the point 1, we have

\[ \frac{1}{x} = 1 - (x - 1) + \frac{1}{\xi^3} (x - 1)^2, \]

where \( \xi \) lies between 1 and \( x \). If \((g_1, g_2) \in \mathcal{E}\), then for all \( i \) \( \|v_0\|_2^3 \|v_i\|_2^3 \geq (1/64) \), and we can therefore assert

\[ (\|v_0\|_2^3 \|v_i\|_2^3 - 1) - 64 (\|v_0\|_2^3 \|v_i\|_2^3 - 1)^2 \leq 1 - Y_i \leq (\|v_0\|_2^3 \|v_i\|_2^3 - 1). \] (A.231)

By Gauss-Lipschitz concentration, we have \( \mathbb{P}(\|v_0\|_2 - \mathbb{E}[\|v_0\|_2] \geq t) \leq 2e^{-cnt^2} \) and \( \mathbb{P}(\|v_0\|_2 - \mathbb{E}[\|v_i\|_2] \geq t) \leq 2e^{-cnt^2} \). Lemma A.5.19 implies that \( 1 - 2/(n - 1) \leq \mathbb{E}[\|v_i\|_2] \leq 1 \) and \( 1 - 2/n \leq \mathbb{E}[\|v_0\|_2] \leq 1 \), so we can conclude when \( n \geq d \) and when \( n \) is larger than a constant that

\[ \|v_0\|_2 - 1 \leq C \sqrt{d/n}; \quad \forall i \in [n], \quad \|v_i\|_2 - 1 \leq C \sqrt{d/n} \]

with probability at least \( 1 - C'ne^{-d} \), by a union bound. Using then the fact that \( \|v_i\|_2 \leq 2 \) for all \( i \) on the event \( \mathcal{E} \) together with the previous estimates and (A.231), we obtain with probability at least \( 1 - C''ne^{-cn} - C'''ne^{-d} \) (via a union bound with the measure of \( \mathcal{E} \)) that for all \( i \),

\[ C \sqrt{d/n} - C' \frac{d}{n} \leq 1 - Y_i \leq C \sqrt{d/n}. \]

As long as \( n \geq d \), we conclude that with the same probability, for all \( i \) we have \( |Y_i - 1| \leq C \sqrt{d/n} \).

We can therefore apply Lemma A.5.34 to get that with probability at least \( 1 - C''ne^{-cn} - C'''ne^{-d} \)

358
we have

\[
\left| \Xi_1(\nu, g_1, g_2) - \mathbb{E}[\Xi_1(\nu, g_1, g_2)] \right| \leq 2 \sum_{i=1}^{n} \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot \nu)}{\sin^3 \nu} - \mathbb{E} \left[ \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot \nu)}{\sin^3 \nu} \right] \\
+ C \sqrt{\frac{d}{n}} + (C')^{1/4} n e^{-cn/4} + (C'')^{1/4} n e^{-d/4},
\]

(A.232)

where we also used the triangle inequality for the \(l_4\) norm to simplify the fourth root term, together with \(n \geq 1\). For \(\nu \in [0, \pi]\), we define \(f_\nu : \mathbb{R} \to \mathbb{R}\) by

\[
f_\nu(g) = \frac{\sigma(g)^3}{\sqrt{2\pi k^2 \sin^3 \nu}} \exp \left( -\frac{1}{2k^2 g^2 \cot^2 \nu} \right),
\]

so that the task that remains is to control \(|\sum_i f_\nu(g_{1i}) - \mathbb{E}[f_\nu(g_{1i})]|\). First, Lemma A.5.36 gives an estimate

\[
f_\nu(g) \leq \frac{C}{n|\cos \nu|^3},
\]

where \(C > 0\) is an absolute constant. When \(0 \leq \nu \leq \pi/4\) or \(3\pi/4 \leq \nu \leq \pi\), we have therefore \(f_\nu(g) \leq C/n\). Meanwhile, if \(\pi/4 \leq \nu \leq 3\pi/4\), we have \(f_\nu(g) \leq C' \sqrt{n} \sigma(g)^3\), so we can conclude \(f_\nu(g) \leq C/n + C' \sqrt{n} \sigma(g)^3\) for all \(\nu\), which shows that \(f_\nu(g)\) is not much larger than \(C' \sqrt{n} \sigma(g)^3\).

Next, let \(\bar{g} \sim \mathcal{N}(0, 1)\), so that \(g = \frac{d}{k} \bar{g}\); we have for any \(t \geq 0\)

\[
\mathbb{P} \left[ C' \sqrt{n} \sigma(g)^3 \geq t \right] = \mathbb{P} \left[ \sigma(\bar{g}) \geq C'' (nt)^{1/3} \right] \leq \exp \left( -\frac{1}{2} (C'')^2 (nt)^{2/3} \right),
\]

where we use the classical estimate \(\mathbb{P}[\bar{g} \geq t] \leq e^{-t^2/2}\), valid for \(t \geq 1\), and accordingly require \(t \geq (C'')^{-3} n^{-1}\). In particular, there is an absolute constant \(C'' > 0\) such that we have

\[
\mathbb{P} \left[ \frac{C' \sqrt{n} \sigma(g)^3}{\sqrt{nd}} \geq \frac{C''}{\sqrt{nd}} \right] = \mathbb{P} \left[ \sigma(\bar{g}) \geq \left( \frac{n}{d} \right)^{1/6} \right] \leq \exp \left( -\frac{1}{2} \left( \frac{n}{d} \right)^{1/3} \right) \leq e^{-d},
\]

where the last inequality holds in particular when \(n \geq 8d^4\) (and this condition implies what is necessary for the second to last to hold when \(d \geq 1\)). Returning to our bound on \(f_\nu\), we note that
when \( n \geq (C/C'')^2 d \), we have that

\[
f_r(g) - \frac{2C''}{\sqrt{nd}} \leq \frac{C}{n} + C' \sqrt{n} \sigma(g)^3 - \frac{2C''}{\sqrt{nd}} \leq C' \sqrt{n} \sigma(g)^3 - \frac{C''}{\sqrt{nd}},
\]

from which we conclude that when our previous hypotheses on \( n \) are in force

\[
\mathbb{P}\left[ f_r(g) \geq \frac{2C''}{\sqrt{nd}} \right] \leq e^{-d}.
\]  

(A.233)

We are going to use this result to control \( |\sum_i f_r(g_{1i}) - \mathbb{E}[f_r(g_{1i})]| \) using a truncation approach. Define \( M = 2C''/\sqrt{nd} \), where \( C'' > 0 \) is the absolute constant in (A.233). We write using the triangle inequality

\[
\left| \sum_{i=1}^n f_r(g_{1i}) - \mathbb{E}[f_r(g_{1i})] \right| \leq \left| \sum_{i=1}^n f_r(g_{1i}) - f_r(g_{1i}) \mathbb{1}_{f_r(g_{1i}) \leq M} \right| \\
+ \left| \sum_{i=1}^n f_r(g_{1i}) \mathbb{1}_{f_r(g_{1i}) \leq M} - \mathbb{E}\left[ f_r(g_{1i}) \mathbb{1}_{f_r(g_{1i}) \leq M} \right] \right| \\
+ \left| \sum_{i=1}^n \mathbb{E}\left[ f_r(g_{1i}) \mathbb{1}_{f_r(g_{1i}) \leq M} \right] - \mathbb{E}[f_r(g_{1i})] \right|.
\]

By (A.233) and a union bound, we have with probability at least \( 1 - ne^{-d} \)

\[
\left| \sum_{i=1}^n f_r(g_{1i}) - f_r(g_{1i}) \mathbb{1}_{f_r(g_{1i}) \leq M} \right| = 0.
\]

Moreover, we calculate

\[
\left| \sum_{i=1}^n \mathbb{E}\left[ f_r(g_{1i}) \mathbb{1}_{f_r(g_{1i}) \leq M} \right] - \mathbb{E}[f_r(g_{1i})] \right| \leq \sum_{i=1}^n \mathbb{E}[f_r(g_{1i}) \mathbb{1}_{f_r(g_{1i}) > M}] \\
\leq \sum_{i=1}^n \mathbb{P}[f_r(g_{1i}) > M]^{1/2} \|f_r(g_{1i})\|_{L^2} \\
\leq Ce^{-d/2}
\]

360
for an absolute constant $C > 0$, using in the second line the Schwarz inequality, and in the third line (A.230) and (A.233). The second term can be controlled with Lemma A.6.3, together with the observation that

$$
\sum_{i=1}^{n} \mathbb{E} \left[ (\mathbb{1}_{f_{\nu}(g_{1i}) \leq M f_{\nu}(g_{1i})} - \mathbb{E}[\mathbb{1}_{f_{\nu}(g_{1i}) \leq M f_{\nu}(g_{1i})]])^2 \right]
$$

$$
= \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}_{f_{\nu}(g_{1i}) \leq M f_{\nu}(g_{1i})}^2] - \mathbb{E}[\mathbb{1}_{f_{\nu}(g_{1i}) \leq M f_{\nu}(g_{1i})}]^2
$$

$$
\leq \sum_{i=1}^{n} \mathbb{E}[f_{\nu}(g_{1i})^2]
$$

$$
\leq C/n,
$$

where the last inequality is due to (A.230). Lemma A.6.3 thus gives for any $t \geq 0$

$$
P \left( \left| \sum_{i=1}^{n} f_{\nu}(g_{1i}) \mathbb{1}_{f_{\nu}(g_{1i}) \leq M} - \mathbb{E}[f_{\nu}(g_{1i}) \mathbb{1}_{f_{\nu}(g_{1i}) \leq M}] \right| \geq t \right) \leq 2 \exp \left( -\frac{t^2/2}{Cn^{-1} + Mt/3} \right).
$$

It follows that there is an absolute constant $C' > 0$ such that

$$
P \left( \left| \sum_{i=1}^{n} f_{\nu}(g_{1i}) - \mathbb{E}[f_{\nu}(g_{1i})] \right| \geq C' \sqrt{\frac{d}{n}} \right) \leq 2e^{-d},
$$

and therefore with probability at least $1 - 2ne^{-d}$ (by a union bound) we have

$$
\left| \sum_{i=1}^{n} f_{\nu}(g_{1i}) - \mathbb{E}[f_{\nu}(g_{1i})] \right| \leq C' \sqrt{\frac{d}{n}} + C''e^{-d/2}.
$$

Combining with (A.232) using a union bound and worst-casing constants in the exponent, we conclude that with probability at least $1 - C'''ne^{-cd} - C''''ne^{-c''''n}$, we have

$$
|\Xi_1(\nu, g_1, g_2) - \mathbb{E}[\Xi_1(\nu, g_1, g_2)]| \leq C' \sqrt{\frac{d}{n}} + C''e^{-cd} + C'''ne^{-c'n}.
$$

Aggregating our hypotheses on $n$, there are absolute constants $C_1, C_2, C_3 > 0$ such that we have
to satisfy $n \geq \max\{Cd, C'd^4, C''\}$. Moreover, to be able to assert $ne^{-c''d} \leq e^{-c''d/2}$, we have to satisfy $d \geq 2/c'' \log n$. Introducing an auxiliary $\tilde{d} > 0$ and setting $d = \tilde{d} \log n$, we have to satisfy $n \geq \max\{C\tilde{d} \log n, C'd^4 \log^4 n, C''\}$ and $\tilde{d} \geq 2/c''$. Choosing $n$ in this way, we can finally conclude that with probability at least $1 - C''n^{-c''\tilde{d}/2} - C'''ne^{-c''n}$, we have

$$|\Xi_1(\nu, g_1, g_2) - \mathbb{E}[\Xi_1(\nu, g_1, g_2)]| \leq C\sqrt{\frac{\tilde{d} \log n}{n}} + C'\tilde{d} + C''n^{-c''n},$$

which is the desired type of bound.

\[\Box\]

**Lemma A.5.40.** In the notation of Lemma A.5.13, there are absolute constants $c, C, C', C'' > 0$ such that for any $\delta \geq 3/2$, we have

$$\mathbb{P}\left[\left|\mathbb{E}_{g_1, g_2}[\Xi_1(\nu, g_1, g_2)] - \mathbb{E}_{g_1, g_2}[\Xi_1(\nu, g_1, g_2)]\right| \geq C + C'n^{1+\delta}\text{-Lipschitz}\right] \geq 1 - 2e^{-cn} - C''n^{-\delta}.$$

**Proof.** Write $f(\nu, g_1) = \mathbb{E}_{g_2}[\Xi_1(\nu, g_1, g_2)]$; it will suffice to differentiate $f$ and $\mathbb{E}[f]$ with respect to $\nu$, bound the derivatives on an event of high probability, and apply the triangle inequality to obtain a high-probability Lipschitz estimate for $|\mathbb{E}_{g_2}[\Xi_1(\nu, g_1, g_2)] - \mathbb{E}_{g_1, g_2}[\Xi_1(\nu, g_1, g_2)]|$.

Define $k = \sqrt{2/n}$. For fixed $(g_1, g_2)$, the function

$$q(\nu, g_1, g_2) = \sum_{i=1}^{\mathcal{N}} \frac{\sigma(g_{1i})^3 \rho(-g_{1i} \cot \nu)}{\psi(||\nu_0||_2)\psi(||\nu'_i||_2)\sin^3 \nu}$$

is differentiable at all but at most $n$ points of $(0, \nu)$, using Lemma A.5.31 to see that the only obstruction to differentiability is the function $\sigma$ in the term $||\nu'_i||_2$; and there has derivative

$$q'(\nu, g_1, g_2) = \sum_{i=1}^{\mathcal{N}} \frac{\sigma(g_{1i})^3}{\sqrt{2\pi k^2} \psi(||\nu_0||_2)} \left(\begin{array}{c} \frac{g_{1i}^2 \cos \nu}{k^2 \psi(||\nu'_i||_2) \sin^6 \nu} \\ \frac{-\psi(||\nu'_i||_2)(\nu'_i, \nu'_i) - \psi(||\nu'_i||_2)\psi(||\nu'_i||_2)\sin^3 \nu}{\psi(||\nu'_i||_2)^2} \\ -3 \frac{\cos \nu}{\psi(||\nu'_i||_2) \sin^3 \nu} \end{array}\right) \exp\left(-\frac{1}{2k^2 g_{1i}^2 \cot^2 \nu}\right).$$

362
The triangle inequality and Lemma A.5.31 yield
\[
|q'(\nu, g_1, g_2)| \leq \frac{4}{\sqrt{2\pi}k^2} \sum_{i=1}^{n} |g_{1i}|^3 \left( \frac{4g_{1i}^2}{k^2 \sin^6 \nu} + \frac{16C\|\dot{\nu}_i\|_2}{\sin^3 \nu} + \frac{12}{\sin^4 \nu} \right) \exp \left( -\frac{1}{2k^2}g_{1i}^2 \cot^2 \nu \right)
\]  
(A.234)
for $C > 0$ an absolute constant. We have $\|\dot{\nu}_i\|_2 \leq \|g_1\|_2 + \|g_2\|_2$ by the triangle inequality, so to obtain a $(\nu, g_2)$-integrable upper bound it suffices to remove the $\nu$ dependence from the previous estimate. We argue as follows: if $0 \leq \nu \leq \pi/4$ or $3\pi/4 \leq \nu \leq \pi$, we have $\cos^2 \nu \geq \frac{1}{2}$, and so for any $p \geq 3$
\[
\frac{\exp \left( -\frac{1}{2k^2}g_{1i}^2 \cot^2 \nu \right)}{\sin^p \nu} \leq \exp \left( \frac{g_{1i}^2}{4k^2 \sin^2 t} \right) \sin^{-p} \nu.
\]  
(A.235)
By Lemma A.5.36, where we put $C = \frac{g_{1i}^2}{4k^2}$ and therefore have to require that $g_{1i} \neq 0$ for all $i \in [n]$ (a set of measure zero in $\mathbb{R}^n$), this yields
\[
|q'(\nu, g_1, g_2)| \leq C\|g_2\|_2,
\]  
(A.236)
where $C > 0$ is a constant depending only on $n$ and $g_1$. In cases where $g_{1i} = 0$ for some $i$, we note that the bound (A.234) is then equal to zero, which also satisfies the estimate (A.236). On the other hand, when $\pi/4 \leq \nu \leq 3\pi/4$, then $\sin t \geq 2^{-1/2}$, and we can assert for any $p \geq 3$
\[
\exp \left( -\frac{1}{2k^2}g_{1i}^2 \cot^2 \nu \right) \leq 2^{p/2}.
\]
By the triangle inequality, this too implies
\[
|q'(\nu, g_1, g_2)| \leq C'\|g_2\|_2,
\]
where $C' > 0$ is a constant depending only on $n$ and $g_1$. Invoking then Lemma A.6.9, we conclude that $q'$ is absolutely integrable over $[0, \pi] \times \mathbb{R}^n$, so that an application of Fubini’s theorem and [219, Theorem 6.3.11] gives the Taylor expansion
\[
f(\nu, g_1) = f(0, g_1) + \int_0^\nu E_{g_2}[q'(t, g_1, g_2)] \, dt.
\]
Next, we show also that \( q' \) is absolutely integrable over \([0, \pi] \times \mathbb{R}^n \times \mathbb{R}^n\), which implies that
\[
E[f(\nu, g_1)] = E[f(0, g_1)] + \int_0^\nu E[q'(t, g_1, g_2)] \, dt
\]
as well. Starting from (A.234), we have
\[
E[|q'(\nu, g_1, g_2)|] \leq E\left[\frac{4}{\sqrt{2\pi}k^2} \sum_{i=1}^n |g_{1i}|^3 \left( \frac{4g_{1i}^2}{k^2 \sin^6 \nu} + \frac{16C\|g_1^i\|_2 + \|g_2^i\|_2}{\sin^3 \nu} + \frac{12}{\sin^4 \nu} \exp \left( -\frac{1}{2k^2g_{1i}^2 \cot^2 \nu} \right) \right) \right],
\]
and the expectation factors over \( g_{1i}, g_1^i, g_2^i \), so we can separately compute the \( g_{1i} \) integrals first. For the first of the three terms on the RHS of the previous expression, we have
\[
E_{g_{1i}} \left[ \frac{|g_{1i}|^5}{\sin^6 \nu} \exp \left( -\frac{1}{2k^2g_{1i}^2 \cot^2 \nu} \right) \right] = \frac{1}{\sqrt{2\pi}k^2} \int_{\mathbb{R}} \frac{|g|^5}{\sin^6 \nu} \exp \left( -\frac{1}{2k^2g^2/\sin^2 \nu} \right) \, dg
= \frac{1}{\sqrt{2\pi}k^2} \int_{\mathbb{R}} |g|^5 \exp \left( -\frac{1}{2k^2g^2} \right) \, dg,
\]
(A.237)
after the change of variables \( g \mapsto g \sin \nu \) in the integral, which is valid whenever \( 0 < \nu < \pi \). To take care of the case where \( \nu = 0 \) or \( \nu = \pi \), we can use the estimate (A.235), valid for \( \nu \) sufficiently close to 0 or \( \pi \), and the assumption \( g_{1i} \neq 0 \) for all \( i \) to conclude that \( \lim_{\nu \downarrow 0} q'(\nu, g_1, g_2) = 0 \) for any such fixed \( (g_1, g_2) \), and by symmetry the analogous result \( \lim_{\nu \uparrow \pi} q'(\nu, g_1, g_2) = 0 \); and whenever for some \( i \) we have \( g_{1i} = 0 \), we use (A.234) to see that the term in the sum involving \( g_{1i} \) poses no problems as \( \nu \downarrow 0 \) or \( \nu \uparrow \pi \) because it is identically 0. Returning to the integral (A.237), we have after a change of variables
\[
\frac{1}{\sqrt{2\pi}k^2} \int_{\mathbb{R}} |g|^5 \exp \left( -\frac{1}{2k^2g^2} \right) \, dg = \frac{k^5}{\sqrt{2\pi}} \int_{\mathbb{R}} |g|^5 \exp \left( -\frac{1}{2g^2} \right) \, dg = Ck^5,
\]
where \( C > 0 \) is an absolute constant, and where we use Lemma A.6.11 for the last equality. The remaining two terms can be treated using the same argument: we get
\[
E_{g_{1i}} \left[ \frac{|g_{1i}|^3}{\sin^3 \nu} \exp \left( -\frac{1}{2k^2g_{1i}^2 \cot^2 \nu} \right) \right] = C'k^3
\]
(after using \(|\sin \nu| \leq 1\) and
\[
\mathbb{E}_{\|g_i\|} \left[ \frac{|g_{1i}|^3}{\sin^4 \nu} \exp \left( -\frac{1}{2k^2 g_{1i}^2 \cot^2 \nu} \right) \right] = C'' k^3
\]
for absolute constants \(C', C'' > 0\). Combining these estimates gives
\[
\mathbb{E}[|q'(\nu, g_1, g_2)|] \leq \frac{C}{n} \sum_{i=1}^{n} \mathbb{E}_{g'_1, g'_2} \left[ \|g'_1\|_2 + \|g'_2\|_2 \right],
\]
and using Lemma A.6.9 (or equivalently Jensen’s inequality) gives finally
\[
\mathbb{E}[|q'(\nu, g_1, g_2)|] \leq C \sqrt{\frac{n-1}{n}} \leq C.
\]
To conclude, we need to show that \(\mathbb{E}_{g_2}[q'(\nu, g_1, g_2)]\) is uniformly bounded by a polynomial in \(n\) with high probability. For this we start from the estimate (A.234) and apply the argument following that, but with more care in tracking the constants: if \(\nu\) is within \(\pi/4\) of either 0 or \(\pi\), we can assert
\[
|q'(\nu, g_1, g_2)| \leq \frac{C}{k} \sum_{i=1}^{n} \frac{C_1 k^4}{|g_{1i}|} + C_2 k^3 (\|g'_1\|_2 + \|g'_2\|_2) + \frac{C_3 k^4}{|g_{1i}|}
\]
whenever \(g_{1i} \neq 0\) for every \(i\) (a set of full measure); and when \(\nu\) is within \(\pi/4\) of \(\pi/2\), we can assert
\[
|q'(\nu, g_1, g_2)| \leq \frac{C}{k} \sum_{i=1}^{n} \frac{C'_i |g_{1i}|^5}{k^2} + C_2 |g_{1i}|^3 (\|g'_1\|_2 + \|g'_2\|_2) + C'_3 |g_{1i}|^3,
\]
where \(C_i, C'_i > 0\) are absolute constants. By Lemma A.6.9, the triangle inequality, and independence, when we consider \(\mathbb{E}_{g_2}[q'(\nu, g_1, g_2)]\), the term \(\mathbb{E}[\|g'_2\|_2]\) is bounded by an absolute constant. Additionally, by Gauss-Lipschitz concentration and Lemma A.6.9, we have that simultaneously for all \(i \|g'_1\|_2 \leq \|g_1\|_2 \leq 2\) with probability at least \(1 - 2e^{-cn}\). Moreover, since \(\|g_1\|_\infty \leq \|g_1\|_2\) we also have control of the magnitude of each \(|g_{1i}|\) on this event, so with proba-
probability at least $1 - 2e^{-cn}$ we have for every $\nu$

$$\left| \mathbb{E}_{g_2}[q'(\nu, g_1, g_2)] \right| \leq \frac{C}{k} \sum_{i=1}^{n} \frac{C_1 k^4}{|g_{1i}|} + C_2 k^3 + \frac{C_3}{k^2} + C_4$$

for absolute constants $C, C_i > 0$. If $X \sim N(0, 1)$, we have for any $t \geq 0$ that $\mathbb{P}(|X| \geq t) \geq 1 - Ct$, where $C > 0$ is an absolute constant; so if $X_i \sim \text{i.i.d.} \ N(0, 1)$, we have by independence and if $t$ is less than an absolute constant $\mathbb{P}[\forall i, |X_i| \geq t] \geq (1 - Ct)^n \geq 1 - C'n t$, where the last inequality uses the numerical inequality $e^{-2t} \leq 1 - t \leq e^{-t}$, valid for $0 \leq t \leq \frac{1}{2}$. From this expression, we conclude that when $0 \leq t \leq cn^{-1/2}$ for an absolute constant $c > 0$, we have

$$\mathbb{P}[\forall \nu \in [0, \pi], |g_{1i}| \geq t] \geq 1 - C n^{3/2} t,$$

so choosing in particular $t = cn^{-(\delta + \frac{3}{2})}$ for any $\delta > 0$, we conclude that $\mathbb{P}[\forall i \in [n], |g_{1i}| \geq cn^{-3/2-\delta}] \geq 1 - C'n^{-\delta}$. Consequently, for any $\delta > 0$ we have with probability at least $1 - C'n^{-\delta} - 2e^{-cn}$

$$\left| \mathbb{E}_{g_2}[q'(\nu, g_1, g_2)] \right| \leq \frac{C}{k} \sum_{i=1}^{n} \frac{C_1 k^4 n^{3/2+\delta}}{|g_{1i}|} + C_2 k^3 + \frac{C_3}{k^2} + C_4,$$

and since $k = \sqrt{2/n}$, this yields $\left| \mathbb{E}_{g_2}[q'(\nu, g_1, g_2)] \right| \leq C_1 n^{1+\delta} + C_2 + C_3 n^{5/2} + C_4 n^{3/2}$ with the same probability. Consequently we can conclude that for any $\delta \geq 3/2$, we have

$$\mathbb{P} \left[ \left| \mathbb{E}_{g_2}[\Xi_1(\nu, g_1, g_2)] - \mathbb{E}_{g_1, g_2}[\Xi_1(\nu, g_1, g_2)] \right| \text{ is } C + C'n^{1+\delta} \text{-Lipschitz} \right] \geq 1 - 2e^{-cn} - C'n^{-\delta}.$$

□

**Lemma A.5.41.** In the notation of Lemma A.5.13, if $d \geq 1$, there are absolute constants $c, c' > 0$ and $C, C' > 0$ and an absolute constant $K > 0$ such that if $n \geq K$, there is an event with probability at least $1 - Ce^{-cn}$ on which

$$\forall \nu \in [0, \pi], |\Xi_2(\nu, g_1, g_2) - \mathbb{E}[\Xi_2(\nu, g_1, g_2)]| \leq C'e^{-cn}.$$
Proof. Let $E$ denote the event $E_{0.5,0}$ in Lemma A.5.16; then by that lemma, $E$ has probability at least $1 - Ce^{-cn}$ as long as $n \geq C'$, where $c, C, C' > 0$ are absolute constants, and for $(g_1, g_2) \in E$, one has for all $v \in [0, \pi]$

$$\Xi_2(v, g_1, g_2) = 0.$$ 

This allows us to calculate, for each $v$,

$$\mathbb{E}[\Xi_2(v, g_1, g_2)] = \mathbb{E}[\mathbb{1}_{E} \Xi_2(v, g_1, g_2)] \leq \mathbb{E}[\mathbb{1}_{E} \Xi_2(v, \cdot)]_{L^2} \leq C' e^{-c'n},$$

after applying Lemma A.5.37 and Lyapunov’s inequality and worst-casing constants. We conclude that with probability at least $1 - Ce^{-cn}$

$$\forall v \in [0, \pi], \ |\Xi_2(v, g_1, g_2) - \mathbb{E}[\Xi_2(v, g_1, g_2)]| \leq C' e^{-c'n}.$$ 

□

Lemma A.5.42. In the notation of Lemma A.5.13, if $d \geq 1$, there are absolute constants $c, c' > 0$ and $C, C' > 0$ and an absolute constant $K > 0$ such that if $n \geq K$, there is an event with probability at least $1 - Ce^{-cn}$ on which

$$\forall v \in [0, \pi], \ |\Xi_3(v, g_1, g_2) - \mathbb{E}[\Xi_3(v, g_1, g_2)]| \leq C' e^{-c'n}.$$ 

Proof. The argument is identical to Lemma A.5.41. Let $E$ denote the event $E_{0.5,0}$ defined in Lemma A.5.16; then by that lemma, $E$ has probability at least $1 - Ce^{-cn}$ as long as $n \geq C'$, where $c, C, C' > 0$ are absolute constants, and for $(g_1, g_2) \in E$, one has for all $v \in [0, \pi]$

$$\Xi_3(v, g_1, g_2) = 0.$$ 

367
This allows us to calculate, for each \( \nu \),

\[
\mathbb{E}[\Xi_3(\nu, g_1, g_2)] = \mathbb{E}[1_{\mathcal{E}} \Xi_3(\nu, g_1, g_2)] \leq \mathbb{E}[1_{\mathcal{E}e}]^{1/2} \|\Xi_3(\nu, \cdot)\|_{L^2} \leq C' e^{-c'n},
\]

after applying Lemma A.5.37 and Lyapunov’s inequality and worst-casing constants. We conclude that with probability at least \( 1 - Ce^{-cn} \)

\[
\forall \nu \in [0, \pi], |\Xi_3(\nu, g_1, g_2) - \mathbb{E}[\Xi_3(\nu, g_1, g_2)]| \leq C' e^{-c'n}.
\]

\[\square\]

**Lemma A.5.43.** In the notation of Lemma A.5.13, if \( d \geq 1 \), there are absolute constants \( c > 0 \) and \( C, C', C'' > 0 \) and absolute constants \( K, K' > 0 \) such that if \( n \geq Kd \log n \) and \( d \geq K' \), there is an event with probability at least \( 1 - Ce^{-cn} - C'n^{-d} \) on which one has

\[
\forall \nu \in [0, \pi], |\Xi_4(\nu, g_1, g_2) - \mathbb{E}[\Xi_4(\nu, g_1, g_2)]| \leq C'' \sqrt{\frac{d \log n}{n}}.
\]

**Proof.** We are going to control the expectation first, showing that it is small; then prove that \( |\Xi_4| \) is small uniformly in \( \nu \). Let \( \mathcal{E} \) denote the event \( \mathcal{E}_{0.5,0} \) in Lemma A.5.16; then by that lemma, \( \mathcal{E} \) has probability at least \( 1 - Ce^{-cn} \) as long as \( n \geq C' \), where \( c, C, C' > 0 \) are absolute constants, and for \((g_1, g_2) \in \mathcal{E}, \) one has for all \( \nu \in [0, \pi] \)

\[
\Xi_4(\nu, g_1, g_2) = -2 \frac{\langle v_0, \hat{v}_\nu \rangle \langle v_\nu, \hat{v}_\nu \rangle}{\|v_0\|_2 \|v_\nu\|_2^3}.
\]

Thus, if we write

\[
\Xi_4(\nu, g_1, g_2) = -2 \mathbb{E}[\Xi_4(\nu, g_1, g_2)] + 1_{\mathcal{E}}(g_1, g_2) \frac{\langle v_0, \hat{v}_\nu \rangle \langle v_\nu, \hat{v}_\nu \rangle}{\|v_0\|_2 \|v_\nu\|_2^3},
\]

we have \( \Xi_4 = \Xi_4 \) for all \( \nu \) whenever \((g_1, g_2) \in \mathcal{E}, \) so that for any \( \nu \)

\[
|\mathbb{E}[\Xi_4(\nu, g_1, g_2)]| = |\mathbb{E}[\Xi_4(\nu, g_1, g_2)] + 1_{\mathcal{E}} \Xi_4(\nu, g_1, g_2)|
\]

368
≤ |E[Ξ_4(ν, g_1, g_2)]| + Ce^{-cn},

where the second line uses the triangle inequality and the Schwarz inequality and Lemma A.5.37 together with the Lyapunov inequality. We proceed with analyzing the expectation of Ξ_4. Using the Schwarz inequality gives

\[ E[(v_0, v_ν)^2(v_ν, v_ν)^2]^1/2 \leq 32E[(v_0, v_ν)^2(v_ν, v_ν)^2]^{1/2}, \]

and the checks at and around (A.208) and (A.209) in the proof of Lemma A.5.15 show that we can apply Lemma A.5.30 to obtain

\[ E[σ(g_{11})σ(g_{11} cos v + g_{21} sin v)(g_{21} cos v - g_{11} sin v)]^2 \leq C/n. \]

But we have using rotational invariance that \( E[σ(g_{11} cos v + g_{21} sin v)(g_{21} cos v - g_{11} sin v)] = 0, \)

which implies

\[ |E[(v_0, v_ν)^2(v_ν, v_ν)^2]| ≤ C/n, \]

from which we conclude

\[ |E[Ξ_4(ν, g_1, g_2)]| ≤ C/\sqrt{n}. \]

Next, we control the deviations of Ξ_4 with high probability. By Lemma A.5.17, there is an event \( E_u \) with probability at least \( 1 - e^{-cn} \) on which \( ||v_ν||_2 ≤ 4 \) for every \( ν \in [0, π] \). Therefore on the event \( E_b = E \cap E_u \), which has probability at least \( 1 - Ce^{-cn} \) by a union bound, we have using Cauchy-Schwarz that for every \( ν \)

\[ |Ξ_4(ν, g_1, g_2)| ≤ 256|⟨v_ν, v_ν⟩|. \]
The coordinates of the random vector $\mathbf{v} \otimes \hat{\mathbf{v}}$ are $\sigma(g_{i1} \cos \nu + g_{i2} \sin \nu)(g_{2i} \cos \nu - g_{1i} \sin \nu)$, and we note

$$\mathbb{E}[\sigma(g_{i1} \cos \nu + g_{i2} \sin \nu)(g_{2i} \cos \nu - g_{1i} \sin \nu)] = -\mathbb{E}[\sigma(g_{i1})g_{2i}] = 0,$$

by rotational invariance. Moreover, (A.209) together with Lemmas A.6.11 and A.5.17 demonstrates subexponential moment growth with rate $C/n$, so Lemma A.6.2 implies for $t \geq 0$

$$\mathbb{P}[\langle \mathbf{v}, \hat{\mathbf{v}} \rangle \geq t] \leq 2e^{-cnt \min\{c't,1\}}.$$

For large enough $n$, this gives

$$\mathbb{P}\left[\langle \mathbf{v}, \hat{\mathbf{v}} \rangle \geq C\sqrt{\frac{d \log n}{n}}\right] \leq 2n^{-2d}.$$

We turn to the uniformization of this pointwise bound. The map

$$\nu \mapsto \sum_{i} \sigma(g_{i1} \cos \nu + g_{i2} \sin \nu)(g_{2i} \cos \nu - g_{1i} \sin \nu)$$

is continuous, and differentiable at all but finitely many points of $[0,\pi]$ (following the zero crossings argument in the proof of Lemma A.5.22) with derivative

$$\nu \mapsto \sum_{i} \left(\sigma(g_{i1} \cos \nu + g_{i2} \sin \nu)(g_{2i} \cos \nu - g_{1i} \sin \nu)^2 - \sigma(g_{i1} \cos \nu + g_{i2} \sin \nu)^2\right),$$

which is evidently integrable using the triangle inequality and Lemma A.6.11. In particular, we can write the derivative as $||\hat{\mathbf{v}} \nu||_{2}^2 - ||\mathbf{v} \nu||_{2}^2$. Thus, by [219, Theorem 6.3.11], to get a Lipschitz estimate on $\nu \mapsto \langle \mathbf{v}, \hat{\mathbf{v}} \rangle$ it suffices to bound the magnitude of the derivative $\nu \mapsto ||\hat{\mathbf{v}} \nu||_{2}^2 - ||\mathbf{v}_0||_{2}^2$. But this is immediate, since on the event $\mathcal{E}_b$ we have $||\hat{\mathbf{v}} \nu||_{2}^2 - ||\mathbf{v}_0||_{2}^2 \leq 20$. It thus follows from Lemma A.5.48 that with probability at least $1 - Ce^{-cn} - C'n^{-2d+1/2}$ we have

$$\forall \nu \in [0,\pi], \langle \mathbf{v}, \hat{\mathbf{v}} \rangle \leq C'' \sqrt{\frac{d \log n}{n}}.$$

(A.238)
As long as $d \geq \frac{1}{2}$, we have that this probability is at least $1 - Ce^{-cn} - C'n^{-d}$, and so the triangle inequality and a union bound yield finally that with probability at least $1 - Ce^{-cn} - C'n^{-d}$

$$\forall \nu \in [0, \pi], \left| \mathbb{E}_4(\nu, g_1, g_2) - \mathbb{E}[\mathbb{E}_4(\nu, g_1, g_2)] \right| \leq C'' \sqrt{\frac{d \log n}{n}}.$$  

□

**Lemma A.5.44.** In the notation of Lemma A.5.13, if $d \geq 1$, there are absolute constants

$$c, c', c'', C, C', C'', C''', C'''' > 0$$

and an absolute constant $K > 0$ such that if $n \geq Kd \log n$, there is an event with probability at least $1 - Ce^{-cn} + C'e^{-d}$ on which one has

$$|\mathbb{E}_5(\nu, g_1, g_2) - \mathbb{E}[\mathbb{E}_5(\nu, g_1, g_2)]| \leq C'' \sqrt{\frac{d}{n}} + C'''e^{-c'd} + C''''e^{-c''n}.$$  

**Proof.** Fix $\nu \in [0, \pi]$. Let $\mathcal{E}$ denote the event $\mathcal{E}_{0,5,0}$ in Lemma A.5.16; then by that lemma, $\mathcal{E}$ has probability at least $1 - Ce^{-cn}$ as long as $n \geq C'$, where $c, C, C' > 0$ are absolute constants, and for $(g_1, g_2) \in \mathcal{E}$, one has for all $\nu \in [0, \pi]$

$$\mathbb{E}_5(\nu, g_1, g_2) = \frac{\langle v_0, v_\nu \rangle \|v_\nu\|_2^2}{\|v_0\|_2 \|v_\nu\|_2^3}.$$  

Thus, if we write

$$\bar{\mathbb{E}}_5(\nu, g_1, g_2) = -1_{\mathcal{E}}(g_1, g_2) \frac{\langle v_0, v_\nu \rangle \|v_\nu\|_2^2}{\|v_0\|_2 \|v_\nu\|_2^3}$$

we have $\bar{\mathbb{E}}_5 = \mathbb{E}_5$ for any $\nu$ whenever $(g_1, g_2) \in \mathcal{E}$, so that by the triangle inequality, for any $\nu$

$$|\mathbb{E}_5(\nu, g_1, g_2) - \mathbb{E}[\mathbb{E}_5(\nu, g_1, g_2)]| \leq |\bar{\mathbb{E}}_5(\nu, g_1, g_2) - \mathbb{E}[\bar{\mathbb{E}}_5(\nu, g_1, g_2)]|$$

$$+ |\mathbb{E}[\bar{\mathbb{E}}_5(\nu, g_1, g_2)] - \mathbb{E}[\bar{\mathbb{E}}_5(\nu, g_1, g_2)]|$$

371
\[
\leq \left\| \widetilde{\Xi}_5(\nu, g_1, g_2) - \mathbb{E}\left[ \widetilde{\Xi}_5(\nu, g_1, g_2) \right] \right\| \\
+ \mathbb{E}\left[ \mathbb{1}_E \| \widetilde{\Xi}_5(\nu, g_1, g_2) - \widetilde{\Xi}_5(\nu, g_1, g_2) \| \right] \\
\leq \left\| \widetilde{\Xi}_5(\nu, g_1, g_2) - \mathbb{E}\left[ \widetilde{\Xi}_5(\nu, g_1, g_2) \right] \right\| + Ce^{-cn},
\]

where the second line uses the triangle inequality, and the third line uses the Schwarz inequality and Lemma A.5.37 together with the Lyapunov inequality.

So, we can proceed analyzing \( \widetilde{\Xi}_5 \). First, we aim to apply Lemma A.5.33 with the choices

\[
X = -\mathbb{1}_E \frac{\langle v_0, v_\nu \rangle}{\|v_0\|_2 \|v_\nu\|_2}; \quad Y = \mathbb{1}_E \frac{\|\dot{v}_\nu\|_2^2}{\|v_\nu\|_2^2},
\]

since \( XY = \widetilde{\Xi}_5(\nu, \cdot, \cdot) \); square-integrability of \( X \) and \( Y \) is evident from the definition of \( \mathbb{1}_E \), and we have \( |X| \leq 1 \) by Cauchy-Schwarz. To control \( Y \), we start by noting

\[
\|Y - 1\|_L^2 \leq 1 + \|Y\|_L^2 \leq 1 + 4\mathbb{E} \left[ \|\dot{v}_\nu\|_2^2 \right]^{1/2} \leq 1 + 4\sqrt{1 + C},
\]

where \( C > 0 \) is an absolute constant; the first inequality is the Minkowski inequality, the second uses the property of \( E \) and drops the indicator by nonnegativity, and the third applies the result Lemma A.5.29, and discards the \( n^{-1} \) factor. For deviations, we start by noting that \( \mathbb{E}[\|v_\nu\|_2^2] = 1 \), and that by Lemmas A.6.2 and A.6.11, we have

\[
\mathbb{P}\left[ \|\dot{v}_\nu\|_2^2 - 1 \geq t \right] \leq 2e^{-cnt\min\{Cr,1\}}.
\]

It follows that there exists an absolute constant \( C' > 0 \) such that, putting \( t = C'\sqrt{d/n} \) and choosing \( n \geq (C'/C)^2d \), we have

\[
\mathbb{P}\left[ \left. \|\dot{v}_\nu\|_2^2 - 1 \geq C' \sqrt{\frac{d}{n}} \right| \leq 2e^{-d}. \quad (A.239)
\]

Moreover, by Lemma A.5.17, we can run a similar argument on \( \|\dot{v}_\nu\|_2^2 \) to get that if \( n \) is larger than
a constant multiple of \( d \)

\[
\mathbb{P} \left[ \| \tilde{v}_\nu \|_2^2 - 1 \right] \geq C \sqrt{\frac{d}{n}} \leq 2e^{-d}. \tag{A.240}
\]

Next, Taylor expansion with Lagrange remainder of the smooth function \( x \mapsto x^{-1} \) on the domain \( x > 0 \) about the point 1 gives

\[
\frac{1}{x} = 1 - (x - 1) + \frac{1}{\xi^3} (x - 1)^2, \tag{A.241}
\]

where \( \xi \) lies between 1 and \( x \). If \( (g_1, g_2) \in E \), then \( \| v_\nu \|_2^6 \geq (1/64) \), and we can therefore assert

\[
1 - \left( \| v_\nu \|_2^2 - 1 \right) \leq \frac{1}{\| v_\nu \|_2^2} \leq 1 - \left( \| v_\nu \|_2^2 - 1 \right) + 64 \left( \| v_\nu \|_2^2 - 1 \right)^2
\]

with probability at least \( 1 - Ce^{-cn} \). Using a union bound together with (A.239) (and changing the constant to \( C \)), we have with probability at least \( 1 - 2e^{-cd} - C' e^{-c'n} \) that

\[
-C \sqrt{\frac{d}{n}} - 64C^2 \frac{d}{n} \leq 1 - \frac{1}{\| v_\nu \|_2^2} \leq C \sqrt{\frac{d}{n}}.
\]

Given that \( n \geq d \), it follows that with the same probability we have

\[
-C(1 + 64C) \sqrt{\frac{d}{n}} \leq 1 - \frac{1}{\| v_\nu \|_2^2} \leq C \sqrt{\frac{d}{n}},
\]

which implies that with probability at least \( 1 - 2e^{-d} - C' e^{-cn} \), we have

\[
\left| 1 - \frac{1}{\| v_\nu \|_2^2} \right| \leq C \sqrt{\frac{d}{n}}.
\]

Now, the triangle inequality gives

\[
\left| \frac{\| \tilde{v}_\nu \|_2^2 - 1}{\| v_\nu \|_2^2} - \frac{1}{\| v_\nu \|_2^2} \right| \leq \frac{1}{\| v_\nu \|_2^2} \left| \| \tilde{v}_\nu \|_2^2 - 1 \right| + \frac{1}{\| v_\nu \|_2^2 - 1}.
\]

373
When \((g_1, g_2) \in \mathcal{E}\), we have \(\|v_\nu\|^2 \geq \frac{1}{4}\), so, by a union bound, with probability at least \(1 - 4e^{-d} - C'e^{-cn}\) we have

\[
\left| \frac{\|v_\nu\|^2}{\|v_\nu\|^2} - 1 \right| \leq 4C\sqrt{\frac{d}{n}},
\]

and since

\[(g_1, g_2) \in \mathcal{E} \implies Y = \frac{\|v_\nu\|^2}{\|v_\nu\|^2},\]

another union bound and the measure bound on \(\mathcal{E}\) let us conclude that with probability at least \(1 - 4e^{-d} - C'e^{-cn}\), we have

\[
|Y - 1| \leq 4C\sqrt{\frac{d}{n}}.
\]

If we choose \(n \geq (1/c)(d + \log C'/4)\), we have \(4e^{-d} + C'e^{-cn} \leq 8e^{-d}\), so the previous bound occurs with probability at least \(1 - 8e^{-d}\). We can now apply Lemma A.5.33 to get with probability at least \(1 - 8e^{-d}\)

\[
\left| \bar{\Xi}_5 - \mathbb{E}[\bar{\Xi}_5] \right| \leq \mathbb{E} \left[1_{\mathcal{E}} \left( \frac{v_0}{\|v_0\|_2}, \frac{v_\nu}{\|v_\nu\|_2} \right) \right] - \mathbb{E} \left[1_{\mathcal{E}} \left( \frac{v_0}{\|v_0\|_2}, \frac{v_\nu}{\|v_\nu\|_2} \right) \right] + C\sqrt{\frac{d}{n}} + C'e^{-d/2}.
\]

Next, we attempt to apply Lemma A.5.33 again, this time to \(X = 1_{\mathcal{E}} \langle v_0, v_\nu \rangle\) and

\[
Y = 1_{\mathcal{E}}(\|v_0\|_2\|v_\nu\|_2)^{-1}.
\]

Using the definition of \(\mathcal{E}\), we have \(|X| \leq 4\) and \(\|Y - 1\|_{L^2} \leq \|Y\|_{L^2} + 1 \leq 5\), where the second bound also leverages the Minkowski inequality; so we need only establish deviations of \(Y\). Applying again (A.241), and using \((g_1, g_2) \in \mathcal{E}\) implies \(\|v_0\|_2\|v_\nu\|_2 \geq \frac{1}{4}\), we get

\[
(\|v_0\|_2\|v_\nu\|_2 - 1) - 64 (\|v_0\|_2\|v_\nu\|_2 - 1)^2 \leq 1 - \frac{1}{\|v_0\|_2\|v_\nu\|_2} \leq (\|v_0\|_2\|v_\nu\|_2 - 1)
\]

with probability at least \(1 - Ce^{-cn}\). Using Lemma A.6.11 and [225, Theorem 3.1.1], we can assert
for any $\nu \in [0, \pi]$ and any $t \geq 0$

$$\mathbb{P}[|\|v_\nu\|_2 - 1| \geq t] \leq 2e^{-cn^2},$$

which implies that there exists an absolute constant $C > 0$ such that for any $d > 0$

$$\mathbb{P}\left[|\|v_\nu\|_2 - 1| \geq C\sqrt{\frac{d}{n}}\right] \leq 2e^{-d}.$$

In particular, when $n \geq d$, we can assert that $\|v_\nu\|_2 \leq 1 + C$ with probability at least $1 - 2e^{-d}$. By the triangle inequality and a union bound, it follows

$$\|v_0\|_2|\|v_\nu\|_2 - 1| \leq \|v_0\|_2|\|v_\nu\|_2 - 1| + \|v_0\|_2 - 1| \leq C\sqrt{\frac{d}{n}}$$

with probability at least $1 - 6e^{-d}$. Then a union bound gives that with probability at least $1 - 6e^{-d} - C'e^{-cn}$, (A.242) leads to

$$-C\sqrt{\frac{d}{n}}\left(1 + 64C\sqrt{\frac{d}{n}}\right) \leq 1 - \frac{1}{\|v_0\|_2\|v_\nu\|_2} \leq C\sqrt{\frac{d}{n}},$$

and using $n \geq d$ and worst-casing constants implies that with the same probability

$$\left|1 - \frac{1}{\|v_0\|_2\|v_\nu\|_2}\right| \leq C\sqrt{\frac{d}{n}}.$$

Then since $(g_1, g_2) \in \mathcal{E} \implies Y = (\|v_0\|_2\|v_\nu\|_2)^{-1}$, another union bound gives that with probability at least $1 - 6e^{-d} - C'e^{-cn}$ we have $|Y - 1| \leq C\sqrt{d/n}$. As in the previous step of the reduction, we can choose $n \geq (1/c)(d + \log C'/6)$ to get that $6e^{-d} + C'e^{-cn} \leq 12e^{-d}$, so that the previous bound occurs with probability at least $1 - 12e^{-d}$. We can thus apply Lemma A.5.33, a union bound, and
our previous work to get that with probability at least $1 - 20e^{-d}$

$$
\left| \Xi_5 - \mathbb{E}[\Xi_5] \right| \leq |\mathbb{1}_E\langle v_0, v_\nu \rangle - \mathbb{E}[\mathbb{1}_E\langle v_0, v_\nu \rangle]| + C\sqrt{\frac{d}{n}} + C'e^{-d/2}.
$$

Whenever $(g_1, g_2) \in E$, we have by the triangle inequality, the Schwarz inequality, and Lemmas A.5.16 and A.5.29 that

$$
|\mathbb{1}_E\langle v_0, v_\nu \rangle - \mathbb{E}[\mathbb{1}_E\langle v_0, v_\nu \rangle]| \leq |\langle v_0, v_\nu \rangle - \mathbb{E}[\langle v_0, v_\nu \rangle]| + |\mathbb{E}[\langle v_0, v_\nu \rangle] - \mathbb{E}[\mathbb{1}_E\langle v_0, v_\nu \rangle]| \leq |\langle v_0, v_\nu \rangle - \mathbb{E}[\langle v_0, v_\nu \rangle]| + Ce^{-cn},
$$

allowing us to drop the indicator. We have $\langle v_0, v_\nu \rangle = \sum_i \sigma(g_{1i})\sigma(g_{1i} \cos \nu + g_{2i} \sin \nu)$, which is a sum of independent random variables; following the argument at and around (A.210), we conclude moreover that these random variables are subexponential with rate $C/n$, where $C > 0$ is an absolute constant. We therefore obtain from Lemma A.6.2 the tail bound

$$
\mathbb{P}[|\langle v_0, v_\nu \rangle - \mathbb{E}[\langle v_0, v_\nu \rangle]| \geq t] \leq 2e^{-cnt\min\{Ct, 1\}},
$$

which, for a suitable choice of absolute constant $C' > 0$ and choosing $n \geq C'd$, yields the deviations bounds

$$
\mathbb{P}
\left|
\langle v_0, v_\nu \rangle - \mathbb{E}[\langle v_0, v_\nu \rangle]
\right| \geq C\sqrt{\frac{d}{n}} \leq 2e^{-d}.
$$

Taking a final union bound (since we assumed throughout that $(g_1, g_2) \in E$) gives that with probability at least $1 - Ce^{-cn} + C'e^{-d}$, one has

$$
|\Xi_5(\nu, g_1, g_2) - \mathbb{E}[\Xi_5(\nu, g_1, g_2)]| \leq C''\sqrt{\frac{d}{n}} + C'''e^{-c'd} + C''''e^{-c''n},
$$

which is sufficient to conclude pointwise concentration as claimed for sufficiently large $n$ after we put $d = d' \log n$ and include extra $\log n$ factors in any points where we need to choose $n$ larger than $d$. \hfill \square
Lemma A.5.45. In the notation of Lemma A.5.13, there are absolute constants $c, C, C', C'' > 0$ such that for any $\delta \geq \frac{1}{2}$, one has

$$P\left[\left| \mathbb{E}_{g_2} \left[ \Xi_5(v, g_1, g_2) \right] - \mathbb{E}_{g_1, g_2} \left[ \Xi_5(v, g_1, g_2) \right] \right| \leq C + C'n^{1+\delta} \right] \geq 1 - C''e^{-cn} - C'''n^{-\delta}$$

as long as $\delta \geq \frac{1}{2}$.

Proof. We will differentiate with respect to $v$ the function

$$f(v, g_1) = -\mathbb{E}_{g_2} \left[ \langle v_0, v \rangle \|v\|_2^2 \psi'(\|v\|_2) \right],$$

and construct an event on which $f'$ has size $\text{poly}(n)$. We need to also differentiate the function $\mathbb{E}[f(\cdot, g_1)]$; for this we will additionally show that $f'(v, \cdot)$ is absolutely integrable over the product $[0, \pi] \times \mathbb{R}^n \times \mathbb{R}^n$, which allows us to apply Fubini’s theorem to move both the $g_1$ and $g_2$ expectations under the $v$ integral in the first-order Taylor expansion we obtain. In particular, the derivative of $\mathbb{E}[f(\cdot, g_1)]$ will in this way be shown to be $\mathbb{E}[f'(\cdot, g_1)]$, so that linearity and the triangle inequality imply a $\text{poly}(n)$ magnitude bound for the derivative of $\mathbb{E}_{g_2}[\Xi_5] - \mathbb{E}[\Xi_5]$.

Define

$$q_i(v, g_1, g_2) = \frac{\langle v_0, v \rangle \psi'(\|v\|_2)(g_{2i} \cos v - g_{1i} \sin v)^2}{\psi(\|v_0\|_2)\psi'(\|v\|_2)^2\|v\|_2^2},$$

so that, for almost all $g_1$,

$$f(v, g_1) = -\sum_{i=1}^n g_2 \left[ q_i(v, g_1, g_2) \sigma(g_{1i} \cos v + g_{2i} \sin v) \right].$$

For each fixed $(g_1, g_2)$ and each $i$, the only obstructions to differentiability of $q_i$ in $v$ arise from the function $\sigma$ (using smoothness of $\psi$ from Lemma A.5.31 and the fact that it is constant whenever $\|v\|$ is small enough that nondifferentiability of $\|\cdot\|_2$ could pose a problem); following the zero-crossings argument of Lemma A.5.22, $q_i$ fails to be differentiable at no more than $n$ points of
which produces a final upper bound that does not depend on \( \mathcal{H} \). Solute integrability of \( \sigma \) from continuity of \( R \) on Lemma A.5.27, so we need to check its remaining hypotheses. First, continuity of \( \mathcal{H} \) using Lemma A.5.37 to see that Fubini’s theorem can be applied. Our aim is now to apply absolute continuity of \( \mathcal{H} \) to show that \( \mathcal{H} \) by the chain rule and the product rule. To conclude absolute continuity of \( \mathcal{H} \), we need to show that \( \mathcal{H} \) is absolutely continuous with a.e. derivative \( \mathcal{H} \). Next, we can write

\[
q_i'(v, g_1, g_2) = \frac{1}{\psi(\|v_0\|_2)} \left( \langle v_0, \dot{v}_v \rangle \psi'(\|v_2\|_2)(g_2 \cos v - g_{1i} \sin v)^2 \right. \\
+ \frac{\langle v_0, v_v \rangle \psi''(\|v_2\|_2)(g_2 \cos v - g_{1i} \sin v)^2}{\psi(\|v_2\|_2)^2}\|v_v\|_2 \\
- \frac{2\langle v_0, v_v \rangle \psi'(\|v_2\|_2)(g_2 \cos v - g_{1i} \sin v)(g_{1i} \cos v + g_{2i} \sin v)}{\psi(\|v_2\|_2)^2}\|v_v\|_2 \\
- \frac{2\psi'(\|v_2\|_2)^2\langle v_v, v_v \rangle \langle v_0, v_v \rangle (g_2 \cos v - g_{1i} \sin v)^2}{\psi(\|v_2\|_2)^3}\|v_v\|_2 \\
- \frac{\psi'(\|v_2\|_2)^2\langle v_v, v_v \rangle \langle v_0, v_v \rangle (g_2 \cos v - g_{1i} \sin v)^2}{\psi(\|v_2\|_2)^2}\|v_v\|^3_2)
\]

by the chain rule and the product rule. To conclude absolute continuity of \( q_i(\cdot, g_1, g_2) \), we need to show that \( q_i' \) is integrable; this follows from Cauchy-Schwarz, the integrability of \( \|v_0\|_2, \|v_v\|_2, \|\dot{v}_v\|_2 \) (Lemma A.5.17), the triangle inequality, and the Lemma A.5.31 estimates \( \psi \geq \frac{1}{4}, |\psi'| \leq C, |\psi''| \leq C' \), and \( |\psi'(x)/x| \leq C'' \) for any \( x \in \mathbb{R} \) (to see this last estimate, note that \( |\psi'| \) is bounded on \( \mathbb{R} \), and use that \( \psi \) is constant whenever \( x \leq \frac{1}{4} \)). Then [219, Theorem 6.3.11] implies that \( q_i(\cdot, g_1, g_2) \) is absolutely continuous with a.e. derivative \( q_i' \). Next, we can write

\[
f(v, g_1) = -\sum_{i=1}^n \mathbb{E}_{g_{2i}} \mathbb{E}_{g_{2i}} [q_i(v, g_1, g_2) \sigma(g_{1i} \cos v + g_{2i} \sin v)]
\]

using Lemma A.5.37 to see that Fubini’s theorem can be applied. Our aim is now to apply Lemma A.5.27, so we need to check its remaining hypotheses. First, continuity of \( q_i(v, \cdot) \) follows from continuity of \( \sigma \), smoothness of \( \psi \), and the fact that the denominator never vanishes. Joint absolute integrability of \( q_i \) and \( q_i' \) follows from our verification of absolute integrability of \( q_i' \) above, which produces a final upper bound that does not depend on \( v \) (which is therefore integrable over \([0, \pi]\) as well); the corresponding result for \( q_i \) follows from Lemma A.5.37. Last, we need the
growth estimate. We have from Lemma A.5.31

\[
|q_i(\nu, \mathbf{g}_1, \mathbf{g}_2)| \leq 32C (g_{2i} \cos \nu - g_{1i} \sin \nu)^2 \leq 32C (|g_{2i}| + |g_{1i}|)^2 \leq 32C |g_{1i}|(1 + |g_{2i}|)^2,
\]

which is evidently quadratic in $|g_{2i}|$ once $|g_{2i}| \geq 1$. Consequently we can apply Lemma A.5.27 to differentiate $f(\cdot, \mathbf{g}_1)$; we get at almost all $\mathbf{g}_1$

\[
f(\nu, \mathbf{g}_1) = -\sum_{i=1}^{n} \left( \mathbb{E}_{g_{2j}:j \neq i} \left[ \mathbb{E}_{g_{2i}} [q_i(0, \mathbf{g}_1, \mathbf{g}_2) \hat{\sigma}(g_{1i})] \right] \right)
\]

\[+ \mathbb{E}_{g_{2j}:j \neq i} \left[ \int_{0}^{\nu} \left( \mathbb{E}_{g_{2i}} \left[ q'_i(t, \mathbf{g}_1, \mathbf{g}_2) \hat{\sigma}(g_{1i} \cos t + g_{2i} \sin t) \right] \right) \right]
\]

\[\cdot \left( -g_{1i} q_i(t, \mathbf{g}_1, \overline{\mathbf{g}}_2)^i \rho(-g_{1i} \cot t) \sin^{-2} t \right) \]

where $\overline{\mathbf{g}}_2^i$ is the vector $\mathbf{g}_2$ but with its $i$-th coordinate replaced by $-g_{1i} \cot t$, and where $\rho$ is the pdf of a $N(0, 2/n)$ random variable. The changes in $\overline{\mathbf{g}}_2^i$ drive updates to the terms in $q_i$ as follows: we have $\sigma(g_{1i} \cos \nu + g_{2i} \sin \nu)$ becoming $0$, and $g_{2i} \cos \nu - g_{1i} \sin \nu$ becoming $-g_{1i} \sin \nu$. Thus, we have

\[-g_{1i} q_i(t, \mathbf{g}_1, \overline{\mathbf{g}}_2^i) \rho(-g_{1i} \cot t) \sin^{-2} t = -g_{1i}^3 (v^i_0, v^i_0) \psi''(||v^i_0||_2) \rho(-g_{1i} \cot t) \]

\[\psi(||v^i_0||_2) \psi(||v^i_0||_2)^2 ||v^i_0||_2 \sin^4 t, \]

where the notation $v^i_0$ is in use in the $\Xi_1$ control section and is defined in Lemma A.5.26, and $v^i_0$ is defined here similarly (the $\mathbb{R}^{n-1}$ vector which is the projection of $v_0$ onto all but the $i$-th coordinates). Using Lemma A.5.31, we can further assert

\[
|g_{1i} q_i(t, \mathbf{g}_1, \overline{\mathbf{g}}_2^i) \rho(-g_{1i} \cot t) \sin^{-2} t| \leq 16C \frac{|g_{1i}|^3 \rho(-g_{1i} \cot t)}{\sin^4 t} \]

(A.244)

where we use that $||v^i_0||_2 \leq ||v_0||_2$. For each fixed $\mathbf{g}_1$ having no coordinates equal to zero, we write $K_i = |g_{1i}| > 0$; if $0 \leq t \leq \pi/4$ or $3\pi/4 \leq t \leq \pi$, we have $\cos^2 t \geq \frac{1}{2}$, and so

\[
\frac{\rho(-g_{1i} \cot t)}{\sin^4 t} \leq \sqrt{\frac{n}{4\pi}} \sin^{-4} t \exp \left( \frac{K_i^2 n}{8} \frac{1}{\sin^2 t} \right).
\]
Using Lemma A.5.36, we have
\[
\frac{\rho(-g_{1i} \cot t)}{\sin^4 t} \leq \sqrt{\frac{n}{4\pi}} \left( \frac{16}{K_i^2 n} \right)^2.
\]

On the other hand, when \( \pi/4 \leq t \leq 3\pi/4 \), then \( \sin t \geq 2^{-1/2} \), and we can assert
\[
\frac{\rho(-g_{1i} \cot t)}{\sin^4 t} \leq 8\sqrt{n/\pi}.
\]

We conclude for any \( t \)
\[
|g_{1i}q_i(t, g_1, \tilde{g}_2^i)\rho(-g_{1i} \cot t)\sin^{-2} t| \leq C/(K_i n^{3/2}) + C' \sqrt{n K_i^3}
\]
(A.245)

for absolute constants \( C, C' > 0 \), and this upper bound is integrable jointly over \( t \) and \( g_2 \). We have checked previously the joint integrability of the \( q'_i \) terms when applying Lemma A.5.27, so we can therefore apply Fubini’s theorem to get \( g_1 \)-a.s.
\[
f(\nu, g_1) = -\mathbb{E}_{g_2} \left[ \sum_{i=1}^n q_i(0, g_1, g_2) \sigma(g_{1i}) \right]
- \int_0^\nu \mathbb{E}_{g_2} \left[ q'_i(t, g_1, g_2) \sigma(g_{1i} \cos t + g_{2i} \sin t) \right]
\left[ -g_{1i}q_i(t, g_1, \tilde{g}_2^i)\rho(-g_{1i} \cot t)\sin^{-2} t \right] dt.
\]

Consequently, to conclude a Lipschitz estimate for \( f(\cdot, g_1) \) it suffices to control the quantity under the \( t \) integral in the previous expression. We will start by controlling the second term using Markov’s inequality. Following (A.244), we calculate
\[
\mathbb{E}_{g_1, g_2} \left[ |g_{1i}q_i(t, g_1, \tilde{g}_2^i)\rho(-g_{1i} \cot t)\sin^{-2} t| \right] \leq 8C \int_{\mathbb{R}} \frac{|g_{1i}|^3 \exp \left( -\frac{n g_{1i} \cos^2 t}{4 \sin^2 t} \right)}{\sin^4 t} dg
\]
\[
= 4Cn \int_{\mathbb{R}} \frac{|g|^3}{\sin^4 t} \exp \left( -\frac{n g^2}{4 \sin^2 t} \right) dg.
\]
\[ \frac{4Cn}{\pi} \int_{\mathbb{R}} |g|^3 \exp \left( -\frac{n}{4} g^2 \right) dg, \]

where the last line follows from the change of variables \( g \mapsto g \sin t \) in the integral. We can evaluate this integral with Lemma A.6.11, which gives a bound

\[ \mathbb{E}_{g_1, g_2} \left[ |g_{1i} q_i(t, g_1, \overline{g}_2^j) \rho(-g_{1i} \cot t) \sin^{-2} t| \right] \leq \frac{128C}{\pi n}, \]

and therefore a bound of \( C' > 0 \) an absolute constant on the sum over \( i \). As a byproduct of this estimate, we can assert that the second term is jointly integrable over \([0, \pi] \times \mathbb{R}^n \times \mathbb{R}^n\), which allows us to apply Fubini’s theorem and obtain the same differentiation result for \( \mathbb{E}[f(\cdot, g_1)] \).

Meanwhile, beginning from (A.245), we can write using the triangle inequality

\[ \left| \mathbb{E}_{g_2} \left[ \sum_{i=1}^{n} g_{1i} q_i(t, g_1, \overline{g}_2^j) \rho(-g_{1i} \cot t) \sin^{-2} t \right] \right| \leq \sum_{i=1}^{n} \frac{C}{|g_{1i}|^{n^{3/2}}} + C' \sqrt{n} |g_{1i}|^3. \]

By Gauss-Lipschitz concentration and Lemma A.6.9, we have that \( \|g_1\|_2 \leq \|g_1\|_2 \leq 2 \) with probability at least \( 1 - 2e^{-cn} \), and since \( \|g_1\|_\infty \leq \|g_1\|_2 \), we conclude with the same probability that \( |g_{1i}| \leq 2 \) simultaneously for all \( i \). Meanwhile, if \( X \sim \mathcal{N}(0, 1) \), we have for any \( t \geq 0 \) that \( \mathbb{P}[|X| \geq t] \geq 1 - Ct \), where \( C > 0 \) is an absolute constant; so if \( X_i \sim \text{i.i.d.} \mathcal{N}(0, 1) \), we have by independence and if \( t \) is less than an absolute constant \( \mathbb{P}[\forall i, |X_i| \geq t] \geq (1 - Ct)^n \geq 1 - C' nt \), where the last inequality uses the numerical inequality \( e^{-2t} \leq 1 - t \leq e^{-t} \), valid for \( 0 \leq t \leq \frac{1}{2} \).

From this expression, we conclude that when \( 0 \leq t \leq cn^{-1/2} \) for an absolute constant \( c > 0 \), we have

\[ \mathbb{P}[\forall i \in [n], |g_{1i}| \geq t] \geq 1 - Cn^{3/2} t, \]

so choosing in particular \( t = cn^{-(\delta + \frac{3}{2})} \) for any \( \delta > 0 \), we conclude that \( \mathbb{P}[\forall i \in [n], |g_{1i}| \geq cn^{-3/2 - \delta}] \geq 1 - C' n^{-\delta} \). Then with probability at least \( 1 - C' n^{-\delta} - 2e^{-cn} \), we have

\[ \left| \mathbb{E}_{g_2} \left[ \sum_{i=1}^{n} g_{1i} q_i(t, g_1, \overline{g}_2^j) \rho(-g_{1i} \cot t) \sin^{-2} t \right] \right| \leq C n^{1+\delta} + C' n^{3/2}, \]
so as long as $\delta \geq \frac{1}{2}$, we have

$$
P \left[ \left\| \frac{1}{g_2} \sum_{i=1}^{n} g_{1i} q_i(t, g_1, g_2) \rho(-g_{1i} \cot t \sin^{-2} t) \right\| \geq Cn^{1+\delta} \right] \leq C' n^{-\delta} + 2e^{-cn}.
$$

Proceeding now to the $q_i'$ term, from the expression (A.243) we get

$$
\sum_{i=1}^{n} q_i'(v, g_1, g_2) \dot{\sigma}(g_{1i} \cos v + g_{2i} \sin v)

= \frac{1}{\psi(||v_0||_2)} \left( \frac{\langle v_0, \dot{v}_v \rangle \psi'(||v||_2) ||\dot{v}_v||_2^2}{\psi(||v||_2)^2 ||v||_2} + \frac{\langle v_0, v_v \rangle \langle v_v, \dot{v}_v \rangle \psi''(||v||_2) ||\dot{v}_v||_2^2}{\psi(||v||_2)^2 ||v||_2^2} \right)

- 2 \frac{\langle v_0, v_v \rangle \psi'(||v||_2) \langle \dot{v}_v, v_v \rangle}{\psi(||v||_2)^2 ||v||_2^3 ||\dot{v}_v||_2^3} - 2 \frac{\langle v_0, v_v \rangle \langle v_v, \dot{v}_v \rangle \langle \dot{v}_v, v_v \rangle ||\dot{v}_v||_2^2}{\psi(||v||_2)^2 ||v||_2^3 ||\dot{v}_v||_2^3}.
$$

Using the triangle inequality, Cauchy-Schwarz, and Lemma A.5.31, we obtain

$$
\frac{\langle v_0, v_v \rangle \psi'(||v||_2) ||\dot{v}_v||_2^2}{\psi(||v_0||_2) \psi(||v_v||_2^2) ||v_v||_2} + \frac{\langle v_0, v_v \rangle \langle v_v, \dot{v}_v \rangle \psi''(||v||_2) ||\dot{v}_v||_2^2}{\psi(||v_0||_2) \psi(||v_v||_2^2) ||v_v||_2^2} \leq C ||\dot{v}_v||_2^3,
$$

(assuming also the fact that $\psi'(x) = 0$ and $\psi''(x) = 0$ whenever $x$ is sufficiently near to 0); and

$$
\frac{\langle v_0, v_v \rangle \psi'(||v||_2) \langle \dot{v}_v, v_v \rangle}{\psi(||v||_2)^2 ||v_v||_2} + \frac{\psi'(||v||_2) \langle v_v, \dot{v}_v \rangle \langle v_0, v_v \rangle ||\dot{v}_v||_2^2}{\psi(||v||_2)^3 ||v_v||_2^3} \leq C ||\dot{v}_v||_2 + C' ||\dot{v}_v||_2^3,
$$

from which we conclude

$$
\left| \sum_{i=1}^{n} q_i'(v, g_1, g_2) \dot{\sigma}(g_{1i} \cos v + g_{2i} \sin v) \right| \leq C ||\dot{v}_v||_2 + C' ||\dot{v}_v||_2^3
$$

for some absolute constants $C, C' > 0$. By Lemma A.5.17, there is an event $E$ of probability at
least $1 - Ce^{-cn}$ on which we have $\|\hat{v}_v\|_2 \leq 4$ for every $v$. Moreover, we have from the triangle inequality that $\|\hat{v}_v\|_2 \leq \|g_1\|_2 + \|g_2\|_2$, which is independent of $v$; and in particular we have

$$\left| \sum_{i=1}^{n} q'_i(v, g_1, g_2) \hat{\sigma}(g_{1i} \cos v + g_{2i} \sin v) \right|^2 \leq \left( C(\|g_1\|_2 + \|g_2\|_2) + C'(\|g_1\|_2 + \|g_2\|_2)^3 \right)^2,$$

which is a polynomial in $\|g_1\|_2$ and $\|g_2\|_2$ by the binomial theorem. Thus, applying independence, Lemma A.6.10, Lemma A.6.11 yields that there is an absolute constant $C'' > 0$ such that

$$\mathbb{E}_{g_1, g_2} \left[ \left( C(\|g_1\|_2 + \|g_2\|_2) + C'(\|g_1\|_2 + \|g_2\|_2)^3 \right)^2 \right] \leq C''.$$

Therefore, as in the framework section of the proof of Lemma A.5.13, we can use the inequality

$$\left| \mathbb{E}_{g_2} \left[ \sum_{i=1}^{n} q'_i(v, g_1, g_2) \hat{\sigma}(g_{1i} \cos v + g_{2i} \sin v) \right] \right| \leq \mathbb{E}_{g_2} \left[ \left| \sum_{i=1}^{n} q'_i(v, g_1, g_2) \hat{\sigma}(g_{1i} \cos v + g_{2i} \sin v) \right| \right],$$

(A.247)

together with the partition

$$\mathbb{E}_{g_2} \left[ \sum_{i=1}^{n} q'_i(v, g_1, g_2) \hat{\sigma}(g_{1i} \cos v + g_{2i} \sin v) \right] \leq C' + \mathbb{E}_{g_2} \left[ 1_{(E')^c} \left| \sum_{i=1}^{n} q'_i(v, g_1, g_2) \hat{\sigma}(g_{1i} \cos v + g_{2i} \sin v) \right| \right],$$

(A.248)

and this last expression can be used to obtain a $g_1$ event of not much smaller probability $1 - Ce^{-cn}$ on which the LHS of (A.248), and hence the LHS of (A.247), is controlled by an absolute constant uniformly in $v$ (in particular, using Markov’s inequality as in the framework section of the proof of Lemma A.5.13). Consequently, one more application of the triangle inequality gives that

$$\mathbb{P}\left[ \mathbb{E}_{g_2} [\Xi_5(v, g_1, g_2)] - \mathbb{E}_{g_1, g_2} [\Xi_5(v, g_1, g_2)] \right] \text{ is } C' n^{1+\delta} \text{-Lipschitz} \geq 1 - C'' e^{-cn} - C''' n^{-\delta}$$

as long as $\delta \geq \frac{1}{2}$. □

**Lemma A.5.46.** In the notation of Lemma A.5.13, if $d \geq 1$, there are absolute constants $c > 0$ and
$C, C', C'' > 0$ and absolute constants $K, K' > 0$ such that if $n \geq Kd \log n$ and $d \geq K'$, there is an event with probability at least $1 - Ce^{-cn} - C'n^{-d}$ on which one has

$$\forall \nu \in [0, \pi], \left| \Xi_6(\nu, g_1, g_2) - \mathbb{E}[\Xi_6(\nu, g_1, g_2)] \right| \leq C'' \sqrt{\frac{d \log n}{n}}.$$ 

**Proof.** The argument is extremely similar to Lemma A.5.43, since both terms have small expectations and deviations essentially determinable by the same mean-zero random variable.

We are going to control the expectation first, showing that it is small; then prove that $|\Xi_6|$ is small uniformly in $\nu$. Let $\mathcal{E}$ denote the event $\mathcal{E}_{0.5.0}$ in Lemma A.5.16; then by that lemma, $\mathcal{E}$ has probability at least $1 - Ce^{-cn}$ as long as $n \geq C'$, where $c, C, C' > 0$ are absolute constants, and for $(g_1, g_2) \in \mathcal{E}$, one has for all $\nu \in [0, \pi]$

$$\Xi_6(\nu, g_1, g_2) = 3\frac{\langle v_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle^2}{\|v_0\|_2^2 \|v_\nu\|_2^5}.$$ 

Thus, if we write

$$\Xi_6^*(\nu, g_1, g_2) = 3\mathbb{1}_{\mathcal{E}}(g_1, g_2)\frac{\langle v_0, v_\nu \rangle \langle v_\nu, \dot{v}_\nu \rangle^2}{\|v_0\|_2^2 \|v_\nu\|_2^5},$$

we have $\Xi_6 = \Xi_6^*$ for all $\nu$ whenever $(g_1, g_2) \in \mathcal{E}$, so that for any $\nu$

$$|\mathbb{E}[\Xi_6(\nu, g_1, g_2)]| = |\mathbb{E}[\Xi_6^*(\nu, g_1, g_2)] + \mathbb{E}[\mathbb{1}_{\mathcal{E}} \Xi_6(\nu, g_1, g_2)]|$$

$$\leq |\mathbb{E}[\Xi_6^*(\nu, g_1, g_2)]| + Ce^{-cn},$$

where the second line uses the triangle inequality and the Schwarz inequality and Lemma A.5.37 together with the Lyapunov inequality. We proceed with analyzing the expectation of $\Xi_6^*$. Using the Schwarz inequality gives

$$|\mathbb{E}[\Xi_6^*(\nu, g_1, g_2)]| \leq 3\mathbb{E}\left[\|v_0\|_2^2 \|v_\nu\|_2^4 \right]^{1/2} \mathbb{E}\left[\frac{\mathbb{1}_{\mathcal{E}}}{\|v_0\|_2^2 \|v_\nu\|_2^5} \right]^{1/2}$$

$$\leq 192 \mathbb{E}\left[\|v_0\|_2^2 \|v_\nu\|_2^4 \right]^{1/2},$$

384
and the checks at and around (A.210) in the proof of Lemma A.5.15 show that we can apply Lemma A.5.30 to obtain

\[
\left| \mathbb{E} \left[ \langle v_0, v_\nu \rangle^2 \langle v_\nu, \dot{v}_\nu \rangle^4 \right] \right. - n^6 \mathbb{E} [\sigma(g_{11}) \sigma(g_{11} \cos \nu + g_{21} \sin \nu)]^2 \mathbb{E} [\sigma(g_{11} \cos \nu + g_{21} \sin \nu) (g_{21} \cos \nu - g_{11} \sin \nu)]^4 \left| \leq \frac{C}{n}, \right.
\]

But we have using rotational invariance that \( \mathbb{E} [\sigma(g_{11} \cos \nu + g_{21} \sin \nu) (g_{21} \cos \nu - g_{11} \sin \nu)] = 0 \), which implies

\[
\left| \mathbb{E} \left[ \langle v_0, v_\nu \rangle^2 \langle v_\nu, \dot{v}_\nu \rangle^4 \right] \right| \leq C/n,
\]

from which we conclude for all \( \nu \)

\[
| \mathbb{E}[\Xi_6(\nu, g_1, g_2)] | \leq C/\sqrt{n}.
\]

Next, we control the deviations of \( \Xi_6 \) with high probability. By Lemma A.5.17, there is an event \( \mathcal{E}_a \) with probability at least \( 1 - e^{-cn} \) on which \( \| \dot{v}_\nu \|_2 \leq 4 \) for every \( \nu \in [0, \pi] \). Therefore on the event \( \mathcal{E}_b = \mathcal{E} \cap \mathcal{E}_a \), which has probability at least \( 1 - Ce^{-cn} \) by a union bound, we have using Cauchy-Schwarz that for every \( \nu \)

\[
| \Xi_6(\nu, g_1, g_2) | \leq 6144|\langle v_\nu, \dot{v}_\nu \rangle|.
\]

Using the high probability deviations bound established in (A.238), it follows that if \( n \) is large enough then with probability at least \( 1 - Ce^{-cn} - C' n^{-2d+1/2} \) we have

\[
\forall \nu \in [0, \pi], \ |\Xi_6(\nu, g_1, g_2) | \leq C'' \sqrt{\frac{d \log n}{n}}.
\]

As long as \( d \geq \frac{1}{2} \), we have that this probability is at least \( 1 - Ce^{-cn} - C' n^{-d} \), and so the triangle
inequality yields finally that with probability at least $1 - C e^{-cn} - C' n^{-d}$

$$\forall \nu \in [0, \pi], \ |\mathbb{E}_6(\nu, g_1, g_2) - \mathbb{E}[\mathbb{E}_6(\nu, g_1, g_2)]| \leq C'' \sqrt{\frac{d \log n}{n}}.$$  

□

**Lemma A.5.47.** Consider the function

$$g(\nu) = -(\pi^2 - [(\pi - \nu) \cos \nu + \sin \nu]^2)(\pi - \nu) \cos \nu - \sin \nu + (\pi - \nu)^2 [(\pi - \nu) \cos \nu + \sin \nu] \sin^2 \nu,$$

which is the negated numerator of $\bar{\varphi}$. Then if $0 \leq \nu \leq \pi/2$, one has a bound

$$\frac{2\pi^2}{3} \nu^3 - \frac{83}{24} \nu^4 \leq g(\nu),$$

and the lower bound is positive if $0 < \nu \leq \pi/2$.

**Proof.** To see that the lower bound is positive under the stated condition, write

$$\frac{2\pi^2}{3} \nu^3 - \frac{83}{24} \nu^4 = \nu^3 \left(\frac{2\pi^2}{3} - \frac{83}{24} \nu\right);$$

the quantity in parentheses is positive in a neighborhood of zero by continuity, and in fact one calculates for its unique zero $\nu_0 = 48\pi^2/249$, and one verifies numerically that $48\pi^2/249 > 1.9 > \pi/2$. We conclude that the bound is positive for $0 < \nu < 1.9$ by continuity.

To establish the bound, we employ Taylor expansion of the numerator, which is a smooth function on $(0, \pi)$ with continuous derivatives of all orders on $[0, \pi]$, in a neighborhood of zero. In our development in the proof of Lemma A.5.5, we showed that the analytic function $-g(\nu) = -(2\pi^2/3) \nu^3 + O(\nu^4)$ near zero, so Taylor’s theorem with Lagrange remainder implies

$$\frac{2\pi^3}{3} \nu^3 + \frac{\nu^4}{24} \inf_{\nu \in [0, \pi/2]} g^{(4)}(\nu) \leq g(\nu),$$
and so it suffices to get suitable bounds on the fourth derivative of \( g \). We will develop the bounds rather tediously. Start by distributing in \( g \) to write

\[
g(\nu) = \nu^3 (-\cos \nu) + \nu^2 \left( 3\pi \cos \nu + \sin \nu \right) + \nu \left( \cos \nu - 2\pi^2 \cos \nu - 2\pi \sin \nu - \cos^3 \nu \right)\]

\[
+ \left( \pi \cos^3 \nu + 2\pi^2 \sin \nu - \sin^3 \nu - \pi \cos \nu \right) .
\]

Using the Leibniz rule, we have for the fourth derivative

\[
g^{(4)}(\nu) = \nu^3 \left( g_3^{(4)}(\nu) \right) + \nu^2 \left( g_2^{(4)}(\nu) + 12g_3^{(3)}(\nu) \right) + \nu \left( g_1^{(4)}(\nu) + 8g_2^{(3)}(\nu) + 36g_3^{(2)}(\nu) \right)\]

\[
+ \left( g_0^{(4)}(\nu) + 4g_1^{(3)}(\nu) + 12g_2^{(2)}(\nu) + 24g_3^{(1)}(\nu) \right) .
\]

To calculate these derivatives, we just need to differentiate \( \sin \), \( \cos \), and their third powers. Write \( c(\nu) = \cos^3(\nu) \) and \( s(\nu) = \sin^3(\nu) \); using the elementary calculations

\[
c^{(1)}(\nu) = 3s(\nu) - 3\sin \nu, \quad c^{(2)}(\nu) = 6\cos \nu - 9c(\nu),
\]

\[
c^{(3)}(\nu) = 21\sin \nu - 27s(\nu), \quad c^{(4)}(\nu) = 60\cos \nu + 81c(\nu);
\]

\[
s^{(1)}(\nu) = 3\cos \nu - 3c(\nu), \quad c^{(2)}(\nu) = 6\sin \nu - 9s(\nu),
\]

\[
c^{(3)}(\nu) = 27c(\nu) - 21\cos \nu, \quad c^{(4)}(\nu) = 60\sin \nu + 81s(\nu),
\]

one can calculate the results

\[
g_3^{(4)}(\nu) = -\cos \nu, \quad g_2^{(4)}(\nu) = 3\pi \cos \nu + \sin \nu,
\]

\[
g_1^{(4)}(\nu) = (61 - 2\pi^2) \cos \nu - 2\pi \sin \nu - 81 \cos^3 \nu,
\]

\[
g_0^{(4)}(\nu) = (2\pi^2 - 60) \sin \nu + 50\pi \cos \nu + 81\pi \cos^3 \nu - 81\sin^3 \nu;
\]
and

\[ g_3^{(3)}(\nu) = -\sin \nu, \quad g_2^{(3)}(\nu) = 3\pi \sin \nu - \cos \nu, \]

\[ g_1^{(3)}(\nu) = (7 - 2\pi^2) \sin \nu + 2\pi \cos \nu - 27 \cos^2 \nu \sin \nu; \]

and

\[ g_3^{(2)}(\nu) = \cos \nu \quad g_2^{(2)}(\nu) = -3\pi \cos \nu - \sin \nu; \]

and finally

\[ g_3^{(1)}(\nu) = \sin \nu. \]

Plugging back into (A.249) and canceling, we get

\[ g^{(4)}(\nu) = \nu^3 \left( -\cos \nu \right) + \nu^2 \left( 3\pi \cos \nu - 11 \sin \nu \right) + \nu \left( 22\pi \sin \nu + (89 - 2\pi^2) \cos \nu - 81 \cos^3 \nu \right) \]

\[ + \left( 27 \sin^3 \nu + 81\pi \cos^3 \nu + 31\pi \cos \nu - (6\pi^2 + 128) \sin \nu \right) \]

\[ h_3(\nu) \]

\[ h_2(\nu) \]

\[ h_1(\nu) \]

\[ h_0(\nu) \]

Since \( \nu > 0 \), we can leverage lower bounds on each \( h_i \) term. We have trivially \( |h_3| \leq 1 \), so that \( |\nu^3 h_3(\nu)| \leq \pi^3/8 \). We will study \( \nu h_1(\nu) + h_0(\nu) \) together to get a better bound. We have

\[ \nu h_1(\nu) + h_0(\nu) = \left( 22\pi \nu - (6\pi^2 + 128) \right) \sin \nu + 27 \sin^3 \nu + \left( (89 - 2\pi^2) \nu + 31\pi \right) \cos \nu \]

\[ + (81\pi - 81\nu) \cos^3 \nu \]

\[ q(\nu) \]

\[ (A.250) \]

using \( \nu \leq \pi/2 \) and \( \cos \geq 0 \) on this domain. We will show that the RHS of the final inequality, denoted \( q \), is a decreasing function of \( \nu \), and is therefore lower bounded by its value at \( \nu = \pi/2 \) on
our interval of interest. We calculate

\[ q'(\nu) = 9\pi \sin \nu + (42 - 8\pi^2) \cos \nu + 22\pi \nu \cos \nu - (89 - 2\pi^2)\nu \sin \nu - 81 \cos^3 \nu. \]

Reordering terms, we can write

\[ q'(\nu) = -81 \cos^3 \nu + \left( \frac{9\pi}{C_1} - \frac{(89 - 2\pi^2)}{C_2} \right) \sin \nu - \left( \frac{(8\pi^2 - 42)}{C_3} - \frac{22\pi}{C_4} \right) \nu \cos \nu. \] (A.251)

We can estimate numerically

\[ 69 \leq C_2 \leq 70; \quad 69 \leq C_4 \leq 70; \quad C_2 > C_4, \]

which shows that \( C_1, C_2, C_3, C_4 > 0 \) and both of the linear prefactors are decreasing functions of \( \nu \). We have on all of \((0, \pi/2)\) by concavity of \( \sin \)

\[ (C_1 - C_2 \nu) \sin \nu \leq \nu \left( C_1 - \frac{2C_2}{\pi} \nu \right), \]

using in particular \( \sin \nu \leq \nu \) and \( \sin \nu \geq (2/\pi)\nu \). Using similarly concavity of \( \cos \) on this domain, in particular the inequalities \( \cos \nu \leq \pi/2 - \nu \) and \( \cos \nu \geq 1 - (2/\pi)\nu \), we have

\[-(C_3 - C_4 \nu) \cos \nu \leq - \left( C_4 \nu^2 - \left( \frac{2C_3}{\pi} + \frac{C_4\pi}{2} \right) \nu + C_3 \right). \]

In total, we have a bound

\[ q'(\nu) \leq -81 \cos^3 \nu - \left( \frac{2C_2}{\pi} + C_4 \right)\nu^2 + \left( \frac{2C_3}{\pi} + \frac{\pi C_4}{2} + C_1 \right) \nu - C_3. \]

We calculate the maximizer of the concave quadratic function of \( \nu \) in the previous bound via
differentiation; plugging in, we get

\[ q'(\nu) \leq -81 \cos^3 \nu + \frac{\left( \frac{2C_3}{\pi} + \frac{\pi C_4}{2} + C_1 \right)^2}{4 \left( \frac{2C_3}{\pi} + C_4 \right)} - C_3. \]

A numerical estimate gives

\[ \frac{\left( \frac{2C_3}{\pi} + \frac{\pi C_4}{2} + C_1 \right)^2}{4 \left( \frac{2C_3}{\pi} + C_4 \right)} - C_3 \leq 20, \]

and using that \(-\cos^3\) is strictly decreasing for \(\nu < \pi\), we can therefore guarantee \(q' \leq 0\) as long as \(\nu \leq \cos^{-1} \sqrt{20/81}\). Writing \(c = \cos^{-1} \sqrt{20/81}\), we estimate numerically \(0.90 \geq c \geq 0.89\), so that this bound is nonvacuous. For \(\nu \geq c\), we apply again concavity of \(\cos\) to develop the lower bound

\[ \cos \nu \geq \left( \frac{\pi/2 - \nu}{\pi/2 - c} \right) \cos c, \quad \nu \in [c, \pi/2]. \]

Using this to estimate the \(-\cos^3\) term in our upper bound for \(q'\), we obtain a bound

\[ q'(\nu) \leq -20 \left( \frac{\pi/2 - \nu}{\pi/2 - c} \right)^3 - \left( \frac{2C_2}{\pi} + C_4 \right) \nu^2 + \left( \frac{2C_3}{\pi} + \frac{\pi C_4}{2} + C_1 \right) \nu - C_3, \quad c \leq \nu \leq \pi/2. \]

We define \(D = 20/(\pi/2 - c)^3\), \(A = 2C_2/\pi + C_4\), \(B = 2C_3/\pi + \pi C_4/2 + C_1\), and \(C = C_3\), so that the RHS can be written as \(-D(\pi/2 - \nu)^3 - Av^2 + Bv - C\). Differentiating once and equating to zero results in the quadratic equation

\[ 3D \left( \nu^2 - \frac{2A}{3D} + \pi \right) \nu + \left( \frac{B}{3D} + \pi^2/4 \right) = 0, \]

which has roots \(M/2 \pm \frac{1}{2} \sqrt{M^2 - 4N}\). Numerically estimating the constants, we get that the two roots lie in \([0.99, 1]\) and \([3.3, 3.4]\), so that we need only consider the smaller root. Differentiating once more to determine the class of the critical point, we find for the second derivative at \(M/2 - \)
\[ \frac{1}{2} \sqrt{M^2 - 4N} \]

\[-3D \sqrt{M^2 - 4N} < 0,\]

so that \( M/2 - \frac{1}{2} \sqrt{M^2 - 4N} \) is a maximizer for our cubic bound, and the bound is increasing for arguments less than this point and decreasing for arguments greater than it; we can conclude that the zero in \([3.3, 3.4]\) is a minimizer, so that our bound can be ascertained negative by checking its value at \( M/2 - \frac{1}{2} \sqrt{M^2 - 4N} \). We find using a numerical estimate

\[ -20 \left( \frac{\pi/2 - (M/2 - \frac{1}{2} \sqrt{M^2 - 4N})}{\pi/2 - c} \right)^3 \]

\[ - \left( \frac{2C_2}{\pi} + C_4 \right) (M/2 - \frac{1}{2} \sqrt{M^2 - 4N})^2 \]

\[ + \left( \frac{2C_3}{\pi} + \frac{\pi C_4}{2} + C_1 \right) (M/2 - \frac{1}{2} \sqrt{M^2 - 4N}) - C_3 \leq -1.7 < 0,\]

which proves that \( q' \leq 0 \) on \([c, \pi/2]\). This shows that our lower bound on \( \nu h_1(\nu) + h_0(\nu) \) in (A.250) is nonincreasing on \([0, \pi/2]\), so that we can assert

\[ \nu h_1(\nu) + h_0(\nu) \geq \left( 22\pi(\pi/2) - (6\pi^2 + 128) \right) \sin(\pi/2) + 27 \sin^3(\pi/2) \]

\[ + \left( (89 - 2\pi^2)(\pi/2) + 31\pi \right) \cos(\pi/2) \]

\[ = 5\pi^2 - 101.\]

It remains to bound \( \nu^2 h_2(\nu) = \nu^2(3\pi \cos \nu - 11 \sin \nu) \). On \([0, \pi/2]\), \( \cos \) is decreasing and \( \sin \) is increasing, so \( 3\pi \cos \nu - 11 \sin \nu \) is decreasing here; it is positive at \( \nu = 0 \) and negative at \( \nu = \pi/2 \), so that by continuity it has a unique zero in \((0, \pi/2)\). Denote this zero as \( \nu_0 \); then using that \( \nu^2 \geq 0 \) with no zeros in the interior, we can write

\[ \inf_{0 \leq \nu \leq \nu_0} \nu^2 h_2(\nu) \geq 0,\]
and
\[
\inf_{\nu_0 \leq \nu \leq \pi/2} \nu^2 h_2(\nu) \geq \left( \sup_{\nu_0 \leq \nu \leq \pi/2} \nu^2 \right) \left( \inf_{\nu_0 \leq \nu \leq \pi/2} h_2(\nu) \right) \\
\geq (\pi/2)^2 (3\pi \cos(\pi/2) - 11 \sin(\pi/2)) = -\frac{11\pi^2}{4},
\]

which gives the bound \( \nu^2 h_2(\nu) \geq -11\pi^2/4 \) on \([0, \pi/2]\). Putting it all together, we have
\[
g^{(4)}(\nu) \geq -\frac{11\pi^2}{4} + 5\pi^2 - 101 - \pi^3/8 \geq -83,
\]
where the last inequality follows from a numerical estimate of the constants. \(\square\)

**Lemma A.5.48** (Uniformization). Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a complete probability space. For some \(t \in \mathbb{R}\), \(\delta_t \geq 0\), \(S \subset \mathbb{R}^d\), and event \(E \in \mathcal{F}\), suppose that \(f : S \times \Omega \to \mathbb{R}\) is second-argument measurable and satisfies

1. For all \(x \in S\), \(\mathbb{P}[f(x, \cdot) \leq t] \geq 1 - \delta_t\);
2. For all \(g \in E\), \(f(\cdot, g)\) is \(L\)-Lipschitz;
3. There is \(M > 0\) such that \(\sup_{x \in S} ||x||_2 \leq M\).

Then \(g \mapsto \sup_{x \in S} f(x, g)\) is measurable, and for every \(\varepsilon > 0\), one has
\[
\mathbb{P}\left[ \sup_{x \in S} f(x, \cdot) \leq t + L\varepsilon \right] \geq 1 - \delta_t \left( 1 + \frac{2M}{\varepsilon} \right)^d - \mathbb{P}[E]. \quad (A.252)
\]

**Proof.** Because \(S\) is a subset of the separable metric space \((\mathbb{R}^d, || \cdot ||_2)\) and all sample trajectories \(f(\cdot, g)\) are assumed (Lipschitz) continuous, the supremum in the definition of the map \(g \mapsto \sup_{x \in S} f(x, g)\) can be taken on a countable subset of \(S\), and the resulting function of \(g\) is measurable (e.g., [230, §2.2 p. 45]). By [225, Proposition 4.2.12] and boundedness of \(S\), for every \(\varepsilon > 0\) there exists an \(\varepsilon\)-net of \(S\) having cardinality at most \((1 + 2M/\varepsilon)^d\); denote these nets as \(N_\varepsilon\). Since each \(N_\varepsilon\) is finite, we may also define for each \(x \in S\) a point \(x_\varepsilon\) such that \(||x - x_\varepsilon||_2 \leq \varepsilon\); then
for every $g \in E$, we have $|f(x, g) - f(x, \varepsilon)| \leq L\varepsilon$. We define a collection of events $E_\varepsilon$ by

$$E_\varepsilon = \{g \in \Omega \mid \forall x \in N_\varepsilon, f(x, g) \leq t\}. \quad (A.253)$$

The triangle inequality then implies that if $g \in E_\varepsilon \cap E$, then for all $x \in S$, one has $f(x, g) \leq t + L\varepsilon$. Consequently, several union bounds yield

$$P \left[ \sup_{x \in S} f(x, \cdot) > t + L\varepsilon \right] \leq P \left[ \sup_{x \in N_\varepsilon} f(x, g) \leq t \right] + P[E] \leq \delta_t \left( 1 + \frac{2M}{\varepsilon} \right)^d + P[E], \quad (A.254)$$

as claimed.

\[\square\]

### A.5.4 Deferred Proofs

**Proof of Lemma A.5.5.** The function $\cos^{-1}$ is $C^\infty$ on $(-1, 1)$, and because $f(\nu) := \cos \varphi(\nu)$ is smooth and satisfies $f'(\nu) = (\pi^{-1} \nu - 1) \sin \nu < 0$ if $\nu < \pi$ with $f(0) = 1$ and $f(\pi) = 0$, we see that $\varphi$ is $C^\infty$ on $(0, \pi)$ by the chain rule. This also shows $\varphi(0) = \cos^{-1}(1) = 0$ and $\varphi(\pi) = \cos^{-1}(0) = \pi/2$. Direct calculation gives

$$\varphi(\nu) = \sqrt{\frac{(\pi - \nu)^2 \sin^2 \nu}{\pi^2 - ((\pi - \nu) \cos \nu + \sin \nu)^2}} \quad (A.255)$$

and

$$\varphi(\nu) = \frac{(\pi^2 - [(\pi - \nu) \cos \nu + \sin \nu] \sin \nu)}{(\pi^2 - [(\pi - \nu) \cos \nu + \sin \nu]^2)^{3/2}} - \frac{(\pi - \nu)^2 [(\pi - \nu) \cos \nu + \sin \nu] \sin^2 \nu}{(\pi^2 - [(\pi - \nu) \cos \nu + \sin \nu]^2)^{3/2}} \quad (A.256)$$
Calculating endpoint limits using these expressions will suffice to show the derivatives are continuous on \([0, \pi]\) and give the claimed values there. We have

\[
\lim_{\nu \searrow 0} (\dot{\varphi}(\nu))^2 = \lim_{\nu \searrow 0} \frac{(\pi - \nu)^2 \sin^2 \nu}{\pi^2 - (\pi - \nu) \cos \nu + \sin \nu} = \lim_{\nu \searrow 0} \frac{2(\pi - \nu) \sin \nu [(\pi - \nu) \cos \nu - \sin \nu]}{(-2)[(\pi - \nu) \cos \nu + \sin \nu][(\nu - (\pi - \nu) \sin \nu - \cos \nu]} = \lim_{\nu \searrow 0} \frac{(\pi - \nu) \cos \nu - \sin \nu}{(\pi - \nu) \cos \nu + \sin \nu} = 1,
\]

by L’Hôpital’s rule, whereas a direct evaluation gives

\[
\lim_{\nu \searrow 0} (\dot{\varphi}(\nu))^2 = \frac{0}{\pi^2} = 0.
\]

Continuity of the square root function gives the claimed results for \(\varphi\). Again by direct calculation, we find

\[
\lim_{\nu \searrow 0} (\ddot{\varphi}(\nu))^2 = \frac{0}{\pi^2} = 0.
\]

Since \(\varphi^2\) is meromorphic in a neighborhood of 0 with, as we have shown, a removable singularity at 0, it is actually analytic, and we can calculate further derivatives at 0 by expanding it locally at 0. We use the expansions \(\sin \nu = \nu - \nu^3/6 + O(\nu^5)\) and \(\cos \nu = 1 - \nu^2/2 + \nu^4/24 + O(\nu^6)\) near 0 to calculate

\[
\left(1 - \frac{\nu}{\pi}\right)^2 \sin^2 \nu = \nu^2 \left(1 - \frac{2}{\pi} \nu - \frac{\pi^2 - 3}{3\pi^2} \nu^2 + O(\nu^3)\right)
\]

and

\[
1 - \left(1 - \frac{\nu}{\pi}\right) \cos \nu + \frac{\sin \nu}{\pi}\right)^2 = \nu^2 \left(1 - \frac{2}{3\pi} \nu - \frac{1}{3} \nu^2 + O(\nu^3)\right),
\]

from which it follows

\[
(\dot{\varphi}(\nu))^2 = \left(1 - \frac{2}{\pi} \nu - \frac{\pi^2 - 3}{3\pi^2} \nu^2 + O(\nu^3)\right) \left(1 - \frac{2}{3\pi} \nu - \frac{1}{3} \nu^2 + O(\nu^3)\right)^{-1}.
\]
By the geometric series, we then obtain

$$(\varphi(v))^2 = 1 - \frac{4}{3\pi}v + \frac{1}{9\pi^2}v^2 + O(v^3).$$

Taking the square root of this expression and applying the binomial series, we thus have

$$\varphi(v) = 1 - \frac{2}{3\pi}v - \frac{1}{6\pi^2}v^2 + O(v^3),$$

from which we read off

$$\lim_{v \searrow 0} \varphi(v) = -\frac{2}{3\pi}; \quad \lim_{v \searrow 0} \varphi'(v) = -\frac{1}{3\pi^2}.$$ 

It is clear from the analytical expression for $\varphi$ and the mean value theorem that $\varphi$ is strictly increasing on $[0, \pi]$, since $(\pi - v)\sin v > 0$ if $0 < v < \pi$. To prove strict concavity for $v \in (0, \pi)$, we start by simplifying notation. Consider the function $\varphi_r(v) = \varphi(\pi - v)$, which satisfies by the chain rule $\varphi'_r(v) = \varphi'(\pi - v)$. Because $\varphi_r$ is strictly concave if and only if $\varphi$ is strictly concave, it suffices to prove that $\varphi'(\pi - v) < 0$. We note

$$\varphi'(\pi - v) < 0 \iff (\pi^2 - [v \cos v - \sin v]^2)(-\sin v - v \cos v) < v^2 \sin^2 v(v \sin v - v \cos v).$$

Multiplying both sides of the latter inequality by $\sin v - v \cos v$, dividing through by $(v \cos v - \sin v)^2$ (which is positive on $(0, \pi)$, since it equals $\cos^2 \varphi$ composed with a reversal about $\pi$), and distributing and moving terms to the RHS gives the equivalent condition

$$\frac{\pi^2 v^2 \cos^2 v - \sin^2 v}{(v \cos v - \sin v)^2} < v^2 - \sin^2 v,$$

and canceling once more gives equivalently

$$\frac{v \cos v + \sin v}{v \cos v - \sin v} < \frac{v^2 - \sin^2 v}{\pi^2}.$$  \hspace{1cm} (A.257)

395
Using \( \nu \cos \nu - \sin \nu < 0 \), which follows from its derivative \(-\nu \sin \nu\) being negative on \((0, \pi)\), and writing \(g(\nu) = \pi^{-2}(\nu^2 - \sin^2 \nu)\), we have equivalently \(\nu \cos \nu + \sin \nu > g(\nu)(\nu \cos \nu - \sin \nu)\), and rearranging gives the inequality

\[
(1 - g(\nu)) \nu \cos \nu + g(\nu) \sin \nu > -\sin \nu. \tag{A.258}
\]

Strict concavity of \(\sin\) on \((0, \pi)\) gives \(\sin \nu < \nu\), and \(0 < g(\nu) < 1\) follows after squaring; so the LHS is a convex combination of \(\nu \cos \nu\) and \(\sin \nu\), which in particular satisfies \(|(1 - g(\nu)) \nu \cos \nu + g(\nu) \sin \nu| \leq \max\{|\sin \nu|, |\nu \cos \nu|\}\). As argued before, we have \(\sin \nu - \nu \cos \nu > 0\) if \(\nu \in (0, \pi)\); moreover, because \(\nu > 0\) we have \(\nu \cos \nu > 0\) if \(\nu \in (0, \pi/2)\) and \(\nu \cos \nu < 0\) if \(\nu \in (\pi/2, \pi)\). We can numerically determine \(\sin(5\pi/8) + (5\pi/8) \cos(5\pi/8) > 0\), and given that \(5\pi/8 \geq 1.95 > \pi/2\), it follows

\[|(1 - g(\nu)) \nu \cos \nu + g(\nu) \sin \nu| < |\sin \nu|, \quad 0 < \nu \leq 1.95,\]

which implies (A.258) when \(0 < \nu \leq 1.95\). Recalling that we are arguing for \(\varphi_r\) in this setting, we translate our results back to \(\varphi\) and conclude that \(\varphi(\nu) < 0\) if \(\pi - 1.95 \leq \nu < \pi\). To address the case where \(0 < \nu < \pi - 1.95\), we employ Lemma A.5.47; it allows us to conclude \(\bar{\varphi} < 0\) provided \(0 < \nu \leq \pi/2\), and a numerical estimate gives that \(\pi - 1.95 < \pi/2\), so that we have \(\bar{\varphi} < 0\) for all \(0 < \nu < \pi\). Taking limits in \(\varphi\) gives concavity at the endpoints \(\{0, \pi\}\) as well.

To bound \(\bar{\varphi}\) away from zero on \([0, \pi/2]\), we apply Lemma A.5.47 to assert

\[
\bar{\varphi}(\nu) \leq \frac{-\frac{2}{3\pi} \nu^3 + \frac{83}{24\pi} \nu^4}{(1 - \cos^2 \varphi(\nu))^{3/2}}, \quad 0 < \nu \leq \pi/2.
\]

The numerator in the last expression is nonpositive if \(0 \leq \nu \leq \pi/2\), and using the lower bound in Lemma A.5.14 on \([0, \pi/2]\), we have

\[
\frac{1}{1 - \cos^2 \varphi(\nu)} \geq \frac{1}{1 - \max\{1 - \frac{1}{2} \nu^2, 0\}}, \quad \nu > 0.
\]
From nonpositivity of the numerator, it follows

$$
\phi(\nu) \leq \frac{-\frac{2}{3\pi} \nu^3 + \frac{83}{24\pi^3} \nu^4}{\left(1 - \max^2\{1 - \frac{1}{2} \nu^2, 0\}\right)^{3/2}}, \quad 0 < \nu \leq \pi/2. 
$$

(A.259)

We have $1 - \frac{1}{2} \nu^2 \geq 0$ as long as $0 \leq \nu \leq \sqrt{2}$; so after removing the max, distributing, and cancelling, we have

$$
\phi(\nu) \leq \frac{-\frac{2}{3\pi} + \frac{83}{24\pi^3} \nu}{\left(1 - \frac{1}{4} \nu^2\right)^{3/2}}, \quad 0 < \nu \leq \sqrt{2}.
$$

The denominator of this last expression is nonnegative and has singularities at $\pm 2$, and is clearly even symmetric; so it is maximized on $0 < \nu \leq \sqrt{2}$ at $\sqrt{2}$, and we have

$$
\phi(\nu) \leq \sqrt{8} \left(-\frac{2}{3\pi} + \frac{83}{24\pi^3} \nu\right), \quad 0 < \nu \leq \sqrt{2}.
$$

Taking limits $\nu \searrow 0$, we can assert this bound on $[0, \sqrt{2}]$, and the bound is clearly an increasing function of $\nu$, from which it follows

$$
\sup_{\nu \in [0, \sqrt{2}]} \phi(\nu) \leq \sqrt{8} \left(-\frac{2}{3\pi} + \frac{83\sqrt{2}}{24\pi^3}\right) \leq -0.15,
$$

where the last inequality follows from a numerical estimate of the constants. On the other hand, when $\sqrt{2} < \nu \leq \pi/2$, we have from (A.259) that

$$
\phi(\nu) \leq -\frac{2}{3\pi} \nu^3 + \frac{83}{24\pi^3} \nu^4, \quad \sqrt{2} \leq \nu \leq \pi/2.
$$

If we differentiate the degree four polynomial on the RHS of this bound and solve for critical points, we find a double critical point at $\nu = 0$ and a critical point at $\nu = 12\pi^2/83$; a numerical estimate confirms that this critical point lies in the interior of $[\sqrt{2}, \pi/2]$. The second derivative of the RHS is $-(4/\pi)\nu + 83/(2\pi^3)\nu^2$, and plugging in $\nu = 12\pi^2/83$ gives a value of $-48\pi/83 + 144\pi/83$, which
is positive; hence the RHS is maximized on the boundary, i.e.,
\[ \tilde{\varphi}(\nu) \leq -\frac{2}{3\pi}\nu^3 + \frac{83}{24\pi^3}\nu^4 \leq \max\left\{ -\frac{25/2}{3\pi} + \frac{83}{6\pi^3}, -\frac{\pi^2}{12} + \frac{83\pi}{384} \right\}, \quad \sqrt{2} \leq \nu \leq \pi/2. \]

A numerical estimate shows that the RHS of the last inequality is no larger than \(-0.14\). Since the intervals we have proved a bound over cover \([0, \pi/2]\), this proves the claim with \(c = -0.14\).

The bound \(\tilde{\varphi} < 1\) on \((0, \pi)\) follows from the fact that \(\varphi\) is strictly concave on \((0, \pi)\) and the mean value theorem; we have already shown \(\dot{\varphi} > 0\) in proving strict increasingness of \(\varphi\). Similarly, the proof of strict concavity in the interior has already established \(\ddot{\varphi} < 0\). To obtain the lower bound on \(\ddot{\varphi}\), we use that \(\ddot{\varphi}\) is continuous on \([0, \pi]\) and the Weierstrass theorem to assert that there is \(C \geq 0\) such that \(\ddot{\varphi} \geq -C\) on \([0, \pi]\); because \(\ddot{\varphi}(0) \neq 0\), we actually have \(C > 0\).

For the quadratic model, we use our previous results and Taylor expand \(\varphi\) about \(0\); we get immediately
\[ \varphi(\nu) \geq \nu + \nu^2 \frac{\inf_{\nu \in [0,\pi]} \ddot{\varphi}(\nu)}{2} \geq \nu - (C/2)\nu^2. \]

For the upper bound, we can assert immediately on \([0, \pi/2]\) a bound
\[ \varphi(\nu) \leq \nu - cv^2, \]

where \(c = 0.07\) suffices. To extend the bound to \(\nu \in [\pi/2, \pi]\), we employ a bootstrapping argument; because \(\varphi\) is concave, we have a bound
\[ \varphi(\nu) \leq \varphi(\pi/2) + \varphi(\pi/2)(\nu - \pi/2) \]
\[ = \cos^{-1}\pi^{-1} + \frac{\pi/2}{\sqrt{\pi^2 - 1}}(\nu - \pi/2), \]

where the second line plugs into the formulas for \(\varphi\) and \(\dot{\varphi}\). We will show that the graph of \(\nu - cv^2\) lies entirely above the graph of the RHS of this inequality. This condition is equivalent to
\[ -cv^2 + \left(1 - \frac{\pi/2}{\sqrt{\pi^2 - 1}}\right)\nu + \left(\frac{(\pi/2)^2}{\sqrt{\pi^2 - 1}} - \cos^{-1}\pi^{-1}\right) \geq 0; \]

398
the LHS of this inequality is a concave quadratic with maximizer \( \nu_* = 1/(2c)(1 - \frac{\pi/2}{\sqrt{\pi^2-1}}) \), and numerical estimation of the constants gives \( \nu_* \geq \pi \). Since \( \nu_* \) is outside \([\pi/2, \pi]\) and the quadratic is concave, we conclude that the bound is tightest at the boundary point \( \pi/2 \), and one checks numerically 

\[
-c\pi^2/4 + \left(1 - \frac{\pi/2}{\sqrt{\pi^2-1}}\right)\pi/2 + \left(\frac{(\pi/2)^2}{\sqrt{\pi^2-1}} - \cos^{-1}\pi^{-1}\right) \geq 0.15 > 0,
\]

which establishes that the bound \( \varphi(\nu) \leq \nu - c\nu^2 \) actually holds on all of \([0, \pi]\). This completes the proof of all of the claims. \(\square\)

### A.6 Auxiliary Results

**Lemma A.6.1** (Hoeffding’s Inequality [225, Theorem 2.2.6]). Let \( X_1, \ldots, X_N \) be independent random variables. Assume that \( X_i \in [m_i, M_i] \) for every \( i \). Then for any \( t > 0 \), we have

\[
\mathbb{P}\left[ \sum_{i=1}^{N} (X_i - \mathbb{E}[X_i]) \geq t \right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N} (M_i - m_i)^2}\right).
\]

**Lemma A.6.2** (Bernstein’s inequality [225, Theorem 2.8.1]). Let \( X_1, \ldots, X_N \) be zero-mean independent subexponential random variables. Then, for every \( t \geq 0 \), one has

\[
\mathbb{P}\left[ \left| \sum_{i=1}^{N} X_i \right| \geq t \right] \leq 2 \exp\left(-c \min\left\{\frac{t^2}{\sum_{i=1}^{N} \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right),
\]

where \( c > 0 \) is an absolute constant, and \( \|\cdot\|_{\psi_1} = \inf\{t > 0 \mid \mathbb{E}[e^{t \cdot 1/|t|}] \leq 2\} \) is the subexponential norm.

**Lemma A.6.3** (Bernstein’s inequality for bounded RVs - [225] Thm. 2.8.4). For \( X_1, \ldots, X_n \) independent, zero mean random variables such that \( \forall i : |X_i| < K \), and every \( t \geq 0 \), we have

\[
\mathbb{P}\left[ \left| \sum_{i=1}^{n} X_i \right| \geq t \right] \leq 2 \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right)
\]
where $\sigma^2 = \sum_{i=1}^{n} \mathbb{E}X_i^2$.

**Lemma A.6.4** (Hanson-Wright Inequality [225, Theorem 6.2.1]). Let $g$ be a vector of $n$ i.i.d., mean zero, sub-Gaussian variables and $A$ be an $n \times n$ matrix. Then for any $t > 0$, we have

$$
P \left[ |g^*Ag - \mathbb{E}g^*Ag| \geq t \right] \leq 2 \exp \left( -c \min \left\{ \frac{t^2}{K^4 \|A\|^2_F}, \frac{t}{K^2 \|A\|} \right\} \right)
$$

where $\max_i \|g_i\|_{\psi_2} \leq K$ (with $\| \cdot \|_{\psi_2}$ denoting the sub-Gaussian norm).

**Lemma A.6.5** (Freedman’s Inequality [95, Theorem 1.6]). Let $(\Delta^i, \mathcal{F}^i)$ be a sequence of martingale differences, with

$$
\mathbb{E}[\Delta^i | \mathcal{F}^{i-1}] = 0,
$$

and suppose that

$$
|\Delta^i| \leq R \ \text{a.s.}
$$

Define the quadratic variation

$$
\mathbb{V}^L = \sum_{i=1}^{L} \mathbb{E} \left[ (\Delta^i)^2 \bigg| \mathcal{F}^{i-1} \right].
$$

Then

$$
P \left[ \exists i = 1 \ldots L \text{ s.t. } \sum_{i=1}^{i} \Delta^i > t \ \text{ and } \ \mathbb{V}^i \leq \sigma^2 \right] \leq 2 \exp \left( -\frac{t^2/2}{\sigma^2 + Rt/3} \right).
$$

**Lemma A.6.6** (Moment control Freedman’s [231]). Let $(\Delta^i, \mathcal{F}^i)$ be a sequence of martingale differences, with

$$
\mathbb{E}[\Delta^i | \mathcal{F}^{i-1}] = 0,
$$

and suppose that

$$
\mathbb{E} \left[ (\Delta^i)^k \big| \mathcal{F}^{i-1} \right] \leq \frac{k!}{2} \mathbb{E} \left[ (\Delta^i)^2 \big| \mathcal{F}^{i-1} \right] R^{k-2} \ \forall k, \text{ a.s.}
$$

Set

$$
\mathbb{V}^j = \sum_{i=1}^{j} \mathbb{E} \left[ (\Delta^i)^2 \big| \mathcal{F}^{i-1} \right].
$$
Then
\[ \mathbb{P} \left[ \exists i = 1 \ldots j \text{ s.t. } \sum_{\ell=1}^{i} \Delta_{\ell} > t \text{ and } V^{i} \leq \sigma^{2} \right] \leq 2 \exp \left( -\frac{t^{2}}{2 \sigma^{2} + Rt} \right). \]

**Lemma A.6.7** (Martingales with subgaussian increments). Let \((\Delta^{i}, \mathcal{F}^{i})\) be a sequence of martingale differences, and suppose that
\[
\mathbb{E}[\exp (\lambda \Delta^{i}) | \mathcal{F}^{i-1}] \leq \exp \left( \frac{\lambda^{2} V^{2}}{2} \right), \; \forall \lambda, \text{ a.s.}
\]

Then
\[ \mathbb{P} \left[ \left| \sum_{i=1}^{L} \Delta_{i} \right| > t \right] \leq 2 \exp \left( -\frac{t^{2}}{2LV^{2}} \right). \]

**Proof.** By assumption, \(\mathbb{E}[\Delta^{i}] = 0\) for each \(i \in [L]\). We calculate using standard properties of the conditional expectation
\[
\mathbb{E} \left[ e^{\lambda \sum_{i=1}^{L} \Delta^{i}} \right] = \mathbb{E} \left[ \mathbb{E} \left[ e^{\lambda \sum_{i=1}^{L} \Delta^{i}} | \mathcal{F}^{L-1} \right] \right] \\
= \mathbb{E} \left[ e^{\lambda \sum_{i=1}^{L-1} \Delta^{i}} \mathbb{E} \left[ e^{\lambda \Delta^{L}} | \mathcal{F}^{L-1} \right] \right] \leq e^{\lambda^{2}V^{2}/2} \mathbb{E} \left[ e^{\lambda \sum_{i=1}^{L-1} \Delta^{i}} \right].
\]

Moreover, one has \(\mathbb{E}[e^{\lambda \Delta^{i}} | \mathcal{F}^{0}] = \mathbb{E}[e^{\lambda \Delta^{i}}] \leq e^{\lambda^{2}V^{2}/2}\). An induction therefore implies
\[ \mathbb{E} \left[ e^{\lambda \sum_{i=1}^{L} \Delta^{i}} \right] \leq e^{\lambda^{2}LV^{2}/2}, \]

and the result follows from standard equivalence properties of subgaussian random variables [225, Proposition 2.5.2]. \(\square\)

**Lemma A.6.8** (Azuma-Hoeffding Inequality [94]). Let \((\Delta^{i}, \mathcal{F}^{i})\) be a sequence of martingale differences, and suppose that
\[ |\Delta^{i}| \leq R_{i} \text{ a.s..} \]
Then

\[
P \left[ \sum_{\ell=1}^{L} \Delta_{\ell}^{i} > t \right] \leq 2 \exp \left( \frac{-t^{2}}{2 \sum_{\ell=1}^{L} R_{\ell}^{2}} \right).
\]

**Lemma A.6.9 (Chi and Inverse-Chi Expectations).** Let \( X \sim \chi(n) \) be a chi random variable with \( n \) degrees of freedom, equal to the square root of the sum of \( n \) independent and identically distributed squared \( N(0, 1) \) random variables. Then

\[
\mathbb{E}[X] = \sqrt{2} \frac{\Gamma \left( \frac{1}{2} (n + 1) \right)}{\Gamma \left( \frac{1}{2} n \right)},
\]

and, if \( n \geq 2 \),

\[
\mathbb{E}[X^{-1}] = \frac{1}{\sqrt{2}} \frac{\Gamma \left( \frac{1}{2} (n - 1) \right)}{\Gamma \left( \frac{1}{2} n \right)}.
\]

**Proof.** We use the fact that the density of \( X \) is given by

\[
\rho(x) = 1_{x \geq 0}(x) \frac{1}{2^{n/2-1} \Gamma \left( \frac{1}{2} n \right)} x^{n-1} e^{-x^2/2},
\]

which can be proved easily using the Gaussian law and a transformation to spherical polar co-ordinates [232, Theorem 2.1.3]. The expectation of \( X \) then results from a simple sequence of calculations using the change of variables formula:

\[
\mathbb{E}[X] = \frac{2}{2^{n/2} \Gamma \left( \frac{1}{2} n \right)} \int_{0}^{\infty} x^n e^{-x^2/2} \, dx
\]

\[
= \frac{1}{2^{n/2} \Gamma \left( \frac{1}{2} n \right)} \int_{0}^{\infty} x^{n/2-1/2} e^{-x/2} \, dx
\]

\[
= \frac{\sqrt{2}}{\Gamma \left( \frac{1}{2} n \right)} \int_{0}^{\infty} x^{(n/2+1/2)-1} e^{-x} \, dx
\]

\[
= \sqrt{2} \frac{\Gamma \left( \frac{1}{2} (n + 1) \right)}{\Gamma \left( \frac{1}{2} n \right)}.
\]
Now we study $X^{-1}$. By the change of variables formula, its density is given by

$$
\rho'(x) = 1_{x \geq 0}(x) \frac{1}{2^{n/2-1} \Gamma(\frac{1}{2}n)} x^{-n} e^{-1/(2x^2)}.
$$

A similar sequence of calculations then yields

$$
\mathbb{E}[X^{-1}] = \frac{2}{2^{n/2} \Gamma(\frac{1}{2}n)} \int_0^\infty x^{-n} e^{-1/(2x^2)} \, dx
$$

$$
= \frac{1}{2^{n/2} \Gamma(\frac{1}{2}n)} \int_0^\infty x^{-\frac{1}{2}(n+1)} e^{-1/(2x)} \, dx
$$

$$
= \frac{1}{2^{n/2} \Gamma(\frac{1}{2}n)} \int_0^\infty x^{\frac{1}{2}(n-1)-1} e^{-\frac{1}{2}x} \, dx
$$

$$
= \frac{1}{\sqrt{2} \Gamma(\frac{1}{2}n)} \int_0^\infty x^{\frac{1}{2}(n-1)-1} e^{-x} \, dx
$$

$$
= \frac{\Gamma(\frac{1}{2}(n-1))}{\sqrt{2} \Gamma(\frac{1}{2}n)},
$$

provided $n > 1$. \hfill \Box

**Lemma A.6.10** (Equivalence of $\ell^p$ Norms). Let $1 \leq p \leq q \leq +\infty$. Then for every $x \in \mathbb{R}^n$ one has

$$
\|x\|_q \leq \|x\|_p \leq n^{1/p-1/q} \|x\|_q.
$$

**Lemma A.6.11** (Gaussian Moments). Let $p \geq 1$, and let $g \sim N(0, 1)$ be a standard normal random variable. Then

$$
\mathbb{E}[|g|^p] = 2^{p/2} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)};
\mathbb{E}[|g|^p] = \frac{1}{2} \mathbb{E}[|g|^2],
$$

where $[x]_+ = \max\{x, 0\}$. In particular $\mathbb{E}[|g|^p] \leq p^{p/2}$, so that $g$ is subgaussian and $g^2$ is subexponential.
Appendix B: Proofs for Construction of Certificates

B.1 Details of Figures

B.1.1 Figure 3.1

**V-number experiment.** In each panel, the two curves are projection of curves \( x_+ : [0, 2\pi] \to \mathbb{S}^3 \) and \( x_- : [0, 2\pi] \to \mathbb{S}^3 \). We actually generate the curves as shown in the figure (i.e., in a three-dimensional space), then map them to the sphere using the map

\[
(u, v, w) \mapsto (u, v, w, \sqrt{1 - u^2 - v^2 - w^2})
\]

In this three-dimensional space, the top left panel’s blue curve (denoted \( x_- \) henceforth) and each panel’s red curve (denoted \( x_+ \) henceforth, and which is the same for all panels) are defined by the parametric equations

\[
\begin{align*}
    x_{-1}(t) &= \cos\left(\frac{\pi}{2}\right) \cos(t) (\sin(4t) + 1 + \delta) + \sin\left(\frac{\pi}{2}\right) \sin(t) (\sin(4t) + 1 + \delta) \\
    x_{-2}(t) &= \cos\left(\frac{\pi}{2}\right) \cos(t) (\sin(4t) + 1 + \delta) + \sin\left(\frac{\pi}{2}\right) \sin(t) (\sin(4t) + 1 + \delta) \\
    x_{-3}(t) &= -\sin\left(\frac{\pi}{2}\right) \cos(t) (\sin(4t) + 1 + \delta) + \cos\left(\frac{\pi}{2}\right) \sin(t) (\sin(4t) + 1 + \delta) \\
    x_{+1}(t) &= 4 \sin(t) \\
    x_{+2}(t) &= 4 (\cos(t) - 1) \\
    x_{+3}(t) &= 0
\end{align*}
\]

where \( \delta \) sets the separation between the manifolds and is set here to \( \delta = 0.05 \). We then rescale both curves by a factor .01: the scale of the curves is chosen such that the curvature of the sphere has a negligible effect on the curvature of the manifolds (since the chart mapping we use here distorts...
Figure B.1: The two curve geometry described in Section B.1.1. The different choices of $\mathcal{M}_-$ that lead to different $\mathcal{B}$-number are overlapping. The legend indicates the $\mathcal{B}$-number of the two curves problem obtained by considering the same $\mathcal{M}_+$ but a different $\mathcal{M}_-$ as indicated by the color.

The curves more nearer to the boundary of the unit disk \(\{(u, v, w) \mid u^2 + v^2 + w^2 \leq 1\}\).\(^1\)

From here, we use an “unfolding” process to obtain the blue curves in the other three panels from $x_-$. To do this, points where $|\frac{dx_2}{dt}| = |\frac{dx_3}{dt}|$ are found numerically. There are 8 such points in total, and parts of the curve between pairs of these points are reflected across the line defined by such a pair in the $(x_2, x_3)$ plane. This can be done for any number of pairs between 1 and 4, generating the curves shown. This procedure ensures that aside from the set of 8 points, the curvature at every point along the curve is preserved and there is no discontinuity in the first derivative, while making the geometries loop back to the common center point more. For an additional visualization of the geometry, see Figure B.1.\(^2\)

Given these geometries, in order to compute the certificate norm for the experiment in the top-right panel, we evaluate the resulting curves at 200 points each, chosen by picking equally spaced points in $[0, 2\pi]$ and evaluating the parametric equations. The certificate itself is evaluated

\(^1\)Although this adds a minor confounding effect to our experiments with certificate norm in the top-right panel, it is suppressed by setting the scale sufficiently small, and it can be removed in principle by using an isometric chart for the upper hemisphere instead of the map given above.

\(^2\)For a three-dimensional interactive visualization, see https://colab.research.google.com/drive/1xmpYeLK606DtXOkJEt_apAniEB9fARRv?usp=sharing.
numerically as in Section B.1.2.

**Rotated MNIST digits.** We rotate an MNIST image around its center by \( i \times \pi / 100 \) for integer \( i \) between 0 and 199. We then apply t-SNE [233] using the scikit-learn package with perplexity 20 to generate the embeddings.

B.1.2 Figure 3.2

We give full implementation details for this figure here, mixed with conceptual ideas that underlie the implementation. The manifolds \( \mathcal{M}_+ \) and \( \mathcal{M}_- \) are defined by parametric equations \( \mathbf{x}_+ : [0, 1] \rightarrow \mathbb{S}^2 \) and \( \mathbf{x}_- : [0, 1] \rightarrow \mathbb{S}^2 \); it is not practical to obtain unit-speed parameterizations of general curves, so we also have parametric equations for their derivatives \( \dot{x}_\sigma : [0, 1] \rightarrow \mathbb{R}^2 \). These are important in our setting since for non-unit-speed curves, the chain rule gives for the integral of a function (say) \( f : \mathcal{M}_+ \rightarrow \mathbb{R} \)

\[
\int_{\mathcal{M}_+} f(\mathbf{x}) \, d\mathbf{x} = \int_{[0,1]} (f \circ \mathbf{x}_+(t)) \|\dot{x}_+(t)\|_2 \, dt.
\]

In particular, in our experiments, we want to work with a uniform density \( \rho = (\rho_+, \rho_-) \) on the manifolds, where the classes are balanced. To achieve this, use the previous equation to get that we require

\[
1 = \int_{\mathcal{M}_+} \rho_+(\mathbf{x}) \, d\mathbf{x} + \int_{\mathcal{M}_-} \rho_-(\mathbf{x}) \, d\mathbf{x}
= \int_{\mathcal{M}_+} (\rho_+ \circ \mathbf{x}_+)(t) \|\dot{x}_+(t)\|_2 \, dt + \int_{\mathcal{M}_-} (\rho_- \circ \mathbf{x}_-)(t) \|\dot{x}_-(t)\|_2 \, dt.
\]

A uniform density on \( \mathcal{M} \) is not a constant value—rather, it is characterized by being translation-invariant. It follows that \( \rho_\sigma \) should be defined by

\[
\rho_\sigma \circ \mathbf{x}_\sigma(t) = \frac{1}{2\|\dot{x}_\sigma(t)\|_2^2}.
\]
For the experiment, we solve a discretization of the certificate problem, for which the above ideas will be useful. Consider $\Theta$ in (3.3) for a fixed depth $L$ (and $n = 2$, since width is essentially irrelevant here). By the above discussion, the certificate problem in this setting is to solve for the certificate $g = (g_+, g_-)$

$$f_* = \frac{1}{2} \left( \int_{[0,1]} \Theta(\cdot, x_+(t))g_+ \circ x_+(t) \, dt + \int_{[0,1]} \Theta(\cdot, x_-(t))g_- \circ x_-(t) \, dt \right).$$

Here, we have eliminated the initial random neural network output $f_{\theta_0}$ from the RHS. Aside from making computation easier, this is motivated by fact that the network output is approximately piecewise constant for large depth $L$, and we therefore expect it not to play much of a role here.

Let $M \in \mathbb{N}$ denote the discretization size. Then a finite-dimensional approximation of the previous integral equation is given by the linear system

$$f_* \circ x_\sigma(t_i) = \frac{1}{2M} \left( \sum_{j=1}^{M} \Theta(x_\sigma(t_i), x_+(t_j))g_+ \circ x_+(t_j) + \sum_{j=1}^{M} \Theta(x_\sigma(t_i), x_-(t_j))g_- \circ x_-(t_j) \right)$$

for all $i \in [M]$ and $\sigma \in \{\pm 1\}$, and where $t_i = (i-1)/M$. Of course, $f_* \circ x_\sigma(t) = \sigma$, so the equation simplifies further, and because the kernel $\Theta$ and this target $f_*$ are smooth, there is a convergence of the data in this linear system in a precise sense to the data in the original integral equation as $M \to \infty$. In particular, define a matrix $T^+$ by $T^+_{ij} = \Theta(x_+(t_i), x_+(t_j))$, define a matrix $T^-$ by $T^-_{ij} = \Theta(x_-(t_i), x_-(t_j))$, and define a matrix $T^\pm$ by $T^\pm_{ij} = \Theta(x_+(t_i), x_-(t_j))$, all of size $M \times M$.

Then the $2M \times 2M$ linear system

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{1}{2M} \begin{bmatrix} T^+ & T^\pm \\ (T^\pm)^* & T^- \end{bmatrix} \begin{bmatrix} g_+ \\ g_- \end{bmatrix}$$

is equivalent to the discretization in (B.1). We implement and solve the system in (B.2) using the definitions we have given above, using the pseudoinverse of the $2M \times 2M$ matrix appearing in this expression to obtain $[g_+, g_-]^*$, and plot the results in Figure 3.2, in particular interpreting $(g_\sigma)$, as
the sampled point $g_{\sigma} \circ x_{\sigma}(t_i)$ as in (B.1) when we plot in the left panel of Figure 3.2. Evidently, it would be immediate to modify the experiment to replace the LHS of (B.1) by the error $f \theta_0 - f_*$: the same protocol given above would work, but there would be an element of randomness added to the experiments.

Specifically, in Figure 3.2 we set $M = 900$. When plotting the solution to (B.2), i.e. the vector $[g_+, g_-]^*$, we moreover scale the vector by a factor of 0.3 to facilitate visualization.

**B.2 Key Definitions**

B.2.1 Problem Formulation

The contents of this section will mirror Section 3.2.1, but provide additional technical details that were omitted there for the sake of concision and clarity of exposition. In this sense, we will focus on a rigorous formulation of the problem here, rather than on intuition: we encourage the reader to consult Section 3.2.1 for a more conceptually-oriented problem formulation.

Adopting the model problem studied in Chapter 2, we let $M_+, M_-$, denote two class manifolds, each a smooth, regular, simple closed curve in $S^{n_0-1}$, with ambient dimension $n_0 \geq 3$. We further assume $M$ precludes antipodal points by asking

$$\angle(x, x') \leq \pi/2, \quad \forall x, x' \in M.$$  \hspace{1cm} (B.3)

We denote $M = M_+ \cup M_-$, and the data measure supported on $M$ as $\mu$. We assume that $\mu$ admits a density $\rho$ with respect to the Riemannian measure on $M$, and that this density is bounded from below by some $\rho_{\min} > 0$. We will also write $\rho_{\max} = \sup_{x \in M} \rho(x)$. For background on curves and manifolds, we refer the reader to to [217, 218].

Given $N$ i.i.d. samples $(x_1, \cdots, x_N)$ from $\mu$ and their labels, given by the labeling function $f_* : M \to \{\pm 1\}$ defined by

$$f_*(x) = \begin{cases} +1 & x \in M_+ \\ -1 & x \in M_- \end{cases}.$$
we train a fully-connected network with ReLU activations and $L$ hidden layers of width $n$ and scalar output. We will write $\theta = (W^1, \ldots, W^{L+1})$ to denote an abstract set of admissible parameters for such a network; concretely, the features at layer $\ell \in \{1, 2, \ldots, L\}$ with parameters $\theta$ and input $x$ are written as $\alpha^\ell_\theta(x) = [W^\ell \alpha^{\ell-1}_\theta(x)]_+$, where $[x]_+ = \max\{x, 0\}$ denotes the ReLU (and we adopt in general the convention of writing $[x]_+$ to denote application of the scalar function $[\cdot]_+$ to each entry of the vector $x$), with boundary condition $\alpha^0_\theta(x) = x$, and the network output on an input $x$ is written $f_\theta(x) = W^{L+1} \alpha^L_\theta(x)$. We will also write $\zeta_\theta(x) = f_\theta(x) - f_\star(x)$ to denote the fitting error. We use Gaussian initialization: if $\ell \in \{1, 2, \ldots, L\}$, the weights are initialized as $W^\ell_{ij} \sim \mathrm{i.i.d.} \mathcal{N}(0, \frac{2}{n})$, and the top level weights are initialized as $W^{L+1}_{i} \sim \mathrm{i.i.d.} \mathcal{N}(0, 1)$ in order to preserve the expected feature norm.\footnote{This initialization style is common in practice (it might be referred to as “fan-out initialization” in that context), but less common in the theoretical literature on kernel regime training of deep neural networks, where a less-natural “NTK parameterization” is typically employed. A detailed discussion of these differences, and how to translate results for one parameterization into those for another, can be found (for example) in [221, §A.3].} In the sequel, we will write $\theta_0$ to denote the collection of these initial random parameters, and therefore $f_{\theta_0}$ to denote the initial random network.

Algorithmically, our problem setup follows Section A.1 exactly, but we write $\Theta^{\text{NTK}}$ for the empirical time-zero NTK in this section instead of the notation used there. In the remainder of this section, we introduce several notations for quantities related to the certificate problem which we will refer to throughout these appendices. We let $\Theta$ denote the following approximation to the neural tangent kernel:

$$
\Theta(x, x') = \frac{n}{2} \sum_{\ell=0}^{L-1} \prod_{\ell' = \ell}^{L-1} \left(1 - \frac{\varphi(\ell')(\zeta(x, x'))}{\pi}\right),
$$

where $\varphi(\ell)$ denotes the $\ell$-fold composition of the angle evolution function

$$
\varphi(t) = \cos^{-1}\left( (1 - \frac{t}{\pi}) \cos \frac{t}{\pi} + \sin \frac{t}{\pi} \right)
$$

. We let $\zeta$ denote the following piecewise constant approximation to $\zeta_0$:

$$
\zeta(x) = -f_\star(x) + \int_M f_{\theta_0}(x') \, d\mu(x').
$$
We also use the notation

\[
\zeta^{(\ell)}(t) = \prod_{\ell' = \ell}^{L-1} \left(1 - \frac{\varphi^{(\ell')}(t)}{\pi}\right)
\]

\[
\psi(t) = \frac{n}{2} \sum_{\ell=0}^{L-1} \zeta^{(\ell)}(t)
\]

for convenience. We find it convenient in our analysis to consider \(\psi\) and its “DC component”, i.e., its value at \(\pi\), separately. To this end, we write \(\psi^\circ = \psi - \psi(\pi)\). We also write the subtracted approximate NTK as \(\Theta^\circ(x, x') = \psi^\circ(\angle(x, x'))\). As a consequence, we have

\[
\psi^\circ(\angle(x, x')) = \Theta^\circ(x, x') = \Theta(x, x') - \psi(\pi).
\]  

We use \(\Theta_\mu\) to represent the integral operator with

\[
\Theta_\mu[g](x) = \int_M \Theta(x, x') g(x') \, d\mu(x'),
\]

and similarly for \(\Theta^\circ_\mu\). An omitted subscript/measure will denote the Riemannian measure on \(M\).

### B.2.2 Geometric Properties

We assume our data manifold \(M = M_+ \cup M_-\), where \(M_+\) and \(M_-\) each is a smooth, regular, simple closed curve on the unit sphere \(\mathbb{S}^{n_0-1}\). Because the curves are regular, it is without loss of generality to assume they are unit-speed and parameterized with respect to arc length \(s\), giving parameterizations as maps from \([0, \text{len}(M_\sigma)]\) to \(\mathbb{S}^{n_0-1}\), as we have defined them in Section 3.2.2 of the main body. Throughout the appendices, we will find it convenient to consider periodic extensions of these arc-length parameterizations, which are smooth and well-defined by the fact that our manifolds are smooth, closed curves: for \(\sigma \in \{\pm\}\), we use \(x_\sigma(s) : \mathbb{R} \to \mathbb{S}^{n_0-1}\) to represent these parameterizations of the two manifolds.\(^4\) We require that the two curves are disjoint. Notice

\(^4\)We clarify an abuse of notation we will commit with these parameterizations throughout the analysis, which stems from the fact that the curves are closed (i.e. topologically circles). That is, there is no preferred basepoint (i.e. the
that as the two curves do not self intersect, we have \( x_\sigma(s) = x_{\sigma'}(s') \) if and only if \( \sigma = \sigma' \) and 
\[ s' = s + k \text{len}(M_\sigma) \]
for some \( k \in \mathbb{Z} \). Precisely, our arguments will require our curves to have
‘five orders’ of smoothness, in other words \( x_\sigma(s) \) must be five times continuously differentiable
for \( \sigma \in \{+,-\} \).

For a differentiable function \( h : M \to \mathbb{C}^p \) with \( p \in \mathbb{N} \), we define its derivative \( \frac{d}{ds}h \) as
\[
\frac{d}{ds} h(x) = \left[ \frac{d}{dt} \bigg|_{t} h(x_\sigma(t)) \right]_{x_\sigma(s) = x} = \left[ \lim_{t \to 0} \frac{1}{t} (h(x_\sigma(s + t)) - h(x_\sigma(s))) \right]_{x_\sigma(s) = x} . \tag{B.7}
\]

We call attention to the “restriction” bar used in this notation: it should be read as “let \( s \) and \( \sigma \) be such that \( x_\sigma(s) = x \)” in the definition’s context. This leads to a valid definition in (B.7) because our curves are simple and disjoint, so for any choice \( s, s' \) with \( x_\sigma(s) = x_{\sigma'}(s') = x \), we have \( x_\sigma(s + t) = x_{\sigma'}(s' + t) \) for all \( t \). We will use this notation systematically throughout these appendices. We further denote its \( i \)-th order derivative by \( h^{(i)}(x) \). For \( i \in \mathbb{N} \), we use \( C^i(M) \) to represent the collection of real-valued functions \( h : M \to \mathbb{R} \) whose derivatives \( h^{(1)}, \ldots, h^{(i)} \) exist and are continuous.

In particular, consider the inclusion map \( \iota : M \to \mathbb{R}^{n_0} \), which is the identification \( \iota(x) = x \).

Following the definition as above, we have
\[
\iota^{(i+1)}(x) = \left[ \lim_{t \to 0} \frac{1}{t} (\iota^{(i)}(x_\sigma(s + t)) - \iota^{(i)}(x_\sigma(s))) \right]_{x_\sigma(s) = x} . \tag{B.8}
\]

In the sequel, with abuse of notation we will use \( x^{(i)} \) to represent \( \iota^{(i)}(x) \). For example, we will write expressions such as \( \sup_{x \in M} \|x^{(2)}\|_2 \) to denote the quantity \( \sup_{x \in M} \|\iota^{(2)}(x)\|_2 \). This notation will enable increased concision, and it is benign, in the sense that it is essentially an identification. We call attention to it specifically to note a possible conflict with our notation for the parameterizations
\[ x_{\sigma}(0) \] for the arc length parameterizations (the curves are only defined up to translation): because our primary use for these parameterizations is in the analysis of extrinsic distances between points on the curves, the basepoint will be irrelevant.
and their derivatives $\mathbf{x}^{(i)}$, which are maps from $\mathbb{R}$ to $\mathbb{R}^{m_0}$ (say), rather than maps defined on $\mathcal{M}$. In this context, we also use $\dot{x}$ and $\ddot{x}$ to represent first and second derivatives $x^{(1)}$ and $x^{(2)}$ for brevity. We have $\|\mathbf{x}\|_2 = \|\dot{x}\|_2 = 1$ from the fact that $\mathcal{M} \subset S^{m_0-1}$ and that we have a unit-speed parameterization. This and associated facts are collected in Lemma B.4.3.

For any real or complex-valued function $h$, the integral operator over manifold can be written as

$$
\int_{\mathbf{x} \in \mathcal{M}} h(\mathbf{x}) d\mu(\mathbf{x}) = \sum_{\sigma = \pm} \int_{s=0}^{\text{len}(\mathcal{M}_\sigma)} h(\mathbf{x}_\sigma(s)) \rho(\mathbf{x}_\sigma(s)) ds,
$$

$$
\int_{\mathbf{x} \in \mathcal{M}} h(\mathbf{x}) d\mathbf{x} = \sum_{\sigma = \pm} \int_{s=0}^{\text{len}(\mathcal{M}_\sigma)} h(\mathbf{x}_\sigma(s)) ds.
$$

We have defined key geometric properties in the main body, in Section 3.2.2. Our arguments will require slightly more technical definitions of these quantities, however. In the remainder of this section, we introduce the same definition of angle injectivity radius and $\mathcal{B}$-number with a variable scale, which helps us in proofs in Section B.4.

First, we give a precise definition for the intrinsic distance $d_M$ on the curves. To separate the notions of “close over the sphere” and “close over the manifold”, we use the extrinsic distance (angle) $\angle(\mathbf{x}, \mathbf{x}') = \cos^{-1}(\mathbf{x}, \mathbf{x}')$ to measures closeness between two points $\mathbf{x}, \mathbf{x}'$ over the sphere. The distance over the manifold is measured through the intrinsic distance $d_M(\mathbf{x}, \mathbf{x}')$, which takes $\infty$ when $\mathbf{x}$ and $\mathbf{x}'$ reside on different components $\mathcal{M}_+$ and $\mathcal{M}_-$ and the length of the shortest curve on the manifold connecting the two points when they belong to the same component. More formally, we have

$$
d_M(\mathbf{x}, \mathbf{x}') = \begin{cases} 
\inf \{ |s - s'| \mid \mathbf{x}_{\sigma}(s) = \mathbf{x}, \mathbf{x}_{\sigma}(s') = \mathbf{x}' \} & f_*(\mathbf{x}) = f_*(\mathbf{x}'), \\
+\infty & \text{otherwise},
\end{cases}
$$

(B.9)

where the infimum is taken over all valid $\sigma \in \{+, -\}$ and $(s, s') \in \mathbb{R}^2$. Notice that as the curves $\mathcal{M}_\sigma$ do not intersect themselves, one has $x_\sigma(s_1) = x_\sigma(s_2)$ if and only if $s_1 = s_2 + k \text{len}(\mathcal{M}_\sigma)$ for some $k \in \mathbb{Z}$. Thus for any two points $\mathbf{x}, \mathbf{x}'$ that belong to the same component $\mathcal{M}_\sigma$, the above
infimum is attained: there exist \( s, s' \) such that \( x_\sigma(s) = x, x_\sigma(s') = x' \), and \( d_M(x, x') = |s - s'| \).

**Angle Injectivity Radius** For \( \varepsilon \in (0, 1) \) we define the angle injectivity radius of scale \( \varepsilon \) as

\[
\Delta_\varepsilon = \min \left\{ \frac{\sqrt{\varepsilon}}{\hat{k}}, \inf_{x, x' \in M} \left\{ \angle(x, x') \left| d_M(x, x') \geq \frac{\sqrt{\varepsilon}}{\hat{k}} \right\} \right\}, \quad (B.10)
\]

which is the smallest extrinsic distance between two points whose intrinsic distance exceeds \( \frac{\sqrt{\varepsilon}}{\hat{k}} \) with

\[
\hat{k} = \max \left\{ \kappa, \frac{2}{\pi} \right\}. \quad (B.11)
\]

Observe that for any scale \( \varepsilon \), \( \Delta_\varepsilon \) is smaller than inter manifold separation \( \min_{x \in M, x' \in M_\perp} \angle(x, x') \).

**\( \boxtimes \)-number** For \( \varepsilon \in (0, 1), \delta \in (0, (1 - \varepsilon)] \), we define \( \boxtimes \)-number of scale \( \varepsilon, \delta \) as

\[
\boxtimes_{\varepsilon, \delta}(M) = \sup_{x \in M} N_M \left( \left\{ x' \left| d_M(x, x') \geq \frac{\sqrt{\varepsilon}}{\hat{k}} \text{ and } \angle(x, x') \leq \frac{\delta \sqrt{\varepsilon}}{\hat{k}} \right\} \right), \frac{1}{\sqrt{1 + \kappa^2}} \right) \quad (B.12)
\]

Here, \( N_M(T, \varepsilon) \) is the size of a minimal \( \varepsilon \) covering of \( T \) in the intrinsic distance on the manifold. We call the set \( \left\{ x' \left| d_M(x, x') \geq \frac{\sqrt{\varepsilon}}{\hat{k}}, \angle(x, x') \leq \frac{\delta \sqrt{\varepsilon}}{\hat{k}} \right\} \) appearing in this definition the winding piece of scale \( \varepsilon \) and \( \delta \): it contains points that are far away in intrinsic distance but close in extrinsic distance. We will give it a formal definition in (B.24), where it will play a key role in our arguments.

In the sequel, we denote \( \Delta, \boxtimes(M) \) to be the angle injectivity radius and \( \boxtimes \)-number with the specific instantiations \( \varepsilon = \frac{1}{20} \) and \( \delta = 1 - \varepsilon \). These are key geometric features used in Theorem 3.3.1 and Theorem 1.

**B.2.3 Subspace of Smooth Functions and Kernel Derivatives**

As the behavior of the kernel and its approximation is easier to understand when constrained in a low frequency subspace, we first introduce the notion of low-frequency subspace formed by the Fourier basis on the two curves.
Fourier Basis and Subspace of Smooth Functions

We define a Fourier basis of functions over the manifold as

\[
\phi_{\sigma,k}(x_{\sigma'}(s)) = \begin{cases} 
\frac{1}{\sqrt{\text{len}(M,\sigma)}} \exp \left( \frac{12\pi k s}{\text{len}(M,\sigma)} \right), & \sigma' = \sigma \\
0, & \sigma' \neq \sigma
\end{cases}
\]  

(B.13)

for each \( k = 0, 1, \ldots \), and further define a subspace of low frequency functions

\[
S_{K_+,K_-} = \text{span}_\mathbb{C}\{\phi_{+,0}, \phi_{+,-1}, \phi_{+,1}, \ldots, \phi_{+,K_+}, \phi_{-,K_-}, \phi_{-,0}, \ldots, \phi_{-,K_-}\}
\]  

(B.14)

for \( K_+, K_- \geq 0 \). Using the fact that our curves are unit-speed, one can see that indeed (B.13) defines an orthonormal basis for \( L^2 \) functions on \( M \).

B.3 Main Results

Theorem B.3.1 (Generalization). Let \( M \) be two disjoint smooth, regular, simple closed curves, satisfying \( \angle(x, x') \leq \pi/2 \) for all \( x, x' \in M \). For any \( 0 < \delta \leq 1/e \), choose \( L \) so that

\[
L \geq K \max \left\{ \frac{1}{(\Delta(1 + \kappa^2))^{C^2(M)}}, C_\mu \log^9 \left( \frac{1}{\delta} \right) \log^{24} (C_\mu n_0 \log(\frac{1}{\delta})), e^{C' \max\{\text{len}(M)\kappa, \log(k)\}}, P \right\}
\]

\[
n = K'L^{99} \log^9 (1/\delta) \log^{18} (Ln_0)
\]

\[
N \geq L^{10},
\]

and fix \( \tau > 0 \) such that \( \frac{C''}{nL^2} \leq \tau \leq \frac{\tau}{nL} \). Then with probability at least \( 1 - \delta \), the parameters obtained at iteration \( \lfloor L^{39/44}/(n\tau) \rfloor \) of gradient descent on the finite sample loss yield a classifier that separates the two manifolds.

The constants \( c, C, C', C'', K, K' > 0 \) are absolute, and \( C_\mu = \frac{\max\{\rho^{19} \rho^{-19}(1+\rho_{\max})^{12}\}}{(\min\{\mu(M_e), \mu(M_\zeta)\})^{11/2}} \). \( P \) is a polynomial \( \text{poly}\{M_3, M_4, M_5, \text{len}(M), \Delta^{-1}\} \) of degree at most 36, with degree at most 12 when viewed as a polynomial in \( M_3, M_4, M_5 \) and \( \text{len}(M) \), and of degree at most 24 as a polynomial in \( \Delta^{-1} \).
Proof. The proof is an application of Theorem A.2.1; we note that the conditions on \( n, L, \delta, N, \) and \( \tau \) imply all hypotheses of this theorem, except for the certificate condition. We will complete the proof by showing that the certificate condition is also satisfied, under the additional hypotheses on \( L \) and with a suitable choice of \( q_{\text{cert}} \).

First, we navigate a difference in the formulation of the two curves’ regularity properties between our work and the analysis of Chapter 2, from which Theorem A.2.1 is drawn. Theorem A.2.1 includes a condition \( L \geq C \kappa_{\text{ext}}^2 C_{\lambda} \) for some absolute constant \( C \), where \( \kappa_{\text{ext}} = \sup_{x \in M} \| \dot{x} \|_2^2 \) is a bound on the extrinsic curvature (we will discuss \( C_{\lambda} \) momentarily). In our context, we have \( M_2 = \kappa_{\text{ext}} \), and following Lemma B.4.3 (using that our curves are unit-speed spherical curves), we get that it suffices to require \( L \geq (1 + \kappa^2) C_{\lambda} \) instead. In turn, we can pass to \( \hat{k} \): since this constant is lower-bounded by a positive number and is larger than \( \kappa \), it suffices to require \( L \geq \kappa^2 C_{\lambda} \). As for \( C_{\lambda} \), this is a constant related to the angle injectivity radius \( \Delta \), and is defined by \( C_{\lambda} = K_{\lambda}^2/\kappa_{\lambda}^2 \), where these two constants satisfy

\[
\forall s \in (0, c_{\lambda}/\kappa_{\text{ext}}], (x, x') \in M_\star \times M_\star, \star \in \{+,-\} : \angle(x, x') \leq s \Rightarrow d_M(x, x') \leq K_{\lambda}s.
\]

We will relate this constant to constants in our formulation. Consider any \( x, x' \in M \). If \( \angle(x, x') \leq \frac{\Delta}{2} \) then from the definition of \( \Delta \) we have \( d_M(x, x') \leq \frac{\sqrt{\varepsilon}}{\hat{k}} \) and hence by (B.49) we find \( d_M(x, x') \leq \angle(x, x') \). If on the other hand \( \angle(x, x') > \frac{\Delta}{2} \), then a trivial bound gives

\[
d_M(x, x') \leq \text{len}(M) = \frac{2\text{len}(M)}{\Delta} < \frac{2\text{len}(M)}{\Delta} \angle(x, x'). \tag{B.15}
\]

We can thus choose \( c_{\lambda} = 1 \), \( K_{\lambda} = \max \left\{ 1, \frac{2\text{len}(M)}{\Delta} \right\} \) to satisfy (B.15), giving \( C_{\lambda} = \max \left\{ 1, \frac{4\text{len}(M)}{\Delta^2} \right\} \).

Thus the requirement \( L > C \kappa_{\text{ext}}^2 C_{\lambda} \) of Theorem A.2.1 is automatically satisfied if

\[
L \geq \max \left\{ P, e^{C \text{len}(M)\hat{k}} \right\}
\]

for a suitable exponent \( C \), where \( P \) is the polynomial in the hypotheses of our result, and so our
hypotheses imply this condition.

Next, we establish the certificate claim. Write $\Theta^{\text{NTK}}$ for the network’s neural tangent kernel, as defined in Section B.2.1, and $\Theta^{\text{NTK}}_\mu$ for the associated Fredholm integral operator on $L^2_\mu$. In addition, write $\zeta_0 = f_{\theta_0} - f_*$ for the initial random network error. Because we have modified some exponents in the constant $C_\mu$, and added conditions on $L$, all hypotheses of Theorem B.3.2 are satisfied: invoking it, we have that there exists $g : \mathcal{M} \rightarrow \mathbb{R}$ satisfying

$$\|g\|_{L^2_\mu} \leq C\frac{\|\zeta\|_{L^2_\mu}}{\rho_{\min n}}$$

and

$$\|\Theta_\mu[g] - \zeta\|_{L^2_\mu} \leq \frac{\|\zeta\|_{L^\infty}}{L}.$$ 

By these bounds, the triangle inequality, the Minkowski inequality, and the fact that $\mu$ is a probability measure, we have

\begin{align*}
\|\Theta^{\text{NTK}}_\mu[g] - \zeta_0\|_{L^2_\mu} &\leq \|\Theta - \Theta^{\text{NTK}}\|_{L^\infty(M \times M)} \|g\|_{L^2_\mu} + \|\Theta_\mu[g] - \zeta\|_{L^2_\mu} + \|\zeta - \zeta_0\|_{L^2_\mu} \\
&\leq C\|\Theta - \Theta^{\text{NTK}}\|_{L^\infty(M \times M)} \frac{\|\zeta\|_{L^\infty(M)}}{n \rho_{\min}} + \frac{\|\zeta\|_{L^\infty}}{L} + \|\zeta - \zeta_0\|_{L^\infty(M)}. \quad (B.16)
\end{align*}

An application of Theorem B.6.1 gives that on an event of probability at least $1 - e^{-cd}$

$$\|\Theta - \Theta^{\text{NTK}}\|_{L^\infty(M \times M)} \leq Cn/L$$

if $d \geq K \log(nn_0 \text{len}(M))$ and $n \geq K'd^4L^5$. In translating this result, we use that in the context of the two curve problem, the covering constant $C_M$ appearing in Theorem A.2.2 is bounded by a constant multiple of $\text{len}(M)$ (this is how we obtain Theorem B.6.1 and some other results in Section B.6). An application of Lemma A.4.11 gives

$$\mathbb{P}\left[\|\zeta_0 - \zeta\|_{L^\infty(M)} \leq \frac{\sqrt{2d}}{L}\right] \geq 1 - e^{-cd}$$

416
\[ \mathbb{P} \left[ \| \zeta_0 \|_{L^\infty(M)} \leq \sqrt{d} \right] \geq 1 - e^{-cd} \]

as long as \( n \geq Kd^4L^5 \) and \( d \geq K' \log(nn_0 \text{ len}(M)) \), where we use these conditions to simplify the residual that appears in Lemma A.4.11. In particular, combining the previous two bounds with the triangle inequality and a union bound and then rescaling \( d \), which worsens the constant \( c \) and the absolute constants in the preceding conditions, gives

\[ \mathbb{P} \left[ \| \zeta \|_{L^\infty(M)} \leq \sqrt{d} \right] \geq 1 - 2e^{-cd}. \]

Combining these bounds using a union bound and substituting into (B.16), we get that under the preceding conditions, on an event of probability at least \( 1 - 3e^{-cd} \) we have

\[
\| \Phi^\text{NTK}_\mu[g] - \zeta_0 \|_{L^2_\mu} \leq \frac{C\sqrt{d}}{L} \left( 2 + \frac{1}{\rho_{\min}} \right) \\
\leq \frac{C\sqrt{d}}{L} \max\{\rho_{\min}, \rho^{-1}_{\min}\},
\] (B.17)

where we worst-case the density constant in the second line, and in addition, on the same event, we have by the norm bound on the certificate \( g \)

\[
\| g \|_{L^2_\mu} \leq C \frac{\sqrt{d}}{n\rho_{\min}}.
\] (B.18)

To conclude, we simplify the preceding conditions on \( n \) and turn the parameter \( d \) into a parameter \( \delta > 0 \) in order to obtain the form of the result necessary to apply Theorem A.2.1. We have in this one-dimensional setting

\[
\text{len}(M) \leq \frac{\text{len}(M_\mu)}{\mu(M_\mu)} + \frac{\text{len}(M_-)}{\mu(M_-)} \leq \frac{2}{\rho_{\min}} \leq 2 \max\{\rho_{\min}, \rho^{-1}_{\min}\}.
\]
where the second inequality here uses simply

$$\mu(M_+) = \int_{M_+} \rho_+(x) \, dx \geq \text{len}(M_+) \rho_{\text{min}}$$

(say). Because \( n \geq 1 \) and \( n_0 \geq 3 \) and \( \max\{\rho_{\text{min}}, \rho_{\text{min}}^{-1}\} \geq 1 \), it therefore suffices to instead enforce the condition on \( d \) as \( d \geq K \log(nn_0 C_\mu) \), where \( C_\mu \) is the constant defined in the lemma statement.

But note from our hypotheses here that we have \( n \geq L \) and \( L \geq C_\mu \); so in particular it suffices to enforce \( d \geq K \log(nn_0) \) for an adjusted absolute constant. Choosing \( d \geq (1/c) \log(1/\delta) \), we obtain that the previous two bounds (B.18) and (B.17) hold on an event of probability at least \( 1 - 3\delta \). When \( \delta \leq 1/e \), given that \( n_0 \geq 3 \) we have \( nn_0 \geq e \) and \( \max\{\log(1/\delta), \log(nn_0)\} \leq \log(1/\delta) \log(nn_0) \), so that it suffices to enforce the requirement \( d \geq K \log(1/\delta) \log(nn_0) \) for a certain absolute constant \( K > 0 \). We can then substitute this lower bound on \( d \) into the two certificate bounds above to obtain the form claimed in (A.4) in Theorem A.2.1 with the instantiation \( q_{\text{cert}} = 1 \), and this setting of \( q_{\text{cert}} \) matches the choice of \( C_\mu \) that we have enforced in our hypotheses here. For the hypothesis on \( n \), we substitute this lower bound on \( d \) into the condition on \( n \) to obtain the sufficient condition \( n \geq K' L^5 \log^4(1/\delta) \log^4(nn_0) \). Using a standard log-factor reduction (e.g. Lemma A.2.14) and possibly worsening absolute constants, we then get that it suffices to enforce \( n \geq K' L^5 \log^4(1/\delta) \log^4(Ln_0 \log(1/\delta)) \), which is redundant with the (much larger) condition on \( n \) that we have enforced here. This completes the proof.

\( \square \)

**Theorem B.3.2 (Certificates).** Let \( M \) be two disjoint smooth, regular, simple closed curves, satisfying \( \angle(x, x') \leq \pi/2 \) for all \( x, x' \in M \). There exist constants \( C, C', C'', C''' \) and a polynomial \( P = \text{poly}(M_3, M_4, M_5, \text{len}(M), \Delta^{-1}) \) of degree at most 36, with degree at most 12 in \( (M_3, M_4, M_5, \text{len}(M)) \) and degree at most 24 in \( \Delta^{-1} \), such that when

\[
L \geq \max \left\{ \exp(C' \text{len}(M) \hat{\kappa}), \left( \frac{1}{\Delta \sqrt{1 + \kappa^2}} \right)^{C''(M)}, C''', \kappa^{10}, P, \kappa^{12} \right\},
\]

418
then for \( \zeta \) defined in (B.5), there exists a certificate \( g : M \to \mathbb{R} \) with

\[
\|g\|_{L^2_\mu} \leq \frac{C\|\zeta\|_{L^2_\mu}}{\rho_{\min}^n \log L}
\]

such that

\[
\|\Theta_\mu[g] - \zeta\|_{L^2_\mu} \leq \|\zeta\|_{L^\infty} L^{-1}.
\]

Proof. See [114, Theorem D.2].

\[ \Box \]

### B.4 Proof for the Certificate Problem

In this section, we prove a key localization lemma that captures the interaction between the network’s depth \( L \) and the data geometry, and is one of the main ingredients in the proof of Theorem B.3.2, our result on certificate construction. We omit the complete proof of Theorem B.3.2, which can be found in [114]. There are two main technical difficulties in establishing this result.

First, \( \Theta \) contains a very large constant term: \( \Theta = \Theta^\circ + \psi(\pi)\mathbb{I}^* \). This renders the operator \( \Theta \) somewhat ill-conditioned. Second, the eigenvalues of \( \Theta^\circ \) are not bounded away from zero: because the kernel is sufficiently regular, it is possible to demonstrate high-frequency functions \( h \) for which \( \|\Theta^\circ[h]\|_{L^2} \ll \|h\|_{L^2} \).

The proof in [114] handles these technical challenges sequentially: in the next section of this thesis, in Section B.4.1, we restrict attention to the DC subtracted kernel \( \Theta^\circ \) and a subspace \( S \) containing low-frequency functions, and show that the restriction \( P_S \Theta^\circ P_S \) to \( S \) is stably invertible over \( S \). In [114, Section E.2], it is argued that the solution \( g \) to \( P_S \Theta^\circ[g] = \zeta \) is regularized enough that \( \Theta^\circ[g] \approx \zeta \), i.e., the restriction to \( S \) can be dropped. Finally, in [114, Section E.3], the move from the DC subtracted kernel \( \Theta^\circ \) without density to the full kernel \( \Theta_\mu \) is justified.
B.4.1 Invertibility Over a Subspace of Smooth Functions

**Proof Sketch and Organization.** In this section, we solve a restricted version of the certificate problem for DC subtracted kernel $\Theta^\circ$, over a subspace $S$ of low-frequency functions defined in (B.14). Namely, for $\zeta \in S$, we demonstrate the existence of a small norm solution $g \in S$ to the equation

$$P_S \Theta^\circ [g] = \zeta.$$  \hspace{1cm} (B.19)

This equation involves the integral operator $\Theta^\circ$, which acts via

$$\Theta^\circ [g](x) = \int_{x' \in M} \Theta^\circ(x, x') g(x') dx'.$$  \hspace{1cm} (B.20)

We argue that this operator is invertible over $S$, by decomposing this integral into four pieces, which we call the **Local**, **Near**, **Far**, and **Winding** components. The formal definitions of these four components follow: for parameters $0 < \varepsilon < 1$, $r > 0$, and $\delta > 0$, we define

**[Local]**:  \hspace{1cm} $L_r(x) = \{x' \in M \mid d_M(x, x') < r\},$ \hspace{1cm} (B.21)

**[Near]**: \hspace{1cm} $N_{r,\varepsilon}(x) = \left\{x' \in M \left| r \leq d_M(x, x') \leq \frac{\sqrt{\varepsilon}}{k}\right\} \right.$, \hspace{1cm} (B.22)

**[Far]**: \hspace{1cm} $F_{\varepsilon,\delta}(x) = \left\{x' \in M \left| d_M(x, x') \geq \frac{\sqrt{\varepsilon}}{k}, \ \angle(x, x') > \frac{\delta \sqrt{\varepsilon}}{k}\right\} \right.$, \hspace{1cm} (B.23)

**[Winding]**: \hspace{1cm} $W_{\varepsilon,\delta}(x) = \left\{x' \in M \left| d_M(x, x') \geq \frac{\sqrt{\varepsilon}}{k}, \ \angle(x, x') \leq \frac{\delta \sqrt{\varepsilon}}{k}\right\} \right.$. \hspace{1cm} (B.24)

It is easy to verify that for any choice of these parameters and any $x \in M$, these four pieces cover $M$: i.e., $L_r(x) \cup N_{r,\varepsilon}(x) \cup F_{\varepsilon,\delta}(x) \cup W_{\varepsilon,\delta}(x) = M$. Intuitively, the **Local** and **Near** pieces contain points that are close to $x$, in the intrinsic distance on $M$. The **Far** component contains points that are far from $x$ in intrinsic distance, and far in the extrinsic distance (angle). The **Winding** component contains portions of $M$ that are far in intrinsic distance, but close in extrinsic distance.

Intuitively, this component captures parts of $M$ that “loop back” into the vicinity of $x$. 

420
Parameter choice. The specific parameters $r, \varepsilon, \delta$ will be chosen with an eye towards the properties of both $M$ and $\Theta^\circ$. The parameter $\varepsilon \in (0, \frac{3}{4})$ is a scale parameter, which controls $r = r_\varepsilon$ such that

1. $r$ is large enough to enable the local component $L_r(x)$ to dominate the kernel’s behavior;

2. $r$ is not too large, so the kernel stays sharp and localized over the local component $L_r(x)$.

Specifically, we choose

\[ a_\varepsilon = (1 - \varepsilon)^3(1 - \varepsilon/12), \quad (B.25) \]
\[ r_\varepsilon = 6\pi L^{-\frac{a_\varepsilon}{a_\varepsilon + 1}}. \quad (B.26) \]

Notice that when $\varepsilon \downarrow 0$, we have $r_\varepsilon \approx L^{-1/2}$. So with a smaller choice of $\varepsilon$ we may get a larger local component with the price of a larger constant dependence.

We further choose $\delta$ to ensure that the Near and Far components overlap. To see that this is possible, note that at the boundary of the Near component, $d_M(x, x') = \sqrt{\varepsilon/\hat{k}}$; from Lemma B.4.4, we have

\[ \angle(x, x') \geq d_M(x, x') - \hat{k}^2 d_M^3(x, x'), \quad (B.27) \]

so at this point $\angle(x, x') \geq (1 - \varepsilon)\sqrt{\varepsilon/\hat{k}}$. Thus as long as $\delta < 1 - \varepsilon$, Near and Far overlap.

Kernel as main and residual. The kernel $\Theta^\circ(x, x')$ is a decreasing function of $\angle(x, x')$: $\Theta^\circ$ is largest over the Local component, smaller over the Near and Winding components, and smallest over the Far component. By choosing the scale parameter $r_\varepsilon$ as in (B.26), we define an operator $M_\varepsilon$ which captures the contribution of the Local component to the kernel:

\[ M_\varepsilon[f](x) = \int_{x' \in L_{r_\varepsilon}(x)} \psi^2(\angle(x, x')) f(x') dx'. \quad (B.28) \]

Because $\angle(x, x')$ is small over $L_{r_\varepsilon}(x)$ when $r_\varepsilon$ is chosen to be small compared to inverse curvature $1/\hat{k}$, on this component, $d_M(x, x') \approx \angle(x, x')$ (which we formalize in Lemma B.4.4). We will use
this property to argue that $M_\varepsilon$ can be approximated by a self-adjoint convolution operator, defined as

$$
\widehat{M}_\varepsilon[f](x) = \int_{s' = s - r_\varepsilon}^{s + r_\varepsilon} \psi^\circ(|s - s'|) f(x_{\sigma}(s')) ds'_{x_{\sigma}(s)} = x.
$$

(B.29)

The restriction is valid because for any choice of $\sigma$ and $s$ such that $x_{\sigma}(s) = x$, the RHS has the same value. On the other hand, given that we require $0 < \varepsilon < \frac{3}{4}$, (B.25) and (B.26) show that when $L$ is chosen larger than a certain absolute constant, we have $r_\varepsilon \leq \pi$, assuring $|s' - s|$ falls in the domain of $\psi^\circ$, which makes this operator well-defined. We will always assume such a choice has been made in the sequel, and in particular include it as a hypothesis in our results.

Notice that $\widehat{M}_\varepsilon$ is an invariant operator: it commutes with the natural translation action on $M$. As a result, it diagonalizes in the Fourier basis defined in (B.13) (i.e., each of these functions is an eigenfunction of $\widehat{M}_\varepsilon$). See Lemma B.4.6 and its proof for the precise formulation of these properties. This enables us to study its spectrum on the subspace of smooth functions defined in (B.14) at the specific scale $\varepsilon$, defined as

$$
S_\varepsilon = S_{K_\varepsilon, +}, K_\varepsilon, -
$$

(B.30)

with $K_{\varepsilon, \sigma} = \left[ \frac{\varepsilon^{1/2} \text{len}(M_\sigma)}{2\pi r_\varepsilon} \right]$ for $\sigma \in \{+, -\}$. In this way, we will establish that $\widehat{M}_\varepsilon$ is stably invertible on $S_\varepsilon$.

In the remainder of the section, we show the diagonalizability and restricted invertibility of $\widehat{M}_\varepsilon$ in Lemma B.4.6, and control the $L^2$ to $L^2$ operator norm of all four components of $\Theta^\circ$ in Lemma B.4.7, Lemma B.4.8, Lemma B.4.9 and Lemma B.4.10. Then we show $\Theta^\circ$ is stably invertible using these results by a Neumann series construction (Lemma B.4.2) and finally prove the main theorem for this section in Theorem B.4.1.

\[Notice that although $\Theta$ and $\zeta$ are real objects, our subspace $S_\varepsilon$ contains complex-valued functions. In the remainder of Section B.4, we will work with complex objects for convenience, which means our constructed certificate candidates can be complex-valued. This will not affect our result because (intuitively) the fact that $\Theta$ and $\zeta$ are real makes the imaginary component of the certificate is redundant, and removing it with a projection onto the subspace of real-valued functions will give us the same norm and residual guarantees for the certificate problem.\]
Theorem B.4.1. For any $\varepsilon \in (0, \frac{3}{4})$, $\delta \in (0, 1-\varepsilon]$, there exist an absolute constant $C$ and constants $C_{\varepsilon}, C'_{\varepsilon, \delta}, C''_{\varepsilon}$ depending only on the subscripted parameters such that if

$$L \geq \max \left\{ \exp \left( C_{\varepsilon, \delta}' \text{len}(\mathcal{M}) \varepsilon \right), \left( 1 + \frac{1}{\Delta_{\varepsilon} \sqrt{1 + \kappa}} \right)^{C_{\varepsilon, \delta}''}, \left( \varepsilon^{-1/2} 2\pi k \right)^{\frac{a_{\varepsilon} + 1}{a_{\varepsilon}}, C_{\varepsilon}} \right\},$$

where $a_{\varepsilon}, r_{\varepsilon}$ as in (B.25) and (B.26) and we set subspace $S_{\varepsilon}$ and the invariant operator $\widehat{M}_{\varepsilon}$ as in (B.30) and (B.29), we have $P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}}$ is invertible over $S_{\varepsilon}$, and

$$\left\| \left( P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}} \right)^{-1} P_{S_{\varepsilon}} \left( \Theta^\circ - \widehat{M}_{\varepsilon} \right) P_{S_{\varepsilon}} \right\|_{L^2 \to L^2} \leq 1 - \varepsilon.$$

Moreover, for any $\zeta \in S_{\varepsilon}$, the equation $P_{S_{\varepsilon}} \Theta^\circ [g] = \zeta$ has a unique solution $g_{\varepsilon} [\zeta] \in S_{\varepsilon}$ given by the convergent Neumann series

$$g_{\varepsilon} [\zeta] = \sum_{\ell=0}^{\infty} (-1)^{\ell} \left( P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}} \right)^{-1} P_{S_{\varepsilon}} \left( \Theta^\circ - \widehat{M}_{\varepsilon} \right) P_{S_{\varepsilon}} \left( P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}} \right)^{-1} \zeta,$$  \hspace{1cm} (B.31)

which satisfies

$$\| g_{\varepsilon} [\zeta] \|_{L^2} \leq \frac{C \| \zeta \|_{L^2}}{\varepsilon n \log L}. \hspace{1cm} (B.32)$$

Proof. We construct $g \in S_{\varepsilon}$ satisfying $P_{S_{\varepsilon}} \Theta^\circ [g] = \zeta$ by equivalently writing

$$P_{S_{\varepsilon}} \Theta^\circ [g] = \left( P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}} + P_{S_{\varepsilon}} \left( \Theta^\circ - \widehat{M}_{\varepsilon} \right) P_{S_{\varepsilon}} \right) [g].$$

Under our hypotheses, Lemma B.4.2 implies the invertibility of $P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}}$ with

$$\lambda_{\min} \left( P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}} \right) \geq \frac{1}{1 - \varepsilon} \left\| \Theta^\circ - \widehat{M}_{\varepsilon} \right\|_{L^2 \to L^2}, \hspace{1cm} (B.33)$$

where $\lambda_{\min} \left( P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}} \right)$ is the minimum eigenvalue of the self-adjoint operator $P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}} : S_{\varepsilon} \to S_{\varepsilon}$ as shown in Lemma B.4.6. In particular, $P_{S_{\varepsilon}} \widehat{M}_{\varepsilon} P_{S_{\varepsilon}}$ is invertible, and the system we seek
to solve can be written equivalently as

\[
\left( P_{S_e} \hat{M}_e P_{S_e} \right)^{-1} \zeta = \left( \text{Id}_{S_e} + \left( P_{S_e} \hat{M}_e P_{S_e} \right)^{-1} P_{S_e} \left( \Theta^\circ - \hat{M}_e \right) P_{S_e} \right) [g].
\]

where the LHS of the last system is in \( S_e \). Next, we argue that the operator that remains on the RHS of the last equation is invertible. Noting that

\[
\left( P_{S_e} \hat{M}_e P_{S_e} \right)^{-1} P_{S_e} \left( \Theta^\circ - \hat{M}_e \right) P_{S_e} \leq \lambda_{\min} \left( P_{S_e} \hat{M}_e P_{S_e} \right)^{-1} \left\| \Theta^\circ - \hat{M}_e \right\|_{L^2 \rightarrow L^2}
\]

\[
\leq 1 - \varepsilon
\]

(B.34)

using both Lemma B.4.6 and (B.33), we have by the Neumann series that

\[
\left( \text{Id}_{S_e} + \left( P_{S_e} \hat{M}_e P_{S_e} \right)^{-1} P_{S_e} \left( \Theta^\circ - \hat{M}_e \right) P_{S_e} \right)^{-1} = \sum_{i=0}^{\infty} \left( -1 \right)^i \left( P_{S_e} \hat{M}_e P_{S_e} \right)^{-1} P_{S_e} \left( \Theta^\circ - \hat{M}_e \right) P_{S_e}^i.
\]

Thus we know \( g_\varepsilon[\zeta] \) in (B.31) serves as the solution to the equation \( P_{S_e} \Theta^\circ [g] = \zeta \).

Furthermore, from Lemma B.4.6 when \( L \geq C_\varepsilon \), we have

\[
\left\| \left( P_{S_e} \hat{M}_e P_{S_e} \right)^{-1} \zeta \right\|_{L^2} \leq \lambda_{\min} \left( P_{S_e} \hat{M}_e P_{S_e} \right)^{-1} \left\| \zeta \right\|_{L^2}
\]

\[
\leq \frac{1}{cn \log L} \left\| \zeta \right\|_{L^2}.
\]

Combining this bound with (B.34) and the triangle inequality in the series representation (B.31),
we obtain the claimed norm bound in (B.32):

\[ \| g_\varepsilon [\xi] \|_{L^2} \leq \sum_{\ell=0}^{\infty} (1 - \varepsilon)^\ell \left\| \left( P_{S_\varepsilon} \widetilde{M}_\varepsilon P_{S_\varepsilon} \right)^{-1} \xi \right\|_{L^2} \leq \frac{C \| \xi \|_{L^2}}{\varepsilon n \log L}. \]

\[ \square \]

**Lemma B.4.2.** Let \( \varepsilon \in (0, \frac{3}{4}) \), \( \delta \in (0, 1 - \varepsilon] \), and let \( a_\varepsilon \), \( \widetilde{M}_\varepsilon \) and \( S_\varepsilon \) be as in (B.25), (B.29) and (B.30). There are constants \( C_\varepsilon, C'_{\varepsilon, \delta}, C''_{\varepsilon} \) depending only on the subscripted parameters such that if

\[ L \geq \max \left\{ \exp \left( C'_{\varepsilon, \delta} \text{len}(M) \right), \left( 1 + \frac{1}{\Delta_\varepsilon \sqrt{1 + \kappa^2}} \right)^{C''_{\varepsilon, \delta}(M)}, \left( \varepsilon^{-1/2} 12\pi \right)^{\frac{a_\varepsilon + 1}{2}}, C_\varepsilon \right\}, \]

we have \( P_{S_\varepsilon} \widetilde{M}_\varepsilon P_{S_\varepsilon} \) is invertible over \( S_\varepsilon \) with

\[ \lambda_{\min} \left( P_{S_\varepsilon} \widetilde{M}_\varepsilon P_{S_\varepsilon} \right) \geq \frac{1}{1 - \varepsilon} \| \Theta^\circ - \widetilde{M}_\varepsilon \|_{L^2 \to L^2}. \]

where \( \lambda_{\min} \left( P_{S_\varepsilon} \widetilde{M}_\varepsilon P_{S_\varepsilon} \right) \) is defined in Lemma B.4.6.

**Proof.** From triangle inequality for the \( L^2 \to L^2 \) operator norm, we have

\[ \left\| \Theta^\circ - \widetilde{M}_\varepsilon \right\|_{L^2 \to L^2} \leq \left\| \Theta^\circ - M_\varepsilon \right\|_{L^2 \to L^2} + \left\| M_\varepsilon - \widetilde{M}_\varepsilon \right\|_{L^2 \to L^2}. \]

To bound the first term, we define

\[ M_\varepsilon(x, x') = \mathbb{1}_{d_M(x, x') < r_\varepsilon} \psi^\circ(\zeta(x, x')). \]

Then it is a bounded symmetric kernel \( M \times M \to \mathbb{R} \), and following (B.28), \( M_\varepsilon \) is its associated
Fredholm integral operator. We can thus apply Lemma B.4.5 and get
\[
\|\Theta - M_\varepsilon\|_{L^2 \to L^2} \leq \sup_{x \in M} \int_{x' \in M} |\Theta(x, x') - M_\varepsilon(x, x')|dx'.
\]

Because the Near, Far and Winding pieces cover \( M \setminus L_{r, \varepsilon}(x) \), we have
\[
\int_{x' \in M \setminus L_{r, \varepsilon}(x)} |\Theta(x, x')|dx' \leq \int_{x' \in N_{r, \varepsilon}(x)} |\Theta(x, x')|dx' + \int_{x' \in W_{r, \varepsilon}(x)} |\Theta(x, x')|dx' + \int_{x' \in F_{r, \varepsilon}(x)} |\Theta(x, x')|dx'.
\]

From Lemma B.4.7, Lemma B.4.8, Lemma B.4.9 and Lemma B.4.10, we know that there exist constants \( C_2, C_3, C_4 \) and for any \( \varepsilon'' \leq 1 \) exist numbers \( C_{\varepsilon''}, C'_{\varepsilon''} \) such that when \( L \geq C_{\varepsilon''} \) and \( L \geq \left( \frac{\varepsilon^{-1/2} 12\pi \hat{k}}{\varepsilon} \right) \), we have
\[
\|\Theta - \overline{M_\varepsilon}\|_{L^2 \to L^2} \leq \sup_{x \in M} \int_{x' \in N_{r, \varepsilon}(x)} |\Theta(x, x')|dx' + \sup_{x \in M} \int_{x' \in W_{r, \varepsilon}(x)} |\Theta(x, x')|dx' + \sup_{x \in M} \int_{x' \in F_{r, \varepsilon}(x)} |\Theta(x, x')|dx' + \|M_\varepsilon - \overline{M_\varepsilon}\|_{L^2 \to L^2}
\]
\[
\leq \frac{3\pi n}{4(1-\varepsilon)} (1 + \varepsilon'') \log \left( \frac{\sqrt{\varepsilon}}{k r_\varepsilon} \right) + C'_{\varepsilon''} n
\]
\[
+ C_2 \text{len}(M) n \frac{\hat{k}}{\delta \sqrt{\varepsilon}}
\]
\[
+ C_3 \Theta_{\varepsilon, \delta}(M) n \log \left( 1 + \frac{1}{\Delta \varepsilon} \right)
\]
\[
+ C_4 (1-\varepsilon)^{-2} k^2 n r_\varepsilon^2.
\]

Meanwhile, from Lemma B.4.6 there exists constant \( C_\varepsilon, C_1 \) such that when \( L \geq C_\varepsilon \)
\[
(1-\varepsilon) \lambda_{\min}(P_{S_\varepsilon} \overline{M_\varepsilon} P_{S_\varepsilon}) \geq (1-\varepsilon)^2 \frac{3\pi n}{4} \log \left( 1 + \frac{L - 2}{3\pi r_\varepsilon} \right) - C_1 (1-\varepsilon) n r_\varepsilon \log^2 L.
\]
We will treat all named constants appearing in the previous two equations as fixed for the remainder of the proof. We argue that the first term in this expression is large enough to dominate each of the terms in (B.35) and the residual term in (B.36).

Set $\varepsilon' = \frac{\varepsilon}{24}$. We will choose $\varepsilon'' = \frac{\varepsilon'}{1 - 2\varepsilon'} < 1$, so that both $\varepsilon'$ and $\varepsilon''$ depend only on $\varepsilon$. Then, since $r_\varepsilon = 6\pi L^{-\frac{a_\varepsilon}{a_\varepsilon + 1}}$, when $L > 4$, we have

$$\frac{L - 2}{3\pi} r_\varepsilon = 2(L - 2)L^{-\frac{a_\varepsilon}{a_\varepsilon + 1}} > L^{\frac{1}{a_\varepsilon + 1}}. \tag{B.37}$$

Since moreover $a_\varepsilon = (1 - \varepsilon)^3(1 - \frac{\varepsilon}{12}) = (1 - \varepsilon)^3(1 - 2\varepsilon')$, we have $a_\varepsilon \varepsilon'' = \varepsilon'(1 - \varepsilon)^3$, and therefore $(1 + \varepsilon'')a_\varepsilon = (1 - \varepsilon')(1 - \varepsilon)^3$. Thus

$$(1 - \varepsilon')(1 - \varepsilon)^2 \frac{3\pi n}{4} \log \left(1 + \frac{L - 2}{3\pi} r_\varepsilon\right) = \frac{3\pi n}{4(1 - \varepsilon)} (1 + \varepsilon'') \log \left(1 + \frac{L - 2}{3\pi} r_\varepsilon\right) \geq \frac{3\pi n}{4(1 - \varepsilon)} (1 + \varepsilon'') \log \left(L^{a_\varepsilon} \right), \tag{B.38}$$

where in the last bound we use $\sqrt{\varepsilon \hat{k}^{-1}} \leq 6\pi$, given that $\varepsilon < 1$ and $\hat{k} \leq \pi/2$. The RHS at the end of this chain of inequalities is the first term of the RHS of the last bound in (B.35). Since the LHS has a leading coefficient of $(1 - \varepsilon')$, we can conclude provided we can split the remaining $\varepsilon'$ across the remaining terms.

Next, we will cover the negative term in (B.36) and the second and fifth terms in (B.35). Using (B.37), we have

$$\frac{\varepsilon'}{3} (1 - \varepsilon)^2 \frac{3\pi n}{4} \log \left(1 + \frac{L - 2}{3\pi} r_\varepsilon\right) \geq \frac{\varepsilon'}{3} (1 - \varepsilon)^2 \frac{3\pi n}{4} \frac{1}{a_\varepsilon + 1} \log(L). \tag{B.39}$$

There exists a constant $C_\varepsilon$ such that when $L \geq C_\varepsilon$, we have for the RHS

$$\frac{\varepsilon'}{3} (1 - \varepsilon)^2 \frac{3\pi n}{4} \frac{1}{a_\varepsilon + 1} \log(L) \geq (C_1 + C_4 + C'_{\varepsilon''}) n.$$
In particular, we can take
\[ C_\varepsilon \geq \exp \left( \frac{C_1 + C_4 + C_\varepsilon'}{\varepsilon' (1 - \varepsilon)^2 \frac{3\pi}{4} \frac{1}{a_\varepsilon + 1}} \right). \]

Next, there exists another constant \( C_\varepsilon > 0 \) such that when \( L \geq C_\varepsilon \), we have \( r_\varepsilon \log^2 L \leq 1 \), whence by the previous bound
\[ \varepsilon' \left( 1 - \varepsilon \right)^2 \frac{3\pi}{4} \frac{1}{a_\varepsilon + 1} \log(L) \geq (C_1 r_\varepsilon \log^2 L + C_4 + C_\varepsilon')n. \]

Finally, notice that when \( L \geq \left( \varepsilon^{-1/2} 12\pi \hat{k} \right)^{\frac{a_\varepsilon + 1}{a_\varepsilon}} \), we have
\[ r_\varepsilon = 6\pi L^{-\frac{a_\varepsilon}{a_\varepsilon + 1}} \leq \frac{\sqrt{\varepsilon}}{2\hat{k}}, \]
so \( r_\varepsilon \hat{k} \leq \sqrt{\varepsilon}/2 \), and since \( \varepsilon \in (0, 3/4) \), we have
\[ (1 - \varepsilon)C_1nr_\varepsilon \log^2 L + C_4(1 - \varepsilon)^{-2} k r_\varepsilon^2 + C_\varepsilon'n \leq \left( C_1 r_\varepsilon \log^2 L + 3C_4 + C_\varepsilon' \right)n, \]
where we used that \( \varepsilon \mapsto \varepsilon (1 - \varepsilon)^{-2} \) is increasing. Combining our previous bounds, this gives
\[ \varepsilon' \left( 1 - \varepsilon \right)^2 \frac{3\pi}{4} \log \left( 1 + \frac{L - 2}{3\pi} r_\varepsilon \right) \geq (1 - \varepsilon)C_1nr_\varepsilon \log^2 L + C_4(1 - \varepsilon)^{-2} k r_\varepsilon^2 + C_\varepsilon'n, \quad \text{(B.40)} \]
as desired.

For the remaining two terms, define
\[ C'_{\varepsilon,\delta} = \frac{(a_\varepsilon + 1)C_2}{(1 - \varepsilon)^2 \varepsilon' \frac{3\pi}{4} \delta \sqrt{\varepsilon}}, \quad C''_{\varepsilon} = \frac{(a_\varepsilon + 1)C_3}{(1 - \varepsilon)^2 \varepsilon' \frac{3\pi}{4}}. \]

We will use the estimate (B.39) as our base. Then when
\[ L \geq \max \left\{ \exp \left( C'_{\varepsilon,\delta} \operatorname{len}(\mathcal{M}) \hat{k} \right), \left( 1 + \frac{1}{\Delta_\varepsilon \sqrt{1 + k^2}} \right)^{C''_{\varepsilon,\delta}(\mathcal{M})} \right\}, \]

428
we have
\[
\frac{\varepsilon'}{3}(1 - \varepsilon)^2 \frac{3\pi n}{4} \log \left(1 + \frac{L - 2}{3\pi r_\varepsilon}\right) \geq C_2 \text{len}(\mathcal{M}) n \frac{\hat{k}}{\delta \sqrt{\varepsilon}}. \tag{B.41}
\]
\[
\frac{\varepsilon'}{3}(1 - \varepsilon)^2 \frac{3\pi n}{4} \log \left(1 + \frac{L - 2}{3\pi r_\varepsilon}\right) \geq C_3 \text{len}(\mathcal{M}) n \log \left(1 + \frac{1}{\Delta_\varepsilon \sqrt{1 + \kappa^2}}\right). \tag{B.42}
\]
Combining (B.38), (B.40), (B.41), (B.42) completes the proof. \hfill \square

**Lemma B.4.3.** For any \( x \in \mathcal{M} \), we have
\[
\langle x, \dot{x} \rangle = \langle \ddot{x}, \dddot{x} \rangle = \langle x, x^{(3)} \rangle = 0, \tag{B.43}
\]
\[
\langle x, \dddot{x} \rangle = -1, \tag{B.44}
\]
\[
\langle x, x^{(4)} \rangle = -\langle \dddot{x}, x^{(3)} \rangle = \|\dddot{x}\|^2_2,
\]
\[
\langle \dddot{x}, x^{(3)} \rangle = -\frac{1}{3} \langle \ddot{x}, x^{(4)} \rangle,
\]
\[
\|P_{x^\perp} \dddot{x}\|^2_2 = \|\dddot{x}\|^2_2 - 1,
\]
\[
M_2 = \sqrt{1 + \kappa^2} \leq M_4, \tag{B.45}
\]
\[
M_2 < 2\hat{k}, \tag{B.46}
\]
\[
\frac{1}{\hat{k}} \leq \min\{\text{len}(\mathcal{M}_-), \text{len}(\mathcal{M}_+)\}, \tag{B.47}
\]
where we use above the notation introduced near (B.8).

**Proof.** As our curve is defined over sphere and has unit speed, we have
\[
\|x\|^2_2 = \|\dot{x}\|^2_2 = 1.
\]
Taking derivatives on both sides, we get
\[
\langle x, \dot{x} \rangle = \langle \ddot{x}, \dddot{x} \rangle = 0.
\]
Continuing to take higher derivatives, we get the following relationships:

\[
\|\dot{x}\|^2 + \langle x, \dot{x} \rangle = 0, \\
3 \langle x, \dot{x} \rangle + \langle x, x^{(3)} \rangle = 0, \\
3\|\ddot{x}\|^2 + 4 \langle \dot{x}, x^{(3)} \rangle + \langle x, x^{(4)} \rangle = 0, \\
\|\ddot{x}\|^2 + \langle \ddot{x}, x^{(3)} \rangle = 0, \\
3 \langle \dddot{x}, x^{(3)} \rangle + \langle \dddot{x}, x^{(4)} \rangle = 0.
\]

which gives us by plugging in the previous constraints

\[
\langle x, \dddot{x} \rangle = -1, \\
\langle x, x^{(3)} \rangle = 0, \\
\langle \dot{x}, x^{(3)} \rangle = -\|\dddot{x}\|^2, \\
\langle x, x^{(4)} \rangle = \|\dddot{x}\|^2, \\
\langle \dddot{x}, x^{(3)} \rangle = -\frac{1}{3} \langle \dddot{x}, x^{(4)} \rangle.
\]

As a consequence, the intrinsic curvature \(\|P_{x+\dddot{x}}\|_2\) and extrinsic curvature \(\|\dddot{x}\|_2\) are related by

\[
\|P_{x+\dddot{x}}\|_2^2 = \left\| \left( I - xx^* \right) \dddot{x} \right\|_2^2 \\
= \langle x, \dddot{x} \rangle^2 + \langle \dddot{x}, \dddot{x} \rangle - 2 \langle x, \dddot{x} \rangle^2 \\
= \|\dddot{x}\|^2 - 1.
\]

Thus we know

\[
M_2 = \sup_{x \in M} \|\dddot{x}\|_2 \\
= \sup_{x \in M} \sqrt{1 + \|P_{x+\dddot{x}}\|_2^2}
\]

430
\[
\sqrt{1 + \sup_{x \in M} \|P_x \cdot \bar{x}\|_2^2} \\
= \sqrt{1 + \kappa^2} \\
\leq \sqrt{\left(\frac{\pi}{2}\right)^2 + \kappa^2} \\
< 2\hat{\kappa}.
\]

Furthermore, the above shows that \( M_2 \geq 1 \), so we have

\[
M_2 \leq M_2^2 = \sup_{x \in M} \|\bar{x}\|_2^2 \\
= \sup_{x \in M} \langle x, x^{(4)} \rangle \\
\leq M_4,
\]

using one of our previously-derived relationships in the second line and Cauchy-Schwarz in the third. Finally, for any point \( x = x_\sigma(s) \), as \( x_\sigma(s + \text{len}(M_\sigma)) = x_\sigma(s) \), we have

\[
0 = x_\sigma(s + \text{len}(M_\sigma)) - x_\sigma(s) = \int_{s' = s}^{s + \text{len}(M_\sigma)} \dot{x}_\sigma(s') ds' \\
= \text{len}(M_\sigma) \dot{x}_\sigma(s) + \int_{s' = s}^{s + \text{len}(M_\sigma)} \int_{s'' = s}^{s'} \dddot{x}_\sigma(s'') ds'' ds'
\]

which leads to

\[
\text{len}(M_\sigma) = \|\text{len}(M_\sigma) \dot{x}_\sigma(s)\|_2 \\
= \left\| \int_{s' = s}^{s + \text{len}(M_\sigma)} \int_{s'' = s}^{s'} \dddot{x}_\sigma(s'') ds'' ds' \right\|_2 \\
\leq \int_{s' = s}^{s + \text{len}(M_\sigma)} \int_{s'' = s}^{s'} M_2 ds'' ds' \\
= \frac{\text{len}(M_\sigma)^2}{2} M_2 < \text{len}(M_\sigma)^2 \hat{\kappa},
\]

completing the proof, where the first line uses the unit-speed property, the second uses the previous
relation, the third uses Jensen’s inequality (given that $\|\cdot\|_2$ is convex and 1-homogeneous), and the last line comes from (B.46).

Lemma B.4.4. Let $\hat{\kappa} = \max \{\kappa, \frac{2}{\pi}\}$. For $\sigma \in \{\pm\}$ and $|s' - s| \leq \frac{1}{\hat{\kappa}}$, we have

\[ |s - s'| - \hat{\kappa}^2 |s - s'|^3 \leq \angle(x_\sigma(s), x_\sigma(s')) \leq |s - s'|. \tag{B.48} \]

As a consequence, for $|s - s'| \leq \frac{\sqrt{\varepsilon}}{\hat{\kappa}}$,

\[ (1 - \varepsilon)|s - s'| \leq \angle(x_\sigma(s), x_\sigma(s')) \leq |s - s'|. \tag{B.49} \]

In particular, for any two points $x, x' \in \mathcal{M}_\sigma$, choosing $s, s'$ such that $x_\sigma(s) = x, x_\sigma(s') = x'$, and $|s - s'| = d_M(x, x')$, we have when $d_M(x, x') \leq \frac{\sqrt{\varepsilon}}{\hat{\kappa}}$,

\[ (1 - \varepsilon)d_M(x, x') \leq \angle(x, x') \leq d_M(x, x'). \]

Proof. We prove (B.48) first.

The upper bound is direct from the fact that $\mathcal{M}$ is a pair of paths in the sphere and $\angle(x, x')$ is the length of a path in the sphere of minimum distance between points $x, x'$, and then using the fact that the distance $|s' - s| \geq d_M(x_\sigma(s), x_\sigma(s'))$ from (B.9).

The lower bound requires some additional estimates. We fix $s, s'$ satisfying our assumptions; as both $|s - s'|$ and $\angle(x_\sigma(s), x_\sigma(s'))$ are symmetric functions of $(s, s')$, it suffices to assume that $s' \geq s$. Define $t = s' - s$, then by assumption we have $0 \leq t \leq \frac{1}{\hat{\kappa}} \leq \frac{\pi}{2}$. As $\cos^{-1}$ is strictly decreasing on $[-1, 1]$, we only need to show that

\[ \langle x_\sigma(s), x_\sigma(s + t) \rangle \leq \cos(t - \hat{\kappa}^2 t^3). \tag{B.50} \]
Using the second order Taylor expansion at \( s \), we have

\[
x_{\sigma}(s + t) = x_{\sigma}(s) + \dot{x}_{\sigma}(s) + \int_{a=s}^{s+t} \int_{b=s}^{b} \ddot{x}_{\sigma}(b) db \, da
\]

and so

\[
\langle x_{\sigma}(s), x_{\sigma}(s + t) \rangle = \left\langle x_{\sigma}(s), x_{\sigma}(s) + \dot{x}_{\sigma}(s) + \int_{a=s}^{s+t} \int_{b=s}^{b} \dot{x}_{\sigma}(b) db \, da \rightangle
\]

\[
= \|x_{\sigma}(s)\|^2_2 + \langle x_{\sigma}(s), \dot{x}_{\sigma}(s) \rangle + \left\langle x_{\sigma}(s), \int_{a=s}^{s+t} \int_{b=s}^{b} \ddot{x}_{\sigma}(b) db \, da \right\rangle
\]

\[
= 1 + \int_{a=s}^{s+t} \int_{b=s}^{b} \langle x_{\sigma}(s), \ddot{x}_{\sigma}(b) \rangle db \, da \quad (B.51)
\]

where we use properties established in Lemma B.4.3, in particular (B.43) in the last line. Take second order Taylor expansion at \( b \) for \( x_{\sigma}(s) \), we have similarly

\[
x_{\sigma}(s) = x_{\sigma}(b) + \dot{x}_{\sigma}(b) + \int_{c=s}^{b} \int_{d=c}^{b} \ddot{x}_{\sigma}(d) dd \, dc.
\]

From (B.43) and (B.44), we have \( \langle x_{\sigma}(b), \ddot{x}_{\sigma}(b) \rangle = -1 \) and \( \langle \dot{x}_{\sigma}(b), \ddot{x}_{\sigma}(b) \rangle = 0 \). Thus uniformly for \( b \in [s, s + t] \)

\[
\langle x_{\sigma}(s), \ddot{x}_{\sigma}(b) \rangle = -1 + \left( \int_{c=s}^{b} \int_{d=c}^{b} \ddot{x}_{\sigma}(d) dd \, dc, \ddot{x}_{\sigma}(b) \right)
\]

\[
= -1 + \int_{c=s}^{b} \int_{d=c}^{b} \langle \ddot{x}_{\sigma}(d), \ddot{x}_{\sigma}(b) \rangle dd \, dc
\]

\[
\leq -1 + \int_{c=s}^{b} \int_{d=c}^{b} \|\ddot{x}_{\sigma}(d)\|_2 \|\ddot{x}_{\sigma}(b)\|_2 dd \, dc
\]

\[
\leq -1 + \int_{c=s}^{b} \int_{d=c}^{b} M_2^2 dd \, dc
\]

\[
\leq -1 + \frac{M_2^2}{2} (b - s)^2,
\]
where in the third line we use Cauchy-Schwarz. Plugging this last bound into (B.51), it follows

\[ \langle x_{\sigma}(s), x_{\sigma}(s + t) \rangle \leq 1 + \int_{a=s}^{s+t} \int_{b=s}^{a} (-1 + (M_2^2/2)(b - s)^2) db \, da \]

\[ = 1 - \frac{t^2}{2} + \frac{M_2^2}{2} \int_{a=s}^{s+t} (b - s)^2 \, db \, da \]

\[ = 1 - \frac{t^2}{2} + \frac{M_2^2}{4} t^4 \]

\[ = 1 - \frac{t^2}{2} + \frac{1 + \kappa^2}{4} t^4, \]  \hfill (B.52)

with an application of Lemma B.4.3 in the final equality. To conclude, we derive a suitable estimate for \( \cos(t - \hat{\kappa}^2 t^3) \). Because \( 0 \leq t \leq \hat{\kappa}^{-1} \), we have that \( t^{-1}(t - \hat{\kappa}^2 t^3) \in [0, 1] \), and because \( t \leq \hat{\kappa}^{-1} \leq \pi/2 \), we can apply concavity of \( \cos \) on \( [0, \pi/2] \) to obtain

\[ \cos(t - \hat{\kappa}^2 t^3) \geq \frac{t - \hat{\kappa}^2 t^3}{t} \cos(t) + \left(1 - \frac{t - \hat{\kappa}^2 t^3}{t}\right) \cos(0). \]

Next, the estimate \( \cos(x) \geq 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \frac{x^6}{6!} \) for all \( x \), a consequence of Taylor expansion, gives

\[ (1 - \hat{\kappa}^2 t^2) \cos(t) + \hat{\kappa}^2 t^2 \geq (1 - \hat{\kappa}^2 t^2) \left(1 - \frac{t^2}{2} + \frac{t^4}{4!} - \frac{t^6}{6!}\right) + \hat{\kappa}^2 t^2 \]

\[ = 1 - \frac{t^2}{2} + \frac{t^4}{4!} + \frac{\hat{\kappa}^2 t^4}{2} - \frac{t^6}{6!} + \frac{\hat{\kappa}^2 t^6}{4!} + \frac{\hat{\kappa}^2 t^8}{6!} \]

after distributing. Because \( \hat{\kappa} \geq \kappa \), we can split terms and write

\[ \frac{t^4}{4!} + \hat{\kappa}^2 t^4 / 2 \geq \frac{1 + \kappa^2}{4!} t^4 + \hat{\kappa}^2 t^4 / 4, \]

and then grouping terms in the preceding estimates gives

\[ \cos(t - \hat{\kappa}^2 t^3) \geq 1 - \frac{t^2}{2} + \frac{1 + \kappa^2}{4!} t^4 + \hat{\kappa}^2 t^4 \left(\frac{1}{4} - \frac{t^2}{4!} + \frac{t^4}{6!} - \frac{t^2}{6!\hat{\kappa}^2}\right). \]
By way of (B.50) and (B.52), we will therefore be done if we can show that

\[
\frac{1}{4} - \left( \frac{1}{4!} + \frac{1}{6! \hat{k}^2} \right) t^2 + \frac{t^4}{6!} \geq 0.
\]

This is not hard to obtain: for example, we can prove the weaker but sufficient bound

\[
1 - \frac{1}{3!} \left( 1 + \frac{1}{30 \hat{k}^2} \right) t^2 \geq 0
\]

by noticing that because \( t \leq \hat{k}^{-1} \), it suffices to show

\[
\frac{1}{\hat{k}^2} \left( 1 + \frac{1}{30 \hat{k}^2} \right) \leq 6,
\]

and because the LHS of the previous line is an increasing function of \( \hat{k}^{-1} \) and moreover \( \hat{k}^{-1} \leq \pi/2 \), this bound follows by verifying that indeed \( (\pi/2)^2(1 + (1/30)(\pi/2)^2) \leq 6 \). Because \( s, s' \) were arbitrary we have thus proved (B.48).

For the remaining claims, (B.49) follows naturally from the fact that when \( |s - s'| \leq \frac{\sqrt{\varepsilon}}{\hat{k}} \), we have \( |s' - s| - \hat{k}^2 |s - s'|^3 \geq (1 - \varepsilon)|s - s'| \). The final claim is a restatement of (B.49) under the additional stated hypotheses.

\[\square\]

Invertibility of \( \overline{M} \) over \( S \).

**Lemma B.4.5** (Young’s inequality for Fredholm operators). Let \( K : M \times M \to \mathbb{R} \) satisfy \( K(x, x') = K(x', x) \) for all \( (x, x') \in M \times M \) and \( \sup_{(x, x') \in M \times M} |K(x, x')| < +\infty \), and let \( K \) denote its Fredholm integral operator (defined as \( g \mapsto K[g] = \int_M K(\cdot, x') g(x') dx' \)). For any \( 1 \leq p \leq +\infty \), we have

\[
\|K\|_{L^p \to L^p} \leq \sup_{x \in M} \int_{x' \in M} |K(x, x')| dx'.
\]

**Proof.** The proof uses the M. Riesz convexity theorem for interpolation of operators [234, §V, Theorem 1.3], which we need here in the form of a special case: it states that for all \( 1 \leq p \leq +\infty \),
one has

\[ \|K\|_{L^p \to L^p} \leq \|K\|_{L^\infty \to L^\infty}^{1/p} \|K\|_{L^1 \to L^1}^{1-1/p}. \]  

(B.53)

To proceed, we will bound the two operator norm terms on the RHS. We have

\[ \|K\|_{L^1 \to L^1} = \sup_{\|g\|_{L^1} = 1} \int_{x \in M} \left| \int_{x' \in M} K(x, x') g(x') dx' \right| dx \]

\[ \leq \sup_{\|g\|_{L^1} = 1} \int_{x \in M} \int_{x' \in M} |K(x, x')||g(x')| dx' dx \]

\[ = \sup_{\|g\|_{L^1} = 1} \left( \int_{x' \in M} \left( \int_{x \in M} |K(x, x')| dx \right) \right) |g(x')| dx' \]

\[ \leq \sup_{\|g\|_{L^1} = 1} \left( \sup_{x' \in M} \left( \int_{x \in M} |K(x, x')| dx \right) \right) \]

\[ = \sup_{x \in M} \int_{x' \in M} |K(x, x')| dx' \]  

(B.54)

The first inequality above uses the triangle inequality for the integral. In the third line, we rearrange the order of integration using Fubini’s theorem, given that \( g \) is integrable and \( K \) is bounded on \( \mathcal{M} \times \mathcal{M} \). In the fourth line, we use \( L^1-L^\infty \) control of the integrand (i.e., Hölder’s inequality), and in the final line we use that \( \|g\|_{L^1} = 1 \) along with symmetry of \( K \) and nonnegativity of the integrand to re-index and remove the outer absolute value. On the other hand, \( L^1-L^\infty \) control and the triangle inequality give immediately

\[ \|K\|_{L^\infty \to L^\infty} = \sup_{x \in M, \|g\|_{L^\infty} = 1} \left| \int_{x' \in M} K(x, x') g(x') dx' \right| \]

\[ \leq \sup_{x \in M} \int_{x' \in M} |K(x, x')| dx'. \]

These two bounds are equal; plugging them into (B.53) thus proves the claim. \( \square \)

**Lemma B.4.6.** Let \( \epsilon \in (0, \frac{3}{4}) \), \( r_\epsilon, S_\epsilon \) and \( \tilde{M}_\epsilon \) be as defined in (B.26), (B.14) and (B.29). Then \( \tilde{M}_\epsilon \) diagonalizes in the Fourier orthonormal basis (B.13). Write \( \lambda_{\min} \left( P_{S_\epsilon} \tilde{M}_\epsilon P_{S_\epsilon} \right) \) for the minimum eigenvalue of the operator \( P_{S_\epsilon} \tilde{M}_\epsilon P_{S_\epsilon} : S_\epsilon \to S_\epsilon \). Then there exist constants \( c, C \) and a constant
such that when \( L \geq C_{\varepsilon} \), we have

\[
\lambda_{\min} \left( P_{S,\varepsilon} \widetilde{M}_{\varepsilon} P_{S,\varepsilon} \right) \geq (1 - \varepsilon) \frac{3\pi n}{4} \log \left( 1 + \frac{L - 2}{\frac{3\pi}{L^2}} r_{\varepsilon} \right) - C n r_{\varepsilon} \log^2 L,
\]

\[
\geq cn \log L.
\]

As a consequence, \( P_{S,\varepsilon} \widetilde{M}_{\varepsilon} P_{S,\varepsilon} \) is invertible over \( S_{\varepsilon} \), and

\[
\left\| \left( P_{S,\varepsilon} \widetilde{M}_{\varepsilon} P_{S,\varepsilon} \right)^{-1} \right\|_{L^2 \to L^2} = \lambda_{\min}^{-1} \left( P_{S,\varepsilon} \widetilde{M}_{\varepsilon} P_{S,\varepsilon} \right).
\]

**Proof.** Choose \( L \gtrsim 1 \) to guarantee that \( \widetilde{M}_{\varepsilon} \) is well-defined. We use \( \psi^\circ \) to denote the DC subtracted skeleton, as defined in (B.6), and \((\phi_{\sigma,k}, \sigma, k)\) the (intrinsic) Fourier basis on \( M \), as defined in (B.13).

For any Fourier basis function \( \phi_{\sigma,k} \), we have

\[
\begin{align*}
\widetilde{M}_{\varepsilon} [\phi_{\sigma,k}] (x_{\sigma}(s)) &= \int_{s' = -r_{\varepsilon}}^{s + r_{\varepsilon}} \psi^\circ (|s - s'|) \phi_{\sigma,k} (x_{\sigma}(s')) \, ds' \\
&= \int_{s' = -r_{\varepsilon}}^{s' = r_{\varepsilon}} \psi^\circ (|s'|) \exp \left( \frac{i 2\pi k s'}{\text{len}(M_{\sigma})} \right) \, ds' \phi_{\sigma,k} (x_{\sigma}(s)) \\
&= \phi_{\sigma,k} (x_{\sigma}(s)) \int_{s' = -r_{\varepsilon}}^{r_{\varepsilon}} \psi^\circ (|s'|) \cos \left( \frac{2\pi k \text{len}(M_{\sigma})}{2\pi r_{\varepsilon}} \right) \, ds',
\end{align*}
\]

which shows that each Fourier basis function is an eigenfunction of \( \widetilde{M}_{\varepsilon} \); because these functions form an orthonormal basis for \( L^2(M) \) (by classical results from Fourier analysis on the circle), \( \widetilde{M}_{\varepsilon} \) diagonalizes in this basis. Moreover, because \( S_{\varepsilon} \) is the span of Fourier basis functions, \( P_{S,\varepsilon} \) also diagonalizes in this basis, and hence so does \( P_{S,\varepsilon} \widetilde{M}_{\varepsilon} P_{S,\varepsilon} \). Because \( \widetilde{M}_{\varepsilon} \) is self-adjoint and \( P_{S,\varepsilon} \) is an orthogonal projection, \( P_{S,\varepsilon} \widetilde{M}_{\varepsilon} P_{S,\varepsilon} \) is self-adjoint; and because \( \text{dim}(S_{\varepsilon}) < +\infty \), the operator \( P_{S,\varepsilon} \widetilde{M}_{\varepsilon} P_{S,\varepsilon} \) has finite rank, and therefore has a well-defined minimum eigenvalue, which we denote as in the statement of the lemma. As \( K_{\kappa, \sigma} = \lfloor \frac{\varepsilon^{1/2} \text{len}(M_{\sigma})}{2\pi r_{\varepsilon}} \rfloor \), we have for any \( |k_{\sigma}| \leq K_{\kappa, \sigma} \) and any \( |s'| \leq r_{\varepsilon} \),

\[
1 \geq \cos \left( \frac{2\pi k_{\sigma} s'}{\text{len}(M_{\sigma})} \right) \geq 1 - \left( \frac{2\pi k_{\sigma} s'}{\text{len}(M_{\sigma})} \right)^2 \geq 1 - \varepsilon.
\]
Then for $\sigma \in \{+, -\}$ and $|k| \leq K_{e, \pm},$

$$\widehat{M}_e[\phi_{\sigma, k}](x_{\sigma}(s)) = \phi_{\sigma, k}(x_{\sigma}(s)) \int_{s'=-r_e}^{r_e} \psi^{\circ}(|s'|) \cos \left( \frac{2\pi ks'}{\text{len}(M_{\sigma})} \right) ds'$$

$$\geq (1 - \varepsilon) \phi_{\sigma, k}(x_{\sigma}(s)) \int_{s'=-r_e}^{r_e} \psi^{\circ}(|s'|) ds',$$

and so

$$\lambda_{\min} \left( P_{S_e} \widehat{M}_e P_{S_e} \right) \geq 2(1 - \varepsilon) \int_{0}^{r_e} \psi^{\circ}(s) ds.$$

From Lemma B.5.7, we have if $L \geq 1$

$$2 \int_{s=0}^{r_e} \psi^{\circ}(s) ds \geq \frac{3\pi n}{4} \log \left( 1 + \frac{L - 2}{3\pi} r_e \right) - Cn r_e \log^2 L.$$

In particular, as $r_e = 6\pi L^{-\frac{a_e}{a_e + 1}},$ there exists a constant $C_e$ such that when $L \geq C'_e,$ we have

$$Cn r_e \log^2 L \leq \frac{\varepsilon}{4} \frac{3\pi n}{4} \log \left( 1 + \frac{L - 2}{3\pi} r_e \right),$$

and thus

$$\lambda_{\min} \left( P_{S_e} \widehat{M}_e P_{S_e} \right) \geq (1 - \varepsilon) \left( 1 - \frac{\varepsilon}{4} \right) \frac{3\pi n}{4} \log \left( 1 + \frac{L - 2}{3\pi} r_e \right)$$

$$\geq \left( 1 - \frac{5\varepsilon}{4} \right) \frac{3\pi n}{4} \log \left( L^{1 - \frac{a_e}{a_e + 1}} \right)$$

$$= \left( 1 - \frac{5\varepsilon}{4} \right) \frac{3\pi n}{4} \frac{1}{a_e + 1} \log L$$

$$\geq \left( 1 - \frac{5}{4} \cdot \frac{3}{4} \right) \frac{3\pi n}{4} \cdot \frac{1}{2} \log L$$

$$\geq cn \log L$$

$$> 0,$$

where we used $L \geq 1$ in the second inequality, and $\varepsilon < 3/4$ and $a_e \leq 1$ in the third inequality. So
\( P_{S_e} \tilde{M}_e P_{S_e} \) is invertible over \( S_e \), with

\[
\left( P_{S_e} \tilde{M}_e P_{S_e} \right)^{-1} [h] = \sum_{\sigma = \pm} \sum_{k = 0}^{K_{e,\sigma}} \left( \int_{s = -r_e}^{r_e} \psi^\circ(|s|) \cos \left( \frac{2\pi k s}{\text{len}(\mathcal{M}_{\sigma})} \right) ds \right)^{-1} \phi_{\sigma,k} \phi^\circ_{\sigma,k} h. \tag{B.55}
\]

The final claim is a consequence of the fact that \( P_{S_e} \tilde{M}_e P_{S_e} \) is self-adjoint and finite-rank.

\[ \square \]

**Lemma B.4.7.** Let \( \varepsilon \in (0, \frac{3}{4}) \), \( a_e, r_e, M_e \) and \( \tilde{M}_e \) be as defined in (B.25), (B.26), (B.28) and (B.29). There exist constants \( C, C' \), such that when \( L \geq C_\varepsilon \) and \( L \geq \left( e^{-1/2}12\pi \hat{k} \right)^{\frac{a_e+1}{a_e}} \), we have

\[
\left\| M_e - \tilde{M}_e \right\|_{L^2 \to L^2} \leq (1 - \varepsilon)^{-2} C' \kappa^2 n r_e^2.
\]

**Proof.** We choose \( L \geq 1 \) to guarantee that \( \tilde{M}_e \) is well-defined for all \( 0 < \varepsilon < 3/4 \). We would like to use Lemma B.4.5 to bound \( \| M_e - \tilde{M}_e \|_{L^2 \to L^2} \), and thus we define two (suggestively-named) bounded symmetric kernels \( \mathcal{M} \times \mathcal{M} \to \mathbb{R} \):

\[
M_e(x, x') = \mathbb{1}_{d_M(x, x') < r_e} \psi^\circ(\angle(x, x'))
\]

and

\[
\tilde{M}_e(x, x') = \mathbb{1}_{d_M(x, x') < r_e} \psi^\circ(d_M(x, x')).
\]

From (B.28), \( M_e \) is indeed \( M_e \)'s associated Fredholm integral operator. To show that under our constraints for \( L, \tilde{M}_e \) is also \( \tilde{M}_e \)'s associated integral operator, we first notice that following (B.9), for any \( x, x' \in \mathcal{M} \), \( d_M(x, x') < r_e \) if and only if there exist \( \sigma, s \) and \( s' \) such that \( x = x_{\sigma}(s) \), \( x' = x_{\sigma}(s') \) and \( |s' - s| < r_e \). This means for any fixed \( x \), if we let \( \sigma \) and \( s \) be chosen such that \( x = x_{\sigma}(s) \), then \( L_{r_e}(x) = \{ x_{\sigma}(s') | |s' - s| < r_e \} \). Furthermore, as \( L \geq \left( e^{-1/2}12\pi \hat{k} \right)^{\frac{a_e+1}{a_e}} \), by (B.47) in Lemma B.4.3 we have \( r_e \leq \frac{\varepsilon}{2\pi} < \min \{ \text{len}(\mathcal{M}_+), \text{len}(\mathcal{M}_-) \}/2 \). Under this condition, we can unambiguously express the intrinsic distance \( d_M \) in terms of arc length at the local scale: for any \( x' \in L_{r_e}(x) \), there is a unique \( s' \) such that \( |s' - s| \leq r_e \). To see this, note that for any other
parameter choice that attains the infimum in (B.9) \( s'' = s' + k \text{len}(\mathcal{M}_e) \) with integer \( k \neq 0 \), the triangle inequality implies \( |s'' - s| \geq |r_e - k \text{len}(\mathcal{M}_e)| \), and one has \( |r_e - k \text{len}(\mathcal{M}_e)| > r_e \) for every \( k \neq 0 \) if \( 0 < r_e < \text{len}(\mathcal{M}_e)/2 \). Then for \( x' \in L_{r_e}(x) \) and any \( s' \in [s - r_e, s + r_e] \) such that \( x_\sigma(s') = x' \), we have \( d_M(x, x') = |s - s'| \). Combining all these points, \( \mathcal{M}_e \)'s associated Fredholm integral operator \( H \) can be written as:

\[
H[f](x_\sigma(s)) = \int_{d_M(x_\sigma(s), x') < r_e} \psi^\circ (d_M(x_\sigma(s), x')) f(x') dx'
\]

\[
= \int_{x' \in \mathcal{M}_e(s') < s} \psi^\circ (d_M(x_\sigma(s), x')) f(x') dx'
\]

\[
= \int_{s' = s - r_e}^{s' = s + r_e} \psi^\circ (d_M(x_\sigma(s), x_\sigma(s'))) f(x_\sigma(s')) ds'
\]

\[
= \mathcal{M}_e[f](x_\sigma(s)),
\]

which means \( \mathcal{M}_e \) is indeed \( \mathcal{M}_e \)'s associated integral kernel.

We can now apply Lemma B.4.5 and and get

\[
\|M_e - \mathcal{M}_e\|_{L^2 \rightarrow L^2} \leq \sup_{x \in \mathcal{M}} \int_{x' \in \mathcal{M}} \|M_e(x, x') - \mathcal{M}_e(x, x')\| dx'
\]

\[
= \sup_{x \in \mathcal{M}} \int_{x' \in L_{r_e}(x)} \|M_e(x, x') - \mathcal{M}_e(x, x')\| dx'
\]

\[
= \sup_{s, \sigma} \int_{s' = s - r_e}^{s' = s + r_e} \left| \psi^\circ(\angle(x_\sigma(s), x_\sigma(s'))) - \psi^\circ(|s - s'|) \right| ds'. \quad (B.56)
\]

Here, we recall that \( r_e < \pi/4 \) (because \( \hat{k} \leq \pi/2 \)), so there is no issue with these evaluations and the domain of \( \psi^\circ \) being \([0, \pi]\). Note that from (B.48), when \( |s - s'| \leq r_e \leq \frac{\sqrt{\epsilon}}{\hat{k}} \), we have

\[
\angle(x_\sigma(s), x_\sigma(s')) \geq |s - s'| - \hat{k}^2 |s - s'|^3
\]

\[
\geq (1 - \epsilon)|s - s'|. \quad (B.57)
\]

As \( \psi^\circ \) is nonnegative, strictly decreasing and convex by Lemma B.6.3, we know both \( \psi^\circ \) and \( |\psi^\circ| \)
are decreasing. Also, by the upper bound in Lemma B.4.4, we have that

\[ \psi^\circ (\angle (x_{\sigma}(s), x_{\sigma}(s'))) - \psi^\circ (|s - s'|) \geq 0, \]

so we can essentially ignore the absolute value in the integrand in (B.56). We can then calculate

\[
\int_{s' = s - r_E}^{s + r_E} \psi^\circ (\angle (x_{\sigma}(s), x_{\sigma}(s'))) - \psi^\circ (|s - s'|) \, ds' \leq \int_{s' = s - r_E}^{s + r_E} \psi^\circ (|s - s'| - k^2|s - s'|^3) - \psi^\circ (|s - s'|) 
\]

\[
= \int_{t = -r_E}^{r_E} \psi^\circ (|t| - k^2|t|^3) - \psi^\circ (|t|) \, dt \]

\[
\leq k^2 \int_{t = -r_E}^{r_E} |t|^3 |\dot{\psi}^\circ (|t| - k^2|t|^3)| \, dt \]

\[
\leq k^2 \int_{t = -r_E}^{r_E} |t|^3 |\dot{\psi}^\circ ((1 - \varepsilon)|t|)| \, dt \]

\[
= 2(1 - \varepsilon)^{-4} k^2 \int_{t = 0}^{(1-\varepsilon)r_E} t^3 |\dot{\psi}^\circ (t)| \, dt.
\]

Above, the first line comes from (B.57) and the fact that \( \psi^\circ \) is strictly decreasing, the fourth and fifth line comes from the fact that \(|\dot{\psi}^\circ|\) is decreasing and (B.57). The last line uses symmetry and a linear transformation. Note that from (B.57) we always have \(|t| - k^2|t|^3\) nonnegative when \(|t| \leq r_E\) and thus all above formulas are well defined. From Lemma B.5.10, we know that there exists \( C, C' \) such that when \( L \geq C \), we have

\[
\int_{0}^{(1-\varepsilon)r_E} t^3 |\dot{\psi}^\circ (t)| \, dt \leq C' n (1 - \varepsilon)^2 r_E^2,
\]

and plugging all bounds back to (B.56) we get

\[
\left\| M_E - \overline{M}_E \right\|_{L^2 \rightarrow L^2} \leq (1 - \varepsilon)^{-2} C' k^2 n r_E^2
\]
as claimed. □

**Lemma B.4.8.** Let \( \varepsilon \in (0, \frac{1}{3}) \), \( r_\varepsilon \) and \( N_{r_\varepsilon, \varepsilon} \) as defined in (B.26) and (B.22). For any \( 0 < \varepsilon'' \leq 1 \), there exist numbers \( C_{\varepsilon''}, C'_{\varepsilon''} \) such that when \( L \geq C_{\varepsilon''} \), we have

\[
\sup_{x \in M} \int_{x' \in N_{r_\varepsilon, \varepsilon}(x)} \psi^\circ \left( \angle(x, x') \right) ds' \leq \frac{3\pi n}{4(1 - \varepsilon)} (1 + \varepsilon'') \log \left( \frac{\sqrt{\varepsilon}}{k r_\varepsilon} \right) + C'_{\varepsilon''} n.
\]

**Proof.** For \( x \in M \), assume the parameters are chosen such that the corresponding near piece is nonempty, for otherwise the claim is immediate. Recalling (B.22), for any \( x' \in N_{r_\varepsilon, \varepsilon}(x) \), we have \( d_M(x, x') \leq \sqrt{\varepsilon}/k \). From Lemma B.4.4, this implies \( \angle(x, x') \geq (1 - \varepsilon) d_M(x, x') \). Let \( \sigma, s \) be such that \( x_\sigma(s) = x \). Notice by the discussion following the definition of the intrinsic distance in (B.9) that the near component \( N_{r_\varepsilon, \varepsilon}(x) \) is contained in the set \( \{ x_{\sigma}(s') \mid |s' - s| \in [r_\varepsilon, \sqrt{\varepsilon}/k] \} \). And from Lemma B.6.3, \( \psi^\circ \) is strictly decreasing, thus we have

\[
\int_{x' \in N_{r_\varepsilon, \varepsilon}(x)} \psi^\circ \left( \angle(x, x') \right) ds' \leq \int_{s' = r_\varepsilon}^{s' = s + r_\varepsilon} \psi^\circ \left( \angle(x_\sigma(s), x_{\sigma}(s')) \right) ds'
\]

\[
+ \int_{s' = s - \sqrt{\varepsilon}/k}^{s' = s - r_\varepsilon} \psi^\circ \left( \angle(x_\sigma(s), x_{\sigma}(s')) \right) ds'
\]

\[
\leq \int_{s' = r_\varepsilon}^{s' = s + r_\varepsilon} \psi^\circ ((1 - \varepsilon)|s' - s|) ds'
\]

\[
+ \int_{s' = s - \sqrt{\varepsilon}/k}^{s' = s - r_\varepsilon} \psi^\circ ((1 - \varepsilon)|s' - s|) ds'
\]

\[
= 2 \int_{t = r_\varepsilon}^{\frac{\sqrt{\varepsilon}}{k}} \psi^\circ ((1 - \varepsilon)t) dt
\]

\[
= \frac{2}{1 - \varepsilon} \int_{t = (1 - \varepsilon)r_\varepsilon}^{\frac{\sqrt{\varepsilon}}{k}} \psi^\circ (t) dt,
\]

where in the last line we apply a linear change of variables. We also note that in the above integrals \( |s' - s| \leq k^{-1} \leq \pi/2 \), so there are no issues above with the domain of \( \psi^\circ \) being \([0, \pi]\). From Lemma B.5.9, for any \( 0 < \varepsilon'' \leq 1 \), there exist numbers \( C_{\varepsilon''}, C'_{\varepsilon''} \) such that if \( L \geq C_{\varepsilon''} \), then \( r_\varepsilon \)
satisfies the condition in (B.74) and we have

\[
\sup_{x \in M} \int_{x' \in N_{r,\varepsilon}(x)} \psi^\circ(\angle(x, x')) ds' \leq \frac{2}{1 - \varepsilon} \int_{r' = (1 - \varepsilon)r_{\varepsilon}} \psi^\circ(|t|) dt \\
\leq \frac{2}{1 - \varepsilon} (1 + \varepsilon'') \frac{3\pi n}{8} \log \left( \frac{1 + (L - 3)(1 - \varepsilon)\sqrt{\varepsilon}/(3\pi)}{1 + (L - 3)(1 - \varepsilon)r_{\varepsilon}/(3\pi)} \right) \\
+ C_{\varepsilon''} n \\
\leq \frac{3\pi n}{4(1 - \varepsilon)(1 + \varepsilon'')} \log \left( \frac{\sqrt{\varepsilon}}{kr_{\varepsilon}} \right) + C_{\varepsilon''} n.
\]

\[\square\]

**Lemma B.4.9.** Let \( \varepsilon \in (0, \frac{3}{4}) \), \( \delta \in (0, 1 - \varepsilon] \). Let \( W_{\varepsilon,\delta} \) as in (B.24). There exist constants \( C, C' \) such that when \( L \geq C \), for any \( x \in M \),

\[
\int_{x' \in W_{\varepsilon,\delta}(x)} \psi^\circ(\angle(x, x')) ds' \leq \bigotimes_{\varepsilon,\delta}(M) C' n \log \left( 1 + \frac{1 + k^2}{\Delta_{\varepsilon}} \right).
\]

**Proof.** To bound the integral, we rely on the observation that for each ‘curve segment’ inside the winding component, the angle \( \angle(x, x') \) cannot stay small for the whole segment, and thus we can avoid worst case control for the angle as we have employed for the far component in Lemma B.4.10.\(^6\) We will begin by constructing a specific finite cover of curve segments for the winding component, then we will bound the integral over each curve segment by providing a lower bound for the angle function.

As \( M \) is compact with bounded length, from the definition in (B.12) we know \( \bigotimes_{\varepsilon,\delta}(M) \) is a finite number for any choice of \( \varepsilon, \delta \). From the definition of the winding component (B.24), for any point \( x \in M \), we can cover \( W_{\varepsilon,\delta}(x) \) by at most \( \bigotimes_{\varepsilon,\delta}(M) \) closed balls in the intrinsic distance on the manifold with radii no larger than \( 1/\sqrt{1 + k^2} \). Topologically, each ball in the intrinsic distance of radii \( r \) is a curve segment of length \( 2r \); thus, \( W_{\varepsilon,\delta}(x) \) can be covered by at most \( 2\bigotimes_{\varepsilon,\delta}(M) \) curve segments, each with length no larger than \( 1/\sqrt{1 + k^2} \). Formally, this implies that for each

\(^6\)Within the lemma, a curve segment means \( \{x_{\sigma}(s)|s \in [s_1, s_2]\} \subseteq M_{\sigma} \) for certain \( \sigma, s_1 \) and \( s_2 \) with \( |s_1 - s_2| < \text{len}(M_{\sigma}) \), and we call \( |s_1 - s_2| \) the length of the curve segment.
\( x \in M \), there exists a number \( N(x) \leq 2^{\Theta_x(M)} \) and for each \( i \in \{1, \cdots, N(x)\} \), there exist a sign \( \sigma_i(x) \in \{\pm\} \) and a nonempty interval \( I_i(x) = [s_{1,i}(x), s_{2,i}(x)] \) with length no greater than \( \frac{1}{\sqrt{1+x^2}} \) and strictly less than \( \operatorname{len}(\mathcal{M}_{\sigma_i(x)}) \) such that

\[
W_{\epsilon, \delta}(x) \subseteq \bigcup_{i=1}^{N(x)} X_i(x)
\]

where \( X_i(x) = \{x_{\sigma_i(x)}(s) \mid s \in I_i(x)\} \subset M \) with \( X_i(x) \cap W_{\epsilon, \delta}(x) \neq \emptyset \). For the purpose of minimum coverage, we can further assume without loss of generality that for each \( x \) and each \( i \), the boundary points \( x_{\sigma_i(x)}(s_{1,i}(x)) \) and \( x_{\sigma_i(x)}(s_{2,i}(x)) \) belong to \( W_{\epsilon, \delta}(x) \): we can always set \( p_{1,i}(x) = \inf \{s \mid s \in [s_{1,i}(x), s_{2,i}(x)], x_{\sigma_i(x)}(s) \in W_{\epsilon, \delta}(x)\} \) and \( p_{2,i}(x) = \sup \{s \mid s \in [s_{1,i}(x), s_{2,i}(x)], x_{\sigma_i(x)}(s) \in W_{\epsilon, \delta}(x)\} \), then the curve segment associated with \( \sigma_i(x) \) and interval \([p_{1,i}(x), p_{2,i}(x)]\) still covers \( X_i(x) \cap W_{\epsilon, \delta}(x) \). As \( W_{\epsilon, \delta}(x) \) is closed, we have the boundary points \( x_{\sigma_i(x)}(p_{1,i}(x)), x_{\sigma_i(x)}(p_{2,i}(x)) \in W_{\epsilon, \delta}(x) \) and as \( X_i(x) \) intersect with \( W_{\epsilon, \delta}(x) \), the definition above is well defined.

We will next increase the number of sets in these coverings, so that they are guaranteed not to fall into any of the “local pieces” at \( x \): although by the definitions (B.21) and (B.24) the local and winding pieces at any \( x \) are disjoint, it may be the case that when we pass to the covering sets \( (X_i(x))_{i \in [N(x)]} \), we overlap with the local piece. In particular, consider a “local piece” \( L_{\sqrt{\nu/k}}(x) \) defined as in (B.21), which from the definition does not intersect with \( W_{\epsilon, \delta}(x) \). For each \( i \), as the boundary points of \( X_i(x) \) fall in \( W_{\epsilon, \delta}(x) \), these boundary points do not belong to \( L_{\sqrt{\nu/k}}(x) \). And as \( L_{\sqrt{\nu/k}}(x) \) is topologically connected and one dimensional, if \( X_i(x) \) intersects with \( L_{\sqrt{\nu/k}}(x) \), it must contains the whole local piece. As \( X_i(x) \) itself is a curve segment and one dimensional, and \( L_{\sqrt{\nu/k}}(x) \) is open, removing \( L_{\sqrt{\nu/k}}(x) \) would leave two curve segments with smaller length. Then these two curve segments lie in \( M \setminus L_{\sqrt{\nu/k}}(x) \), and cover \( X_i(x) \setminus L_{\sqrt{\nu/k}}(x) \). In other words, for any \( x \in M \), there exists \( N'(x) \leq 4^{\Theta_x(M)} \) and for \( i \in \{1, \cdots, N'(x)\} \), there exist signs \( \sigma'_i(x) \in \{\pm\} \)

444
and intervals $I'_i(x) = [s'_{1,i}(x), s'_{2,i}(x)]$ with length no greater than $\frac{1}{\sqrt{1+x^2}}$ such that

$$W_{\varepsilon, \delta}(x) \subseteq \bigcup_{i=1}^{N'(x)} X'_i(x),$$

where $X'_i(x) = \{x_{\sigma'_i}(x) | s \in I'_i(x)\} \subset M \setminus L_{\sqrt{\varepsilon}/k}(x)$ with $X'_i(x) \cap W_{\varepsilon, \delta}(x) \neq \emptyset$. We therefore have

$$\int_{x' \in W_{\varepsilon, \delta}(x)} \psi^\circ \left( \angle(x, x') \right) ds' \leq \sum_{i=1}^{N'(x)} \int_{s \in I'_i(x)} \psi^\circ \left( \angle(x, x_{\sigma'_i}(x)) \right) ds. \quad (B.58)$$

We next derive additional properties of the pieces $X'_i(x)$ that will allow us to obtain suitable estimates for the integrals on the RHS of (B.58). As each $X'_i(x)$ is a compact set, we let

$$s^*_i(x) \in \arg \min_{s \in I'_i(x)} \angle(x, x_{\sigma'_i}(x))$$

and denote $x^*_i(x) = x_{\sigma'_i}(x)(s^*_i(x))$. Below we will abbreviate $x^*_i(x), s^*_i(x)$ and $\sigma'_i(x)$ as $x^*_i, s^*_i$ and $\sigma'_i$ when the base point $x$ is clear. We further abbreviate $\dot{x}^*_i = \dot{x}_{\sigma'_i}(x)(s^*_i(x))$. As $X'_i(x)$ intersects with the winding component, we have $\angle(x, x^*_i) \leq \frac{\delta \sqrt{\varepsilon}}{k} < \frac{\pi}{2}$. And as $X'_i(x) \cap L_{\sqrt{\varepsilon}/k}(x) = \emptyset$, we have $d_M(x, x^*_i) \geq \sqrt{\varepsilon}/\hat{k}$. This means $x^*_i \in W_{\varepsilon, \delta}(x)$ from (B.24). As $\cos$ is strictly decreasing from 0 to $\pi$ and $s^*_i$ minimizes $\angle(x, x_{\sigma'_i}(s))$, it also maximizes $\langle x, x_{\sigma'_i}(s) \rangle$. For any $s \in I'_i(x)$, from the second order Taylor expansion of $x_{\sigma'_i}(s)$ around $x^*_i$ we have

$$\langle x, x^*_i \rangle \geq \langle x, x_{\sigma'_i}(s) \rangle$$

$$= \langle x, x^*_i \rangle + (s - s^*_i) \langle x, \dot{x}^*_i \rangle + \left\langle x, \int_{a=s^*_i}^{a} \int_{b=s^*_i}^{a} x^{(2)}_{\sigma'_i}(b) \, db \, da \right\rangle$$

$$\geq \langle x, x^*_i \rangle + (s - s^*_i) \langle x, \dot{x}^*_i \rangle - \frac{(s - s^*_i)^2}{2} M_2,$$

with the last line following from Cauchy-Schwarz. In the previous equations, we are of course using the convention that for a real-valued function $f$ and numbers $a < b$, the notation $\int_b^a f(x) \, dx$ denotes the integral $-\int_a^b f(x) \, dx$. We are going to use this bound to reprove a classical first-order
optimality condition for interval-constrained problems. We split into cases depending on where
the point \( s^*_i \) lies: if \( s^*_i \) is not the right end point \( s'_{2,i} \), by taking \( s \) approaching \( s^*_i \) from above, we
would have \( \langle x, \dot{x}_i^* \rangle \leq 0 \). Similarly, if \( s^*_i \) is not the left end point \( s'_{1,i} \), by taking \( s \) approaching \( s^*_i \)
from below, we would have \( \langle x, \dot{x}_i^* \rangle \geq 0 \). This gives

\[
\langle x, \dot{x}_i^* \rangle = \begin{cases} 
  \leq 0 & s^*_i = s_{2,i} \\
  \geq 0 & s^*_i = s_{1,i} \\
  = 0 & \text{otherwise}
\end{cases}
\]

which implies

\[
(s - s^*_i) \langle x, \dot{x}_i^* \rangle \leq 0, \quad \forall s \in I'_i(x). \tag{B.59}
\]

We use again the Taylor expansion at \( s^*_i \) and get

\[
\left\| x_{\sigma_i}(s) - x_i^* - (s - s^*_i) \dot{x}_i^* \right\|_2 \leq \left\| \int_{a=s^*_i}^{s} \int_{b=s^*_i}^{a} x^{(2)}_{\sigma_i}(b) \, db \, da \right\|_2 \\
\leq \frac{(s - s^*_i)^2}{2} M_2 \\
= \frac{1}{2} (1 + \kappa^2)^{1/2} (s - s^*_i)^2 \tag{B.60}
\]

with an application of (B.45) in the last line. Moreover, we have

\[
\|x - x_i^*\|_2 = 2 \sin \left( \frac{\angle(x, x_i^*)}{2} \right) \\
\geq \frac{4}{\pi} \sin \left( \frac{\pi}{4} \right) \angle(x, x_i^*) \\
= \frac{2\sqrt{2}}{\pi} \angle(x, x_i^*) \\
\geq \frac{2\sqrt{2}}{\pi} \Delta_x, \tag{B.61}
\]
where the first line is a trigonometric identity, the first inequality uses \( \angle(x, x^*_i) < \pi/2 \) together with the fact that \( \sin \) function is concave from 0 to \( \pi \) and thus \( \sin(at) \geq a \sin(t) \) for \( a \in [0, 1] \) and \( t \in [0, \pi] \) (applied to \( a = \angle(x, x^*_i)/(\pi/2) \) and \( t = \pi/4 \)), and the last line follows directly from the definition of \( \Delta_\varepsilon \) in (B.10). Making use of the preceding estimates, for any \( s \in I'_i(x) \) we can finally calculate

\[
\|x_{\sigma'_i}(s) - x\|^2 = \|x^*_i - x + (s - s^*_i)\hat{x}^*_i + (x_{\sigma'_i}(s) - x^*_i - (s - s^*_i)\hat{x}^*_i)\|^2 \\
\geq \|x^*_i - x + (s - s^*_i)\hat{x}^*_i\|^2 - \|x_{\sigma'_i}(s) - x^*_i - (s - s^*_i)\hat{x}^*_i\|^2 \\
\geq \|x^*_i - x\|^2 + \|(s - s^*_i)\hat{x}^*_i\|^2 - 2 \langle x, (s - s^*_i)\hat{x}^*_i \rangle \\
- \left( \frac{1}{2} (1 + \kappa^2)^{1/2} (s - s^*_i)^2 \right)^2 \\
\geq \left( \frac{2\sqrt{2}}{\pi \Delta_\varepsilon} \right)^2 + (s - s^*_i)^2 - \frac{1}{4} (1 + \kappa^2) (s - s^*_i)^4 \\
\geq \left( \frac{2\sqrt{2}}{\pi \Delta_\varepsilon} \right)^2 + \frac{3}{4} (s - s^*_i)^2 \\
\geq \left( \frac{2\Delta_\varepsilon}{\pi} + \frac{\sqrt{3}}{2\sqrt{2}} |s - s^*_i| \right)^2. \tag{B.62}
\]

Above, the second line uses the triangle inequality, the third line uses the parallelogram identity plus Lemma B.4.3 (first term) and (B.60) (second term), the fourth line comes from (B.61) and (B.59), and the fifth line comes from our construction that the length of each interval \( I'_i(x) \) is no greater than \( 1/\sqrt{1 + \kappa^2} \) and therefore the same is true of \( |s - s^*_i| \). The last line is an application of inequality of arithmetic and geometric means. Additionally, for any \( x, x' \) of unit norm, one has

\[
\angle(x, x') \geq 2 \sin \left( \frac{\angle(x, x')}{2} \right) \\
= \|x - x'\|_2.
\]
Combining this and (B.62), for all \( s \in I'_i(x) \) we have

\[
\angle(x_{\sigma'_i}(s), x) \geq \|x_{\sigma'_i}(s) - x\|_2 \geq \frac{2}{\pi} \Delta_e + \frac{\sqrt{3}}{2\sqrt{2}} |s - s'_i| \\
\geq \frac{1}{\sqrt{3}} \Delta_e + \frac{1}{\sqrt{3}} |s - s'_i|,
\]

where the last line just worst-cases constants for simplicity. From Lemma B.6.3, \( \varphi^{o} \) is nonnegative and strictly decreasing, so

\[
\int_{s \in I'_i(x)} \varphi^{o}(\angle(x, x_{\sigma'_i}(s))) \, ds = \int_{s = s'_i}^{s'_{2,i}(x)} \varphi^{o}(\angle(x, x_{\sigma'_i}(s))) \, ds + \int_{s = s'_{1,i}(x)}^{s'_i} \varphi^{o}(\angle(x, x_{\sigma'_i}(s))) \, ds \\
\leq \int_{s = s'_i}^{s'_{2,i}(x)} \varphi^{o}\left(\frac{1}{\sqrt{3}} \Delta_e + \frac{1}{\sqrt{3}} |s - s'_i|\right) \, ds \\
+ \int_{s = s'_{1,i}(x)}^{s'_i} \varphi^{o}\left(\frac{1}{\sqrt{3}} \Delta_e + \frac{1}{\sqrt{3}} s\right) \, ds \\
\leq 2 \int_{s = 0}^{1/\sqrt{1 + \kappa^2}} \varphi^{o}\left(\frac{1}{\sqrt{3}} \Delta_e + \frac{1}{\sqrt{3}} s\right) \, ds \\
= 2\sqrt{3} \int_{t = 1/\sqrt{3} \Delta_e}^{1/\sqrt{3} \Delta_e} \varphi^{o}(t) \, dt
\]

where again, the second to third line comes from the fact that our intervals has length at most \( 1/\sqrt{1 + \kappa^2} \). From (B.73) in Lemma B.5.9 and a summation over all \( N'(x) \leq 4 \Theta_{\infty, \delta}(M) \) segments in the covering, there exists constant \( C' \) such that when \( L \geq C' \),

\[
\sum_{i=1}^{N'(x)} \int_{s \in I'_i(x)} \varphi^{o}(\angle(x, x_{\sigma'_i}(s))) \, ds \leq \Theta_{\infty, \delta}(M) C' n \log\left(\frac{1 + (L - 3) \left(\frac{1}{\sqrt{3}} \Delta_e + \frac{1}{\sqrt{3} \sqrt{1 + \kappa^2}}\right) / (3\pi)}{1 + (L - 3) \frac{1}{\sqrt{3}} \Delta_e / (3\pi)}\right) \\
\leq \Theta_{\infty, \delta}(M) C' n \log\left(1 + \frac{1}{\Delta_e} \frac{\sqrt{1 + \kappa^2}}{\Delta_e}\right).
\]

Recalling our bound (B.58), we can thus take a supremum over \( x \in M \) and conclude. \( \square \)

**Lemma B.4.10.** Let \( \varepsilon \in (0, 1) \), \( \delta \in (0, 1 - \varepsilon] \). Let \( F_{e, \delta} \) as in (B.23). There exist constants \( C, C' \)
such that when \( L \geq C \), we have for any \( x \in M \),

\[
\int_{x' \in F_{e,\delta}(x)} \psi^\circ \left( \angle(x,x') \right) ds' \leq C' \text{len}(M)n \frac{\delta}{\delta \sqrt{\varepsilon}}.
\]

**Proof.** We have the simple bound from Lemma B.5.8 and decreasingness of \( \psi^\circ \) from Lemma B.6.3, that there exists constant \( C' \), with

\[
\int_{x' \in F_{e,\delta}(x)} \psi^\circ \left( \angle(x,x') \right) ds' \leq \text{len}(M)\psi^\circ \left( \frac{\delta \sqrt{\varepsilon}}{\delta} \right)
\]

\[
\leq \text{len}(M)C'n \frac{L-3}{1 + (L-3)\frac{\delta \sqrt{\varepsilon}}{\delta} / (3\pi)}
\]

\[
\leq \text{len}(M)C'n \frac{\delta}{\delta \sqrt{\varepsilon}},
\]

as claimed. \( \square \)

### B.5 Bounds for the Skeleton Function \( \psi \)

In this section, we are going to provide sharp bounds on the “skeleton” function \( \psi \) and its higher-order derivatives. We recall that the angle evolution function is defined as

\[
\varphi(t) = \arccos \left( \left(1 - \frac{t}{\pi} \right) \cos t + \frac{1}{\pi} \sin t \right), \quad t \in [0, \pi].
\]

Define \( \varphi^{[0]} = \text{Id}, \varphi^{[\ell]} \) as \( \varphi \)'s \( \ell \)-fold composition with itself (which will be referred to as the iterated angle evolution function). Then the skeleton is defined as

\[
\psi(t) = \frac{n}{2} \sum_{\ell=0}^{L-1} \xi_{\ell}(t),
\]

where

\[
\xi_{\ell}(t) = \prod_{\ell'=\ell}^{L-1} \left(1 - \frac{1}{\pi} \varphi^{[\ell']} \left(t \right) \right), \quad \ell = 0, \cdots, L-1.
\]
To analyze the function $\psi$, we will establish in this section several “sharp-modulo-constants” estimates that connect $\psi$ to a much simpler function, derived using the local behavior of $\varphi$ at 0 and its consequences for the iterated compositions $\varphi^{[\ell]}$ that appear in the definition of $\psi$. In particular, let us define $\widehat{\varphi} : [0, \pi] \to [0, \pi]$ by $\widehat{\varphi}(t) = t/(1 + t/(3\pi))$, so that

$$\widehat{\varphi}^{[\ell]}(t) = \frac{t}{1 + \ell t/(3\pi)},$$

and moreover define

$$\widehat{\xi}_\ell(t) = \prod_{\ell' = \ell}^{L-1} \left(1 - \frac{\widehat{\varphi}^{[\ell']} (t)}{\pi}\right), \quad \widehat{\psi}(t) = \frac{n}{2} \sum_{\ell=0}^{L-1} \widehat{\xi}_\ell(t).$$

We will prove that $\widehat{\varphi}^{[\ell]}$ provides a sharp approximation to $\varphi^{[\ell]}$ (Lemmas B.5.2 and B.5.3), and then work out a corresponding sharp approximation of $\widehat{\psi}$ to $\psi$ (Lemmas B.5.7 and B.5.9). We will then derive estimates for the low-order derivatives of $\psi$ in Section B.5.4. Unfortunately, it is impossible to obtain $L^1$ estimates for $\psi$ in terms of $\widehat{\psi}$ that are sharp enough to facilitate operator norm bounds for $\Theta_\mu$, which would let us construct certificates for an operator with kernel $\widehat{\psi}$ rather than the NTK $\Theta_\mu$; but the estimates we derive in this section will be nonetheless sufficient to enable our localization and certificate construction arguments in Section B.4.

We note that bounds similar to a subset of the bounds in this section have been developed in an $L$-asymptotic, large-angle setting by [115]. The bounds we develop here are non-asymptotic and hold for all angles, and are established using elementary arguments that we believe are slightly more transparent. We reuse (and restate in Section B.6) some estimates from Section A.3.1 here, but the majority of our estimates will be fundamentally improved (a representative example is Lemma B.5.3).

Throughout this section, we use $\varphi, \widehat{\varphi}, \varphi$ to represent first, second and third derivatives of $\varphi$ (see Lemma B.6.3 for basic regularity assertions for this function and its iterated compositions) and likewise for $\xi$ and $\psi$. In particular, for example, in our notation the function $\varphi^{[\ell]}$ refers to the derivative of $\varphi^{[\ell]}$, not the $\ell$-fold iterated composition of $\varphi$. Although this leads to an abuse of
notation, the concision it enables in our proofs will be of use.

B.5.1 Sharp Lower Bound for the Iterated Angle Evolution Function

Lemma B.5.1. One has

\[ \varphi(t) \leq \frac{t}{1 + t/(3\pi)}, \quad t \in [0, \pi]. \]

Proof. As \( \cos \) is monotonically decreasing in \([0, \pi)\), it is the same as proving

\[ (1 - \frac{t}{\pi}) \cos(t) + \frac{\sin t}{\pi} - \cos \frac{t}{1 + \frac{t}{3\pi}} \geq 0 \]

We have the gradient as

\[ -(1 - \frac{t}{\pi}) \sin t + \sin(\frac{t}{1 + t/(3\pi)}) \left( \frac{1}{(1 + t/(3\pi))^2} \right) \]
\[ \geq -(1 - \frac{t}{\pi}) \sin t + \sin t \left( \frac{1}{(1 + t/(3\pi))^3} \right) \]
\[ \geq \left( -(1 - \frac{t}{\pi}) + \frac{1}{(1 + \frac{t}{3\pi})^3} \right) \sin t \]
\[ \geq 0 \]

For the first inequality, we use the estimate

\[ \sin(ax) \geq x \sin a; \quad 0 \leq x \leq 1, \quad 0 \leq a \leq \pi, \tag{B.63} \]

which is easily established using concavity of \( \sin \) on \([0, \pi]\) and the secant line characterization, and for the final inequality, we use the estimate \( 1 - 3a \leq \frac{1}{(1+a)^3} \) for any \( a > -1 \), which follows from convexity of \( a \mapsto (1 + a)^{-3} \) on this domain and the tangent line characterization (at \( a = 0 \)). Since at \( t = 0 \), we have the inequality holds, we know it holds for the whole interval \([0, \pi]\) by the mean value theorem. \( \square \)
Lemma B.5.2 (Corollary of Lemma A.3.9). If \( \ell \in \mathbb{N}_0 \), one has the "fluid" estimate for the iterated angle evolution function

\[
\varphi^{[\ell]}(t) \leq \frac{t}{1 + \ell t/(3\pi)}.
\]

Proof. Follow the argument of Lemma A.3.9, but use Lemma B.5.1 as the basis for the argument instead of Lemma B.6.2. \( \square \)

Lemma B.5.3. There exists an absolute constant \( C_0 > 0 \) such that for all \( \ell \in \mathbb{N} \)

\[
\varphi^{[\ell]} - \varphi^{[\ell]} \leq C_0 \frac{\log(1 + \ell)}{\ell^2}. \tag{B.64}
\]

As a consequence, there exist absolute constants \( C, C', C'' > 0 \) such that for any \( 0 < \varepsilon \leq 1/2 \), if \( L \geq C\varepsilon^{-2} \) then for every \( t \in [0, C'\varepsilon^2] \) one has

\[
\hat{\varphi}^{[L]}(t) \leq \frac{1 + \varepsilon}{(1 + Lt/(3\pi))^2},
\]

and for every \( t \in [0, \pi] \) one has

\[
\hat{\varphi}^{[L]}(t) \leq \frac{C''}{(1 + Lt/(3\pi))^2}.
\]

Finally, we have for \( \ell > 0 \)

\[
\xi_\ell(t) \leq (1 + e^{6C_0} \frac{\log(1 + \ell)}{\ell}) \xi_\ell(t), \tag{B.65}
\]

and if \( L \geq 3 \)

\[
\psi(t) \leq \hat{\psi}(t) + 4ne^{6C_0} \log^2 L.
\]

Proof. Fix \( L \in \mathbb{N} \) arbitrary. We prove (B.64) first, then use it to derive the remaining estimates. The main tool is an inductive decomposition: start by writing

\[
\hat{\varphi}^{[L]}(t) - \varphi^{[L]}(t) = \hat{\varphi} \circ \hat{\varphi}^{[L-1]}(t) - \varphi \circ \varphi^{[L-1]}(t)
\]
\[ = \hat{\varphi} \circ \varphi^{[L-1]}(t) - \hat{\varphi} \circ \varphi^{[L-1]}(t) + \hat{\varphi} \circ \varphi^{[L-1]}(t) - \varphi \circ \varphi^{[L-1]}(t), \]

and then use the definition of \( \hat{\varphi} \) to simplify the first term on the RHS of the final equation (via direct algebraic manipulation) to

\[ \hat{\varphi} \circ \varphi^{[L-1]}(t) - \hat{\varphi} \circ \varphi^{[L-1]}(t) = \frac{\varphi^{[L-1]}(t) - \varphi^{[L-1]}(t)}{\left(1 + \frac{1}{\pi} \varphi^{[L-1]}(t)\right) \left(1 + \frac{1}{\pi} \varphi^{[L-1]}(t)\right)}. \]

This gives an expression for the difference \( \hat{\varphi}^{[L]}(t) - \varphi^{[L]}(t) \) as an affine function of the previous difference \( \hat{\varphi}^{[L-1]}(t) - \varphi^{[L-1]}(t) \):

\[ \hat{\varphi}^{[L]}(t) - \varphi^{[L]}(t) = \frac{\varphi^{[L-1]}(t) - \varphi^{[L-1]}(t)}{\left(1 + \frac{1}{\pi} \varphi^{[L-1]}(t)\right) \left(1 + \frac{1}{\pi} \varphi^{[L-1]}(t)\right)} + \hat{\varphi} \circ \varphi^{[L-1]}(t) - \varphi \circ \varphi^{[L-1]}(t), \]

and unraveling inductively, we obtain

\[ \varphi^{[L]}(t) - \varphi^{[L]}(t) = \sum_{\ell=0}^{L-1} \left( \prod_{\ell'=\ell+1}^{L-1} \frac{1}{\left(1 + \frac{1}{\pi} \varphi^{[\ell']}\right) \left(1 + \frac{1}{\pi} \varphi^{[\ell']}\right)} \right) \left( \hat{\varphi} - \varphi \right) \circ \varphi^{(\ell)}(t), \]

where for concision we write \((\hat{\varphi} - \varphi)(t) = \hat{\varphi}(t) - \varphi(t)\). Note that all the product coefficients in this expression are nonnegative numbers. Denoting by \( \hat{C}_1 \) the constant attached to \( t^3 \) in the result Lemma B.5.11 and defining \( C_1 = \max \{ \hat{C}_1, 1 \} \), Lemma B.5.11 gives

\[ \varphi^{[L]}(t) - \varphi^{[L]}(t) \leq C_1 \sum_{\ell=0}^{L-1} \left( \prod_{\ell'=\ell+1}^{L-1} \frac{1}{\left(1 + \frac{1}{\pi} \varphi^{[\ell']}\right) \left(1 + \frac{1}{\pi} \varphi^{[\ell']}\right)} \right) \left( \varphi^{[\ell]}(t) \right)^3. \quad (B.66) \]

To prove (B.64), we will use a two-stage approach:

1. (First pass) First, we will control only the first factor in the product term in (B.66) using Lemma B.5.11, given that \( \varphi \geq 0 \) allows us to upper bound by the product term without the second factor. The resulting bound on the LHS of (B.66) will be weaker (in terms of its
dependence on $L$) than (B.64).

2. (Second pass) After completing this control, we will have obtained a lower bound on $\varphi^{[L]}$; we can then return to (B.66) and use this lower bound to get control of both factors in the product term, which will allow us to sharpen our previous analysis and establish the claimed bound (B.64).

**First pass.** We have

$$\prod_{\ell' = 1}^{L-1} \frac{1}{1 + \frac{1}{3\pi} \varphi^{[\ell']} (t)} = \frac{1 + \frac{(\ell+1)t}{3\pi}}{1 + \frac{Lt}{3\pi}}.$$  

Tossing the product term involving $\varphi^{[\ell]}$ and applying Lemma B.5.2 in (B.66), we thus have a bound

$$\varphi^{[L]} (t) - \varphi^{[L]} (t) \leq \frac{C_1}{1 + \frac{Lt}{3\pi}} \sum_{\ell=0}^{L-1} \frac{t^3}{1 + \frac{\ell t}{3\pi}} + \frac{C_1 t / (3\pi)}{1 + \frac{Lt}{3\pi}} \sum_{\ell=0}^{L-1} \frac{t^3}{1 + \frac{\ell t}{3\pi}}.$$  

For the first term in this expression, we calculate using an estimate from the integral test

$$\sum_{\ell=0}^{L-1} \frac{t^3}{1 + \frac{\ell t}{3\pi}} \leq t^3 + \int_0^L \frac{t^3}{1 + \frac{\ell t}{3\pi}} \, d\ell = t^3 + \frac{3\pi t^2}{1 + \frac{Lt}{3\pi}} \left(1 - \frac{1}{1 + \frac{Lt}{3\pi}}\right) = t^3 + \frac{Lt^3}{1 + \frac{Lt}{3\pi}},$$

and for the second term, we calculate similarly

$$\sum_{\ell=0}^{L-1} \frac{t^3}{1 + \frac{\ell t}{3\pi}} \leq t^3 + \int_0^L \frac{t^3}{1 + \frac{\ell t}{3\pi}} \, d\ell = t^3 + \frac{3\pi t^2}{2} \left(1 - \frac{1}{(1 + \frac{Lt}{3\pi})^2}\right) = t^3 + \frac{Lt^3}{1 + \frac{Lt}{3\pi}} \left(1 + \frac{Lt}{3\pi}\right)^2.$$
Combining these results gives

\[
\varphi^{[L]}(t) - \varphi^{[L]}(t) \leq \frac{C_1 t^3}{(1 + \frac{Lt}{3\pi})^3} + \frac{C_1 L t^3}{(1 + \frac{Lt}{3\pi})^2} + \frac{C_1 L t^4}{(1 + \frac{Lt}{3\pi})^2 + (1 + \frac{Lt}{3\pi})^2} + \frac{C_1 L t^4}{(1 + \frac{Lt}{3\pi})^2}
\]

\[
\leq \frac{3\pi C_1 t}{L} \left( 3\pi + 2t + \frac{1}{3\pi t^2} \right).
\]

(B.67)

This bound gives us a nontrivial estimate as far out as \( t = \pi \), but the result is weaker there than what we need. We can proceed with a bootstrapping approach to improve our result for large angles. To begin, we have shown via (B.67)

\[
\varphi^{[L]}(t) \geq \varphi^{[L]}(t) - \frac{16\pi^2 C_1 t}{L}.
\]

Let us write \( t_0 = C/\sqrt{L} \), where \( C > 0 \) is a constant we will optimize below, and define

\[
\bar{\varphi}_L(t) = \begin{cases} 
\varphi^{[L]}(t) - \frac{16C_1 \pi^2 t}{L} & 0 \leq t < t_0 \\
\varphi^{[L]}(t_0) - \frac{16C_1 \pi^2 t_0}{L} & t_0 \leq t \leq \pi.
\end{cases}
\]

The notation here is justified by noticing that \( \varphi^{[L]} \) is concave and nondecreasing, so that our previous estimates imply \( \varphi^{[L]} \geq \bar{\varphi}_L \). It follows

\[
\bar{\varphi}^{[L]} - \varphi^{[L]} \leq \bar{\varphi}^{[L]} - \bar{\varphi}_L.
\]

Our previous bound (B.67) is an increasing function of \( t \), and sufficient for \( 0 \leq t \leq t_0 \). For \( t \geq t_0 \), we have

\[
\bar{\varphi}^{[L]}(t) - \bar{\varphi}_L(t) \leq \frac{16C_1 \pi^2 t_0}{L} + \bar{\varphi}^{[L]}(t) - \bar{\varphi}^{[L]}(t_0),
\]

455
and we can calculate using increasingness of $\varphi^{[L]}$

$$\varphi^{[L]}(t) - \varphi^{[L]}(t_0) \leq \frac{\pi}{1 + L/3} - \frac{C}{\sqrt{L + CL/(3\pi)}}$$

$$= \frac{\pi\sqrt{L} - C}{(1 + L/3)(\sqrt{L} + CL/(3\pi))} \leq \frac{9\pi^2}{CL^{3/2}},$$

whence the bound

$$\varphi^{[L]} - \varphi^{[L]} \leq \frac{\pi}{L^{3/2}} \left( 16\pi C_1 C + \frac{9\pi}{C} \right)$$

$$\leq \frac{24\pi^2 \sqrt{C_1}}{L^{3/2}} \quad \text{(B.68)}$$

valid on the entire interval $[0, \pi]$; the final inequality corresponds to the choice $C = \frac{3}{4\sqrt{C_1}}$.

**Second pass.** To start, with an eye toward the unused product term in (B.66), we have from (B.68)

$$1 + \frac{1}{3\pi} \varphi^{[L]}(t) \geq 1 + \frac{1}{3\pi} \varphi^{[L]}(t) - \frac{8\pi \sqrt{C_1}}{L^{3/2}}.$$

Using the numerical inequality $e^{-2x} \leq 1 - x$, valid for $0 \leq x \leq 1/2$ at least, we have if $L \geq (256\pi^2 C_1)^{1/3}$

$$1 + \frac{1}{3\pi} \varphi^{[L]}(t) \geq \exp \left( - \frac{16\pi \sqrt{C_1}}{L^{3/2}} \right) \left( 1 + \frac{1}{3\pi} \varphi^{[L]}(t) \right).$$

Applying this bound to terms in the second product term in (B.66) with index

$$\ell \geq \left( 256\pi^2 C_1 \right)^{1/3} \equiv r(C_1),$$

456
we therefore have \(^7\)

\[
\psi^{[L]}(t) - \varphi^{[L]}(t) \leq C_1 \sum_{\ell = 0}^{L-1} \left( \varphi^{[\ell]}(t) \right)^3 \left( \prod_{\ell' = \max\{r(C_1), \ell+1\}}^{L-1} \frac{1}{1 + \frac{1}{3\pi} \psi^{[\ell']} (t)} \right) \\
\leq C_1 \sum_{\ell = 0}^{L-1} \left( \varphi^{[\ell]}(t) \right)^3 \exp \left( 16\pi \sqrt{C_1} \sum_{\ell' = \max\{r(C_1), \ell+1\}}^{L-1} \frac{1}{(\ell')^{3/2}} \right) \\
\times \left( \prod_{\ell' = \max\{r(C_1), \ell+1\}}^{L-1} \frac{1}{1 + \frac{1}{3\pi} \psi^{[\ell']} (t)} \right)^2 \\
= C_1 e^{16\pi \sqrt{C_1} \zeta(3/2)} \sum_{\ell = 0}^{L-1} \left( \varphi^{[\ell]}(t) \right)^3 \left( 1 + \frac{\max\{r(C_1), \ell+1\} t}{3\pi} \right)^2 \\
= C_1 e^{16\pi \sqrt{C_1} \zeta(3/2)} \left( 1 + \frac{(r(C_1) - 1) t}{3\pi} \right)^2 \sum_{\ell = 0}^{L-1} \left( \varphi^{[\ell]}(t) \right)^3 \\
+ \sum_{\ell = r(C_1) - 1}^{L-1} \left( \varphi^{[\ell]}(t) \right)^3 \left( 1 + \frac{(\ell + 1) t}{3\pi} \right)^2.
\]

Now, since \(\varphi \leq \psi\), we have

\[
\left( 1 + \frac{(r(C_1) - 1) t}{3\pi} \right)^2 \sum_{\ell = 0}^{L-1} \left( \varphi^{[\ell]}(t) \right)^3 \leq r(C_1)^2 t^3,
\]

and

\[
\sum_{\ell = r(C_1)}^{L-1} \left( \varphi^{[\ell]}(t) \right)^3 \left( 1 + \frac{(\ell + 1) t}{3\pi} \right)^2 \leq 2r^3 \sum_{\ell = 0}^{L-1} \frac{1}{1 + \frac{\ell t}{3\pi}} \\
\leq 2r^3 + 2r^3 \int_0^L \frac{1}{1 + \frac{\ell t}{3\pi}} d\ell \\
= 2r^3 + 6\pi t^2 \log(1 + Lt/3\pi),
\]

\(^7\)Although it has a different meaning in our argument at large, here and in some subsequent bounds \(\zeta(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}\) denotes the Riemann zeta function. In this setting, we have \(\zeta(3/2) \leq e\).
whence

\[\varphi^{[L]}(t) - \varphi^{[L]}(t) \leq \frac{C_1 e^{16\pi \sqrt{C_1} \zeta(3/2)}}{(1 + \frac{L}{3\pi})^2} \left( (2 + r(C_1)^2) t^3 + 6\pi t^2 \log(1 + Lt/3\pi) \right)\]

\[\leq \frac{9\pi^2 C_1 e^{16\pi \sqrt{C_1} \zeta(3/2)}}{L^2} \left( (2 + r(C_1)^2) t + 6\pi \log(1 + Lt/3\pi) \right)\]

\[\leq 54\pi^3 C_1 (2 + r(C_1)^2) e^{16\pi \sqrt{C_1} \zeta(3/2)} \frac{\log(1 + L)}{L^2}.\]

In the final line, we are simply shuffling constants using \(t \leq \pi\). This completes the proof of (B.64).

**Derived estimates.** The remaining claims can be derived from the main claim we have just established; we will do so now. Below, we write \(C_0 = 54\pi^3 C_1 (2 + r(C_1)^2) e^{16\pi \sqrt{C_1} \zeta(3/2)}\). We will also assume \(\ell \geq 1\).

We prove the claim about \(\xi_\ell\) first. First, notice that for nonnegative numbers \(a, b\), one has \(1 - a + b \leq e^{2b} (1 - a)\) provided \(a \leq 1/2\). Since \(\varphi \leq \pi/2\), we have for each \(\ell > 0\)

\[\xi_\ell(t) \leq \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{\varphi^{[\ell']}(t)}{\pi} + \frac{C_0 \log(1 + \ell')}{\pi (\ell')^2} \right)\]

\[\leq \exp \left( 2C_0 \sum_{\ell' = \ell}^{L-1} \frac{\log(1 + \ell')}{(\ell')^2} \right) \tilde{\xi}_\ell(t).\]

By the integral test estimate, we have for \(\ell > 0\)

\[\sum_{\ell' = \ell}^{L-1} \frac{\log(1 + \ell')}{(\ell')^2} \leq \frac{\log(1 + \ell)}{\ell^2} + \int_{\ell}^{L} \frac{\log(1 + \ell')}{(\ell')^2} d\ell'\]

\[\leq \frac{\log(1 + \ell)}{\ell^2} + \log \left( \frac{1 + \frac{1}{\ell}}{1 + \frac{1}{L}} \right) + \frac{\log(1 + \ell) / \ell - \log(1 + L) / L}{\ell}\]

\[\leq \frac{3 \log(1 + \ell)}{\ell},\]

where we applied \(\log(1 + x) \leq x\) for all \(x > -1\), whence for \(\ell > 0\)

\[\xi_\ell(t) \leq e^{6C_0 \log(1+\ell)/\ell} \tilde{\xi}_\ell(t).\]
In particular, using the fact that \( \log(1 + \ell)/\ell \leq 1 \) and the estimate \( e^{cx} \leq 1 + xe^c \) for \( x \in [0, 1] \) (by convexity of the exponential function), we obtain

\[
\xi_\ell(t) \leq \left(1 + e^{6C_0 \log(1 + \ell)/\ell}\right) \tilde{\xi}_\ell(t),
\]

as claimed. The proof of the second inequality is very similar: first, repeated application of the chain rule gives

\[
\varphi^{[L]} = \prod_{\ell=0}^{L-1} \varphi \circ \varphi^{[\ell]}.
\]

Using the expression

\[
\varphi(t) = \frac{(1 - t/\pi) \sin t}{\sin \varphi(t)},
\]

we can exploit a telescopic cancellation in the preceding expression for \( \varphi^{[L]} \), obtaining

\[
\varphi^{[L]} = \frac{\sin t}{\sin \varphi^{[L]}(t)} \prod_{\ell=0}^{L-1} \left(1 - \frac{\varphi^{[\ell]}(t)}{\pi}\right).
\]

As the form of this upper bound is identical to the one we controlled for \( \xi_\ell \), only with a different constant factor, we can now apply the first part of that argument to the present setting, obtaining a bound

\[
\varphi^{(L)} \leq \frac{\sin t}{\sin \varphi^{(L)}(t)} \exp \left(\sum_{\ell=1}^{L-1} \varphi^{[\ell]} - \varphi^{[\ell]}\right) \exp \left(-\frac{1}{\pi} \sum_{\ell=0}^{L-1} \frac{t}{1 + \ell t/(3\pi)}\right)
\]

where in simplifying we also used that \( \varphi^{[0]} = \tilde{\varphi}^{[0]} \). To proceed, we split the first sum, obtaining for any index \( 1 \leq \ell_* \leq L - 1 \)

\[
\sum_{\ell=1}^{L-1} \varphi^{[\ell]} - \varphi^{[\ell]} = \sum_{\ell=1}^{\ell_* - 1} (\varphi^{[\ell]} - \varphi^{[\ell]}) + \sum_{\ell=\ell_*}^{L-1} (\tilde{\varphi}^{[\ell]} - \varphi^{[\ell]})
\]

\[
\leq C_t \sum_{\ell=1}^{\ell_* - 1} \frac{1}{\ell} + 3C_0 \frac{\log(1 + \ell_*)}{\ell_*} + \ell_* \sum_{\ell=1}^{\ell_*} \frac{1}{\ell} + 3C_0 \frac{\log(1 + \ell_*)}{\ell_*}
\]

\[
\leq C_t \log(\ell_*) + 3C_0 \frac{\log(1 + \ell_*)}{\ell_*}
\]

459
\[ \leq C \log(1 + \ell_*) \left( t + \frac{1}{\ell_*} \right), \]

where in the second line the bound on the first sum used (B.67), and the second used the estimate we proved in the previous section and the integral test estimate above; in the third line we estimated the harmonic series with the integral test; and in the fourth line we worst-cased. Next, for any \( t \leq 1/\ell_* \), we have by the above

\[ \sum_{\ell=1}^{L-1} \hat{\varphi}^{[\ell]} - \varphi^{[\ell]} \leq C \log(1 + \ell_*)/\ell_*, \]

and because the RHS approaches 0 as \( \ell_* \to \infty \), for any \( 0 < \varepsilon \leq 1 \) there is an integer \( N(\varepsilon) > 0 \) such that for all \( \ell_* \geq N \) we have

\[ \sum_{\ell=1}^{L-1} \hat{\varphi}^{[\ell]} - \varphi^{[\ell]} \leq \log(1 + \varepsilon). \]

In particular, obtaining a lower bound for the RHS by concavity of \( \log \), it is sufficient to take \( \ell_* \geq C\varepsilon^{-2} \) for a suitably large absolute constant \( C > 0 \). To ensure there exists such a value of \( \ell_* \), it suffices to choose \( L \geq C\varepsilon^{-2} \) and therefore \( t \leq C'\varepsilon^2 \). In particular, plugging this estimate into our previous bound, we have shown that for any \( \varepsilon > 0 \), if \( L \geq C'\varepsilon^{-2} \) then for all \( t \leq C\varepsilon^2 \) we have

\[ \hat{\varphi}^{[L]} \leq (1 + \varepsilon) \frac{\sin t}{\sin \varphi^{(L)}(t)} \exp \left( -\frac{1}{\pi} \sum_{\ell=0}^{L-1} \frac{t}{1 + \ell t/(3\pi)} \right). \]

We then calculate by an estimate from the integral test

\[ \sum_{\ell=0}^{L-1} \frac{t}{1 + \ell t/(3\pi)} \geq \int_0^L \frac{t}{1 + \ell t/(3\pi)} \, d\ell = 3\pi \log(1 + Lt/(3\pi)), \]

which establishes under the previous conditions on \( L \) and \( t \) that

\[ \hat{\varphi}^{[L]} \leq \frac{\sin t}{\sin \varphi^{(L)}(t)} \frac{1 + \varepsilon}{(1 + Lt/(3\pi))^3}. \]

460
To conclude, we need to simplify the sin ratio term. Using Lemma B.5.4, for any $0 < \varepsilon' \leq 1/2$, we have for $0 \leq t \leq C\varepsilon'$ that

$$\frac{\sin t}{\sin \varphi^{(L)}(t)} \leq (1 + 2\varepsilon')(1 + Lt/(3\pi)),$$

which suffices to prove the claim for small $t$ after noting $(1 + 2\varepsilon')(1 + \varepsilon) = 1 + 2\varepsilon' + \varepsilon + 2\varepsilon'\varepsilon$, choosing whichever is smaller, and adjusting the preceding conditions on $t$ and $L$ (i.e. the absolute constants in the previous bounds may grow/shrink as necessary). To show the claimed bound on the entire interval $[0, \pi]$, we can follow exactly the argument above, but instead of partitioning the sum of errors $\tilde{\varphi}^{[\ell]} - \varphi^{[\ell]}$ as above we simply use bound the sum of errors as in the bound on $\tilde{\xi}_\ell$ previously to obtain a large constant in the numerator; the sin ratio is controlled in this case using the first conclusion in Lemma B.5.4, which is valid on the whole interval $[0, \pi]$.

Finally, we obtain the estimate on $\psi$ by calculating using the estimate involving $\xi_\ell$ and $\tilde{\xi}_\ell$ that we proved earlier. First, we note that although we required $\ell > 0$ above, the fact that $\tilde{\varphi}^{[0]} = \varphi^{[0]}$ implies that we have an estimate $\xi_0 \leq (1 + \log(2)e^{6C_0})\tilde{\xi}_0$. We therefore have

$$\psi(t) = \frac{n}{2} \sum_{\ell=0}^{L-1} \tilde{\xi}_\ell(t) \leq \frac{n}{2} \sum_{\ell=0}^{L-1} \left(1 + e^{6C_0}\log \frac{1 + \ell}{\ell}\right) \tilde{\xi}_\ell(t)$$

$$\leq \tilde{\psi}(t) + \left(n/2\right)e^{6C_0}\left(\log(2)\tilde{\xi}_0(t) + \sum_{\ell=1}^{L-1} \frac{\log(1 + \ell)}{\ell} \tilde{\xi}_\ell(t)\right).$$

It is easy to see that $\tilde{\xi}_\ell \leq 1$. Hence

$$\psi(t) \leq \tilde{\psi}(t) + \left(n/2\right)e^{6C_0}\left(\log(2) + \sum_{\ell=1}^{L-1} \frac{\log(1 + \ell)}{\ell}\right)$$

$$\leq \tilde{\psi}(t) + ne^{6C_0}\left(\log(2) + \sum_{\ell=2}^{L-1} \frac{\log(\ell)}{\ell}\right)$$

$$\leq \tilde{\psi}(t) + ne^{6C_0}\left(2\log(2) + \int_{\ell=2}^{L-1} \frac{\log(\ell)}{\ell} \, d\ell\right)$$

$$\leq \tilde{\psi}(t) + ne^{6C_0}\left(2\log(2) - (1/2) \log^2 2 + (1/2) \log^2 (L - 1)\right).$$
\[ \leq \hat{\psi}(t) + 4ne^{6c_0 \log^2 L}, \]

where the final bound requires \( L \geq 3. \)

\[ \square \]

**Lemma B.5.4.** For \( \ell \in \mathbb{N}_0, \) one has for \( t \in [0, \pi] \)

\[ \frac{\sin(t)}{\sin(\varphi[\ell](t))} \leq 3(1 + \ell t/(3\pi)) \]

and there exists an absolute constant \( C > 0 \) such that for any \( 0 < \epsilon \leq 1/2, \) if \( 0 \leq t \leq C\epsilon \) one has

\[ \frac{\sin(t)}{\sin(\varphi[\ell](t))} \leq (1 + 2\epsilon)(1 + \ell t/(3\pi)). \]

**Proof.** We prove the bound on \([0, \pi]\) first. Because \( t \mapsto t^{-1} \sin t \) is decreasing on \([0, \pi]\), we apply Lemma A.3.10 to get

\[ \frac{\sin(t)}{\sin(\varphi[\ell](t))} \leq \frac{t}{\varphi[\ell](t)} \leq \frac{t}{t/(1+\ell t/(3\pi))} = 1 + \ell t/(\pi) \leq 3(1 + \ell t/(3\pi)). \]

Now fix \( 0 < \epsilon \leq 1/2. \) We claim that there is an absolute constant \( C > 0 \) such that if \( t \leq C\epsilon, \)

we have

\[ \varphi[\ell](t) \geq (1 - \epsilon)\varphi[\ell](t). \]

Assuming this claim, we have for \( t \leq C\epsilon \)

\[ \frac{\sin(t)}{\sin(\varphi[\ell](t))} \leq \frac{t}{\varphi[\ell](t)} \]

462
\[
\frac{t}{1 - \epsilon \varphi^{[\ell]}(t)} 
\leq \frac{1}{1 - \epsilon} (1 + \ell t/(3\pi)) 
\leq (1 + 2\epsilon)(1 + \ell t/(3\pi)),
\]
which is enough to conclude after rescaling. Now we want to show the claim. Let \( C_0 = \max \{1, C_1\} \) where \( C_1 \) denotes the constant on \( t^3 \) in Lemma B.5.11. We first notice that

\[
\varphi(t) \geq \bar{\varphi}(t) - C_0 t^3 \\
= \frac{t}{1 + t/(3\pi)} - C_0 t^3 \\
= \frac{t}{1 + t/(3\pi)} - \frac{t}{1 + \pi/(3\pi)} \frac{4C_0}{3} t^2 \\
\geq (1 - \frac{4C_0}{3} t^2) \frac{t}{1 + t/(3\pi)}.
\]

We are going to proceed with an induction-like approach. Put \( \epsilon_1 = 4C_0 t^2/3 \), and choose \( t \leq \sqrt{3/(4C_0)} \) so that \( 1 - \epsilon_1 \geq 0 \). Supposing that it holds \( \varphi^{[\ell-1]} \geq (1 - \epsilon_{\ell-1}) \bar{\varphi}^{[\ell-1]}(t) \) for a positive \( \epsilon_{\ell-1} \) such that \( 1 - \epsilon_{\ell-1} \geq 0 \) (we have shown there is such \( \epsilon_1 \) and controlled it), we have by some applications of the induction hypothesis, Lemma B.5.1, and the previous small-\( t \) estimate (we use below that \( t \leq \sqrt{3/(4C_0)} \))

\[
\varphi^{[\ell]}(t) \geq \left( 1 - \frac{4C_0}{3} \left( \varphi^{[\ell-1]}(t) \right)^2 \right) \bar{\varphi}(\varphi^{[\ell-1]}(t)) \\
= \left( 1 - \frac{4C_0}{3} \left( \varphi^{[\ell-1]}(t) \right)^2 \right) \frac{(1 - \epsilon_{\ell-1})}{1 + (\ell - 1)t/(3\pi)} \frac{t}{1 + (\ell - 1)t/(3\pi)} \\
\geq \left( 1 - \frac{4C_0}{3} \left( \varphi^{[\ell-1]}(t) \right)^2 \right) \frac{t^2}{1 + (\ell - 1)t/(3\pi)} \frac{1 + \ell t/(3\pi) - \epsilon_{\ell-1} t/(3\pi)}{1 + (\ell - 1)t/(3\pi)} \\
\geq \left( 1 - \frac{4C_0}{3} \left( \varphi^{[\ell-1]}(t) \right)^2 \right) (1 - \epsilon_{\ell-1}) \frac{i^2}{1 + (\ell - 1)t/(3\pi)} \\
\geq \left( 1 - \frac{4C_0}{3} \left( \varphi^{[\ell-1]}(t) \right)^2 \right) (1 - \epsilon_{\ell-1}) \frac{i^2}{1 + (\ell - 1)t/(3\pi)} - \epsilon_{\ell-1} \frac{i^2}{1 + (\ell - 1)t/(3\pi)} \bar{\varphi}^{[\ell]}(t).
\]

This shows that we can take \( \epsilon_{\ell} = \epsilon_{\ell-1} + (4C_0/3)t^2/(1 + (\ell - 1)t/(3\pi))^2 \) as long as this term is not
larger than 1. Unraveling inductively to check, we get

$$
\epsilon_\ell = \sum_{\ell'=0}^{\ell-1} \frac{4C_0}{3} \frac{t^2}{(1 + \ell' t/(3\pi))^2} \\
\leq \frac{4C_0}{3} t^2 \left( 1 + \int_{\ell'=0}^{\ell-1} \frac{1}{(1 + \ell' t/(3\pi))^2} d\ell' \right) \\
\leq \frac{4C_0}{3} t^2 \left( 1 + \frac{\ell - 1}{1 + (\ell - 1)t/(3\pi)} \right) \\
\leq \frac{4C_0}{3} \left( \pi + \frac{1}{3\pi} \right) t \\
= \frac{16\pi C_0}{3} t.
$$

In particular, the induction is consistent as long as $t \leq 3/(16\pi C_0)$. Note as well that since $C_0 \geq 1$ we have $\sqrt{3/(4C_0)} \geq 3/(16\pi C_0)$. Thus by induction, we know that when $0 < \epsilon < 1$ and $t \leq \frac{3C_0\epsilon}{16\pi}$, we have

$$
\varphi^{[\ell]}(t) \geq (1 - \epsilon)\widehat{\varphi}^{[\ell]}(t)
$$

as claimed. \hfill \square

B.5.2 Sharp Lower Bound for $\psi$

**Lemma B.5.5.** There is an absolute constant $C_0 > 0$ such that

$$
\psi(\pi) \leq \frac{n(L - 1)}{8} + 6\pi ne^{6C_0 \log^2 L}
$$

**Proof.** Following Lemma B.5.3 (worsening constants slightly for convenience), we directly have

$$
\psi(\pi) \leq \widehat{\psi}(\pi) + 6\pi ne^{6C_0 \log^2 L}.
$$
\( \tilde{\psi}(t) \) has a closed form expression, by notice that

\[
\tilde{\xi}_\ell(t) = \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{\tilde{\phi}^{[\ell']}(t)}{\pi} \right) \\
= \prod_{\ell' = \ell}^{L-1} \left( 1 - \frac{t/\pi}{1 + \ell' t/(3\pi)} \right) \\
= \prod_{\ell' = \ell}^{L-1} \frac{1 + (\ell' - 3)t/(3\pi)}{1 + \ell' t/(3\pi)} \\
= \frac{(1 + (\ell - 3)t/(3\pi)) (1 + (\ell - 2)t/(3\pi)) (1 + (\ell - 1)t/(3\pi))}{(1 + (L - 3)t/(3\pi)) (1 + (L - 2)t/(3\pi)) (1 + (L - 1)t/(3\pi))} \tag{B.70}
\]

and

\[
\tilde{\psi}(t) = \frac{n}{2} \sum_{\ell = 0}^{L-1} \tilde{\xi}_\ell(t) \\
= \frac{n}{2} \sum_{\ell = 0}^{L-1} \frac{(3\pi + (\ell - 3)t)(3\pi + (\ell - 2)t)(3\pi + (\ell - 1)t)}{(3\pi + (L - 3)t)(3\pi + (L - 2)t)(3\pi + (L - 1)t)} \\
\]

\[
= \frac{n}{8t} \frac{(3\pi + (L - 3)t)(3\pi + (L - 2)t)(3\pi + (L - 1)t)}{(3\pi + (L - 3)t)(3\pi + (L - 2)t)(3\pi + (L - 1)t)} \\
= \frac{n(L - 4) - \frac{n}{8t} (3\pi - \frac{(3\pi - 4t)(3\pi - 3t)(3\pi - 2t)(3\pi - t)}{(3\pi + (L - 3)t)(3\pi + (L - 2)t)(3\pi + (L - 1)t)})}{8} \tag{B.71}
\]

From the second to the fourth line above, we used a telescopic series cancellation trick to sum.

Then we get the claim as

\[
\psi(\pi) \leq \tilde{\psi}(\pi) + 6\pi ne^{6C_0} \log^2 L \\
= \frac{n(L - 4)}{8} + \frac{3\pi n}{8t} + 6\pi ne^{6C_0} \log^2 L
\]
Lemma B.5.6. When $L \geq 2$, we have for any $r > 0$

$$
\int_0^r \psi(t)dt \geq \frac{n(L-4)}{8} - \frac{3\pi n}{8} r + \frac{3\pi n}{8} \log(1 + \frac{L-2}{3\pi} r)
$$

Proof. From Lemma B.5.2, we have $\varphi^{[\ell]}(t) \leq \frac{1}{1-t/(3\pi)}$. Thus we get

$$
\xi_{\ell}(t) = \prod_{\ell' = \ell}^{L-1} \left(1 - \frac{1}{\pi} \varphi^{[\ell']}(t)\right) \geq \frac{(3\pi + (\ell-3)t)(3\pi + (\ell-2)t)(3\pi + (\ell-1)t)}{(3\pi + (L-3)t)(3\pi + (L-2)t)(3\pi + (L-1)t)}.
$$

As a result, we have

$$
\psi(t) = \frac{n}{2} \sum_{\ell = 0}^{L-1} \xi_{\ell}(t)
$$

$$
\geq \frac{n}{2} \sum_{\ell = 0}^{L-1} \frac{(3\pi + (\ell-3)t)(3\pi + (\ell-2)t)(3\pi + (\ell-1)t)}{(3\pi + (L-3)t)(3\pi + (L-2)t)(3\pi + (L-1)t)}
$$

$$
= \frac{n}{2} \sum_{\ell = 0}^{L-1} \frac{4t}{(3\pi + (L-3)t)(3\pi + (L-2)t)(3\pi + (L-1)t)} \sum_{\ell = 0}^{L-1} (3\pi + (\ell-4)t)(3\pi + (\ell-3)t)(3\pi + (\ell-2)t)(3\pi + (\ell-1)t)
$$

$$
= \frac{n}{8t} (3\pi + (L-4)t) (3\pi + (L-3)t) (3\pi + (L-2)t) (3\pi + (L-1)t)
$$

$$
- \frac{n}{8t} (3\pi + (L-3)t) (3\pi + (L-2)t) (3\pi + (L-1)t)
$$

$$
= \frac{n}{8t} (3\pi + (L-4)t) - \frac{n}{8t} \frac{(3\pi - 4t)(3\pi - 3t)(3\pi - 2t)(3\pi - t)}{(3\pi + (L-3)t)(3\pi + (L-2)t)(3\pi + (L-1)t)}
$$

$$
= \frac{n}{8t} (L-4) - \frac{n}{8t} \left(1 - \frac{(1-4t')(1-3t')(1-2t')(1-t')}{(1+(L-3)t')(1+(L-2)t')(1+(L-1)t')}ight)
$$

$$
\therefore \frac{n(L-4)}{8} + \frac{n}{8t'} \left(1 - \frac{(1-4t')(1-3t')(1-2t')(1-t')}{(1+(L-3)t')(1+(L-2)t')(1+(L-1)t')}ight)
$$
\[ \geq \frac{n(L - 4)}{8} + \frac{n}{8t'} \left( 1 - \frac{(1 - 3t')(1 - 2t')(1 - t')}{(1 + (L - 2)t')^3} \right) \]
\[ + \frac{n}{8t'} \frac{4t'(1 - 3t')(1 - 2t')(1 - t')}{(1 + (L - 3)t')(1 + (L - 2)t')(1 + (L - 1)t')} \]
\[ \geq \frac{n(L - 4)}{8} + \frac{n}{8t'} \left( 1 - \frac{1}{(1 + (L - 2)t')^3} \right) \]
\[ \stackrel{\text{L' = L - 2}}{=} \frac{n(L - 4)}{8} + \frac{n}{8} \frac{3L' + 3L'^2t' + L'^3t'^2}{(1 + L't')^3} \]
\[ = \frac{n(L - 4)}{8} + \frac{n}{8} \left( \frac{L'}{1 + L't'} + \frac{L'}{(1 + L't')^2} + \frac{L'}{(1 + L't')^3} \right). \]

In the third and fourth lines above, we used a splitting and cancellation trick to sum similar to what we used in Lemma B.5.5. In moving from the seventh to the eighth line, we used the inequality \((x - 1)(x + 1) \leq x^2\) after splitting off a term that can be negative for large \(t'\). In moving from the eighth to the ninth line, we used nonnegativity of the third summand and upper bounded the numerator of the term in the second summand. (In both of the previous simplifications, we are using that \(t' \leq 1/3\).) The remaining simplifications obtain a common denominator in the second term and then cancel. Integrating, we thus find

\[ \int_0^r \psi(t) dt \geq \int_0^r \psi(t) dt - \frac{3\pi n}{8} \log(1 + \frac{L - 2}{3\pi}) - Cnr \log^2 L \]

when \(L' \geq 0\)
\[ \Rightarrow \frac{n(L - 4)}{8} + \frac{3\pi n}{8} \log(1 + \frac{L - 2}{3\pi} r). \]

**Lemma B.5.7.** There exists an absolute constant \(C > 0\) such that when \(L \geq 2\), we have for any \(r > 0\)

\[ \int_0^r \psi(t) dt - \psi(\pi) dt \geq \frac{3\pi n}{8} \log(1 + \frac{L - 2}{3\pi}) - Cnr \log^2 L \]

**Proof.** Following Lemma B.5.6 and Lemma B.5.5, we directly get

\[ \int_0^r \psi(t) dt \geq \int_0^r \psi(t) dt - \left( \frac{n(L - 1)}{8} + 6\pi ne^{6C_0} \log^2 L \right) r \]

When \(L' \geq 0\),
\[ \frac{n(L - 4)}{8} + \frac{3\pi n}{8} \log(1 + \frac{L - 2}{3\pi} r). \]
B.5.3 Nearly-Matching Upper Bound

**Lemma B.5.8.** There exist absolute constants $C, C' > 0$ and absolute constants $K, K' > 0$ such that for any $0 < \varepsilon \leq 1$, if $L \geq K\varepsilon^{-3}$ then for any $0 \leq t \leq K'\varepsilon^3$ one has

$$
\psi(t) - \psi(\pi) \leq (1 + \varepsilon) \left( 1 + \frac{18}{1 + (L - 3)t/(3\pi)} + \frac{C \log^2(L)}{n} \right) \frac{L - 3}{8(1 + (L - 3)t/(3\pi))}
$$

and for any $0 \leq t \leq \pi$ one has

$$
\psi(t) - \psi(\pi) \leq C'n \frac{L - 3}{1 + (L - 3)t/(3\pi)}.
$$

**Proof.** We try to control the DC subtracted skeleton $\psi(t) - \psi(\pi)$ by its derivative $\dot{\psi}(t)$, which would require us to control the derivatives $\dot{\xi}_i(t)$ and further $\dot{\phi}^{[\ell]}(t)$. Fix $0 < \varepsilon \leq 1/2$. When $L \geq C_0\varepsilon^{-2}$ for some constant $C_0 > 0$, Lemma B.5.3 provides sharp bound for $\dot{\phi}^{[\ell]}(t)$ with

$$
\dot{\phi}^{[\ell]}(t) \leq \frac{1 + \varepsilon}{(1 + c\ell t)^2} \quad t \in [0, C'\varepsilon^2]
$$

$$
\dot{\phi}^{[\ell]}(t) \leq \frac{C_1}{(1 + c\ell t)^2} \quad t \in [0, \pi]
$$

with absolute constants $C', C_1 > 0$ and $c = 1/(3\pi)$. For notation convenience, define $t_1 = C'\varepsilon^2$ and write

$$
M_t = \begin{cases} 
1 + \varepsilon & 0 \leq t \leq t_1 \\
C_1 & \text{otherwise.}
\end{cases}
$$
We can compactly write the previous two bounds together as

$$\hat{\psi}^{[\ell]}(t) \leq \frac{M_t}{(1 + c\ell t)^2}.$$  

This allows us to separate $\psi(t) - \psi(\pi)$ into two components $\psi(t) - \psi(t_1)$ and $\psi(t_1) - \psi(\pi)$, where we get the correct constant $1 + \epsilon$ in the first component and control the second component by the fact that $\psi$ becomes sharp when $L$ is large, making the difference between $\psi(t_1)$ and $\psi(\pi)$ negligible.

Now, for $\ell \geq 4$, with $c = \frac{1}{3\pi}$, we have

\[
|\hat{\xi}_\ell(t)| = \frac{\xi_\ell(t)}{\pi} \sum_{\ell' = \ell}^{L-1} \frac{\hat{\psi}^{[\ell']}}{1 - \hat{\psi}^{[\ell']}/\pi} \\
\leq \frac{\xi_\ell(t)}{\pi} \sum_{\ell' = \ell}^{L-1} \frac{\hat{\psi}^{[\ell']}}{1 - \frac{t/\pi}{1 + c\ell t}} \\
= \frac{\xi_\ell(t)}{\pi} \sum_{\ell' = \ell}^{L-1} \frac{1 + c\ell' t}{1 + c(\ell' - 3)t} \hat{\psi}^{[\ell']} \\
\leq \frac{\xi_\ell(t)}{\pi} \sum_{\ell' = \ell}^{L-1} \frac{1 + c\ell' t}{1 + c(\ell' - 3)t} \frac{M_t}{(1 + c\ell' t)^2}
\]

Let $c_{1,\ell} = 1 + e^{6C_0 \log(1+\ell)} / \ell$. (B.65) and (B.70) provide control for $\hat{\xi}_\ell(t)$ and we have

\[
|\hat{\xi}_\ell(t)| \leq \frac{c_{1,\ell}}{\pi} \frac{(1 + c(\ell - 3)t)(1 + c(\ell - 2)t)(1 + c(\ell - 1)t)}{(1 + c(L - 3)t)(1 + c(L - 2)t)(1 + c(L - 1)t)} \sum_{\ell' = \ell}^{L-1} \frac{1 + c\ell' t}{1 + c(\ell' - 3)t} \frac{M_t}{(1 + c\ell' t)^2} \\
\leq \frac{c_{1,\ell} M_t}{\pi} \frac{(1 + c(\ell - 3)t)(1 + c(\ell - 2)t)(1 + c(\ell - 1)t)}{(1 + c(L - 3)t)(1 + c(L - 2)t)(1 + c(L - 1)t)} \int_{\ell' = \ell}^{L-1} \frac{1}{1 + c((\ell' - 2)t)^2} d\ell' \\
\leq \frac{c_{1,\ell} M_t}{\pi} \frac{(1 + c(\ell - 3)t)(1 + c(\ell - 2)t)(1 + c(\ell - 1)t)}{(1 + c(L - 3)t)(1 + c(L - 2)t)(1 + c(L - 1)t)} \frac{L - \ell}{(1 + c(\ell - 3)t)(1 + c(L - 3)t)} \\
\leq \frac{M_t}{\pi} \frac{(1 + c(\ell - 2)t)(1 + c(\ell - 1)t)(L - \ell)}{(1 + c(L - 3)t)^2(1 + c(L - 2)t)(1 + c(L - 1)t)} \\
+ \frac{C \log(1 + \ell) / \ell}{\pi} \frac{L}{(1 + c(L - 3)t)^2}.
\]

In moving from the fifth to the sixth line, we used that $(1 + c(\ell - 3)t)(1 + c(\ell)t) = 1 + c(2\ell - 3)t + c^2(\ell - 3)t^2 \geq 1 + c(2\ell - 4)t + c^2(\ell^2 - 4\ell + 4)t^2 = (1 + c(\ell - 2)t)^2$ provided $\ell \geq 4$ and
subsequently the integral test. In the splitting in the last line, we used that \( M_t \) is always bounded by a (very large) absolute constant, and worst-cased (as this term will be sub-leading in \( L \)).

To control derivatives of \( \psi \), we need to control sums of the derivatives above. We will derive some further estimates for this purpose. First, we calculate

\[
\sum_{\ell=1}^{L-1} (1 + c(\ell - 2)t)(1 + c(\ell - 1)t)(L - \ell)
= \sum_{\ell=1}^{L-1} \left( (L - \ell) + (L - \ell)(2\ell - 3)ct + (L - \ell)(\ell - 1)(\ell - 2)c^2t^2 \right)
= \frac{L(L-1)}{2} + \frac{L(L-1)(L-\frac{7}{2})}{3}ct + \frac{L(L-1)(L-2)(L-3)}{12}c^2t^2
\leq \left( \frac{L(L-1)}{2} + \frac{L(L-1)(L-\frac{7}{2})}{3}ct \right) + 1 \left( \frac{L(L-3)}{12} \right)(1 + c(L - 1)t)(1 + c(L - 2)t)
\leq \left( \frac{6\frac{L-1}{L-3} + 4ct(L-1)}{(1 + c(L-1)t)(1 + c(L-2)t)} + 1 \right) \frac{L(L-3)}{12} (1 + c(L - 1)t)(1 + c(L - 2)t).
\]

If \( L \geq 4 \), we can simplify a term in the last line of the previous expression as

\[
\frac{6\frac{L-1}{L-3} + 4ct(L-1)}{(1 + c(L-1)t)(1 + c(L-2)t)} \leq \frac{18}{1 + c(L - 1)t},
\]

and under \( L \geq 4 \) we also have

\[
\sum_{\ell=2}^{L-1} \frac{\log(1 + \ell)}{\ell} \leq \log(L) \int_{\ell=1}^{L-1} \frac{1}{\ell} \leq \log^2 L.
\]

Applying the upper bound from before and adding some terms to the sum (because all terms are nonnegative), we get

\[
\pi \sum_{\ell=4}^{L-1} |\xi_\ell(t)| \leq \left( 1 + \frac{18}{1 + c(L - 1)t} \right) \frac{L(L-3)}{12} \frac{M_t}{(1 + c(L - 3)t)^2} + \frac{CL \log^2 L}{(1 + c(L - 3)t)^2}. \tag{B.72}
\]

From Lemma A.3.17, for \( \ell = \{0, 1, 2, 3\} \), we can bound \( \xi_\ell(t) \leq \frac{1+\ell t}{1+L t/\pi} \). Using that \( \xi_\ell \) is decreasing
for all $\ell \geq 0$ and nonnegative, for $t, t' \in [0, \pi]$, $t' \geq t$, we are now able to control the DC subtracted skeleton as

$$\psi(t) - \psi(t') \leq \frac{n}{2} \frac{4 + (0 + 1 + 2 + 3)t/\pi}{1 + Lt/\pi} - \frac{n}{2} \sum_{\ell=4}^{L-1} \int_{v=t}^{t'} \xi_\ell(v) dv$$

$$\leq \frac{2 + 3t/\pi}{1 + Lt/\pi} n + \frac{n}{2\pi} \int_{v=t}^{t'} \left( 1 + \frac{18}{1 + c(L-1)t} \right) \frac{M_v(L(L-3)/12) + C \log^2 L}{(1 + c(L-3)v)^2} dv$$

$$\leq \frac{2 + 3t/\pi}{1 + Lt/\pi} n + \frac{n}{2\pi} \int_{v=t}^{t'} L \left( 1 + \frac{18}{1+c(L-1)t} \right) \frac{M_v(L(L-3)/12) + C \log^2 L}{(1 + c(L-3)v)^2} dv$$

From the second to the third line, we use the fact that $M_v$ is nondecreasing in $v$. Thus we have

$$\psi(t) - \psi(t') \leq \frac{2 + 3t/\pi}{1 + Lt/\pi} n + \frac{n}{2\pi} L \left( 1 + \frac{18}{1+c(L-1)t} \right) \frac{M_v(L(L-3)/12) + C \log^2 L}{1 + c(L-3)v} \bigg|_{v=t}$$

$$\leq \frac{5n}{1 + Lt/\pi} + \frac{nL}{2} \left( 1 + c(L-3)v \right) \left( 1 + c(L-3\pi) \right)$$

$$\leq \frac{5n}{1 + Lt/\pi} + \frac{L}{1 + c(L-3)v} \frac{M_v(L(L-3)/12) + 12 \log^2 L}{1 + c(L-3)\pi}$$

$$\leq \frac{n}{8(1 + c(L-3)t) \left( 1 + \frac{18}{1 + c(L-1)t} + \frac{12 \log^2 L}{L-3} \right)} + \frac{5n}{1 + Lt/\pi}$$

$$\leq \frac{nM_{L'}}{8(1 + c(L-3)t) \left( 1 + \frac{18}{1 + c(L-1)t} + \frac{C \log^2(L)}{L-3} \right)}$$

In moving from the second to the third line, we simplified/rearranged and used that $C_1 \geq 1$. In moving from the fourth to the fifth line, we replace the numerator of $L$ in the leading term with...
$L - 3 + 3$, then expand and simplify. In particular we have

$$
\psi(t_1) - \psi(\pi) \leq \frac{nC_1}{8} \frac{1}{ct_1} \left( 19 + \frac{C \log^2(L)}{L} \right)
$$

$$
\leq \frac{C_1 C'(19 + C) n}{8 ce^2}
$$

and

$$
\psi(t) - \psi(t_1) \leq \frac{n(1 + \epsilon)}{8} \frac{1}{ct} \left( 1 + \frac{18}{1 + c(L - 3)t} + \frac{C \log^2(L)}{L} \right)
$$

$$
\leq \frac{n(1 + \epsilon)}{8} \frac{L - 3}{1 + c(L - 3)t} \left( 1 + \frac{18}{1 + c(L - 3)t} + \frac{C \log^2(L)}{L} \right).
$$

This inequality holds for all $t$ because when $t \geq t_1$, the left hand side is negative. Notice that when $t \leq \epsilon^3/(2C_1 C'(19 + C))$ and $L - 3 \geq C_1 C'(19 + C)/(c \epsilon^3)$, we would have

$$
\frac{C_1 C'(19 + C) n}{8 ce^2} = \epsilon \frac{n}{8} \frac{L - 3}{2c(L - 3)\epsilon^3/(2C_1 C'(19 + C))}
$$

$$
\leq \epsilon \frac{n}{8} \frac{L - 3}{1 + c(L - 3)\epsilon^3/(2C_1 C'(19 + C))}
$$

$$
\leq \frac{n \epsilon}{8} \frac{L - 3}{1 + c(L - 3)t}.
$$

Thus when $L \geq C_1 C'(19 + C)/(c \epsilon^3) + 3$, for $t \leq \epsilon^3/(2C_1 C'(19 + C))$,

$$
\psi(t) - \psi(\pi) = \psi(t) - \psi(t_1) + \psi(t_1) - \psi(\pi)
$$

$$
\leq \frac{n(1 + 2\epsilon)}{8} \frac{L - 3}{1 + c(L - 3)t} \left( 1 + \frac{18}{1 + c(L - 3)t} + \frac{C \log^2(L)}{L} \right).
$$

Combining the results and notice that we can absorb the factor of 2 into constants by defining $\epsilon' = 2\epsilon$ would give us the claim.

□

**Lemma B.5.9.** There exist absolute constants $C, C' > 0$ and $K, K' > 0$ such that for any $0 < \epsilon \leq 1$,
if \( L \geq K\varepsilon^{-3}, \) for any \( 0 \leq a \leq b \leq \pi, \) one has

\[
\int_a^b (\psi(t) - \psi(\pi)) dt \leq Cn \log \left( \frac{1 + (L - 3)b/(3\pi)}{1 + (L - 3)a/(3\pi)} \right). \tag{B.73}
\]

And if \( r > 0 \) satisfies \( r \leq K'\varepsilon^3, \) one further has

\[
\int_r^b (\psi(t) - \psi(\pi)) dt \leq (1 + \varepsilon) \frac{3\pi n}{8} \log \left( \frac{1 + (L - 3)b/(3\pi)}{1 + (L - 3)r/(3\pi)} \right) + C'n \log(\pi/(K'\varepsilon^3)). \tag{B.74}
\]

Proof. (B.73) follows directly from Lemma B.5.8 and integration. To achieve an upper bound for integral from \( r \) to \( b, \) we cut the integral at \( t_1 = K'\varepsilon^3 \) and apply bounds from Lemma B.5.8 separately. Specifically, set \( b' = \min\{b, t_1\}, \) from Lemma B.5.8 and (B.73) we would have

\[
\int_r^b (\psi(t) - \psi(\pi)) dt
\]

\[
= \int_r^{b'} (\psi(t) - \psi(\pi)) dt + \int_{b'}^b (\psi(t) - \psi(\pi)) dt
\]

\[
\leq (1 + \varepsilon) \left( 1 + \frac{C \log^2(L)}{L} \right) \frac{3\pi n}{8} \log \left( \frac{1 + (L - 3)b'//(3\pi)}{1 + (L - 3)r/(3\pi)} \right)
\]

\[
+ \int_r^b \frac{18n(1 + \varepsilon)}{8} \frac{L - 3}{(1 + (L - 3)t/(3\pi))^2} dt + C'n \log \left( \frac{1 + (L - 3)b'//(3\pi)}{1 + (L - 3)b'/(3\pi)} \right)
\]

\[
\leq (1 + \varepsilon) \left( 1 + \frac{C \log^2(L)}{L} \right) \frac{3\pi n}{8} \log \left( \frac{1 + (L - 3)b'//(3\pi)}{1 + (L - 3)r/(3\pi)} \right) + \frac{9n}{2(3\pi)^2} + C'n \log(b/b')
\]

\[
\leq (1 + \varepsilon) \frac{3\pi n}{8} \log \left( \frac{1 + (L - 3)b'//(3\pi)}{1 + (L - 3)r/(3\pi)} \right) + \frac{2C \log^2(L)}{L} \frac{3\pi n}{8} \log(1 + (L - 3)\pi/(3\pi))
\]

\[
+ \frac{n}{2\pi^2} + C'n \log(\pi/t_1)
\]

\[
\leq (1 + \varepsilon) \frac{3\pi n}{8} \log \left( \frac{1 + (L - 3)b'//(3\pi)}{1 + (L - 3)r/(3\pi)} \right) + \frac{C \log^3(L)}{L} + C'n \log(\pi/(K'\varepsilon^3)).
\]

(B.74) then follows by setting \( L \geq K\varepsilon^{-3} \) for some \( K > 0. \) \qed
B.5.4 Higher Order Derivatives of $\psi$

**Lemma B.5.10.** There exist absolute constants $C, C'$ such that when $L \geq C$, we have for any $r \in [0, \pi]$,

$$\max_{t \geq r} |\dot{\psi}(t)| \leq \frac{C'n}{r^2} \quad (B.75)$$

and we can control the integration

$$\int_{t=0}^{r} r^3 |\ddot{\psi}(t)| dt \leq C'n r^2 \quad (B.76)$$

**Proof.** From (B.72) (we control $M_t \leq C$ for an absolute constant $C > 0$ in this context, so that we do not need to deal with the conditions on $\varepsilon$ that appear there) and Lemma A.3.18, we have

$$|\dot{\psi}(t)| \leq \frac{\sum_{\ell=0}^{3} |\dot{\xi}_\ell(t)| + Cn}{2} \left(1 + \frac{18}{1 + (L-1)t/(3\pi)} \right) \left(1 + \frac{L(L-3)}{12} \frac{1}{(1 + (L-3)t/(3\pi))^2} + \frac{L \log^2 L}{(1 + (L-3)t/(3\pi))^2} \right)$$

$$\leq \frac{n}{2} \frac{12L}{1 + Lt/\pi} + \frac{Cn}{2} \left(1 + \frac{18}{1 + (L-1)t/(3\pi)} \right) \left(1 + \frac{L(L-3)}{12} \frac{1}{(1 + (L-3)t/(3\pi))^2} + \frac{L \log^2 L}{(1 + (L-3)t/(3\pi))^2} \right)$$

$$\leq \frac{6\pi n}{t} + \frac{Cn}{2} \left( \frac{L(L-3)}{(L-3)^2 t^2} + \frac{L \log^2 L}{(L-3)^2 t^2} \right)$$

$$\leq \frac{Cn}{t^2}.$$

This directly get us (B.75) and (B.76). \hfill \square

B.5.5 Additional Proofs for Some Bounds

**Lemma B.5.11.** There exists an absolute constant $C_1 > 0$ such that

$$\tilde{\varphi}(t) - \varphi(t) \leq C_1 t^3.$$
Proof. From Lemma B.6.2, $\varphi$ is 3 times continuously differentiable on $(0, \pi)$, and

$$
\varphi(0) = 0, \quad \varphi'(0) = 1, \quad \varphi''(0) = -\frac{2}{3\pi}.
$$

It is easy to check that

$$
\tilde{\varphi}(0) = 0, \quad \tilde{\varphi}'(0) = 1, \quad \tilde{\varphi}''(0) = -\frac{2}{3\pi}.
$$

Since the Taylor expansions of these two functions around 0 agree to third order, and both are 3 times continuously differentiable on $(0, \pi)$, we obtain by Lagrange’s remainder theorem that for any $t \in [0, \pi)$,

$$
\tilde{\varphi}(t) - \varphi(t) = \int_0^t \left( \tilde{\varphi}'(s) - \varphi'(s) \right) \frac{s^2}{2} \, ds \leq C_1 t^3
$$

for some finite constant $C_1 = \sup_{t \in [0, \pi)} |\tilde{\varphi}'(t) - \varphi'(t)|$. At $t = \pi$ we have $\tilde{\varphi}(\pi) - \varphi(\pi) = \frac{\pi}{1+\pi/3} - \frac{\pi}{3} \leq 0$ hence the same bound holds for $t \in [0, \pi]$. \qed

Lemma B.5.12. One has

$$
\varphi(t) \geq \frac{1}{2}, \quad t \in [0, \pi/2]
$$

$$
|\varphi(t)| \leq 4, \quad t \in [0, \pi]
$$

Proof. We know $\varphi$ is monotonically increasing and concave on $[0, \pi]$, thus for $t \in [0, \pi/2]$,

$$
\varphi(t) \geq \varphi\left(\frac{\pi}{2}\right) = \frac{1/2}{\sin(\varphi(\pi/2))} \geq \frac{1}{2}.
$$

Using Lemma A.3.10 we also have for $t \in [0, \pi]$, $\varphi(t) \leq \varphi(0) = 1$,

$$
\varphi(t) \geq \frac{t}{1+t/\pi} \geq \frac{t}{2},
$$

475
and the first bound here can be used to obtain

\[ t - \varphi(t) \leq \frac{t^2}{1 + t/\pi} \leq \frac{t^2}{\pi}. \]

Thus since \( \varphi \leq \pi/2 \)

\[ \cos t \sin \varphi(t) - \phi(t) \sin t \cos \varphi(t) \geq \cos t \sin \varphi(t) - \sin t \cos \varphi(t) \]
\[ \geq -\sin(t - \varphi(t)), \]

and in particular, using the expression for \( \bar{\varphi} \) from Lemma B.6.2

\[ -\bar{\varphi}(t) = -(1 - \frac{t}{\pi}) \frac{\cos t \sin \varphi(t) - \phi(t) \sin t \cos \varphi(t)}{\sin^2 \varphi(t)} + \frac{\sin t}{\pi \sin(\varphi(t))} \]
\[ \leq (1 - \frac{t}{\pi}) \frac{\sin(t - \varphi(t))}{\sin^2 \varphi(t)} + \frac{2}{\pi} \]
\[ \leq \frac{t^2}{\pi} \frac{1}{\sin^2(t/2)} + \frac{2}{\pi} \]
\[ \leq \frac{t^2}{(t/\pi)^2} + \frac{2}{\pi} \]
\[ \leq 4. \]

\[ \square \]

**Lemma B.5.13.** There exist constants \( c_3, c_4 > 0 \) such that \( \bar{\varphi}(t) < -c_3 \) for \( t \in [0, \frac{\pi}{2}] \) and \( |\bar{\varphi}| \leq c_4 \) for \( t \in [0, \pi] \).

**Proof.** The existence of \( c_3 \) follows from Lemma B.6.2 directly. The existence of \( c_4 \) follows from smoothness of \( \varphi \) on \((0, \pi)\) and the fact that \( \bar{\varphi}(0) = -\frac{1}{3\pi^2}, \bar{\varphi}(\pi) = \frac{2}{\pi} \) both exist. \[ \square \]

**Lemma B.5.14.** There exists an absolute constant \( C > 0 \) such that for any \( 0 < t \leq \pi \) and \( \ell \in \mathbb{N}_0 \), one has

\[ \sum_{\ell' = \ell}^{\infty} \bar{\varphi}^{(\ell')}(t) \leq \frac{C}{t} \frac{1}{1 + \ell t/(3\pi)} \]
Proof. Using Lemma B.5.3, we have

\[ \phi^{[\ell]}(t) \leq \frac{C}{(1 + \ell t/(3\pi))^2}. \]

We can then calculate

\[
\sum_{\ell' = \ell}^{\infty} \phi^{[\ell']} (t) \leq C \sum_{\ell' = \ell}^{\infty} \frac{1}{(1 + \ell' t/(3\pi))^2} \leq C \left( \frac{1}{1 + \ell t/(3\pi)} + \int_{\ell' = \ell}^{\infty} \frac{1}{(1 + \ell' t/(3\pi))^2} \, d\ell' \right) \\
\leq C \left( \frac{1}{1 + \ell t/(3\pi)} + \frac{3\pi/t}{1 + \ell t/(3\pi)} \right) \\
\leq \frac{C}{1 + \ell t/(3\pi)},
\]

as claimed.

\[ \square \]

B.6 Auxiliary Results

B.6.1 Concentration of the Initial Random Network and Its Gradients

Theorem B.6.1 (Corollary of Theorem A.2.2 and Lemma A.3.8). Let \( M \) be a two curve problem instance. For any \( d \geq K \log(nn_0 \, \text{len}(M)) \), if \( n \geq K'd^4 L^5 \) then one has on an event of probability at least \( 1 - e^{-cd} \)

\[ \| \Theta - \Theta_{\text{NTK}} \|_{L^\infty(M \times M)} \leq Cn/L, \]

where \( c, C, K, K' > 0 \) are absolute constants.

B.6.2 Basic Estimates for the Infinite-Width Neural Tangent Kernel

Lemma B.6.2. One has

1. \( \varphi \in C^\infty(0, \pi) \), and \( \varphi, \phi, \hat{\varphi}, \text{ and } \check{\varphi} \) extend to continuous functions on \([0, \pi]\);

2. \( \varphi(0) = 0 \) and \( \varphi(\pi) = \pi/2 \); \( \phi(0) = 1 \), \( \hat{\varphi}(0) = -2/(3\pi) \), and \( \check{\varphi}(0) = -1/(3\pi^2) \); and
\[ \dot{\varphi}(\pi) = \ddot{\varphi}(\pi) = 0; \]

3. \( \varphi \) is concave and strictly increasing on \([0, \pi]\) (strictly concave in the interior);

4. \( \ddot{\varphi} < -c < 0 \) for an absolute constant \( c > 0 \) on \([0, \pi/2]\);

5. \( 0 < \varphi < 1 \) and \( 0 > \ddot{\varphi} \geq -C \) on \((0, \pi)\) for some absolute constant \( C > 0 \);

6. \( \nu(1 - C_1 \nu) \leq \varphi(\nu) \leq \nu(1 - c_1 \nu) \) on \([0, \pi]\) for some absolute constants \( C_1, c_1 > 0 \).

**Proof.** Combine the results in Lemma A.5.5 with Lemma B.5.13 to obtain the conclusion. \( \square \)

**Lemma B.6.3** (Corollaries of Lemma B.6.2). One has:

1. The function \( \varphi \) is smooth on \((0, \pi)\), and (at least) \( C^3 \) on \([0, \pi]\).

2. For each \( \ell = 0, 1, \ldots, L \), the functions \( \varphi^{(\ell)} \) are nonnegative, strictly increasing, and concave (positive and strictly concave on \((0, \pi))\).

3. If \( 0 \leq \ell < L \), the functions \( \xi_{\ell} \) are nonnegative, strictly decreasing, and convex (positive and strictly convex on \((0, \pi))\).

4. The function \( \psi \) is smooth on \((0, \pi)\), \( C^3 \) on \([0, \pi]\), and is nonnegative, strictly decreasing, and convex.
Appendix C: Proofs and Experimental Details for Resource-Efficient Invariant Networks

We collect further details and proofs relevant to the material in Chapter 4. For further details, we refer the reader to the preprint [235].

C.1 Implementation and Experimental Details

C.1.1 Implementation Details for Parametric Transformations of the Image Plane

Our implementation of parametric image deformations revolves around the specific definition of interpolation we have made:

\[ y \circ \tau = \sum_{(k,l) \in \mathbb{Z}^2} y_{kl}(\tau_0 \circ k \mathbf{1}) \circ \phi(\tau_1 \circ l \mathbf{1}), \]

and the identification of the image \( y \in \mathbb{R}^{m \times n} \) with a function on \( \mathbb{Z}^2 \) with support in \( \{0, \ldots, m-1\} \times \{0, \ldots, n-1\} \). Although we use the notation \( \circ \) for interpolation in analogy with the usual notation for composition of functions, this operation is significantly less well-structured: although we can define interpolation of motion fields \( \tau_0 \circ \tau_1 \), it is impossible in general to even have associativity of \( \circ \) (let alone inverses), so that in general \( (x \circ \tau_0) \circ \tau_1 \neq x \circ (\tau_0 \circ \tau_1) \). This failure is intimately linked to the existence of parasitic interpolation artifacts when computing and optimizing with interpolated images, which we go to great lengths to avoid in our experiments. On the other hand, there does exist a well-defined identity vector field: from our definitions, we can read off the canonical definition of the identity transformation, from which definitions for other parametric

---

\(^1\)These conventions are not universal, although they seem most natural from a mathematical standpoint—for example, PyTorch thinks of its images as lying on a grid in the square \([-1, +1] \times [-1, +1]\) instead, with spacing and offsets depending on the image resolution and other implementation-specific options. In our released code, we handle conversion from our notation to this notation.
transformations we consider here follow. Defining (with a slight abuse of notation)

\[ m = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ m-1 \end{bmatrix}; \quad n = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ n-1 \end{bmatrix}, \]

we have from the definition of the cubic convolution interpolation kernel \( \phi \) that

\[ y \circ (m1^* \otimes e_0 + n^* \otimes e_1) = y. \]

One can then check that the following linear embedding of the affine transformations, which we will write as \( \text{Aff}(2) = \text{GL}(2) \times \mathbb{R}^2 \), leads to the natural vector field analogue of affine transformations on the continuum \( \mathbb{R}^2 \) (c.f. Section C.2):

\[ \text{Aff}(2) \cong \text{span}\left\{ m1^* \otimes e_0, n^* \otimes e_0, m1^* \otimes e_1, n^* \otimes e_1, 1_{m,n} \otimes e_0, 1_{m,n} \otimes e_1 \right\}. \quad (C.1) \]

Of course, these vector fields can be any size—they need not match the size of the image. As we mention in Section 4.3.3, we always initialize our networks with the identity transform; in the basis above, this corresponds to the vector \((1, 0, 0, 1, 0, 0)\) (i.e., this is like a row-wise flattening of the affine transform’s matrix \( A \in \text{GL}(2) \), concatenated with \( b \)).

Next we turn to computation of the proximal operator, which we need for unrolling (see Section 4.3.3). Given (4.6) and the fact that (C.1) is a subspace, we can compute the proximal operator for \( \text{Aff}(2) \) given an orthonormal basis for \( \text{Aff}(2) \). It is then unfortunate that the natural basis vectors that we have used in the expression (C.1) are not orthogonal: we have \( \langle m1^*, n^* \rangle = \langle m, 1 \rangle \langle n, 1 \rangle \gg 0 \), for example. To get around this, in practice we apply a tech-
nique we refer to as *centering* of transformations. Indeed, notice that for any $c \in \mathbb{R}^2$, we have

$$\text{Aff}(2) \equiv \text{span}\{(m - c_01)^* \otimes e_0, (n - c_11)^* \otimes e_0, (m - c_01)^* \otimes e_1, (n - c_11)^* \otimes e_1, 1_{m,n} \otimes e_0, 1_{m,n} \otimes e_1\} + 1_{m,n} \otimes c.$$  

(C.2)

In the continuum, applying an affine transform in this way corresponds to the mapping $x \mapsto A(x - c) + b + c$, hence the name: the image plane is shifted to have its origin at $c$ for the purposes of applying the transform. When we implement affine transforms as suggested by (C.2), we choose $c$ to make the basis vectors orthogonal this necessitates that $c = ((m - 1)/2, (n - 1)/2)$. Then we are able to write down a concrete expression for the projection operator in these coordinates:

$$\text{proj}_{\text{Aff}(2)}(\tau) = \left(\left(m - \frac{m-1}{2}\right)^* \tau 1, 1^* \tau_0 \left(n - \frac{n-1}{2}\right), \left(m - \frac{m-1}{2}\right)^* \tau_1 1, 1^* \tau_0 1, 1^* \tau_1 1\right).$$  

(C.3)

The low-rank structure of the basis vectors implies that this transformation can be computed quite rapidly. Although it may seem we have undertaken this discussion for the sake of mathematical rigor, in our experiments we observe significant computational benefits to centering by the prescription above. For example, when computing with (4.10), using a non-orthogonal basis for the affine transforms (or a center that is not at the center of the region being transformed) often leads to skewing artifacts in the final transformation recovered. We also notice slower convergence.

Finally, for our experiments in Section 4.4 with the rigid motion model $\text{SE}(2)$, some additional discussion is required. This is because the orthogonal transformations $\text{SO}(2)$ are not a linear subspace, like the affine transforms (C.1), but a smooth manifold (diffeomorphic to a circle). For these transformations, we modify the formula (4.1) by differentiating in a parameterization of

\[\text{proj}_{\text{Aff}(2)}(\tau) = \left(\left(m - \frac{m-1}{2}\right)^* \tau 1, 1^* \tau_0 \left(n - \frac{n-1}{2}\right), \left(m - \frac{m-1}{2}\right)^* \tau_1 1, 1^* \tau_0 1, 1^* \tau_1 1\right).\]

In practice, our choice of step size is made to scale each element in this basis to be orthonormal (in particular, applying different steps to the matrix and translation parameters of the transformation)—strictly speaking the projection in (C.3) is not the orthogonal projection because this extra scaling has not been applied. We do not specify this scaling here because its optimal value often depends on the image content: for example, see the step size prescriptions in Theorem 4.5.1.
SE(2): concretely, we use

$$\text{SO}(2) \cong \left\{ \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \middle| \theta \in [0, 2\pi] \right\}.$$ 

Writing $F : \mathbb{R} \rightarrow \mathbb{R}^{m \times n \times 2}$ for this parameterization composed with our usual vector field representation (C.2) for subgroups of the affine transforms, we modify the objective (4.1) to be $\min_\theta \varphi(y \circ F(\theta))$. A simple calculation then shows that gradients in this parameterization are obtainable from gradients with respect to the affine parameterization as

$$\nabla_\theta [\varphi(y \circ F)](\theta) = \nabla_A [\varphi(y \circ \tau, b)] \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} -\sin \theta & -\cos \theta \\ \cos \theta & -\sin \theta \end{pmatrix}.$$ 

This is a minor extra nonlinearity that replaces the proximal operation when we unroll networks as in (4.5) with this motion model. Gradients and projections with respect to the translation parameters are no different from the affine case.

C.1.2 Gradient Calculations for Unrolled Network Architectures

We collect in this section several computations relevant to gradients of the function $\varphi$ (following the structure of (4.1)) in the optimization formulations (4.2), (4.3), (4.4) and (4.10).

**$\tau$ gradients.** All of the costs we consider use the $\ell^2$ error $\| \cdot \|_F$, so their gradient calculations with respect to $\tau$ are very similar. We will demonstrate the gradient calculation for (4.2) to show how (4.6) is derived; the calculations for other costs follow the same type of argument. To be concise, we will write $\nabla_\tau \varphi$ for the gradient with respect to $\tau$ of the relevant costs $\varphi(y \circ \tau)$.

**Proposition C.1.1.** Let $\varphi$ denote the $\| \cdot \|_F$ cost in (4.2). One has

$$\nabla_\tau \varphi(\tau) = \sum_{k=0}^{c-1} (g_{\sigma^2} \ast \mathcal{P}_{\Omega} \{ y \circ (\tau - x_0) \}) \otimes I_2 \circ (dy_k \circ \tau).$$
Proof. The cost separates over channels, so by linearity of the gradient it suffices to assume $c = 1$. We proceed by calculating the differential of $\varphi(y \circ \tau)$ with respect to $\tau$. We have for $\Delta$ of the same shape as $\tau$ and $t \in \mathbb{R}$

$$
\frac{\partial}{\partial t} \bigg|_{t=0} \varphi(y \circ (\tau + t\Delta)) = \langle P_\Omega \left[ g_{\sigma^2} * (y \circ (\tau - x_o)) \right] \otimes 1_2, P_\Omega \left[ g_{\sigma^2} * ((dy \circ \tau) \otimes \Delta) \right] \rangle,
$$

where $dy \in \mathbb{R}^{m \times n \times 2}$ is the Jacobian matrix of $y$, defined as (here $\phi$ denotes the derivative of the cubic convolution interpolation kernel $\phi$)

$$
dy_0 = \sum_{(k,l) \in \mathbb{Z}^2} y_{kl} \phi(m^* - k1) \odot \phi(1n^* - l1), \quad dy_1 = \sum_{(k,l) \in \mathbb{Z}^2} y_{kl} \phi(m^* - k1) \odot \phi(1n^* - l1),
$$

and where for concision we are writing $g_{\sigma^2} * dy$ to denote the filtering of each of the two individual channels of $dy$ by $g_{\sigma^2}$. Using three adjoint relations ($P_\Omega$ is an orthogonal projection, hence self-adjoint; the adjoint of convolution by $g_{\sigma^2}$ is cross-correlation with $g_{\sigma^2}$; elementwise multiplication is self-adjoint) and a property of the tensor product, the claim follows. □

**Convolutional representation of cost-smoothed formulation (4.3).** The cost-smoothed formulation (4.3) can be directly expressed as a certain convolution with $g_{\sigma^2}$, leading to very fast convolution-free inner loops in gradient descent implementation. To see this, write

$$
\|P_\Omega \left[ y \circ (\tau + \tau_{0,\Delta}) - x_o \right]\|_F^2
= \|P_\Omega \left[ y \circ (\tau + \tau_{0,\Delta}) \right]\|_F^2 + \|P_\Omega \left[ x_o \right]\|_F^2 + 2\langle P_\Omega \left[ y \circ (\tau + \tau_{0,\Delta}) \right], P_\Omega \left[ x_o \right] \rangle
= \left( [y \circ (\tau + \tau_{0,\Delta})]^{\odot 2}, P_\Omega \left[ 1 \right] \right) + \|P_\Omega \left[ x_o \right]\|_F^2 + 2\langle y \circ (\tau + \tau_{0,\Delta}), P_\Omega \left[ x_o \right] \rangle,
$$

using self-adjointness of $P_\Omega$ and the fact that it can be represented as an elementwise multiplication, and writing $[ \cdot ]^{\odot 2}$ for elementwise squaring. Thus, denoting the $\| \cdot \|_F$ cost in (4.3) by $\varphi(\tau)$,
\( \varphi \) can be written as

\[
2\varphi(\tau) = \left( \sum_{\Lambda} (g_{\sigma^2})_\Lambda \left[ y \circ (\tau + \tau_{0,\Lambda}) \right] \right) \odot \mathcal{P}_\Omega [1] + \langle g_{\sigma^2}, 1 \rangle \| \mathcal{P}_\Omega [x_o] \|_F^2
+ 2 \left( \sum_{\Lambda} (g_{\sigma^2})_\Lambda \cdot y \circ (\tau + \tau_{0,\Lambda}), \mathcal{P}_\Omega [x_o] \right).
\]

This can be expressed as a cross-correlation with \( g_{\sigma^2} \):

\[
2\varphi(\tau) = \langle g_{\sigma^2} \ast [y \circ \tau] \odot \mathcal{P}_\Omega [1] \rangle + \langle g_{\sigma^2}, 1 \rangle \| \mathcal{P}_\Omega [x_o] \|_F^2 + 2 \langle g_{\sigma^2} \ast (y \circ \tau), \mathcal{P}_\Omega [x_o] \rangle.
\]

and taking adjoints gives finally

\[
2\varphi(\tau) = \langle [y \circ \tau] \odot g_{\sigma^2} \ast \mathcal{P}_\Omega [1] \rangle + \langle g_{\sigma^2}, 1 \rangle \| \mathcal{P}_\Omega [x_o] \|_F^2 + 2 \langle y \circ \tau, g_{\sigma^2} \ast \mathcal{P}_\Omega [x_o] \rangle.
\]

This gives a convolution-free gradient step implementation for this cost (aside from pre-computing the fixed convolutions in the cost), and also yields a useful interpretation of the cost-smoothed formulation (4.3), and its disadvantages relative to the background-modeled formulation (4.4).

**Filter gradient for complementary smoothing formulation (4.10).** Relative to the standard registration model formulation (4.2), the complementary smoothing spike registration formulation (4.10) contains an extra complicated transformation-dependent gaussian filter. We provide a key lemma below for the calculation of the gradient with respect to the parameters of the complementary smoothing cost in “standard parameterization” (see the next paragraph below). The full calculation follows the proof of Proposition C.1.1 with an extra “product rule” step and extra adjoint calculations.

**Proposition C.1.2.** Given fixed \( \sigma^2 > \sigma_{0}^2 > 0 \), define \( \Sigma(A) = \sigma^2 I - \sigma_{0}^2 (A^*A)^{-1} \), and define

\[
g(A) = \sqrt{\det(A^*A)} g_{\Sigma(A)},
\]
where the filter is $m \times n$ and the domain is the open set \( \{ A \mid \sigma I - \sigma_0^2 (A^*A)^{-1} > 0 \} \). Then for any fixed \( V \in \mathbb{R}^{m \times n} \), one has
\[
\nabla_A [\langle V, g \rangle] (A) = \sigma_0^2 A^{-*} \left( \Sigma(A)^{-1} \sum_{i,j} V_{ij} g(A)_{ij} w_{ij}^* \right) \Sigma(A)^{-1} - \langle g(A), V \rangle \Sigma(A)^{-1} \right) (A^*A)^{-1} \\
+ \langle g(A), V \rangle A^{-*},
\]
where \( A^{-*} = (A^{-1})^* \).

Proof. For \((i, j) \in \{0, \ldots, m-1\} \times \{0, \ldots, n-1\} \), let \( w_{ij} = (i - \lfloor m/2 \rfloor, j - \lfloor n/2 \rfloor) \). Then we have
\[
g(A) = \frac{1}{2\pi} \sum_{i,j} e_{ij} \exp \left( -\frac{1}{2} w_{ij}^* \Sigma(A)^{-1} w_{ij} - \frac{1}{2} \log \det \Sigma(A) + \frac{1}{2} \log \det A^*A \right).
\]

Let \( dg \) denote the differential of \( A \mapsto g(A) \). By the chain rule, we have for any \( \Delta \in \mathbb{R}^{2 \times 2} \)
\[
\langle V, dg_A(\Delta) \rangle = \frac{1}{2} \sum_{i,j} V_{ij} g(A)_{ij} \frac{\partial}{\partial t} \left|_{t=0} \right. \left[ -w_{ij}^* \Sigma(A + t\Delta)^{-1} w_{ij} - \log \det \Sigma(A + t\Delta) \\
+ \log \det(A + t\Delta)^* (A + t\Delta) \right].
\]

We need the differential of several mappings here. We will use repeatedly that if \( X \in \text{GL}(2) \) and \( W \in \mathbb{R}^{2 \times 2} \), one has
\[
d[ X \mapsto \langle W, X^{-1} \rangle ]_X(\Delta) = -\langle \Delta, X^{-*}WX^{-*} \rangle. \tag{C.4}
\]

Applying (C.4) and the chain rule, we get
\[
d[ \langle W, \Sigma \rangle ]_A(\Delta) = \sigma_0^2 \langle (A^*A)^{-1} W (A^*A)^{-1}, A^*A + A^*A \rangle \\
= \sigma_0^2 \langle A^{-*} (W + W^*) (A^*A)^{-1}, \Delta \rangle. \tag{C.5}
\]
In particular, using the chain rule and (C.4) and (C.5) gives

$$\frac{\partial}{\partial t} \left|_{t=0} \right. \left[ w_{ij}^* \Sigma(A + t\Delta)^{-1} w_{ij} \right] = -2\sigma_0^2 \left\langle A^{-*} \Sigma(A)^{-1} w_{ij} w_{ij}^* \Sigma(A)^{-1} (A^* A)^{-1}, \Delta \right\rangle. \quad (C.6)$$

Next, using the Leibniz formula for the determinant, we obtain

$$d[\log \det X(\Delta)] = \langle X^{-s}, \Delta \rangle. \quad (C.7)$$

The chain rule and (C.7) and (C.5) thus give

$$\frac{\partial}{\partial t} \left|_{t=0} \right. \left[ \log \det \Sigma(A + t\Delta) \right] = 2\sigma_0^2 \left\langle A^{-*} \Sigma(A)^{-1} (A^* A)^{-1}, \Delta \right\rangle, \quad (C.8)$$

and similarly

$$\frac{\partial}{\partial t} \left|_{t=0} \right. \left[ \log \det (A + t\Delta)^*(A + t\Delta) \right] = 2\langle A^{-*}, \Delta \rangle. \quad (C.9)$$

Combining (C.6), (C.8) and (C.9), we have

$$\langle V, dg_A(\Delta) \rangle = \sum_{i,j} V_{ij} g(A)_{ij} \left( \sigma_0^2 A^{-*} \left( \Sigma(A)^{-1} w_{ij} w_{ij}^* \Sigma(A)^{-1} - \Sigma(A)^{-1} \right) (A^* A)^{-1} + A^{-*}, \Delta \right),$$

and the claim follows by distributing and reading off the gradient. \(3\) \(\square\)

**Differentiating costs in “inverse parameterization”**. Our theoretical study of spike alignment in Section C.2 and our experiments on the discretized objective (4.10) in Section 4.5.2 suggest strongly to prefer “inverse parameterization” relative to standard parameterization of affine transformations for optimization. By this, we mean the following: given a cost \(\varphi(\tau_{A,b})\) optimized over affine transformations \((A,b)\), one optimizes instead \(\varphi(\tau_{A^{-1},-A^{-1}b})\). This nomenclature is motivated by, in the continuum, the inverse of the affine transformation \(x \mapsto Ax + b\) being \(x \mapsto A^{-1}(x - b)\). Below, we show the chain rule calculation that allows one to easily obtain

---

3After distributing, the sum over \(i, j\) in the first factor can be computed relatively efficiently using a Kronecker product.
Proposition C.1.3. Let \( \varphi : \mathbb{R}^{2 \times 2} \times \mathbb{R}^2 \to \mathbb{R} \), and let \( F(A, b) = (A^{-1}, -A^{-1}b) \) denote the inverse parameterization mapping, defined on \( \text{GL}(2) \times \mathbb{R}^2 \). One has

\[
\nabla_A [\varphi \circ F](A, b) = -A^{-*} (\nabla_A [\varphi] \circ F(A, b)) A^{-*} + A^{-*} (\nabla_b [\varphi] \circ F(A, b)) (A^{-1}b)^*,
\]

\[
\nabla_b [\varphi \circ F](A, b) = -A^{-*} (\nabla_b [\varphi] \circ F(A, b)),
\]

where \( A^{-*} = (A^{-1})^* \).

Proof. Let \( d[\varphi \circ F] \) denote the differential of \( \varphi \circ F \) (and so on). We have for \( \Delta_A \) and \( \Delta_b \) the same shape as \( A \) and \( b \)

\[
dF_{A,b}(\Delta_A, \Delta_b) = \frac{\partial}{\partial t} \bigg|_{t=0} \left( (A + t\Delta_A)^{-1}, -(A + t\Delta_A)^{-1}(b + t\Delta_b) \right)
\]

\[
= \left( -A^{-1}\Delta_A A^{-1}, -\left( A^{-1}\Delta_b - A^{-1}\Delta_A A^{-1}b \right) \right)
\]

where the asserted expression for the derivative through the matrix inverse follows from, say, the Neumann series. Now, the chain rule and the definition of the gradient imply

\[
d[\varphi \circ F]_{A,b}(\Delta_A, \Delta_b) = \langle \nabla_A [\varphi \circ F](A, b), -A^{-1}\Delta_A A^{-1} \rangle
\]

\[
+ \langle \nabla_b [\varphi \circ F](A, b), -\left( A^{-1}\Delta_b - A^{-1}\Delta_A A^{-1}b \right) \rangle,
\]

and the claim follows by distributing and taking adjoints in order to read off the gradients from the previous expression. \( \square \)

We remark that centering, as discussed in Section C.1.1, can be implemented identically to the standard parameterization case when using inverse parameterization.
C.1.3 Additional Experiments and Experimental Details

General details for experiments. We use normalized cross correlation (NCC) and zero normalized cross correlation (ZNCC) for measuring the performance of registration on textured and spike data respectively. Specifically, for two multichannel images $X, Y \in \mathbb{R}^{m \times n \times c}$, let \( \hat{X} \) and \( \hat{Y} \) be the channel-wise mean-subtracted images from $X$ and $Y$. The quantities NCC and ZNCC are defined as

$$\text{NCC}(X, Y) = \frac{\langle X, Y \rangle}{\|X\|_F \|Y\|_F}$$

and

$$\text{ZNCC}(X, Y) = \frac{\langle \hat{X}, \hat{Y} \rangle}{\|\hat{X}\|_F \|\hat{Y}\|_F}.$$ 

Figure 4.1 Experimental Details

In the experiment comparing the complexity of optimization and covering-based methods for textured motif detection shown in Figure 4.1, the raw background image used has dimension 2048 x 1536, and a random section of dimension 600 x 500 is selected every run. The crab template is first placed at the center of the selected section, then a random transformation is applied to generate the scene $y$. Translation consists of random amounts on both $x$ and $y$ directions uniformly in $[-5, 5]$ pixels. Euclidean transforms in addition apply a rotation with angle uniformly from $[-\frac{\pi}{4}, \frac{\pi}{4}]$. Similarity transforms in addition applies a scaling uniformly from $[0.8, 1.25]$. Generic affine transforms are parameterized by a transformation matrix $A \in \mathbb{R}^{2 \times 2}$ and offset vector $b \in \mathbb{R}^2$, with the singular values of $A$ uniformly from $[0.8, 1.25]$ and the left and right orthogonal matrices being rotation matrices with angle uniformly in $[-\frac{\pi}{4}, \frac{\pi}{4}]$. For each of the 4 modes of transform, 10 random images are generated. The optimization formulation used is (4.4), with $x_o$ the crab body motif shown in Figure 4.2(a). The optimization-based method uses a multi-scale scheme, which uses a sequence of decreasing values of $\sigma$ and step sizes, starting at $\sigma = 5$ and step size $0.005\sigma$ (except for affine mode which starts at $\sigma = 10$), with $\sigma$ halved every 50 iterations until stopping criteria over the ZNCC is met, where ZNCC is calculated over the motif support $\Omega$ only. For each value of $\sigma$, a dilated support $\Omega$ is used, which is the dilation of $\Omega$ two $\sigma$ away from the support of the motif. The background model covers the region up to $5\sigma$ away from the motif. The background $\beta$ is initialized as a gaussian-smoothed version of the difference between the initialized image and the ground truth motif, and then continuously updated in the optimization. For the first 5 iterations
of every new scale $\sigma$, only the background is updated while the transformation parameters are held constant. The covering-based method samples a random transform from the corresponding set of transforms used in each try.

Figure 4.4 Experimental Details

In the experiment of verifying the convergence of multichannel spike registration as shown in Figure 4.4, the motif consists of 5 spikes placed at uniformly random positions in a $61 \times 81$ image. To allow the spike locations to take non-integer values, we represent each spike as a gaussian density with standard deviation $\sigma_0 = 3$ centered at the spike location, and evaluated on the grid. A random affine transformation of the motif is generated as the scene. As a result, we are able to use this $\sigma_0$-smoothed input in (4.10) without extra smoothing, and we can compensate the variance of the filter applied to $x_0$ in the formulation to account for the fact that we already smoothed by $\sigma_0$ when generating the data. The smoothing level in the registration is chosen according to equation (4.16) in Theorem 4.5.1. Due to the discretization effect and various artifacts, the step sizes prescribed in Theorem 4.5.1 will lead to divergence, so we reduce the step sizes by multiplying a factor of 0.5.

Further Experimental Details

The beach background used for embedding the crab template throughout the experiments is CC0-licensed and available online: https://www.flickr.com/photos/scotnelson/28315592012. Our code and data are available at https://github.com/sdbuch/refine.

C.1.4 Canonized Object Preprocessing and Calibration for Hierarchical Detection

The hierarchical detection network implementation prescription in Section 4.4.3 assumes the occurrence maps $x_v$ for $v \in V \setminus \{1, \ldots, K\}$ are given; in practice, these are first calculated using the template $y_o$ and its motifs, by a process we refer to as extraction. Simultaneously, to extract
these occurrence maps and have them be useful for subsequent detections it is necessary to have appropriate choices for the various hyperparameters involved in the network: we classify these as ‘registration’ hyperparameters (for each \( v \in V \), the step size \( \nu_v \); the image, scene, and input smoothing parameters \( \sigma_v^2, \sigma_0^2, \) and \( \sigma_{\text{in}}^2 \); the number of registration iterations \( T_v \); and the vertical (“height”) and horizontal (“width”) stride sizes \( \Delta H_v \) and \( \Delta W_v \)) or ‘detection’ hyperparameters (for each \( v \in V \), the suppression parameter \( \alpha_v \) and the threshold parameter \( \gamma_v \)). We describe these issues below, as well as other relevant implementation issues.

**Hyperparameter selection.** We discuss this point first, because it is necessary to process the ‘leaf’ motifs before any occurrence maps can be extracted. In practice, we ‘calibrate’ these hyperparameters by testing whether detections succeed or fail given the canonized template \( \mathbf{y}_o \) as input to the (partial) network. Below, we first discuss hyperparameters related to visual motifs (i.e., the formulation (4.12)), then hyperparameters for spiky motifs (i.e., the formulation (4.13)).

**Stride density and convergence speed:** The choice of these parameters encompasses a basic computational tradeoff: setting \( T_v \) larger allows to leverage the entire basin of attraction of the formulations (4.12) and (4.13), enabling more reliable values of \( \min_{\lambda \in \Lambda_v} \text{loss}(v, \lambda) \) and the use of larger values of \( \Delta H_v \) and \( \Delta W_v \); however, it requires more numerical operations (convolutions and interpolations) for each stride \( \lambda \in \Lambda_v \). In our experiments we err on the side of setting \( T_v \) large, and tune the stride sizes \( \Delta W_v \) and \( \Delta H_v \) over multiples of 4 (setting them as large as possible while being able to successfully detect motifs). The choice of the step sizes \( \nu_v \) is additionally complicated by the smoothing and motif-dependence of this parameter. As we describe in Section 4.3.3, we treat the step sizes taken on each component of \((A, b)\) independently, and in our experiments use a small multiple (i.e., \(10^{-1}\) for spiky motifs and \(10^{-4}\) for textured motifs) of the theoretical prescriptions in Section 4.5 for spike alignment that we discuss later in this section.

**Smoothing parameters:** The smoothing level \( \sigma_v^2 \) in (4.12) increases the size of the basin of attraction when set larger. For this specific formulation, we find it more efficient to expand
the basin by striding, and enforce a relatively small value of $\sigma_v^2 = 9$ for all visual motifs. Without input smoothing, we empirically observe that the first-round-multiscale cost-smoothed formulation (4.12) is slightly unstable with respect to high-frequency content in $y$: this motivates us to introduce this extra smoothing with variance $\sigma_{in}^2 = 9/4$, which removes interpolation artifacts that hinder convergence. We find the multiscale smoothing mode of operation described in Section 4.4.3 to be essential for distinguishing between strides $\lambda$ which have “failed” to register the motif $x_v$ and those that have succeeded, through the error loss($\lambda, v$): in all experiments, we run the second-phase multiscale round for (4.12) as described in Section 4.4.3, for 256 iterations and with $\sigma^2 = 10^{-2}$ and $\sigma_{in}^2 = 10^{-12}$. We describe the choice of $\sigma_{0,v}^2$ below, as it is more of a spike registration hyperparameter (c.f. (4.14)).

**Detection parameters:** The scale parameters $\alpha_v$ are set based on the size of the basin of attraction around the true transformation of $x_v$, and in particular on the scale of loss($\lambda, v$) at “successes” and ”failures” to register. In our experiments, we simply set $\alpha_v = 1$ for visual motifs. The choice of the threshold parameter $\gamma_v$ is significantly more important: it accounts for the fact that the final cost loss($\lambda, v$) at a successful registration is sensitive to both the motif $x_v$ and the background/visual clutter present in the input $y$. In our experiments in Section 4.4.4, we tune the parameters $\gamma_v$ on a per-motif basis by calculating loss($\lambda, v$) for embeddings $y_o \circ \tau_0$ for $\tau_0 \in \text{SO}(2)$ up to some maximum rotation angle in visual clutter, classifying each $\lambda$ as either a successful registration or a failure, and then picking $\gamma_v$ to separate the successful runs for all rotation angles from the failing runs. For the motifs and range of rotation angles we consider, we find that such a threshold always exists. However, at larger rotation angles we run into issues with the left and right eye motifs being too similar to each other, leading to spurious registrations and the non-existence of a separating threshold. In practice, this calibration scheme also requires a method of generating visual clutter that matches the environments one intends to evaluate in. The calibrated threshold parameters used for our experiments in Section 4.4.4 are available in our released implementation.
**Hyperparameters for spiky motifs:** The same considerations apply to hyperparameter selection for spiky motifs (i.e., the formulation (4.13)). However, the extra structure in such data facilitates a theoretical analysis that corroborates the intuitive justifications for hyperparameter tradeoffs we give above and leads to specific prescriptions for most non-detection hyperparameters, allowing them to be set in a completely tuning-free fashion. We present these results in Section 4.5. For detection hyperparameters, we follow the same iterated calibration process as for visual motifs, with scale parameters $\alpha_v = 2.5 \cdot 10^5$ (typical values of the cost (4.13) are much smaller than those of the cost (4.12), due to the fact that the gaussian density has a small $L^2$ norm). For the occurrence map smoothing parameters $\sigma^2_{0,\nu}$, our network construction above necessitates setting these parameters to be the same for all $\nu \in V$; we find empirically that a setting $\sigma^2_{0,\nu} = 9$ is sufficient to avoid interpolation artifacts. Finally, the bounding box masks $\Omega_\nu$ are set during the extraction process (see below), and are dilated by twice the total size of the filters $g_{\sigma_\nu^2}$. In practice, when implementing gaussian filters, we make the image size square, with side lengths $6\sigma$ (rounded to the next largest odd integer).

**Occurrence map extraction.** Although the criteria above (together with the theoretical guidance from Section 4.5) are sufficient to develop a completely automatic calibration process for the various hyperparameters above, in practice we perform calibration and occurrence map extraction in a ‘human-in-the-loop’ fashion. The extraction process can be summarized as follows (it is almost identical to the detection process described in Section 4.4.3, with a few extra steps implicitly interspersed with calibration of the various hyperparameters):

1. **Use the canonized template as input:** We set $y_o$ as the network’s input.

2. **Process leaf motifs:** Given suitable calibrated settings of the hyperparameters for leaf motifs $\nu \in V$, perform detection and generate all occurrence maps $\omega_\nu$ via (4.14).

3. **Extract occurrence motifs at depth $\text{diam}(G) - 1$:** For each $\nu$ with $d(\nu) = \text{diam}(G) - 1$, we follow the assumptions made in Section 4.4.1 (in particular, that each visual motif occurs...
only once in \( y_o \) and \( G \) is a tree) and after aggregating the occurrence map from \( v \)'s child nodes via (4.11), we extract \( x_v \) from \( y_v \) by cropping to the bounding box for the support of \( y_v \). Technically, since (4.14) uses a gaussian filter, the support will be nonzero everywhere, and instead we threshold at a small nonzero value (e.g. 1/20 in our experiments) to determine the “support”.

4. **Continue to the root of \( G \):** Perform registration to generate the occurrence maps for nodes at depth \( \text{diam}(G) - 1 \), then continue to iterate the above steps until the root node is reached and processed.

Note that the extracted occurrence motifs \( x_v \) for \( v \in V \setminus \{1, \ldots, K\} \) depend on proper settings of the registration and detection hyperparameters: if these parameters are set imprecisely, the extracted occurrence maps will not represent ideal detections (e.g. they may not be close to a full-amplitude gaussian at the locations of the motifs in \( y_o \) as they should, or they may not suppress failed detections enough).

**Other implementation issues.** The implementation issue of centering, discussed previously in Section C.1.1, is relevant to the implementation of the unrolled solvers for (4.13) and (4.14). We find that a useful heuristic is to center the transformation \( \tau \) at the location of the center pixel of the embedded motif \( x_v \) (i.e., for a stride \( \Lambda \in \Lambda_v \), at the coordinates \( \Lambda + ((m_v - 1)/2, (n_v - 1)/2) \)). To implement this centering, the locations of the detections in the spike map definition (4.14) need to have the offsets \( ((m_v - 1)/2, (n_v - 1)/2) \) added.

The network construction in Section 4.4.3 relies on the extraction process described above to employ an identical enumeration strategy in the traversal of the graph \( G \) as the detection process (i.e., assuming that nodes are ordered in increasing order above). In our implementation described in Section 4.4.4, we instead label nodes arbitrarily when preparing the network’s input, and leave consistent enumeration of nodes during traversal to the NetworkX graph processing library [236].
C.2 Proof of Theorem 4.5.1

We consider a continuum model for multichannel spike alignment, motivated by the higher-level features arising in the hierarchical detection network developed in Section 4.4: signals $X$ are represented as elements of $\mathbb{R}^{R \times R \times C}$, and are identifiable with $C$-element real-valued vector fields on the (continuous, infinite) image plane $\mathbb{R}^2$. In this setting, we write $\|X\|_{L^2}^2 = \sum_{i=1}^C \|X_i\|_{L^2}^2$ for the natural product norm (in words, the $\ell^2$ norm of the vector of channelwise $L^2$ norms of $X$). For $p \in \mathbb{R}^2$, let $\delta_p \in \mathbb{R}^{R \times R}$ denote a Dirac distribution centered at $p$, defined via

$$\int_{\mathbb{R}^2} \delta_p(x) f(x) \, dx = f(p)$$

for all Schwartz functions $f$ [234, §I.3]. This models a ‘perfect’ spike signal. For $p \in \mathbb{R}^2$ and $M \in \mathbb{R}^{2 \times 2}$ positive semidefinite, let $g_{p,M}$ denote the gaussian density on $\mathbb{R}^2$ with mean $p$ and covariance matrix $M$. Consider a target signal

$$X_o = \sum_{i=1}^C \delta_{v_i} \otimes e_i, \quad (C.10)$$

and an observation

$$X = \sum_{i=1}^C \delta_{u_i} \otimes e_i \quad (C.11)$$

satisfying

$$v_i = A_* u_i + b_* \quad (C.12)$$

for some $(A_*, b_*) \in \text{GL}(2) \times \mathbb{R}^2$. These represent the unknown ground-truth affine transform to be recovered. Consider the objective function

$$\varphi_{L^2, \sigma}(A, b) \equiv \frac{1}{2C} \left\| g_{0, \sigma^2 I - \sigma_0^2 (A^* A)^{-1}} \left( \det^{1/2}(A^* A) \left( g_{0, \sigma_0^2 I} * X \right) \circ \tau_{A, b} \right) - g_{0, \sigma^2 I} * X_o \right\|_{L^2}^2,$$

where $A^*$ denotes the transpose, convolutions are applied channelwise, and for a signal $S \in \mathbb{R}^{R \times R \times c}$, $S \circ \tau_{A,b}(u,v) = S(a_{11} u + a_{12} v + b_1, a_{21} u + a_{22} v + b_2)$. We study the following “in-
verse parameterization” of this function:

\[ \varphi_{L^2,\sigma}^{\text{inv}}(A, b) \equiv \varphi_{L^2,\sigma}(A^{-1}, -A^{-1}b). \]

We analyze the performance of gradient descent for solving the optimization problem

\[ \min_{A, b} \varphi_{L^2,\sigma}^{\text{inv}}(A, b). \]

Under mild conditions, local minimizers of this problem are global. Moreover, if \( \sigma \) is set appropriately, the method exhibits linear convergence to the truth:

**Theorem C.2.1 (Multichannel Spike Model, Affine Transforms, \( L^2 \)).** Consider an instance of the multichannel spike model (C.10)-(C.11)-(C.12), with \( U = [u_1, \ldots, u_c] \in \mathbb{R}^{2 \times c} \). Assume that the spikes \( U \) are centered and nondegenerate, so that \( U1 = 0 \) and \( \text{rank}(U) = 2 \). Then gradient descent

\[
A_{k+1} = A_k - \nu t_A \nabla_A \varphi_{L^2,\sigma}^{\text{inv}}(A_k, b_k),
\]

\[
b_{k+1} = b_k - \nu t_b \nabla_b \varphi_{L^2,\sigma}^{\text{inv}}(A_k, b_k)
\]

with smoothing

\[ \sigma^2 \geq \frac{\max_i \|u_i\|^2}{s_{\min}(U)^2} \left( \frac{1}{s_{\max}(U)^2} \right)^2 + c \|b_*\|^2 \]

and step sizes

\[
t_A = \frac{c}{s_{\max}(U)^2},
\]

\[
t_b = 1,
\]

\[
\nu = 8\pi a^4,
\]
from initialization $A_0 = I, b_0 = 0$ satisfies

$$t_A^{-1} \|A_k - A_*\|_F^2 + \|b_k - b_*\|_2^2 \leq \left( 1 - \frac{1}{2k} \right)^{2k} \left( t_A^{-1} \|I - A_*\|_F^2 + \|b_*\|_2^2 \right), \quad (C.13)$$

where

$$\kappa = \frac{s_{\max}(U)^2}{s_{\min}(U)^2},$$

with $s_{\min}(U)$ and $s_{\max}(U)$ denoting the minimum and maximum singular values of the matrix $U$.

Proof. Below, we use the notation $\|M\|_{\ell_p \to \ell_q} = \sup_{\|x\|_{\ell_p} \leq 1} \|Mx\|_{\ell_q}$. We begin by rephrasing the objective function in a simpler form: by properties of the gaussian density,

$$\varphi_{L_2,\sigma}(A, b) \, = \, \frac{1}{2c} \sum_{i=1}^{c} \left\| g_{A^{-1}(u_i - b), \sigma^2 I} - g_{v_i, \sigma^2 I} \right\|_{L_2}^2,$$

whence by an orthogonal change of coordinates

$$\varphi_{L_2,\sigma}(A, b) \, = \, \frac{1}{c} \sum_{i=1}^{c} \psi \left( \frac{1}{2} \|A^{-1}(u_i - b) - v_i\|_2^2 \right),$$

where

$$\psi(t^2/2) \, = \, \frac{1}{2} \left\| g_{(t e_1, \sigma^2 I) - g_{(0, \sigma^2 I)} \right\|_{L_2}^2 \, = \, \frac{1}{4\pi \sigma^2} - \left\langle g_{(t e_1, \sigma^2 I), g_{(0, \sigma^2 I)} \right\rangle \, = \, \frac{1}{4\pi \sigma^2} - \frac{1}{(2\pi \sigma^2)^2} \left( \int_{\mathbb{R}} e^{-s^2/\sigma^2} ds \right) \left( \int_{\mathbb{R}} e^{-(s-t)^2/2\sigma^2} e^{-s^2/2\sigma^2} ds \right) \, = \, \frac{1}{4\pi \sigma^2} - \frac{2^{-1/2}}{(2\pi \sigma^2)^{3/2}} \int_{\mathbb{R}} e^{-(s-t)^2/2\sigma^2} e^{-s^2/4\sigma^2} ds \, = \, \frac{1}{4\pi \sigma^2} \left( 1 - \exp \left( -\frac{t^2/2}{2\sigma^2} \right) \right).$$

So

$$\varphi_{L_2,\sigma}^{inv}(A, b) = \varphi_{L_2,\sigma}(A^{-1}, -A^{-1} b) = \frac{1}{c} \sum_{i=1}^{c} \psi \left( \frac{1}{2} \|A u_i + b - v_i\|_2^2 \right).$$
Differentiating, we obtain

\[ \nabla_A \varphi_{L^2,\sigma}^{\text{inv}}(A, b) = \frac{1}{c} \sum_{i=1}^{c} \psi \left( \frac{1}{2} \| \delta_i \|_2^2 \right) \delta_i u_i^* \]

\[ \nabla_b \varphi_{L^2,\sigma}^{\text{inv}}(A, b) = \frac{1}{c} \sum_{i=1}^{c} \psi \left( \frac{1}{2} \| \delta_i \|_2^2 \right) \delta_i, \]

where for concision

\[ \delta_i = Au_i + b - v_i \]

\[ = (A - A_*)u_i + b - b_* . \]

In these terms, we have the following expression for a single iteration of gradient descent:

\[ A^* = A - \frac{t_A}{c} \sum_{i=1}^{c} v\psi \left( \frac{1}{2} \| \delta_i \|_2^2 \right) \delta_i u_i^* \]

\[ = A - \frac{t_A}{c} \sum_{i=1}^{c} v\psi \left( \frac{1}{2} \| \delta_i \|_2^2 \right) (A - A_*)u_i u_i^* - \frac{t_A}{c} \sum_{i=1}^{c} v\psi \left( \frac{1}{2} \| \delta_i \|_2^2 \right) (b - b_*)u_i^* \]

\[ = A - \frac{vt_A}{c} (A - A_*) U \Psi U^* \] - \frac{vt_A}{c} (b - b_*) \Psi^* U^* \]

and

\[ b^* = b - \frac{t_b}{c} \sum_{i=1}^{c} v\psi \left( \frac{1}{2} \| \delta_i \|_2^2 \right) \delta_i \]

\[ = b - \frac{t_b}{c} (b - b_*) \langle 1, \nu \psi \rangle - \frac{vt_b}{c} (A - A_*) U \Psi , \]

where above, we have set

\[ \Psi = \begin{bmatrix} \psi \left( \frac{1}{2} \| \delta_1 \|_2^2 \right) \\ \vdots \\ \psi \left( \frac{1}{2} \| \delta_c \|_2^2 \right) \end{bmatrix} \in \mathbb{R}^{c \times c} , \quad \Psi = \begin{bmatrix} \psi \left( \frac{1}{2} \| \delta_1 \|_2^2 \right) \\ \vdots \\ \psi \left( \frac{1}{2} \| \delta_c \|_2^2 \right) \end{bmatrix} \in \mathbb{R}^c . \quad (C.14) \]
Writing $\Delta_A = A - A_\star$, $\Delta_b = b - b_\star$, we have

\[
\begin{align*}
\Delta_A^+ &= \Delta_A \left( I - \frac{v_A}{c} U \Psi^* U^* \right) - \frac{v_A}{c} \Delta_b \psi^* U^* \\
\Delta_b^+ &= \left( 1 - \frac{t_b}{c} \langle 1, \nu \psi \rangle \right) \Delta_b - \Delta_A \frac{v_b}{c} U \psi.
\end{align*}
\]

To facilitate a convergence proof, we modify this equation to pertain to scaled versions of $\Delta_A$, $\Delta_b$:

\[
\begin{align*}
t^{-1/2}_A \Delta_A^+ &= \left( t^{-1/2}_A \Delta_A \right) \left( I - \frac{t_A}{c} U (\nu \Psi) U^* \right) - \frac{t^{1/2}_A t^{1/2}_b}{c} (t^{-1/2}_b \Delta_b) (\nu \psi)^* U^* \\
t^{-1/2}_b \Delta_b^+ &= \left( 1 - \frac{t_b}{c} \langle 1, \nu \psi \rangle \right) \left( t^{-1/2}_b \Delta_b \right) - \left( t^{-1/2}_A \Delta_A \right) \frac{t^{1/2}_b t^{1/2}_A}{c} U (\nu \psi).
\end{align*}
\]

In matrix-vector form, and writing $\tilde{\Delta}_A = t^{-1/2}_A \Delta_A$ and $\tilde{\Delta}_b = t^{-1/2}_b \Delta_b$, we have

\[
\begin{align*}
\begin{bmatrix} \vec(\tilde{\Delta}_A) \\ \tilde{\Delta}_b \end{bmatrix}^+ &= \left( I_6 - \left[ \frac{t_A}{c} U (\nu \Psi) U^* \otimes I_2 \quad \frac{t^{1/2}_A t^{1/2}_b}{c} (U(\nu \psi)) \otimes I_2 \right] \right) \begin{bmatrix} \vec(\tilde{\Delta}_A) \\ \tilde{\Delta}_b \end{bmatrix} \\
&= \left( I_6 - \left[ \frac{t_A}{c} U (\nu \Psi) U^* \quad \frac{t^{1/2}_A t^{1/2}_b}{c} (U(\nu \psi)) \right] \otimes I_2 \right) \begin{bmatrix} \vec(\tilde{\Delta}_A) \\ \tilde{\Delta}_b \end{bmatrix} \\
&= M \begin{bmatrix} \vec(\tilde{\Delta}_A) \\ \tilde{\Delta}_b \end{bmatrix}, \quad (C.15)
\end{align*}
\]

where in this context $\otimes$ denotes the Kronecker product of matrices. Since $I_6 = I_4 \otimes I_2$, and because the eigenvalues of a Kronecker product of symmetric matrices are the pairwise products of the eigenvalues of each factor, we have

\[
\| M \|_{\ell^2 \to \ell^2} = \left\| I - \left[ \frac{t_A}{c} U (\nu \Psi) U^* \quad \frac{t^{1/2}_A t^{1/2}_b}{c} (U(\nu \psi)) \right] \otimes I_2 \right\|_{\ell^2 \to \ell^2}.
\]
By our choice of \( t_A \) and \( t_b \), and the assumption \( U1 = 0 \), we can write

\[
\begin{bmatrix}
\frac{t_A}{c} U (v \tilde{\Psi}) U^* & \frac{t_A}{c} (U(v \tilde{\Psi})) \\
\frac{t_A}{c} b & \langle 1, v \tilde{\Psi} \rangle
\end{bmatrix}
= \begin{bmatrix}
\frac{t_A}{c} U (v \tilde{\Psi} - I) U^* + \frac{UU^*}{\|U\|_{\ell^2-\ell^2}} & \frac{t_A}{c} (U(v \tilde{\Psi} - I)) \\
\frac{t_A}{c} b & \langle 1, v \tilde{\Psi} - 1 \rangle + 1
\end{bmatrix}
\]

and so by the triangle inequality for the operator norm

\[
\|M\|_{\ell^2-\ell^2} \leq \left\| I - \left[ \begin{array}{c}
\frac{UU^*}{\|U\|_{\ell^2-\ell^2}} \\
1
\end{array} \right] \right\|_{\ell^2-\ell^2} + \left\| \left[ \begin{array}{c}
\frac{U}{\|U\|_{\ell^2-\ell^2}} (v \tilde{\Psi} - I) U^* \\
\frac{1}{\|U\|_{\ell^2-\ell^2}} (v \tilde{\Psi} - 1)
\end{array} \right] \right\|_{\ell^2-\ell^2} \leq 1 - \frac{1}{\kappa} + 2\|v \tilde{\Psi} - 1\|_{\ell^\infty},
\]

(C.16)

since

\[
\left\| \begin{bmatrix}
\frac{U}{\|U\|_{\ell^2-\ell^2}} (v \tilde{\Psi} - I) U^* & \frac{1}{\|U\|_{\ell^2-\ell^2}} (v \tilde{\Psi} - 1) \\
\langle \frac{1}{\kappa}, v \tilde{\Psi} - 1 \rangle
\end{bmatrix} \right\|_{\ell^2-\ell^2} \leq \left\| \begin{bmatrix}
\frac{U}{\|U\|_{\ell^2-\ell^2}} (v \tilde{\Psi} - 1) U^* & \frac{1}{\|U\|_{\ell^2-\ell^2}} (v \tilde{\Psi} - 1) \\
\langle \frac{1}{\kappa}, v \tilde{\Psi} - 1 \rangle
\end{bmatrix} \right\|_{\ell^2-\ell^2} + \left\| \begin{bmatrix}
0 & \frac{U}{\|U\|_{\ell^2-\ell^2}} (v \tilde{\Psi} - 1) U^* \\
\langle \frac{1}{\kappa}, v \tilde{\Psi} - 1 \rangle
\end{bmatrix} \right\|_{\ell^2-\ell^2}
\]

and by Hölder’s inequality

\[
|\langle \frac{1}{\kappa}, v \tilde{\Psi} - 1 \rangle| \leq \|v \tilde{\Psi} - 1\|_{\ell^\infty}, \quad \frac{1}{\|U\|_{\ell^2-\ell^2}} \|v \tilde{\Psi} - 1\|_{\ell^2} \leq \|v \tilde{\Psi} - 1\|_{\ell^\infty}.
\]

**Inductive argument for** (C.13). We begin by noting that since \( A_0 = I, b_0 = 0 \), (C.13) holds for \( k = 0 \). Now assume that it is true for 0, 1, \ldots, \( k - 1 \). If we can verify that

\[
2\|v \tilde{\Psi} - 1\|_{\ell^\infty} \leq \frac{1}{2\kappa},
\]

(C.17)
then by (C.15) and (C.16) together with \( t_b = 1 \), we have

\[
\left\| t_A^{-1/2} \text{vec}(A_k - A_\bullet) \right\|_F^2 \leq \left( 1 - \frac{1}{2\kappa} \right)^2 \left\| t_A^{-1/2} \text{vec}(A_{k-1} - A_\bullet) \right\|_F^2.
\]

Applying the inductive hypothesis, we obtain (C.13) for iteration \( k \). So, once we can show that under the inductive hypothesis, (C.17) holds, the result will be established.

We begin by showing that under the inductive hypothesis, the errors \( \delta_i \) are all bounded. Indeed, by the parallelogram law

\[
\| \delta_i \|_2^2 = \| A_{k-1} u_i + b_{k-1} - v_i \|_2^2
\]
\[= \| (A_{k-1} - A_\bullet) u_i + (b_{k-1} - b_\bullet) \|_2^2 \]
\[\leq 2 \frac{\| A_{k-1} - A_\bullet \|_F^2}{t_A} \| u_i \|_2^2 + 2 \| b_{k-1} - b_\bullet \|_2^2,
\]

and so applying the inductive hypothesis to bound

\[
t_A^{-1} \| A_{k-1} - A_\bullet \|_F^2 + \| b_{k-1} - b_\bullet \|_2^2 \leq t_A^{-1} \| I - A_\bullet \|_F^2 + \| b_\bullet \|_2^2,
\]

we obtain for all \( i \)

\[
\| \delta_i \|_2 \leq \sqrt{2} \times \sqrt{t_A^{-1} \| A_\bullet - I \|_F^2 + \| b_\bullet \|_2^2} \times \max \left\{ t_A^{1/2} \| u_i \|_2, 1 \right\}
\]
\[\leq \sqrt{2} \times \sqrt{t_A^{-1} \| A_\bullet - I \|_F^2 + \| b_\bullet \|_2^2} \times \frac{\sqrt{c} \| U \|_{\ell^1 \to \ell^2}}{\| U \|_{\ell^2 \to \ell^2}}.
\]

(C.18)

Since

\[
\psi(s) = \frac{1}{4\pi\sigma^2} \left( 1 - \exp\left( -\frac{s}{2\sigma^2} \right) \right),
\]

we have

\[
\psi(s) = \frac{1}{8\pi\sigma^4} \exp\left( -\frac{s}{2\sigma^2} \right).
\]
and for all $s \geq 0$

$$|1 - \nu \psi(s)| = |1 - 8\pi \sigma^4 \psi(s)| \leq \frac{s}{2\sigma^2}$$

by the standard exponential convexity estimate. Plugging in our bound (C.18), we obtain for all $i$

$$|1 - \nu \psi\left(\frac{1}{2} \Vert \delta_i \Vert_2^2\right)| \leq \frac{t_A^{-1} \Vert A \star - I \Vert_F^2 + \Vert b \star \Vert_2^2}{2\sigma^2} \times \frac{c \Vert U \Vert_{\ell^2_{\to \ell^2}}^2}{\Vert U \Vert_{\ell^2_{\to \ell^2}}^2}.$$

Under our choice of $t_A$ and hypotheses on $\sigma$, this is bounded by $\frac{1}{4k}$. \qed