Trade-Offs and Opportunities in High-Dimensional Bayesian Modeling

Collin Andrew Cademartori

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

2024

© 2024

Collin Andrew Cademartori

All Rights Reserved

Abstract

Trade-Offs and Opportunities in High-Dimensional Bayesian Modeling

Collin Andrew Cademartori

With the increasing availability of large multivariate datasets, modern parametric statistical models makes increasing use of high-dimensional parameter spaces to flexibly represent complex data generating mechanisms. Yet, ceteris paribus, increases in dimensionality often carry drawbacks across the various sub-problems of data analysis, posing challenges for the data analyst who must balance model plausibility against the practical considerations of implementation. We focus here on challenges to three components of data analysis: computation, inference, and model checking. In the computational domain, we are concerned with achieving reasonable scaling of the computational complexity with the parameter dimension without sacrificing the trustworthiness of our computation. Here, we study a particular class of algorithms - the vectorized approximate message passing (VAMP) iterations - which offer the possibility of linear per-iteration scaling with dimension. These iterations perform approximate inference for a class of Bayesian generalized linear regression models, and we demonstrate that under flexible distributional conditions, the estimation performance of these VAMP iterations can be predicted to high accuracy with probability decaying exponentially fast in the size of the regression problem. In the realm of statistical inference, we investigate the relationship between parameter dimension and identification. We develop formal notions of weak identification and model expansion in the Bayesian setting and use this to argue for a very general tendency for dimensionality-increasing model expansion to weaken the identification of model parameters. We draw two substantive

conclusions from this formalism. First, the negative association between dimensionality and identification can be weakened or reversed when we construct prior distributions that encode sufficiently strong dependence between parameters. Absent such prior information, we derive bounds which indicate that decreasing identification is usually unavoidable with sufficient inflation of the dimension without increasing the severity of the third challenge we consider: that of dimensionality to model checking.

We divide the topic of model checking into two sub-problems: fitness testing and correctness testing. Using our model expansion formalism, we show again that both of these problems tend to become more difficult as the model dimension grows. We propose two extensions of the posterior predictive p-value - certain conditional and joint p-values, which are designed to address these challenges for fitness and correctness testing respectively. We demonstrate the potential of these p-values to allow successful model checking that scales with dimensionality theoretically and with examples.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I thank my advisors Professors Cynthia Rush and Andrew Gelman for their guidance and for countless enlightening conversations over the past six years. I am especially grateful for their continuous support while I endeavored, and at times struggled substantially, to develop my thinking into an independent research program. I must thank Cynthia in particular for her encouragement and guidance early in my studies and for her invaluable advice in so many meetings since, without which I certainly could not have produced this dissertation. I also must thank Andrew for teaching me how to think both broadly and boldly about statistics, which has served as indispensable inspiration as I have learned how to sharpen my vague questions into viable research directions.

Next, I thank the other members of my dissertation committee - Professors Liam Paninski, Bianca Dumitrascu, and Aki Vehtari - whose work and teaching have shaped my understanding of our field. Their support of my work has been a continual source of encouragement.

I also wish to thank my friends outside of the academy - Joe Schwalb, Briggs Woolley, Liv Rigdon, Davis Cowart, Ivy Karlsgodt, and Sy Chounchaisit - for much needed support, laughter, and diversion from statistics. Much of this work began during the height of the COVID-19 pandemic, and I am certain that I could not have succeeded in developing my ideas during that difficult time without their camaraderie.

I especially thank my partner, Beverly Yang, for her unwavering belief in me and for her constant support and care. I am deeply grateful for her companionship, a source of both joy and stability through even the most difficult moments.

Finally, and most importantly, I thank my parents, Lorin Oswald and Kevin Cademartori, for their love and support over so many years. From the very beginning, they encouraged me to pursue my own interests to the fullest, while making me feel no expectation other than that I lead a life that I can be proud of and that makes me happy. Any success that I find in life, of which this dissertation is now a part, I credit first and foremost to this enormous gift.

Dedication

To my grandfather, Charles A. Oswald III.

Chapter 1: Desirable and disastrous dimensionality

The technology-driven proliferation of both large, complex datasets and powerful computational hardware in recent decades has created a greater impetus for employing increasingly complex statistical models in data analysis tasks. In the realm of parametric probabilistic modeling, this increasing complexity is reflected in the dimension of the statistical model's parameter space. In practice, large, multiplex datasets may compel such large parameter spaces for various reasons, for instance:

1. A problem may be naturally high dimensional, as when we consider a regression task with a large and undifferentiated pool of potential covariates. In such cases, there may be no principled way to locate a lower dimensional model which does not carry an unjustifiable collection of assumptions.

2. Bringing more data to bear on a problem may necessitate an increase in parameter dimension, as when data from distinct sub-population are combined by introducing hierarchical parameters to model the inter-population (dis)similarity. This is particularly relevant in an age of large interrelated data sets collected without direct control of the data generating processes.

3. Related to the last two cases, complexity may be *a priori* preferable when there is reason to suspect that there is no simple description of the true data generating mechanism. This may occur in highly multi-causal domains, or generally when simplifying conditional independence relations are unlikely to hold, as in the presence of many unobserved variables.

In such situations, the data analyst must be guarded against the challenges that frequently arise as the parameter dimension grows. This thesis is concerned with both understanding the prevalence

and nature of dimensionality-induced challenges as well as with methods for mitigating or overcoming such challenges in practice. We consider the setting of Bayesian modeling in particular and focus on three categories of difficulties: computational, evidential, and inferential.

1.1 Computational concerns

A Bayesian model is nothing more than a specification of a joint distribution $q(\mathbf{y}, \boldsymbol{\theta})$ over observed data $\mathbf{y} \in \mathcal{Y}$ and unobserved quantities $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Such a model is almost always specified as a sampling distribution $f(\mathbf{y} \mid \boldsymbol{\theta})$ (or likelihood, when treating it as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{y}$) and a prior distribution $q(\boldsymbol{\theta})$. Conceptually, it is often convenient to think of $f(\mathbf{y} \mid \boldsymbol{\theta})$ as specifying a (random) process by which observed quantities $\mathbf{y}$ are generated from unobserved quantities, and to think of $q(\boldsymbol{\theta})$ as specifying the distribution of plausible values for the unobserved quantities $\boldsymbol{\theta}$.

In Bayesian inference, we are commonly interested in computing the posterior distribution of the model, $q(\boldsymbol{\theta} \mid \mathbf{y})$. In practice, we will never know this distribution *completely* (insofar as we cannot know *every* quantity which might depend on this distribution). Instead, we are usually content with the computation of certain functionals of $q(\boldsymbol{\theta} \mid \mathbf{y})$, including (potentially) means, variances, correlations, quantiles, or marginal densities. Computational approaches can be divided into three categories: exact, asymptotically exact, and approximate.

1.1.1 Exact inference is powerful but brittle

Exact inference refers to methods that compute the relevant functional of the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$ exactly. The canonical exact inference method is the method of conjugate families. In this setting, the prior and sampling distributions $q(\boldsymbol{\theta})$ and $f(\mathbf{y} \mid \boldsymbol{\theta})$ are taken from parametric families

$$\mathcal{Q} = \{q_{\lambda}(\cdot) \mid \boldsymbol{\lambda} \in \boldsymbol{\Lambda} \subset R^k\}, \qquad \mathcal{F} = \{f_{\sigma}(\cdot \mid \cdot) \mid \boldsymbol{\sigma} \in \boldsymbol{\Sigma} \subset \mathbb{R}^m\}. \tag{1.1}$$

These families are said to be conjugate if, whenever $q \in \mathcal{Q}$ and $f \in \mathcal{F}$, then $q(\cdot \mid \mathbf{y}) \in \mathcal{Q}$. In this case, there is some value of the hyperparameters $\widetilde{\boldsymbol{\lambda}}$ such that $q(\cdot \mid \mathbf{y}) = q_{\widetilde{\lambda}}(\cdot)$, and these

hyperparameters are determined as a function of the prior hyperparameters, sampling distribution hyperparameters, and the observed data. That is, there is a map $\phi : \mathbf{\Lambda} \times \mathbf{\Sigma} \times \mathcal{Y} \to \mathbf{\Lambda}$ such that

$$\widetilde{\boldsymbol{\lambda}} = \phi(\boldsymbol{\lambda}, \boldsymbol{\sigma}, \mathbf{y}). \tag{1.2}$$

When we can apply this method to our model, then posterior inference proceeds in two steps. First, compute $\widetilde{\boldsymbol{\lambda}}$ using (1.2). This determines the posterior distribution as $q_{\widetilde{\lambda}}$. Second, use the properties of $q_{\widetilde{\lambda}}$ to compute the desired posterior functionals. For example, when $\mathcal{F}$ is the Poisson family, then the family of gamma distributions $\mathcal{Q}$ with shape and rate hyperparameters $\boldsymbol{\lambda} = (\alpha, \beta)$ is conjugate. Then posterior inference consists of computing the posterior hyperparameters $\widetilde{\lambda} = (\widetilde{\alpha}, \widetilde{\beta})$. If we are interested in the posterior mean and variance, for example, then these can be computed for the posterior gamma distribution as $\widetilde{\alpha}/\widetilde{\beta}$ and $\widetilde{\alpha}/\widetilde{\beta}^2$ respectively.

In some cases, the desired posterior functionals may not be given in terms of known closed-form functions of the hyperparameters $\widetilde{\boldsymbol{\lambda}}$. However, in most cases where conjugacy is useful, we can at least readily sample from $q_{\widetilde{\lambda}}$ with very low per-sample complexity. Then most common functionals can be computed using large sample approximations. For example, sample average approximate posterior expectations, and histograms approximate marginal densities.

The primary limitation of conjugacy in Bayesian inference lies in the relative paucity of known conjugate families of priors and sampling distributions. Consequently, conjugate models will rarely be available in sufficiently complex modeling settings. For instance, exact conjugate inference is possible for linear regression models when the prior on the regression coefficients and noise variance is taken from the normal-inverse gamma family. However, this conjugacy relationship is highly brittle to modifications in the regression assumptions. In particular, if the error terms are taken to be heteroskedastic or correlated, if nonlinearity is introduced between the linear combination and the outcome as in generalized linear models, or if data exhibits nontrivial missingness or hierarchical structure, then the conjugate structure is often lost entirely.

1.1.2 Asymptotically exact inference is flexible but superlinear

Whereas conjugate families allowed us to sample from the exact posterior distribution, it is often fruitful to trade this exactness for flexibility. This is enabled by the class of Markov Chain Monte Carlo (MCMC) methods, which generate asymptotically exact samples from the posterior distribution while imposing far fewer requirements on the form of the model relative to conjugate methods. The asymptotically exact property means that if we run such a method for long enough, then the distribution from which the next sample is generated is arbitrarily close to the true posterior distribution. More formally, an MCMC method generates a sequence of samples $\{\boldsymbol{\theta}^{(s)}\}_{s=1}^{S}$ where each sample is taken from a corresponding distribution $\boldsymbol{\theta}^{(s)} \sim q^{(s)}(\cdot)$. The distributions $q^{(s)}$ are determined in sequence from some initial proposal distribution $q^{(0)}$ by the action of a Markov chain. In particular, given a conditional distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$ which we refer to as the transition kernel, the distributions $q^{(s)}$ are determined sequentially as the marginal distributions of

$$q^{(s,s-1)}(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^{(s-1)}) = q^{(s-1)}(\boldsymbol{\theta}^{(s-1)})\pi(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{\theta}^{(s-1)}). \tag{1.3}$$

When configured correctly, and assuming appropriate conditions on the model, these will converge to the true posterior in distribution, i.e.

$$\delta_{\mathrm{TV}}\left(q^{(s)}, q(\cdot \mid \mathbf{y})\right) \to 0, \tag{1.4}$$

where δ_{TV} is the total variation distance defined by

$$\delta_{\mathrm{TV}}\left(p, q\right) = \sup_{E} \left|\mathbb{P}_p\left(\boldsymbol{\theta} \in E\right) - \mathbb{P}_q\left(\boldsymbol{\theta} \in E\right)\right|, \tag{1.5}$$

where the supremum is taken over all measurable subsets of $\boldsymbol{\Theta}$. Such convergence results follow from the convergence of the underlying Markov chains that MCMC algorithms simulate (see, e.g. [79] for details). This total variation convergence justifies, for example, treating a finite-sample

proportion

$$\frac{1}{S}\sum_{s=1}^{S}\mathbb{1}_E(\boldsymbol{\theta}^{(s)}) \tag{1.6}$$

as a reliable estimate of $\mathbb{P}_{q(\boldsymbol{\theta}|\mathbf{y})}(E)$ for S sufficiently large regardless of the choice of measurable $E \subset \boldsymbol{\Theta}$.

The benefit of this approach is that, if the distributions $q^{(s)}$ have converged to some stable limit, then this limit must be the true posterior distribution. Consequently, MCMC methods are often first run for a warmup period during which the distributions $q^{(s)}$ converge to $q(\cdot \mid \mathbf{y})$. Once convergence is (approximately) achieved, the subsequent post-warmup samples are collected and treated as a sample from the posterior distribution.

Such algorithms apply very widely to parametric models, but they are subject to some common limitations. In general, there are no guarantees on the rate of the convergence in (1.4), and in practice this convergence can be intractably slow for sufficiently complex models. Furthermore, because the samples $\boldsymbol{\theta}^{(s)}$ are typically positively correlated, MCMC algorithms may need to run for a large number of iterations in order to achieve a sample which adequately represents the (desired functionals of the) posterior distribution.

Perhaps most important in the context of high-dimensional modeling, however, is the scaling of per-iteration complexity with parameter dimension. This scaling will generally depend on the structure of the posterior distribution, but it is shown in [16] that we cannot generally expect better scaling than $O(d^{5/4})$ in general for the Hamiltonian Monte Carlo (HMC) algorithm. Although HMC is widely popular and enjoys better per-iteration dimension scaling than most alternatives, this scaling remains super-linear in d.

1.1.3 Approximate inference is fast but potentially unreliable

Continuing along the path from exact to asymptotically exact inference, approximate inference methods trade any guarantee of (eventual) exactness for greater computational flexibility and better scaling. These methods typically approximate the true posterior distribution $q(\cdot \mid \boldsymbol{\theta})$ by one or more distributions $q^*(\cdot)$ taken from a class of distributions $\mathcal{Q}^*$ which provides nice computational

properties. The resulting approximations q^* may remain arbitrarily far from the true posterior no matter how long the approximate inference algorithm is run, giving up the asymptotic exactness of MCMC algorithms.

Many approximate inference algorithms can be characterized by the choice of approximating family $\mathcal{Q}^*$ along with the measure of discrepancy between $q^*(\cdot)$ and $q(\cdot \mid \mathbf{y})$ which is minimized to find the best approximation $q^* \in \mathcal{Q}^*$. The most widely-used family of approximating distributions is the mean-field family, which has the form

$$\mathcal{Q}^* = \left\{ q^* = \prod_{i=1}^{d} q_{\lambda_i}^{(i)}(\boldsymbol{\theta}_i) \text{ and } \{\boldsymbol{\lambda}_i\}_{i=1}^{d} \subset \boldsymbol{\Lambda} \right\}, \tag{1.7}$$

where d is the dimension of the parameter space $\boldsymbol{\Theta}$, and $\{q_{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ is some parametric family of one-dimensional distributions. Critically, the mean-field family assumes independence between all components of the parameter vector $\boldsymbol{\theta}$ and enforces a specific parametric form for the marginal distributions.

Once a corresponding discrepancy measure $D(\cdot, \cdot)$ has been fixed, the approximate inference algorithm proceeds by iteratively updating $q^* \in \mathcal{Q}^*$ in order to minimize $D(q^*(\cdot) \mid q(\cdot \mid \mathbf{y}))$. Specific instances of the approximate inference scheme arise by choosing a combination of $\mathcal{Q}^*$ and D so as to make the corresponding minimization problem computationally tractable. The popular mean-field variational inference algorithm minimizes (a bound on) the KL divergence $D(q^*(\cdot) \mid q(\cdot \mid \mathbf{y}))$ from the approximate distribution to the posterior when $\mathcal{Q}^*$ is the mean-field family, where the KL divergence is given by

$$D(p \mid q) = \mathbb{E}_{p(\boldsymbol{\theta})} \log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}. \tag{1.8}$$

The KL divergence $D(p \mid q)$ quantifies discrepancy between distributions p and q insofar as it is nonnegative and $D(p \mid q) = 0$ if and only if $p = q$ p-almost everywhere. Furthermore, Pinsker's

inequality tells us that the KL divergence controls the total variation distance, in that

$$\delta_{\mathrm{TV}}(p, q) \leq \sqrt{\frac{1}{2} D(p \mid q)}. \tag{1.9}$$

Thus, the KL divergence provides an analytically convenient target for optimization while controlling a directly interpretable quantification of distance between distributions.

While the distributional families $\mathcal{Q}^*$ employed by approximate methods often make unrealistically strong simplifying assumptions (e.g. the mutual independence enforced by the mean-field family), it is precisely this enforced simplicity that opens up the possibility of per-iteration computational complexity scaling that is linear in the parameter dimension d for such algorithms. Furthermore, in practice, the number of iterations required for convergence of such approximate methods may be more than an order of magnitude smaller compared to MCMC methods.

Expectation propagation algorithms are another class of approximate inference procedures which can be applied to minimize the reverse KL divergence (i.e. from the posterior to the approximate distribution). Minimizing the KL divergence in this direction gives some potentially beneficial properties depending on the posterior functionals of interest. For instance, mean-field variational inference has been observed to underestimate marginal variances when the components of $\boldsymbol{\theta}$ are not independent in the true posterior distribution. Moreover, this can be directly related to the KL divergence which the variational algorithm minimizes. (See [13] for a review of the properties of variational inference and comparison to other algorithms.) In many cases, by flipping the direction of the divergence, expectation propagation avoids this problem (with appropriate choice of approximating family $\mathcal{Q}^*$) [68, 69]. However, while mean-field variational inference can be applied with little consideration for the underlying model, expectation propagation procedures often require a more careful, model-specific construction of the family $\mathcal{Q}^*$.

Yet more model specific, the approximate message passing (AMP) algorithms were developed to rapidly compute sparse coefficient estimates in high-dimensional regression models [28]. Many variants of the original AMP algorithm have been developed with applicability to a wider class

of regression models and with better robustness to different distributional structures. The "vectorized" variants introduced in [77, 84], for instance, can achieve convergence more reliably under a very flexible class of design matrices which can be arbitrarily poorly conditioned. These vectorized AMP algorithms can also be derived as a special case of expectation propagation, which we demonstrate in chapter 2.

The primary drawback of approximate methods in general comes from the potentially unbounded error in the approximation q^* to the true posterior. While this approximation will minimize the chosen discrepancy $D(q^*(\cdot), q(\cdot \mid \mathbf{y}))$, the size of the minimized objective is typically unknown and difficult to estimate. This stands in contrast to MCMC methods, for example, for which diagnostics can indicate convergence with high practical reliability. The AMP family of algorithms offers a partial counterexample to this trend. In certain limits as both sample size and number of regressors increase, the estimation error of many AMP algorithms converge to the predictions of a computable scalar iteration referred to as the state evolution [28, 77]. Furthermore, in sufficiently nice cases, this convergence has been shown to be exponentially fast [81]. Thus this state evolution can provide a reliable characterization of the overall estimation error, and hence also the approximation error, which is one component thereof. Taken as a whole, the primary computational challenge presented by high-dimensional models is this: to find algorithms which simultaneously scale tractably with the problem dimension and provide trustworthy approximations of the true posterior functionals of interest.

1.2 Evidential concerns

When we apply the inferences or predictions obtained from a statistical model to draw a substantive conclusion or take some action, we may reasonably wish to know if that conclusion or action is justified. Such justification depends not only on the outputs of the model, but also on the relationship between the model and the world. In particular, justifying an interpretation of a model often requires some kind of rigorous test of the fitness of the model (i.e. how well it can recapitulate features of our measured data) or the correctness of the model (i.e. whether the data

can falsify any of the modeling assumptions).

Because we find that model checking is often treated in the statistical literature as a unified enterprise with a single goal, we dwell on this distinction between fitness and correctness testing for a moment. A model may be incorrect but provide adequate fitness to all checked aspects of the data, simply because the data are insufficient to determine the structure of the model uniquely. In such a case, depending on the intended application of the model, we may be in a position to trust the model's predictions about observable quantities but not necessarily the corresponding inferences about unobservables.

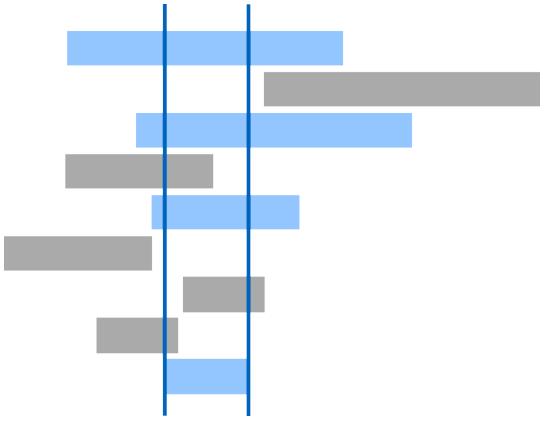

Figure 1.1: A schematic representation of the distinction between model correctness and fitness. Wider rectangles represent models with greater sampling variation. Blue rectangles represent correctly-specified models (which completely overlap the most specific correct model).

On the other hand, because statistical models are partially stochastic, there will not be a unique correct model in any given case. This is just a mathematical way to say any given "data generating process" in the world is usually many different processes operating at different scales. The knowledge of graduates of a given university is generated (partly) by the quality of the university's classes. But the quality of the university's classes is generated (partly) by the educational policies of the nation in which the university resides (which also affect the other universities within the same nation). Statistically, if $\mathbf{y} \sim f(\cdot \mid \boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim q(\cdot)$, then it is also true that $\mathbf{y} \sim f(\cdot) = \mathbb{E}_{q(\boldsymbol{\theta})} f(\cdot \mid \boldsymbol{\theta})$. Thus, both $f(\cdot \mid \boldsymbol{\theta})$ and $f(\cdot)$ are correct models of the process generating $\mathbf{y}$. However, we may still be able to say that $f(\cdot \mid \boldsymbol{\theta})$ fits the data $\mathbf{y}$ better than $f(\cdot)$, since the former may have lower

variance and may more typically produce data resembling $\mathbf{y}$ in relevant features.

This distinction is illustrated diagrammatically in Figure 1.1. Each rectangle represents a different model. The width of the rectangle represents the specificity of the model, with wider rectangles representing models with a higher variance sampling distribution (like $f(\cdot)$ above). The blue highlighted rectangles represent correct models, which are those completely overlapping the most specific correct model at the bottom. As the figure suggests, some incorrect models (such as that represented by the third rectangle from the bottom) may be incorrect but also exhibit better fitness to the data than correct but less specific models (such as that represented by the topmost rectangle). Essentially, this correctness-fitness distinction is nothing more than the bias-variance tradeoff in statistical inference rephrased in the setting of model checking.

1.2.1 Assessing model fitness in the context of uncertainty

Assessments of model fitness often begin by selecting one or more features of the data as represented by test statistics $T_1, \ldots, T_k$, where $T_i : \mathcal{Y} \to \mathbb{R}$ for $1 \leq i \leq k$. Common choices of test statistics include measures of central tendency and spread, measures of association, and measures of extreme values. Once the features of the data have been fixed, it must be determined whether the model is capable of reproducing the observed values of the test statistics $T_i(\mathbf{y})$. Such a determination is not entirely straightforward, however, for two reasons:

1. The data $\mathbf{y}$ are assumed to be a random realization from a true data generating distribution $f(\cdot \mid \boldsymbol{\theta}^*)$. Thus, even if the model is entirely correct, we would not expect another data set generated from this model to reproduce the observed $T_i(\mathbf{y})$ values exactly for any $1 \leq i \leq k$, and we even expect *some* values not to match even approximately if k is sufficiently large.

2. The parameter $\boldsymbol{\theta}^*$ which corresponds to the true data generating distribution is itself unknown. Thus, in any comparison of model to data, we wish to account for our underlying the uncertainty about the distribution $f(\cdot \mid \boldsymbol{\theta}^*)$.

If we take a frequentist approach to inference, we may obviate the second issue by simply

substituting a point estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}^*$. We then usually resolve the issue of randomness in the data by comparing the distribution of possible test statistic values under the model to the observed value. More specifically, given $\boldsymbol{\theta}$, we can compare the distribution of $T_i(\mathbf{y}_{\text{rep}})$ where $\mathbf{y}_{\text{rep}} \sim f(\cdot \mid \boldsymbol{\theta})$ to the observed value of $T_i(\mathbf{y})$. (Here, $\mathbf{y}_{\text{rep}}$ represents "replicated" data; i.e. data generated from the same process as $\mathbf{y}$.) If the latter is an extreme value relative to the distribution of $T_i(\mathbf{y}_{\text{rep}})$, then we can conclude that observing a value similar in magnitude to $T_i(\mathbf{y})$ would be extremely unlikely if the data were truly drawn from $f(\cdot \mid \boldsymbol{\theta})$. An easily applied method for making this procedure quantitative is to first fix a numerical measure that summarizes the comparison of $T_i(\mathbf{y})$ to the distribution of $T_i(\mathbf{y}_{\text{rep}})$ given $\boldsymbol{\theta}$, such as a tail probability or p-value:

$$p_T(\boldsymbol{\theta}) = \mathbb{P}_{f(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})}\left(|T(\mathbf{y}_{\text{rep}})| > |T(\mathbf{y})|\right). \tag{1.10}$$

We then regard $T_i(\mathbf{y})$ as too extreme if $p_T(\boldsymbol{\theta})$ is sufficiently small. Taking $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ then completes the fitness assessment in the frequentist case.

In the Bayesian case, we must still contend with our uncertainty about $\boldsymbol{\theta}$. A common solution is to simply average our comparisons (e.g. the values $p_T(\boldsymbol{\theta})$) over the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$, representing our average fitness assessment accounting for our uncertainty about the true data generating distribution. If we used a p-value to summarize the fitness given each $\boldsymbol{\theta}$, then the resulting posterior expectation of these is referred to as the posterior predictive p-value:

$$p_T^{\text{post}} = \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})}\left[p_T(\boldsymbol{\theta})\right] \tag{1.11}$$

This approach is simple and interpretable in sufficiently low-dimensional models, but it becomes harder to extract useful information from as the model dimension increases. This can be seen, for instance, in the fact observed by [67] that the posterior predictive p-value tends towards 0.5 as the model dimension grows (when $\mathbf{y}$ is drawn from its marginal, or prior predictive, distribution $f(\mathbf{y})$). It may be tempting to conclude that this is a good thing, that it indicates that larger models are simply more likely to fit any given feature of the data than smaller models. In any given

model, this may in fact be the case.

However, the posterior predictive p-value may approach 0.5 with increasing dimension even though the higher dimensional models do not fit the data better overall. To see why, we need only examine the distribution of $p_T(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \sim q(\cdot \mid \mathbf{y})$. As we show in chapter 3, as the dimension of the model increases, the posterior variance of $p_T(\boldsymbol{\theta})$ tends to increase as well. We also demonstrate that this fact is sufficient to explain the contraction of p^{post} towards 0.5, regardless of whether the fitness of $f(\cdot \mid \boldsymbol{\theta})$ is improving with dimension for a majority of $\boldsymbol{\theta}$ under the posterior.

In words, we may say that the posterior predictive p-value tends to indicate better model fitness (for any given test statistic T) as model dimension grows because higher-dimensional models tend to be more equivocal about the true data generating process. Specifically, the posterior will commonly support some sampling distributions under which the observed $T(\mathbf{y})$ is fit well and others (potentially a majority) under which it is fit poorly. But because of this variation in the fitness, it is harder for the average value (i.e. the posterior predictive p-value) to reach extreme values. Consequently, higher dimensional models tend to produce checks of fitness that are more ambiguous, and provide a correspondingly weaker basis for judging the model.

1.2.2 Assessing model correctness in the context of uncertainty

Relative to fitness checking, testing the modeling assumptions for correctness may provide stronger justification for conclusions about unobservables on the basis of model inferences, with the caveat that all such testing is subject to the fundamental asymmetry of falsification, namely that a model may be definitively falsified but never entirely confirmed. Nevertheless, a model which is not falsified after a strong test of its assumptions is generally regarded as more trustworthy than one which has not been so rigorously tested.

Assessments of model correctness generally hinge on frequency calculations conditional on model truth - i.e., how often would an observed quantity lie in some set if the observed data from which the quantity is calculated were truly generated by the model. Formally, if $C : \mathcal{Y} \to \mathcal{Q} \subset \mathbb{R}$ is a map which quantifies concordance between our model and data such that smaller values indicate

less concordance, then we can assess the correctness of the model using the probability

$$\pi = \mathbb{P}_{h(\mathbf{y}_{\text{rep}})}\left(C(\mathbf{y}_{\text{rep}}) \leq C(\mathbf{y})\right), \tag{1.12}$$

where $h(\cdot)$ represents a distribution that we expect $\mathbf{y}$ to be drawn from if the model is correct. (We will see momentarily that this is not unique.) If $\mathbf{y} \sim h(\mathbf{y})$, then π is uniformly distributed (by the probability integral transform). Thus, if we decide to treat our model as incorrect if π is below some threshold $\pi_0 \in (0, 1)$, then π_0 is exactly the probability of erroneously considering the model incorrect (i.e. the Type I error rate).

As with tests of fitness, it is common to take the concordance quantity C to be a p-value. Depending on the specific choice of C, the probability π may not be easily computable, in which case it is common to reject the model when $C(\mathbf{y}) < \alpha$ for some level $\alpha > 0$. This may be justified by noting that $\pi \to 0$ as $C(\mathbf{y}) \to 0$.

In classical frequentist testing, correctness testing often comes for free from fitness testing. Suppose we fix some test statistic T and a point estimate $\hat{\boldsymbol{\theta}}$. Then a frequentist may test the *fitness* of the model $f(\cdot \mid \hat{\boldsymbol{\theta}})$ by computing the p-value $p_T(\hat{\boldsymbol{\theta}})$. This immediately yields a corresponding test of model *correctness*. To obtain this test, we first ask for the frequency $\pi(\alpha)$ with which $p_T(\hat{\boldsymbol{\theta}}) < \alpha$ for any $\alpha > 0$ under the assumption that $f(\cdot \mid \hat{\boldsymbol{\theta}})$ is the true data generating process. But this is just to say that $p_T(\hat{\boldsymbol{\theta}})$ is calculated under the same distribution from which $\mathbf{y}$ is generated, and so the probability integral transform again tells us that $p_T(\hat{\boldsymbol{\theta}})$ is uniformly distributed. Critically, we have in this case that $\pi(\alpha) = \alpha$. Consequently, if we set some rejection threshold $\alpha = 0.05$ and find that $p_T(\hat{\boldsymbol{\theta}}) < \alpha$, then we know that this would occur with frequency 0.05 if the model were correct, allowing us to test and reject the model with known error rates.

However, we must emphasize that the preceding analysis was entirely conditional on the particular value of our point estimate $\hat{\boldsymbol{\theta}}$, and is thus simultaneously a test of both (i) the modeling assumptions carried by the parametric family $\mathcal{F} = \{f(\cdot \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and (ii) the particular choice of estimator $\hat{\boldsymbol{\theta}}$. In the Bayesian setting, we lack a single point estimator and must accommodate the

uncertainty intrinsic within the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$. Beneficially, the resulting test can be said to test the correctness of only the modeling assumptions (free from any additional choice of estimation procedure). But once we lose the convenience of a single point estimator, the relevant frequency calculations for tests of correctness no longer come for free from fitness calculations as in the frequentist case.

We may try to construct a Bayesian test of correctness by starting with a fitness test as in the frequentist example. For instance, we may begin by computing a posterior predictive p-value p_T^{post} and then attempt to find the frequency π with which $p_T^{\text{post}} < \alpha$ if the model were correct (where now model correctness means the correctness of the complete Bayesian specification, including the sampling family $\mathcal{F}$ and prior $q(\cdot)$, but excluding any particular choice of estimator). In terms of the above definitions, this amounts to taking $C = p^{\text{post}}$ and

$$h(\mathbf{y}) = f(\mathbf{y}) = \mathbb{E}_{q(\boldsymbol{\theta})} f(\mathbf{y} \mid \boldsymbol{\theta}), \tag{1.13}$$

the marginal or prior predictive distribution of the data under the model.

In the frequentist case, since $\pi = \alpha$, we had that the frequentist p-value $p_T(\hat{\boldsymbol{\theta}})$ was uniformly distributed under the frequentist null hypothesis, i.e. that $f(\cdot \mid \hat{\boldsymbol{\theta}})$ is correct. The posterior predictive p-value, on the other hand, typically has a distribution more concentrated around 0.5 compared to uniform when $\mathbf{y} \sim f(\mathbf{y})$. This means that if we would like to find an α for which $p_T^{\text{post}} < \alpha$ with known probability, say π_0, then we will generally need to select $\alpha > \pi_0$. Conversely, for any fixed value of α, say $\alpha = 0.05$ as above, the frequency with which $p_T^{\text{post}} < \alpha$ will typically be smaller than α if the model is correct.

This becomes an issue as soon as we consider the power of the resulting test of model correctness. In tests of model correctness, there is usually no pre-specified alternative to the model under consideration, and so the power cannot be formally defined in the usual way. But for our purposes, it suffices to recognize that as π_0 gets smaller, we are less likely to mistakenly conclude that the model is incorrect, but we are also probably less likely to detect that the model is in fact incorrect

for any particular fixed alternative. Consequently, if we ignore the distinction between α and π_0 and choose α based on common convention, we may end up with a value of π_0 small enough that our power against any particular relevant alternative model specifications becomes unpleasantly small as well.

In practice, we may hope to overcome this issue by simply adopting a more conservative stance to correctness testing of Bayesian models overall. Most simply, this might take the form of a policy of taking $\alpha \gg \pi_0$. While such a policy may perform well for this or that particular model, it is unlikely to be a sustainable strategy in the context of growing model dimension, for the same reason that high dimensionality posed a challenge for fitness testing. Specifically, as the parameter dimension grows, the frequency distribution of p_T^{post} when $\mathbf{y} \sim f(\mathbf{y})$ tends to contract around 0.5. Thus, in order to maintain $\pi_0 \geq 0.05$, for instance, the corresponding value of α will likely need to be taken closer and closer to 0.5 as the parameter dimension grows. In particular, we can expect no stable relation between π_0 and α which allows us to maintain sufficient power as dimension increases absent further information.

In summary, the fundamental challenge that high dimensionality poses for model checking is to find methods of assessment and testing which are rich enough to inform us about which aspects of data are poorly fit, and which are powerful enough to represent a real challenge to our modeling assumptions, so that passing the test may justify interpretation of the model.

1.3 Inferential concerns

Even if our computational approach succeeds in providing close approximations of posterior functionals and our model assessments warrant action on the basis of those posterior quantities, the posterior itself still must be informative enough to support nontrivial conclusions. For instance, if the posterior distribution is identical to the prior distribution, then we will have learned nothing from the observed data. Assuming that we constructed this model because our prior information was an insufficient basis to draw relevant conclusions or take relevant actions, this would represent a failure of inferential informativeness.

1.3.1 Weak marginal identification as a challenge to inference

Of course, the posterior could only recapitulate the prior distribution exactly if the likelihood function

$$l(\boldsymbol{\theta}) \doteq \boldsymbol{\theta} \mapsto f(\mathbf{y} \mid \boldsymbol{\theta}) \tag{1.14}$$

were identically constant over the support of the prior. This would be an extreme form of what is referred to an nonidentification in the frequentist literature. More generally, we may say that a model is nonidentified if the likelihood function $l(\boldsymbol{\theta})$ has no unique maximum. Most frequently when this occurs, there is an infinite subset $\boldsymbol{\Theta}^* \subset \boldsymbol{\Theta}$ such that $l(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} l(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$. In the constant case above, we have $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}$ (if the prior is supported over the whole parameter space). By definition, nonidentification means that the maximum likelihood estimator is not well-defined, and thus poses an existential challenge to frequentist inference.

For Bayesian models, on the other hand, nonidentification in the frequentist sense does not pose any existential issues for inference. As long as the prior distribution is proper, a nonidentified model will still yield a proper and well-defined posterior distribution. This had led some to claim that issues of identification do not affect Bayesian methods tout court.

To take a somewhat less trivial example, if $\boldsymbol{\Theta} = \mathbb{R}^d$ and $\boldsymbol{\Theta}^*$ is an affine subspace of $\boldsymbol{\Theta}$ of dimension $k \ll d$, then the posterior and prior distributions conditional on $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$ will be identical, but the posterior distribution may still be substantially more informative than the prior in directions orthogonal to $\boldsymbol{\Theta}^*$. From the perspective of posterior informativeness, however, we need not focus on subspaces along which the posterior is *exactly* equal to the prior. For instance, for a given component of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_1$, if the variance of $q(\boldsymbol{\theta}_1 \mid \mathbf{y})$ is 90% of the prior variance (i.e. the variance of $q(\boldsymbol{\theta}_1)$), then the posterior provides nearly as little usable information as in the exactly nonidentified case.

Consequently, the frequentist notion of nonidentification may be replaced by a notion of weak identification in the Bayesian setting. Rather than posing a challenge to the existence of estimators, this Bayesian notion of identification relates to our ability to learn nontrivial information from the

data given the model specification. We may try to formalize this notion in terms of a function $\Delta(\mathbf{y})$ which measures posterior learning relative to the prior insofar as $\mathbb{E}_{f(\mathbf{y})}\Delta(\mathbf{y}) \geq 0$ (i.e. we expect to learn a nonnegative amount of information if the model is correct) and $\Delta(\mathbf{y})$ monotonically increases with some appropriately defined notion of the additional information present in the posterior over the prior. We may be interested in defining such a Δ over the full posterior or over only certain marginal distributions, such as the marginal posterior distributions of the scalar components of $\boldsymbol{\theta}$.

One natural way to define such Δ is to start with a distributional quantity that is expected to decrease when passing from prior to posterior. If we focus only on the scalar marginals, then we might take the variance as such a quantity since it is always true that

$$\mathbb{E}_{f(\mathbf{y})}\text{Var}\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \leq \text{Var}\left(\boldsymbol{\theta}\right). \tag{1.15}$$

We may then define an appropriate Δ as $\Delta(\mathbf{y}) = \text{Var}\left(\boldsymbol{\theta}\right) - \text{Var}\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$. If we wish to define Δ for the entire posterior distribution, we may use the entropy functional $H(q) = -\mathbb{E}_{q(\boldsymbol{\theta})} \log q(\boldsymbol{\theta})$, for example, since this also obeys the required expected decrease:

$$\mathbb{E}_{f(\mathbf{y})}H(q(\cdot \mid \mathbf{y})) \leq H(q(\cdot)), \tag{1.16}$$

We then define $\Delta(\mathbf{y}) = H(q(\cdot)) - H(q(\cdot \mid \mathbf{y}))$. In terms of any such choice of Δ, we may say that (an appropriate marginal of) the posterior distribution is ϵ-weakly identified if $\Delta < \epsilon$.

1.3.2 Challenges of avoiding weak identification in high dimensions

The relation between our notion of weak identification and parameter dimension is not entirely straightforward, as the level of identification in any given case will depend on much more than the problem dimension. On one hand, we know that regression models, for example, become nonidentified in the frequentist sense as soon as the number of predictors exceeds the number of observations. As noted above, this creates weak identification along some subspaces in the

corresponding posterior for a Bayesian linear regression. On the other hand, it is not inevitable that increasing parameter dimension weakens posterior identification in general. There is no reason that a larger model could not exhibit a stronger connection between observables and unobservables on average and lead to better posterior identification overall.

Thus, the strongest relationship that we may expect between dimension and identification would be something like a trend contingent on other factors. Nevertheless, the preceding still leaves open whether such a trend exists (in either direction) and under what conditions such a trend would hold. To gain intuition, it is useful to return to the regression example. In that case, the poor identification comes from the increasing correlation between (observed measurements) of the predictor variables as the number of these predictors increases. This increasing correlation creates a kind of competition between distinct predictors to explain the same variation in the outcome. When different predictors compete in this way, it becomes difficult to disentangle the effects of each predictor separately on the outcome, creating weak identification for the corresponding regression coefficients.

The regression case also has the special property that we can associate parameters (the coefficients) to predictors which are shared across models, allowing us to compare the identification of the effect of a fixed predictor (i.e. its coefficient) across lower- and higher-dimensional models. In other words, the presence of a common predictor variable justified the comparison of parameters across models. In order to think about the effect of increasing dimension on parameters in general, we need a generic notion of a parameter being shared between models to enable similar comparisons. While there is no unique way to identify parameters between models, the following notion of model expansion captures many commonly encountered cases, including the regression example.

Definition 1 (Model Expansion)**.** *A model $q(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ defined with additional parameter $\boldsymbol{\lambda} \in \overline{\mathbb{R}}^k$ is an expansion of base model $q_{\text{base}}(\mathbf{y}, \boldsymbol{\theta})$ if*

$$q_{\text{base}}(\mathbf{y}, \boldsymbol{\theta}) = q(\mathbf{y}, \boldsymbol{\theta} \mid \boldsymbol{\lambda}_0) \text{ for some } \boldsymbol{\lambda}_0 \in \overline{\mathbb{R}}^k, \tag{1.17}$$

where $\overline{\mathbb{R}} = [-\infty, \infty]$. *If* $\boldsymbol{\lambda}_0 = \pm\infty$*, then the right-hand side is interpreted as a pointwise limit (in terms of an appropriate notion of distributional convergence) as* $\boldsymbol{\lambda}_0 \to \pm\infty$.

In words, q is an expansion of q_{base} if it embeds q_{base} as a conditional distribution. When we have a model expansion of this form, we can reasonably compare the identification of the parameters $\boldsymbol{\theta}$ between base model and expansion and speak of these parameters as being "shared" by the two models. This definition also covers many instances of what might be colloquially referred to as model expansion. For example:

- Let $q_{\text{base}}(\mathbf{y}, \boldsymbol{\theta})$ be a generalized linear model with response vector $\mathbf{y}$ and parameters $\boldsymbol{\theta}$ including predictor coefficients and additional parameters. Adding a new predictor and corresponding coefficient λ with independent prior is then an expansion since $q_{\text{base}}(\mathbf{y}, \boldsymbol{\theta}) = q(\mathbf{y}, \boldsymbol{\theta} \mid \lambda = 0)$.

- Let $q_{\text{base}}(\mathbf{y}, \theta)$ be an exchangeable Poisson model over the data $\{y_i\}$ with θ the Poisson rate. We can extend this model with a parameter $\lambda > 0$ to account for possible overdispersion in the data. This is commonly modeled with a negative binomial distribution

$$f(y \mid \theta, \lambda) = \binom{y + \lambda - 1}{y} \left(\frac{\theta}{\theta + \lambda}\right)^y \left(\frac{\lambda}{\theta + \lambda}\right)^\lambda.$$

  If λ is assigned an independent prior, then since $q_{\text{base}}(\mathbf{y}, \boldsymbol{\theta}) = \lim_{\lambda \to \infty} q(\mathbf{y}, \boldsymbol{\theta} \mid \lambda)$, this is again an expansion.

We will show in chapter 4 that model expansion of the above form tends to decrease identification. In particular, let Δ_q and Δ_{base} denote an identification function (as defined in the last section) for the *marginal* models $q(\mathbf{y}, \boldsymbol{\theta})$ and $q_{\text{base}}(\mathbf{y}, \boldsymbol{\theta})$ over only the shared parameters $\boldsymbol{\theta}$ respectively, where

$$q(\mathbf{y}, \boldsymbol{\theta}) = \int q(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\lambda}) d\boldsymbol{\lambda}. \tag{1.18}$$

We show that Δ_q tends to be smaller than Δ_{base}, increasing the risk of ϵ-weak identification for any fixed $\epsilon > 0$ in the higher-dimensional model when Δ is taken to be the variance- or entropy-based

identification function defined in the previous section. As we will see, the reason for this is essentially the same as in the regression case. Specifically, as we add parameters, it is unlikely that the variation in the data explained by the new parameters is entirely independent of the variation explained by the previous parameters. When there is some dependence between these parameters, as when regression predictors correlate and compete to explain the same outcomes, posterior identification suffers.

The tendency for Δ_q to fall relative to Δ_{base} is established through an identity of the form

$$\Delta_q = \Delta_{\text{base}} + C - B. \tag{1.19}$$

The term C represents a comparison between Δ_{base} and $\Delta_{q(\cdot|\boldsymbol{\lambda})}$ for $\boldsymbol{\lambda} \sim q(\boldsymbol{\lambda})$. Since $\Delta_{\text{base}} = \Delta_{q(\cdot|\boldsymbol{\lambda}_0)}$, we can think of $\boldsymbol{\lambda}$ as indexing a family of models and C as measuring identification in that family relative to a fixed point $\boldsymbol{\lambda}_0$. Depending on how the family is constructed, C could be arbitrarily large or small.

On the other hand, B is a term with is nonnegative so long as the joint prior $q(\boldsymbol{\theta}, \boldsymbol{\lambda})$ does not enforce too much dependence between the shared and expanded parameters. We therefore refer to as a bias term. The size of this term is directly related to the dimension k of the space of expanded parameters $\boldsymbol{\Lambda}$ and determines the severity of the downward trend in identification relative to model dimension. As we will show in chapter 4, examining the form of the terms C and B will point towards general modeling strategies which may help guard against weakening identification when applicable.

Notably, the choice of prior distribution plays an important role. As mentioned above, when $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are sufficiently strongly associated in the joint prior $q(\boldsymbol{\theta}, \boldsymbol{\lambda})$, then the weakening effect of expansion on identification (established through the term B in (1.19)) may be substantially attenuated or reversed. Importantly, this improvement to identification does not require any condition on the marginal distributions of the prior, and therefore can achieved with even very weak prior information about the individual scalar parameters separately.

Indeed, this is essentially the strategy employed by sparsifying priors in Bayesian regression, which also provide intuition for this property. When a regression possesses two highly correlated predictors $\mathbf{x}$ and $\mathbf{x}'$ with corresponding coefficients β and β', then large positive values of β can be approximately canceled by large negative values of β', resulting in larger uncertainty about the value of β relative to a model with only $\mathbf{x}$. However, a sparsifying prior will place lower probability on models with more coefficients far from 0. Such a prior will strongly favor cases where either β or β' is far from 0, but not both, enforcing strong negative dependence between the two coefficients. Since this essentially imposes a "soft" choice between including $\mathbf{x}$ or $\mathbf{x}'$ in the model, it enables us to retain marginal identification properties closer to the smaller models without explicitly excluding either predictor a priori (and therefore without needing strong prior information about β or β' marginally).

Thus, the fundamental challenge of high dimensionality to inference is to both find strategies to limit the tendency of model expansion to diminish identification while also leveraging the full posterior distribution to extract as much useful information as possible.

Chapter 2: Challenges to computation

This chapter is taken from [20], with Professor Cynthia Rush, to appear in IEEE Transactions on Information Theory.

As discussed in chapter 1, approximate Bayesian inference schemes gain substantial computational speed and flexibility at the cost of incurring an approximation error that is usually difficult or impossible to quantify. Such algorithms often proceed by approximating the true posterior distribution $q(\cdot \mid \mathbf{y})$ by a distribution $\widetilde{q}$ which has a computationally tractable form. In this chapter, we study the error of a pair of approximate message passing algorithms, which can be viewed as special cases of the expectation propagation (EP) algorithm introduced in [68].

2.1 Introduction

In order to make expectation propagation schemes computationally feasible, we must choose both a family of distributions, which is usually taken to be a parametric family $\mathcal{Q}^* = \{\widetilde{q} = q_{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda} \in \boldsymbol{\Lambda} \subset \mathbb{R}^d\}$, and a corresponding notion of discrepancy $D(q(\cdot \mid \mathbf{y}), \widetilde{q})$ between the true posterior and our approximation which our algorithm will (approximately) minimize. Expectation propagation attempts to minimize a certain KL divergence from the posterior to the approximating distribution. This divergence is given as

$$D(q(\cdot \mid \mathbf{y}), \widetilde{q}) = D\left(q(\cdot \mid \mathbf{y}) \mid \widetilde{q}\right) = \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})} \log\left(\frac{q(\boldsymbol{\theta} \mid \mathbf{y})}{\widetilde{q}(\boldsymbol{\theta})}\right). \tag{2.1}$$

The corresponding approximating family for EP is then constructed in two parts:

1. Choose a factorization of the posterior distribution:

$$q(\boldsymbol{\theta} \mid \mathbf{y}) \propto \prod_{m=1}^{M} q^{(m)}(\boldsymbol{\theta}_m), \tag{2.2}$$

   where the q_m are (proportional to) probability distributions and $\boldsymbol{\theta}_m$ is the vector consisting of those components of $\boldsymbol{\theta}$ on which q_m depends.

2. Approximate each factor $q^{(m)}$ by a distribution $\widetilde{q}^{(m)}$ taken from an approximating family $\mathcal{Q}^*$.

EP then proceeds by iteratively constructing a sequence $\widetilde{q}_k^{(m)}$ for $m = 1, \ldots, M$ and $k = 1, 2, 3, \ldots$ which attempts to minimize a corresponding KL divergence. Of course, our posterior always trivially satisfies the factorization (2.2) with $M = 1$ factor. The power of expectation propagation usually comes in choosing a nontrivial factorization for which the corresponding factor updates for $\widetilde{q}^{(m)}$ are substantially simpler than in the trivial single factor case.

To see how this can work, we must first elaborate the process for updating the individual approximate factors $\widetilde{q}^{(m)}$. First we define the gap distributions, which leave out a single approximate factor for each $1 \leq m \leq M$:

$$q^{\setminus m} = \prod_{m' \neq m} \widetilde{q}^{(m')}. \tag{2.3}$$

In order to update the factor $\widetilde{q}_k^{(m)}$, EP attempts to minimize the KL divergence:

$$\widetilde{q}_{k+1}^{\setminus m} = \arg\min_{\widetilde{q}^{\setminus m}} D\left(q^{(m)} \widetilde{q}_k^{\setminus m} \mid \widetilde{q}_k\right). \tag{2.4}$$

The result of this minimization gives an updated gap distribution $\widetilde{q}_{k+1}^{\setminus m}$. In order to construct the updated approximate factor for m, we simply divide:

$$\widetilde{q}_{k+1}^{(m)} = \frac{\widetilde{q}_k}{\widetilde{q}_{k+1}^{\setminus m}}. \tag{2.5}$$

Of course, for this step to be valid, the quotient on the right hand side above must also lie in the approximating family $\mathcal{Q}^*$. Furthermore, it is still not apparent how the KL divergence (2.4)

can be minimized in practice. Conceptually, this may be simpler than minimizing the global KL divergence (2.1) since the information from the posterior is only incorporated one factor at a time. But in order to address these validity and feasibility questions in detail, we must first be more specific about the form of the approximating family Q^*.

In particular, we will henceforth assume that the approximating family has the particular parametric form of an exponential family:

$$Q^* = \left\{ q_{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda} \in \mathbb{R}^N \right\}, \tag{2.6}$$

where

$$q_{\boldsymbol{\lambda}}(\cdot) \propto \exp\left(\sum_{n=1}^{N} \boldsymbol{\lambda}_n h_n(\cdot) \right), \tag{2.7}$$

and where $h_n : \boldsymbol{\Theta} \to \mathbb{R}$ for $1 \leq n \leq N$ are called the sufficient statistics for the family Q^*. Such exponential families possess a few essential properties which make them natural choices as approximating families for expectation propagation:

- If $q_{\boldsymbol{\lambda}}, q_{\boldsymbol{\lambda}'} \in Q^*$, then $q_{\boldsymbol{\lambda}} q_{\boldsymbol{\lambda}'} = q_{\boldsymbol{\lambda}+\boldsymbol{\lambda}'} \in Q^*$ and $q_{\boldsymbol{\lambda}} / q_{\boldsymbol{\lambda}'} = q_{\boldsymbol{\lambda}-\boldsymbol{\lambda}'} \in Q^*$.
- For any distribution q (not necessarily in Q^*), we have that

$$\widetilde{q} = \arg\min_{\widetilde{q}' \in Q^*} D\left(q \mid \widetilde{q}'\right) \text{ if and only if } \mathbb{E}_q h_n(\boldsymbol{\theta}) = \mathbb{E}_{\widetilde{q}} h_n(\boldsymbol{\theta}) \tag{2.8}$$

  for all $1 \leq n \leq N$ [68].

The first property means that the when the factors $\widetilde{q}^{(m)}$ lie in Q^*, the gap distributions $\widetilde{q}^{\backslash m}$ also lie in Q^*, and the factor update (2.5) is also well-defined.

To see the importance of the second property, note that in most exponential families used in practice, the expected sufficient statistics can be written as (known) functions of the parameters, i.e. for each $1 \leq n \leq N$ there exists $\phi_n : \mathbb{R}^N \to \mathbb{R}$ such that

$$\mathbb{E}_{\widetilde{q}_{\boldsymbol{\lambda}}} h_n(\boldsymbol{\theta}) = \phi_n(\boldsymbol{\lambda}). \tag{2.9}$$

Thus, the second property of exponential families implies that if the expectations $E_n = \mathbb{E}_{\widetilde{q}^{\setminus m} q^{(m)}} h_n(\boldsymbol{\theta})$ can be approximated, then the gap distribution updates (2.4) can be carried out by solving the system of equations $E_n = \phi(\boldsymbol{\lambda})$. Thus, making the expectation propagation scheme work in practice comes down to choosing the posterior factorization and exponential family so that the expectations E_n are computable.

Before presenting some specific instances of this EP scheme, we note that the resulting algorithm proceeds by iteratively minimizing the per-factor KL divergences (2.4). We claimed at the beginning of our discussion that EP attempts to minimize the overall KL divergence (2.8). At this point we may ask whether iteratively minimizing the former will also minimize the latter. Unfortunately the answer in general is no, but in sufficiently nice models EP will succeed in finding an approximation that closely matches the expected sufficient statistics of the true posterior.

2.1.1 Vectorized approximate message passing from EP

We now turn to a specific implementation of these EP updates in the form of the (Generalized) Vector Approximate Message Passing algorithm (G)VAMP for inference in (generalized) linear models. In particular, we consider estimating an unknown coefficient vector or signal $\mathbf{x}_0$, assuming knowledge of both the output $\mathbf{y}$ and matrix $\mathbf{A}$, using models of the form

$$\mathbf{y} \sim f\left(\mathbf{y} \mid \mathbf{A}\mathbf{x}_0\right), \qquad \mathbf{x}_0 \overset{iid}{\sim} q(x_0), \tag{2.10}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ with $M < N$, $q(x_0)$ is a prior distribution on the signal, and $f\left(\mathbf{y} \mid \cdot\right)$ is some output distribution that factors as $f(\mathbf{y} \mid \mathbf{z}) = \prod_{i=1}^{N} f(y_i \mid z_i)$. At times, we will give special consideration to the simple case where $f(\mathbf{y} \mid \cdot)$ is a Gaussian distribution and (2.10) describes the linear model: letting $g(\cdot)$ denote a normal density, the linear regression model in

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}, \qquad \mathbf{w} \overset{iid}{\sim} g(w), \qquad \mathbf{x}_0 \overset{iid}{\sim} q(x_0). \tag{2.11}$$

Algorithm 1 presents VAMP. We define a function $g_2 : \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^N$ as

$$g_2(\mathbf{r}_{2k}, \gamma_{2k}, \mathbf{y}) = (\gamma_w \mathbf{A}^T \mathbf{A} + \gamma_{2k} \mathbf{I})^{-1} (\gamma_w \mathbf{A}^T \mathbf{y} + \gamma_{2k} \mathbf{r}_{2k}), \tag{2.12}$$

where $\gamma_w = 1/\tau_w$ and $\tau_w < \infty$ is the elementwise variance of the noise $\mathbf{w}$. We assume that the noise variance τ_w is known to simplify the analysis, though, practically VAMP can be run without such knowledge [32]. The function g_2 in (2.12) has divergence with respect to the components of its first argument $\mathbf{r}_{2k}$ given by

$$\operatorname{div}\left[g_2(\mathbf{r}_{2k}, \gamma_{2k}, \mathbf{y})\right] = \sum_{i=1}^{N} \frac{\partial}{\partial [\mathbf{r}_{2k}]_i} g_2(\mathbf{r}_{2k}, \gamma_{2k}, \mathbf{y}) = \gamma_{2k} \operatorname{Tr}\left((\gamma_w \mathbf{A}^T \mathbf{A} + \gamma_{2k} \mathbf{I})^{-1} \right),$$

where $\operatorname{Tr}(\cdot)$ is the trace operator, i.e. it sum the diagonal elements of its argument. The function $g_1 : \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^N$ taking input $(\mathbf{r}_{1k}, \gamma_{1k})$ must be specified by the user. The choice of g_1, and

Algorithm 1 VAMP

Require: Number of iterations K, design matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, observed $\mathbf{y} \in \mathbb{R}^M$, and denoiser $g_1(\cdot, \gamma_{1k})$. (Note: $\mathbf{A}$ and $\mathbf{y}$ enter the algorithm through the definition of g_2 in (2.12).)

Initialize $\mathbf{r}_{10}$ and $\gamma_{10} \geq 0$.
for $k \leftarrow 0, \ldots, K$ **do**
  $\hat{\mathbf{x}}_{1k} \leftarrow g_1(\mathbf{r}_{1k}, \gamma_{1k})$, $\quad \alpha_{1k} \leftarrow \operatorname{div}\left[g_1(\mathbf{r}_{1k}, \gamma_{1k})\right]$,
  $\eta_{1k} \leftarrow \gamma_{1k}/\alpha_{1k}$, $\quad \gamma_{2k} \leftarrow \eta_{1k} - \gamma_{1k}$,
  $\mathbf{r}_{2k} \leftarrow \left(\eta_{1k}\hat{\mathbf{x}}_{1k} - \gamma_{1k}\mathbf{r}_{1k}\right)/\gamma_{2k}$,

  $\hat{\mathbf{x}}_{2k} \leftarrow g_2(\mathbf{r}_{2k}, \gamma_{2k}, \mathbf{y})$, $\quad \alpha_{2k} \leftarrow \operatorname{div}\left[g_2(\mathbf{r}_{2k}, \gamma_{2k}, \mathbf{y})\right]$,
  $\eta_{2k} \leftarrow \gamma_{2k}/\alpha_{2k}$, $\quad \gamma_{1(k+1)} \leftarrow \eta_{2k} - \gamma_{2k}$,
  $\mathbf{r}_{1(k+1)} \leftarrow \left(\eta_{2k}\hat{\mathbf{x}}_{2k} - \gamma_{2k}\mathbf{r}_{2k}\right)/\gamma_{1k}$.
end for

its relationship to g_2, can be better understood representing VAMP as an expectation propagation procedure for the regression model (2.11) with, specifically, Gaussian noise $\mathbf{w} \sim N(\mathbf{0}, \tau_w \mathbf{I}_M)$. Recall that the posterior distribution is proportional to the product of the data likelihood and the signal prior, so in our case is given by

$$q(\mathbf{x}_0 \mid \mathbf{y}) \propto f\left(\mathbf{y} \mid \mathbf{A}\mathbf{x}_0, \gamma_w^{-1}\mathbf{I}\right) \cdot \prod_{i=1}^{N} q([\mathbf{x}_0]_i), \tag{2.13}$$

where $f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal density evaluated at $\mathbf{y}$ having mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, so that $f\left(\mathbf{y} \mid \mathbf{A}\mathbf{x}_0, \gamma_w^{-1}\mathbf{I}\right)$ is the data likelihood in the case of Gaussian noise. In (2.11), we assume that $\mathbf{x}_0$ has i.i.d. elements, thus $\prod_{i=1}^N q([\mathbf{x}_0]_i)$ is the signal prior distribution. This provides the factorization of the posterior which will form the basis of our EP algorithm. The corresponding approximating family is taken to be the family of (multivariate) normal distributions with covariance of the form $c\mathbf{I}$ for $c > 0$.

First, expectation propagation approximates the factor $f(\mathbf{y} \mid \mathbf{A}\mathbf{x}_0, \gamma_w^{-1}\mathbf{I})$, by a normal density $\prod_{i=1}^N g([\mathbf{x}_0]_i \mid [\mathbf{r}_1]_i, \tau)$, for some means $[\mathbf{r}_1]_i$ and variance τ. This yields the approximate posterior

$$q(\mathbf{x}_0 \mid \mathbf{y}) \propto \prod_{i=1}^N g([\mathbf{x}_0]_i \mid [\mathbf{r}_1]_i, \tau_1\mathbf{I}) \cdot \prod_{i=1}^N q([\mathbf{x}_0]_i). \tag{2.14}$$

The relevant moment calculations for the distribution (2.14) reduce to one-dimensional integrals over each component $1 \leq i \leq N$ separately. In practice, these integrals can be computed efficiently using quadrature procedures. Importantly, reducing to one-dimensional integrals means that this step has $O(N)$ complexity.

Next, expectation propagation approximates the priors $q([\mathbf{x}_0]_i)$ by independent normal distributions with means $[\mathbf{r}_2]_i$ and equal variances τ_2, giving the approximate posterior

$$q(\mathbf{x}_0 \mid \mathbf{y}) \propto f\left(\mathbf{y} \mid \mathbf{A}\mathbf{x}_0, \gamma_w^{-1}\mathbf{I}\right) \cdot g\left(\mathbf{x}_0 \mid \mathbf{r}_2, \tau_2\mathbf{I}\right). \tag{2.15}$$

Now the model (2.15) is essentially just a conjugate normal regression model, and the relevant moment calculations can be performed analytically. These conjugacy equations will involve matrix operations which will generally have super-linear time complexity. However, as we will see, we can simplify the computations by computing the singular value decomposition of the matrix $\mathbf{A}$ once at the beginning of the algorithm. Fixing the dimension of the data $\mathbf{y}$, this allows the updates for (2.15) to achieve $O(N)$ per-iteration cost.

In the context of EP, the functions g_1 and g_2 defined above perform the factor EP updates for (2.14) and (2.15) respectively. The definitions of these functions depend on both the signal

prior and desired posterior summary. In (2.15), the prior is modeled as Gaussian regardless of the true prior. Since the resulting approximate posterior is again Gaussian, and the MMSE and MAP estimates are identical, (2.12) is the natural choice for g_2, as it calculates the mean of (2.15). However, (2.14) depends on the choice of (i.i.d.) prior $p(x_0)$ and the desired summary, so g_1 must be specified by the user accordingly.

Next we turn to the GVAMP iteration. For GVAMP, the user must supply two denoising functions, g_{x1} and g_{z1}, because now both the prior and the output distributions are user-specified. Indeed, the function g_{x1} is meant to combine an estimate $\mathbf{r}_{2k}$ of $\mathbf{x}_0$ with the prior information $p(x_0)$ to produce a new estimate of $\mathbf{x}_0$. Likewise, the function g_{z1} is meant to combine an estimate $\mathbf{p}_{2k}$ of $\mathbf{z}_0$ with the specified output distribution $f(y \mid z)$ to produce an updated estimate of $\mathbf{z}_0$.

We define two other denoising functions in terms of the singular value decomposition $\mathbf{A} = \mathbf{USV}^T$ with $\mathbf{U} \in \mathbb{R}^{M\times M}$, $\mathbf{V} \in \mathbb{R}^{N\times N}$ orthogonal, and $\mathbf{S} \in \mathbb{R}^{M\times N}$ the rectangular diagonal matrix of singular values. Let $g_{x2} : \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}^N$ and $g_{z2} : \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}^M$ be defined as

$$g_{x2}\left(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \gamma_{2k}, \tau_{2k}\right) = \mathbf{V}\mathbf{D}_k\left(\tau_{2k}\mathbf{S}^T\mathbf{U}^T\mathbf{p}_{2k} + \gamma_{2k}\mathbf{V}^T\mathbf{r}_{2k}\right), \tag{2.16}$$

$$g_{z2}\left(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \gamma_{2k}, \tau_{2k}\right) = \mathbf{U}\mathbf{S}\mathbf{D}_k\left(\tau_{2k}\mathbf{S}^T\mathbf{U}^T\mathbf{p}_{2k} + \gamma_{2k}\mathbf{V}^T\mathbf{r}_{2k}\right). \tag{2.17}$$

where

$$\mathbf{D}_k = \operatorname{diag}(\mathbf{d}_k), \text{ with } [\mathbf{d}_k]_n = \frac{1}{\tau_{2k}s_n^2 + \gamma_{2k}} \text{ and } s_n = \begin{cases} [\mathbf{S}]_{nn}, & n \le M, \\ 0, & M < n \le N. \end{cases}$$

The divergences of these functions with respect to the components of $\mathbf{r}_{2k}$ and $\mathbf{p}_{2k}$, respectively, are

$$\operatorname{div}[g_{x2}\left(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \gamma_{2k}, \tau_{2k}\right)] = \sum_{n=1}^{N} \frac{\gamma_{2k}}{\tau_{2k}s_n^2 + \gamma_{2k}}, \qquad \operatorname{div}[g_{z2}\left(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \gamma_{2k}, \tau_{2k}\right)] = \sum_{n=1}^{N} \frac{\tau_{2k}s_n^2}{\tau_{2k}s_n^2 + \gamma_{2k}}. \tag{2.18}$$

The above formulas follow because $\operatorname{Tr}(\mathbf{AB}) = \operatorname{Tr}(\mathbf{BA})$ when both products are defined. As for

VAMP, the iteration dependence of the denoising is entirely controlled by parameters γ_{2k} and τ_{2k}.

With this setup, we specify the GVAMP algorithm in Algorithm 2. As in the case of VAMP,

Algorithm 2 GVAMP

Require: Number of iterations K, design matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, observed $\mathbf{y} \in \mathbb{R}^M$, and denoisers $g_{x1}(\cdot, \gamma_{1k})$ and $g_{z1}(\cdot, \tau_{1k}, \mathbf{y})$. (Note: $\mathbf{A}$ and $\mathbf{y}$ enter the algorithm through g_{x2}, g_{z2}, and g_{z1} in (2.18).)

Initialize $\mathbf{r}_{10}$, $\mathbf{p}_{10}$, $\gamma_{10} \geq 0$, and $\tau_{10} \geq 0$.
for $k \leftarrow 0, \ldots, K$ **do**

$$\hat{\mathbf{x}}_{1k} \leftarrow g_{x1}(\mathbf{r}_{1k}, \gamma_{1k}), \qquad \alpha_{1k} \leftarrow \operatorname{div}\left[g_{x1}(\mathbf{r}_{1k}, \gamma_{1k})\right],$$
$$\mathbf{r}_{2k} \leftarrow (\hat{\mathbf{x}}_{1k} - \alpha_{1k}\mathbf{r}_{1k})/(1-\alpha_{1k}), \qquad \gamma_{2k} \leftarrow \gamma_{1k}\left(\tfrac{1}{\alpha_{1k}} - 1\right),$$

$$\hat{\mathbf{z}}_{1k} \leftarrow g_{z1}(\mathbf{p}_{1k}, \tau_{1k}, \mathbf{y}), \qquad \beta_{1k} \leftarrow \operatorname{div}\left[g_{z1}(\mathbf{p}_{1k}, \tau_{1k}, \mathbf{y})\right],$$
$$\mathbf{p}_{2k} \leftarrow (\hat{\mathbf{z}}_{1k} - \beta_{1k}\mathbf{p}_{1k})/(1-\beta_{1k}), \qquad \tau_{2k} \leftarrow \tau_{1k}\left(\tfrac{1}{\beta_{1k}} - 1\right),$$

$$\hat{\mathbf{x}}_{2k} \leftarrow g_{x2}(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \gamma_{2k}, \tau_{2k}), \qquad \alpha_{2k} \leftarrow \operatorname{div}\left[g_{x2}(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \gamma_{2k}, \tau_{2k})\right],$$
$$\mathbf{r}_{1(k+1)} \leftarrow (\hat{\mathbf{x}}_{2k} - \alpha_{2k}\mathbf{r}_{2k})/(1-\alpha_{2k}), \qquad \gamma_{1(k+1)} \leftarrow \gamma_{2k}\left(\tfrac{1}{\alpha_{2k}} - 1\right),$$

$$\hat{\mathbf{z}}_{2k} \leftarrow g_{z2}(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \gamma_{2k}, \tau_{2k}), \qquad \beta_{2k} \leftarrow \operatorname{div}\left[g_{z2}(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \gamma_{2k}, \tau_{2k})\right],$$
$$\mathbf{p}_{1(k+1)} \leftarrow (\hat{\mathbf{z}}_{2k} - \beta_{2k}\mathbf{p}_{2k})/(1-\beta_{2k}), \qquad \tau_{1(k+1)} \leftarrow \tau_{2k}\left(\tfrac{1}{\beta_{2k}} - 1\right).$$

end for

we can understand the GVAMP iteration as a (limit) of an EP procedure. In this case, we introduce one additional complication. For purposes of deriving an EP procedure, we replace our generalized linear model in which $\mathbf{z}_0 = \mathbf{A}\mathbf{x}_0$ with a noise injected model where $\mathbf{z}_0 \sim \text{normal}\left(\mathbf{A}\mathbf{x}_0, \gamma_w^{-1}\right)$. Making this adjustment introduces a normal factor into the model which allows for defining an EP iteration depending on γ_w. We will then recover an inference scheme for the original generalized linear model by taking the limit as $\gamma_w \to \infty$. The density of the noise-injected model is given as

$$\prod_{i=1}^{N} q([\mathbf{x}_0]_i) \prod_{j=1}^{M} q([\mathbf{z}_0]_j \mid [\mathbf{A}\mathbf{x}_0]_j, 1/\gamma_w) \prod_{j=1}^{M} f([\mathbf{y}]_j \mid [\mathbf{z}_0]_j). \tag{2.19}$$

Using the same normal approximate family as in the VAMP case, the first EP update computes

moments for the density

$$q_1(\mathbf{r}_{1k} \mid \gamma_{1k}) = \prod_{i=1}^{N} q([\mathbf{x}_0]_i)\, g([\mathbf{r}_{1k}]_i \mid [\mathbf{x}_0]_i, 1/\gamma_{1k}), \tag{2.20}$$

where $g(\cdot \mid \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. Note that we drop the approximate factor corresponding to the last factor in (2.19) above since the "true" factor $\prod_{i=1}^{N} q([\mathbf{x}_0]_i)$ does not involve the $\mathbf{z}_0$, and thus there is nothing to update for this factor. We can view g_{x1} as performing this EP update. The second factor update proceeds from the following model:

$$q_2(\mathbf{r}_{2k}, \mathbf{p}_{2k} \mid \gamma_{2k}, \tau_{2k}) = \prod_{i=1}^{N} g([\mathbf{x}_0]_i \mid [\mathbf{r}_{2k}]_i, 1/\gamma_{2k}) \prod_{j=1}^{M} q([\mathbf{z}_0]_j \mid [\mathbf{A}\mathbf{x}_0]_j, 1/\gamma_w)\, g([\mathbf{p}_{2k}]_j \mid [\mathbf{z}_0]_j, 1/\tau_{2k}). \tag{2.21}$$

We can view g_{x2} and g_{z2} as performing the EP inference for the parameters $\mathbf{x}_0$ and $\mathbf{z}_0$ of this model respectively in the limit as $\gamma_w \to \infty$ (i.e. as the linear constraint is made exact). The final factor update uses the last factor in (2.19):

$$q_3(\mathbf{p}_{1k} \mid \tau_{1k}) = \prod_{j=1}^{M} g([\mathbf{z}_0]_j \mid [\mathbf{p}_{1k}]_j, 1/\tau_{1k})\, f([\mathbf{y}]_j \mid [\mathbf{z}_0]_j). \tag{2.22}$$

The function g_{z1} can be seen as performing the EP update for the parameter $\mathbf{z}_0$ of this model.

2.1.2 Approximate message passing in general

The VAMP and GVAMP iterations can also be viewed as extensions of the original approximate message passing (AMP) iteration. From this perspective, AMP-style algorithms are a class of iterative methods for solving various high-dimensional statistical estimation and inference problems [28, 73, 9, 56, 76]. In its original form, AMP is the following two-step iteration for recovering $\mathbf{x}_0$ from knowledge of $\mathbf{y}$ and $\mathbf{A}$ under the linear model of (2.11) and the assumption that $\mathbf{A}$ is elementwise i.i.d. Gaussian. At iteration $k \geq 0$, the algorithm updates its estimate of the signal $\mathbf{x}_0$

with estimates $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots$. Initializing with $\hat{\mathbf{x}}_{-1} = \mathbf{v}_{-1} = \mathbf{1}$, calculate for $k \geq 0$,

$$\mathbf{r}_k = \hat{\mathbf{x}}_{k-1} + \mathbf{A}^T \mathbf{v}_{k-1}, \qquad \hat{\mathbf{x}}_k = g_k(\mathbf{r}_k), \tag{2.23}$$

$$\mathbf{v}_k = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}_k + \frac{N}{M}\mathbf{v}_{k-1}\mathrm{div}[g_k(\mathbf{r}_k)], \tag{2.24}$$

where $g_k : \mathbb{R} \to \mathbb{R}$, which acts elementwise on its vector input, is the so-called 'denoiser' function, an appropriately-chosen Lipschitz function depending on the estimation procedure to be performed, and $\mathrm{div}[g_k(\mathbf{r})] = \sum_{i=1}^{N} \frac{\partial}{\partial [\mathbf{r}]_i} g_k([\mathbf{r}]_i)$ is the divergence of the denoiser, measuring the sensitivity of g_k at its input. The $\mathbf{v}_k$ update in (2.24) can be interpreted as a corrected residual: the usual residual, $\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}_k$, with a so-called 'Onsager correction' given by $\frac{N}{M}\mathbf{v}_{k-1}\mathrm{div}[g_k(\mathbf{r}_k)]$.

With the presence of the correction term in the residual step, it is possible under certain conditions to characterize the asymptotic distribution of $\mathbf{r}_k$ in the large system limit where $N \to \infty$ and $M/N \to \delta \in (0, 1)$ (see [9, 14]). In particular, for variances τ_k that can be characterized exactly by a scalar recursion referred to as state evolution, in this limit, the elements of the vectors $\mathbf{r}_k$ behave like samples from a Gaussian distribution with mean vector $\mathbf{x}_0$ and covariance matrix $\tau_k \mathbf{I}_N$, denoted henceforth as $N(\mathbf{x}_0, \tau_k \mathbf{I}_N)$, where $\mathbf{x}_0$ is the true signal and $\mathbf{I}_N$ is an $N \times N$ identity matrix.

More formally, it is shown in [9] that for all pseudo-Lipschitz (defined in Section 2.2) loss functions ϕ, in the large system limit, empirical averages converge to deterministic limits,

$$\frac{1}{N}\sum_{i=1}^{N} \phi([\mathbf{r}_k]_i, [\mathbf{x}_0]_i) \to \mathbb{E}\{\phi(X_0 + \tau_k Z, X_0)\}, \tag{2.25}$$

where $Z \sim N(0, 1)$ is independent of $X_0 \sim p(x_0)$. In [81], this distributional convergence is refined with a finite sample analysis showing that these empirical averages exhibit concentration around their limits with rates exponential in N up to $t = O\left(\frac{\log N}{\log \log N}\right)$ iterations.

However, asymptotic guarantees for the AMP algorithm in (2.23)-(2.24) have been primarily studied in settings where the elements of the design matrices $\mathbf{A}$ are i.i.d. sub-Gaussian [9, 7] (though, an active line of recent research studies *universality* properties of AMP algorithms

more generally; see Section 2.1.3) and the concentration results only when the elements are i.i.d. Gaussian. Furthermore, considering the linear models in (2.10)-(2.11), concentration results have only been given for AMP for the model in (2.11), which is limited to the case of Gaussian outputs. In particular, there has been no concentration results developed for generalized versions of AMP, which can handle models such as (2.10), or for AMP algorithms with measurement matrices that are not i.i.d. Gaussian, despite such algorithms playing an important role in high-dimensional statistics [22, 87, 26], wireless communications [45, 11], and many other applications [92, 70, 71, 72, 3, 100].

In this chapter, we will extend the finite sample analysis for AMP for the linear model in (2.11) given in [81] for a suite of AMP-style algorithms that can be used when it is assumed that **(i)** the data follows the generalized linear model in (2.10), **(ii)** the random design matrices $\mathbf{A}$ comes from a class of orthogonally invariant matrices (defined below), and **(iii)** both are assumed. In particular, generalized approximate message passing (GAMP) [76], is an AMP-style algorithm used to perform estimation when $\mathbf{A}$ is assumed to be i.i.d. Gaussian but the data come from the generalized linear regression model in (2.10). Moreover, vector approximate message passing (VAMP) and its generalized extension GVAMP are AMP-style algorithms, recently introduced in [77] and [84], designed to perform estimation for the models (2.11) and (2.10), respectively, but under a much larger class of random design matrices $\mathbf{A}$. In particular, this is the class of *orthogonally invariant* $\mathbf{A}$, which have the property that $\mathbf{AV}$ and $\mathbf{UA}$ have the same distribution as $\mathbf{A}$ for any orthogonal matrices $\mathbf{V} \in \mathbb{R}^{N\times N}$ and $\mathbf{U} \in \mathbb{R}^{M\times M}$ independent of $\mathbf{A}$. Unlike the sub-Gaussian case, this framework allows the random matrices $\mathbf{A}$ to have nearly arbitrarily singular values. In particular, we will only require that the singular values are bounded above by a universal constant as $N \to \infty$.

This work refines the existing GAMP and VAMP [77, 88, 76, 33] asymptotics by providing a finite sample analysis of GVAMP – which works under the assumption of rotationally invariant design matrices and generalized liner models (2.10) – and by extension to GAMP and VAMP as well. In particular, the non-asymptotic result for VAMP shown in this paper is given under only

slightly stronger conditions than are used in the proof of its large system limit asymptotics in [77]. In more detail, in [77], it was shown that empirical averages of pseudo-Lipschitz functions of the VAMP estimates $\frac{1}{N}\sum_{i=1}^{N}\phi\left([\hat{\mathbf{x}}_{1k}]_i, [\mathbf{x}_0]_i\right)$ converge in the large system limit (i.e. as $N \to \infty$) to their expected values under limiting distributions, which can be characterized exactly. Here, we use concentration tools to show that this convergence in fact occurs exponentially fast in N and demonstrate how these results can be extended to the GVAMP and GAMP algorithms as well.

To summarize, this work generalizes the previous finite sample analyses for the AMP algorithm simultaneously in two main ways:

- Our work covers GAMP and GVAMP algorithms for generalized linear problems in the form of 2.10, with output distributions constrained only by a certain log-concavity requirement. This encompasses many common data models not covered by the Gaussian linear model, including, for example, logistic and Poisson regression.

- Our analysis applies to a much less restrictive distributional assumption over the measurement matrix $\mathbf{A}$ than previous finite sample analyses. In particular, our measurement matrix model samples the singular vector matrices uniformly from their orthogonal groups and allows the singular values to be marginally drawn from any arbitrary distribution with bounded support. This setup is substantially more general than the case of i.i.d. Gaussian matrices, and can in particular accommodate arbitrarily poor conditioning of the measurement matrix.

We emphasize that while refining earlier asymptotic results to provide non-asymptotic guarantees is useful to clarify the effect of the iteration number on the problem dimension and the speed of convergence to the state evolution predictions, it also has other application-specific implications; e.g. the finite sample analyses of [81] were used to establish the error exponent for sparse regression codes with AMP decoding for the additive white Gaussian noise channel [82] and the results in this work will be similarly useful in establishing error exponent for sparse regression codes with GAMP or VAMP decoding for more general channel models considered in [45, 11]. Moreover, the analysis in [4] uses a finite-sample style argument as in [81] to study signal recovery in the sparse

spiked matrix estimation model under the assumption of changing signal sparsity with growing dimension.

Structurally, the proof we provide is based on an approach used for the finite sample analysis of AMP in [81]; though, there are critical differences in the required analysis for GVAMP. In short, we develop a suite of concentration of measure tools to prove that, as the algorithm runs, the output retains concentration around its expected values. The idea is that if the algorithm concentrates through iteration k, then we can prove that it will concentrate at iteration $k+1$, with only a slightly degraded rate for the concentration. The differences between our proof and that given in [81] are detailed in Section 2.3.7. At a high level, when studying generalized versions of AMP algorithms, one must employ the output vector $\mathbf{y}$ in the denoising functions, which is not the case in standard AMP. This causes added dependencies within the algorithm, as now the measurement matrix $\mathbf{A}$ is no longer independent of the denoiser functions. On the other hand, studying the algorithm under rotationally invariant assumptions on $\mathbf{A}$ (as opposed to i.i.d. Gaussian assumptions), also complicates the analyses by adding additional dependencies that need to be handled carefully in the concentration arguments (see point **(1)** in Section 2.3.7).

2.1.3 Related work

Recent work [59, 58] extends the finite sample analysis of [81] to a related AMP algorithm for spiked matrix estimation. This work shows rates exponential in N for up to $O\left(\frac{N}{\texttt{poly}\log N}\right)$ iterations by using novel proof methods, improving on the $O\left(\frac{\log N}{\log\log N}\right)$ iteration guarantees found in this paper and in [81]. In this work, we use a proof strategy that is common in the AMP literature: first one studies properties of the random matrix conditional on linear constraints and then inductively in the iteration k, shows concentration (or more traditionally, almost sure convergence) of various objects appearing in the AMP algorithm. This idea was first used to study AMP by Bayati and Montanari [9] but has roots in statistical physics and spin glass theory [14]. The super-exponential dependence on iteration number appears in our concentration rates as a consequence of the inductive aspect of part of the proof. This same dependence is obtained in the finite sample

analysis of the AMP algorithm in [81]. In particular, this dependence accrues through applications of our basic tools for proving concentration (i.e. the concentration for sums and products lemmas in Appendix B). Each iteration introduces a multiplicative factor of the iteration number in the dependence through application of Lemmas 9-11. As this dependence results from the basic structure of the proof, we do not see obvious routes for improvements unless one changes the proof style completely. Indeed, the recent work of [59] uses new analysis ideas to provide tighter control on the rates of concentration in the case of AMP for the spiked matrix problems. In short, they decompose the iterates of the AMP algorithm for the spiked matrix model into a signal component, a Gaussian component, and a residual term and then the concentration result boils down to controlling these residuals. In some sense, instead of studying the AMP algorithm directly and tracking rates of concentration across the iterations, these authors compare their AMP algorithm to an idealized version and demonstrate the concentration to the state evolution values by controlling the difference between the real algorithm and the idealized one across iterations. An intriguing open question is whether their proof technique can be extended to AMP for linear and generalized linear regression, as is studied here.

As we mentioned above, AMP algorithms have been demonstrated both empirically and theoretically to perform well over a wide range of problems and to converge faster than the simple iterative soft thresholding method. However, there are now many variants on the original AMP algorithm. Remaining within the realm of Gaussian design matrices, for example, works have extended the AMP analyses to non-separable denoising functions that do not act componentwise on their vector arguments and can therefore take advantage of correlation between entries of the unknown signal [10, 64], to matrices with independent entries and a blockwise variance structure [49, 27], or to incorporate additional side information on the signal in order to accommodate streaming estimation procedures [62]. As mentioned, GAMP is designed to accommodate general separable output distributions; similarly for GVAMP. In other words, GAMP extends AMP to solve problems like (2.10). MLAMP [65] takes this idea further, accommodating multi-layer problems where separable nonlinear channels are stacked between multiple linear transformations.

In addition to VAMP, there have been other approaches to extending AMP to work under more general assumptions on the measurement matrix beyond i.i.d. Gaussian entries. These include the orthogonal approximate message passing algorithm (OAMP) [63], extensions to matrices with i.i.d. sub-Gaussian entries [8, 24], semi-random matrices [29], and other generalizations of AMP for rotationally invariant matrices [30, 93, 74]. The OAMP algorithm is designed, like VAMP, to accommodate a larger class of design matrices $\mathbf{A}$. In particular, OAMP assumes that $\mathbf{A}$ has distribution which is orthogonally invariant. If $\mathbf{A}$ has singular value decomposition $\mathbf{USV}^T$, then this requires that $\mathbf{U}$ and $\mathbf{V}$ are uniformly distributed and $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{S}$ are mutually independent. Extending finite sample analyses to more general classes of AMP algorithms like those discussed above is an exciting avenue for future work.

The types of generalized versions of AMP we study here, namely GAMP and GVAMP, which are designed for data models such as (2.10), and VAMP, which allows to extend beyond i.i.d. Gaussian measurement matrices, have found many applications in the literature to date. These algorithms play an important role in high-dimensional statistics [22, 87, 26], wireless communications [45, 11], and many other applications [92, 70, 71, 72, 3, 100]. In particular, the finite sample analysis for VAMP and GAMP presented here can be used to find the error rates for the capacity-achieving sparse regression coding schemes introduced in [45, 11], which use VAMP and GAMP as decoders. Moreover, AMP algorithms are often used to establish computational limits of various high-dimensional statistics problems, as they are conjectured to be optimal amongst polynomial-time algorithms in some problem settings. In particular, finite sample analyses like that presented in this work, allow one to study such computational limits in the context of changing sparsity, where the fraction of nonzeros amongst the signal elements approaches zero as the sample size grows (as opposed to the fraction of nonzeros being fixed in the limit). For example, in a very simple Bernoulli-Rademacher signal model of rank-one matrix estimation from noisy observations, the work in [4] considers both information theoretic and computational limits, with the computational limits found via a finite sample analysis of the associated AMP algorithm. We believe the analyses in this work could be used to study computational limits for changing sparsity

regimes in more general problem settings.

As we shall see below, the i.i.d. (sub-)Gaussian assumption is stronger than is needed to characterize the asymptotic behavior of (G)VAMP. And if one restricts attention to VAMP, the orthogonally invariant condition can be relaxed to the less stringent right orthogonally invariance condition. Furthermore, (G)VAMP has been demonstrated empirically to converge faster than GAMP across a range of design matrices (see [77, Section VI]).

We finally mention that part of this work was presented at ISIT 2020 [21]. In particular, [21] provides the statement of the result for VAMP but does not include proof details, nor does it include the extension of these finite sample analyses to generalized versions of VAMP or AMP, namely to GVAMP or GAMP. These extensions are significant contributions beyond the original VAMP result, as they provide finite sample analyses for AMP algorithms for the generalized regression model of (2.10), whereas the VAMP result applied only to (2.11). Applications of the generalized versions of VAMP and AMP were discussed above.

2.1.4 Notation

Throughout, we use lowercase boldface letters like $\mathbf{v}$ to denote vectors and uppercase boldface letters like $\mathbf{V}$ to denote matrices. We use $\mathbf{V}^T$ to denote the transpose of $\mathbf{V}$. For matrices $\mathbf{V}$, we use $\mathbf{V}_{ij}$ to denote the element in the i^{th} row and j^{th} column. Similarly, $\mathbf{V}_{\cdot j}$ is the j^{th} columns of the matrix and $\mathbf{V}_{i\cdot}$ is the i^{th} row. For vectors $\mathbf{v}$, we use $[\mathbf{v}]_i$ to denote the i^{th} element of $\mathbf{v}$.

For general vectors $\mathbf{x}$ and $\mathbf{y}$, $q(\mathbf{x})$ denotes the probability density function of $\mathbf{x}$ and $q(\mathbf{x}|\mathbf{y})$ is the conditional probability density function of $\mathbf{x}$ given $\mathbf{y}$. When $q(\mathbf{x})$ is a (potentially multivariate) normal density, we will denote it by $g(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ the covariance matrix, and we let $N(\mu, \sigma^2)$ denote a Gaussian random variable with mean μ and variance σ^2.

2.1.5 Outline

The remainder of this chapter is organized as follows. In section 2.2, we state our main concentration result, Theorem (1), and introduce the state evolution recursion which exactly characterizes

the distributions of the limiting variables. In section 2.3, we state and prove a concentration result for a more general abstract recursion and then prove Theorem (1) by showing how to recover the (G)VAMP estimates from this abstract recursion. In section 2.3.7, we conclude with a discussion of future work.

2.2 Main result

Before stating our main concentration result for GVAMP, from which results for VAMP and GAMP follow, we first revisit the notion of empirical convergence, we provide some more details on the matrix assumptions required for the VAMP algorithm to perform well, and we spend some time discussing the assumptions under which our main results follow. First, we define the class of functions that act as test functions for assessing convergence and concentration.

Definition 1. *A function $\phi : \mathbb{R}^J \to \mathbb{R}$ is pseudo-Lipschitz of order 2, denoted $\phi \in \mathrm{PL}(2)$, if for vectors $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^J$ and some constant $L \geq 0$, it satisfies $|\phi(\mathbf{v}) - \phi(\mathbf{v}')| \leq L\|\mathbf{v} - \mathbf{v}'\| \left(1 + \|\mathbf{v}\| + \|\mathbf{v}'\|\right)$.*

With this concept, we can then define empirical convergence and concentration of vector sequences to L^2 random variables/distributions in the following way.

Definition 2 (Empirical convergence)**.** *We will say that a sequence of (random or deterministic) vectors $\{\mathbf{v}_n \in \mathbb{R}^J\}_{n=1}^{\infty}$ converges empirically (with 2nd moment) to $\mathbf{V} \in \mathbb{R}^J$ if $\frac{1}{N}\sum_{n=1}^{N} \phi(\mathbf{v}_n) \to \mathbb{E}\{\phi(\mathbf{V})\}$, almost surely for any $\phi \in \mathrm{PL}(2) : \mathbb{R}^J \to \mathbb{R}$. (Note that this implies $\mathbb{E}V^2 < \infty$ since $x \mapsto x^2 \in \mathrm{PL}(2)$).*

Definition 3 (Exponentially fast concentration)**.** *We will say that a random vector sequence $\{\mathbf{v}_n\}_{n=1}^{\infty} \subset \mathbb{R}^J$ concentrates exponentially fast (in N) on $\mathbf{V} \in \mathbb{R}^J$ if for all $\epsilon > 0$ and all $N \geq 0$,*

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{n=1}^{N} \phi(\mathbf{v}_n) - \mathbb{E}\phi(\mathbf{V})\Big| > \epsilon\Big) \leq K\exp(-kN\epsilon^2),$$

for any $\phi \in \mathrm{PL}(2) : \mathbb{R}^J \to \mathbb{R}$, where $K, k > 0$ are universal constants independent of ϕ and N.

If, for any finite $N \geq 0$, the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$ are drawn i.i.d. from a subgaussian distribution, then [81, Lemma B.4] shows that $\{\mathbf{v}_n\}$ concentrates on $\mathbf{v}_1$. In what follows, we simply write that a vector sequence converges empirically to or concentrates exponentially fast on some random variable, without explicitly stating that the random variable has a finite 2nd moment.

The following definitions allow us to characterize those matrices $\mathbf{A}$ to which our results apply.

Definition 4 (Haar Distribution)**.** *A matrix $\mathbf{V} \in \mathbb{R}^{N \times N}$ is Haar distributed on the class of $N \times N$ orthogonal matrices if $\mathbf{V}_0\mathbf{V} \overset{d}{=} \mathbf{V}$ for any non-random orthogonal matrix $\mathbf{V}_0 \in \mathbb{R}^{N \times N}$.*

Definition 5 (Orthogonal Invariance)**.** *For a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, let $\mathbf{s}$ be its vector of singular values and let $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ be its singular value decomposition, where $\mathbf{U} \in \mathbb{R}^{M \times M}$ and $\mathbf{V} \in \mathbb{R}^{N \times N}$ are orthogonal matrices and $\mathbf{S} \in \mathbb{R}^{M \times N}$ is the rectangular, diagonal matrix with $\mathbf{S}_{ii} = \mathbf{s}_i$ for $1 \leq i \leq M$. $\mathbf{A}$ is **orthogonally invariant** if $\mathbf{V}$ and $\mathbf{U}$ are independent of each other and of $\mathbf{S}$ and are **Haar distributed** on the groups of $N \times N$ and $M \times M$ orthogonal matrices, respectively. In particular, $\mathbf{S}$ may be deterministic. This implies that $\mathbf{U}_0\mathbf{A}\mathbf{V}_0 \overset{d}{=} \mathbf{A}$ for any orthogonal $\mathbf{U}_0 \in \mathbb{R}^{M \times M}$ and $\mathbf{V}_0 \in \mathbb{R}^{N \times N}$ which are independent of $\mathbf{A}$.* [1].

2.2.1 Assumptions for the main results

Now we state the assumptions under which we provide our main result for the GVAMP algorithm. We can remove some of these assumptions for the finite sample analysis of VAMP (see Remark 3).

Assumption 0. The initial estimates $\mathbf{r}_{10}$ and $\mathbf{p}_{10}$ are independent of each other, independent of the matrices $\mathbf{U}$ and $\mathbf{V}$, and independent of the ground-truth $\mathbf{x}_0$ and $\mathbf{z}_0$. Further, $\mathbf{r}_{10}$ and $\mathbf{p}_{10}$ concentrate on some random vector (R_{10}, P_{10}).

The output vector $\mathbf{y}$ can be expressed as $\mathbf{y} = h(\mathbf{z}_0, \mathbf{w})$ where $\mathbf{z}_0 = \mathbf{A}\mathbf{x}_0$ is the transformed input, $\mathbf{w}$ is an independent "disturbance" vector, and h is a measurable function. Next, the truth $\mathbf{x}_0$ is also independent of these other quantities. The components of $\mathbf{w}$ must be either (1) sampled i.i.d. from

[1] This definition is slightly stronger than the related condition that $\mathbf{A}\widetilde{\mathbf{U}} \overset{d}{=} \widetilde{\mathbf{V}}\mathbf{A} \overset{d}{=} \mathbf{A}$ for any orthogonal $\widetilde{\mathbf{U}} \in \mathbb{R}^{M \times M}$ and $\widetilde{\mathbf{V}} \in \mathbb{R}^{N \times N}$, independent of $\mathbf{A}$. This latter condition places no restriction on singular vectors corresponding to 0 singular values, so either $\mathbf{U}$ or $\mathbf{V}$ need not be Haar distributed over all columns.

a subgaussian distribution or (2) non-random sequences which are bounded independently of N and which converge in $PL(2)$ sense to a limiting random variable with finite second moment. The components of $\mathbf{x}_0$ must be independent of $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{S}$, must be bounded independently of N, and $\mathbf{x}_0$ must concentrate on a random variable X_0.

The entries of the singular value vector $\mathbf{s} \in \mathbb{R}_+^M$ are taken from a compact interval $[0, S_{\max}]$ with $S_{\max}$ independent of N. They may be deterministic or random, though in the former case they must converge in $PL(2)$ to a random variable with finite second moment, and in the latter case they must be i.i.d. Furthermore, $\mathbf{s}$ is independent of the singular vector matrices $\mathbf{U}$ and $\mathbf{V}$.

The initial precision estimates γ_{10} and τ_{10} are positive for all N and converge to some $\overline{\gamma}_{10}, \overline{\tau}_{10} > 0$, respectively. Finally, we assume that $\delta = \frac{M}{N}$ is bounded away from 0 independent of N.

Assumption 1. The design matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is orthogonally invariant. If $\mathbf{A}$ has singular value decomposition $\mathbf{USV}^T$, then this implies that $\mathbf{V}$ and $\mathbf{U}$ are independently Haar distributed on the group of $N \times N$ and $M \times M$ orthogonal matrices, denoted $\mathsf{O}(N)$ and $\mathsf{O}(M)$, respectively. The Haar property means that $\mathbf{V}'\mathbf{V} \overset{d}{=} \mathbf{V}$ and $\mathbf{U}'\mathbf{U} \overset{d}{=} \mathbf{U}$ for any other $\mathbf{V}' \in \mathsf{O}(N)$ and $\mathbf{U}' \in \mathsf{O}(M)$. Definitions of orthogonal invariance and the Haar distribution along with some useful properties of these kinds of random elements are given in Definitions 4 and 5.

Assumption 2. The log prior $\log q(x)$ and the log likelihood $\log f(y \mid z)$ (as a function of z) are concave and β-smooth, where $q(x)$ and $f(y \mid z)$ are given in (2.10). The latter property just requires that, for some $\beta > 0$,

$$-\frac{\partial^2}{\partial x^2} \log q(x) < \beta, \qquad -\frac{\partial^2}{\partial z^2} \log f(y \mid z) < \beta. \tag{2.26}$$

Furthermore, the second derivatives above are bounded away from 0 over any compact subset of their domains.

Assumption 3. The estimating functions g_{x1} and g_{z1} are separable[2], and both g_{x1} and g_{z1} and their derivatives are uniformly Lipschitz[3] at $\overline{\gamma}_{1k} \in (0, \infty)$ for all $k \geq 0$. When g_{x1} and g_{z1} are the

[2]A function $g : \mathbb{R}^J \times \mathbb{R} \to \mathbb{R}^J$ is *separable* if for $\mathbf{v} \in \mathbb{R}^J$ and $z \in \mathbb{R}$, there exist a function $\widetilde{g} : \mathbb{R}^2 \to \mathbb{R}$ for which $[g(\mathbf{v}, z)]_j = \widetilde{g}([\mathbf{v}]_j, z)$ for all $j \in [J]$.

[3]A function $\phi(\mathbf{w}, c)$ is uniformly Lipschitz at c_0 if there is an open neighborhood U of c_0 and a constant $L > 0$ such

MAP or MMSE estimators for the models (2.14), they are separable. Henceforth, we will slightly abuse notation by writing these functions as taking both vector and scalar input.

Assumption 4. The α_{ik} and β_{ik} are truncated to lie in some interval $[t_{\min}, t_{\max}] \subset (0, 1)$. The γ_{ik} and τ_{ik} are also clipped so that they lie in some intervals $[\gamma_{\min}, \gamma_{\max}]$ with $0 < \gamma_{\min} < \gamma_{\max} < \infty$ and $[\tau_{\min}, \tau_{\max}]$ with $0 < \tau_{\min} < \tau_{\max} < \infty$ respectively.

Assumption 5. We terminate the algorithm according to the following stopping criteria. First, we stop the iteration if $\tau_{1k} < \epsilon_1$ or $\tau_{2k} < \epsilon_2$. To define our second stopping criterion, we consider the residuals $\mathbf{r}_{ik} - \mathbf{x}_0$ for $i = 1, 2$. Let π_{1k}^{in} be the magnitude of the projection of $\mathbf{r}_{1k} - \mathbf{x}_0$ onto the space spanned by the previous residual vectors

$$\{(\mathbf{r}_{1j} - \mathbf{x}_0)\}_{j=0}^{k-1} \cup \{(\mathbf{r}_{2j} - \mathbf{x}_0)\}_{j=0}^{k-1}.$$

Let π_{2k}^{in} be defined similarly as the magnitude of the projection of $\mathbf{r}_{2k} - \mathbf{x}_0$ on all previous residual vectors. Then π_{jk}^{out} for $j = 1, 2$ are defined analogously in terms of the residuals in the output space: $\mathbf{p}_{1k} - \mathbf{z}_0$ and $\mathbf{p}_{2k} - \mathbf{z}_0$. Our analysis will imply that these projection magnitudes will converge to certain limits as $N \to \infty$. We stop the iteration if any of the residual vectors is nearly spanned the previous residuals in this limit, i.e. when $\lim_{N\to\infty} \|\mathbf{r}_{jk} - \mathbf{x}_0\|^2 - \left[\pi_{jk}^{\text{in}}\right]^2 < \epsilon_1'$ or $\lim_{N\to\infty} \|\mathbf{p}_{jk} - \mathbf{z}_0\|^2 - \left[\pi_{jk}^{\text{out}}\right]^2 < \epsilon_2'$.

Remark 1. *We note that the function h in **Assumption 0** accommodates the linear regression model* (2.11) *as well as many forms of the generalized linear model* (2.10)*. In particular, assume that*

1. *$f(y \mid z)$ is continuous and differentiable in z for all y, and*

2. *$f(y \mid z)$ has an invertible quantile function $Q(\cdot \mid z)$ for each z.*

*Then if we set $h(z, w) = Q(w \mid z)$ and take $\mathbf{w} \overset{iid}{\sim} \mathsf{uniform}(0, 1)$, $h(\mathbf{z}_0, \mathbf{w})$ has distribution $f(\mathbf{y} \mid \mathbf{z}_0)$, as required by **Assumption 0**. The function h is also implicitly constrained by the log concavity assumption on $f(y \mid z)$ (**Assumption 2**) as well as by the Lipschitz assumption on g_{z1} (**Assumption***

that $\phi(\cdot, c)$ is $L-$Lipschitz for any fixed $c \in U$ and $|\phi(\mathbf{w}, c_1) - \phi(\mathbf{w}, c_2)| \leq L\,(1 + \|\mathbf{w}\|)\,|c_1 - c_2|$ for all $c_1, c_2 \in U$.

3*) since this denoiser leverages the likelihood* $p(y = h(z, w) \mid \cdot)$ *for estimating* $\mathbf{z}_0$*. With* h *equal to the quantile function as above,* h *is invertible in* w *for each* z*, and log concavity of* $f(y \mid z)$ *w.r.t.* z *is equivalent to log-concavity of the function* $\frac{\partial}{\partial z} h^{-1}(w; z)$ *where* $h^{-1}(w; z)$ *is the inverse of* $h(z, w)$ *with respect to* w *for* z *fixed.*

Remark 2. *The truncation in* ***Assumption 4*** *prevents the* $\mathbf{r}$ *and* $\mathbf{p}$ *updates from blowing up and thus assists with convergence. From the point of view of concentration, the truncation allows us to control the exponential rate of convergence. Truncating the* α_{ik} *and* β_{ik} *is not necessary if the second derivatives in* (A.125) *are also lower bounded by some* $\alpha > 0$ *(in which case the densities are also* α*-strongly log-concave).*

Remark 3. *For VAMP, some simplifications of the above assumptions are possible. Specifically:*

- *In* ***Assumption 0****, the requirement that* $\mathbf{y} = h(\mathbf{z}_0, \mathbf{w})$ *is automatic for VAMP since we can take* $h(z, w) = z + w$ *with* $\mathbf{w} \sim \text{normal}(0, \sigma)$.
- *In* ***Assumption 1****, we only need that* $\mathbf{A}$ *is right orthogonally invariant, which only imposes the Haar assumption on* $\mathbf{V}$ *(while* $\mathbf{U}$ *may be taken to be an arbitrary deterministic matrix).*
- *In* ***Assumption 2****, the latter inequality in* (A.125) *is automatic in the VAMP case (since the output is normal by definition).*
- *For VAMP, the requirements of* ***Assumption 3*** *apply to the estimating functions* g_1 *and* g_2.

Remark 4. ***Assumption 2*** *can be replaced with the weaker condition that the quantities* $\overline{\alpha}_{1k}$*,* $\overline{\alpha}_{2k}$*,* $\overline{\beta}_{1k}$*,* $\overline{\alpha}_{2k}$ *(introduced in Section 2.2.3 below) can be contained within the truncation interval* $[t_{\min}, t_{\max}]$ *defined in* ***Assumption 4****. See, e.g., Lemma 1 in [78] for a proof that* ***Assumption 2*** *implies this weaker condition.*

Remark 5. *We note that these assumptions do exclude our analysis from applying to (G)VAMP using a denoiser* g_1 *having a derivative that is piecewise constant and, thus, not uniformly Lipschitz, like the soft-thresholding function. However, if the derivative is bounded (like in the case*

of the soft-thresholding function), the analysis will still apply to appropriately constructed smooth approximations to the function that would provide qualitatively similar inference.

For VAMP, our assumptions are only slightly more restrictive than those used for the asymptotic analyses in [77], with the main differences being that we assume (i) exponentially fast concentration rather than convergence of input quantities and (ii) clipping and truncating of scalar iterates in Assumption 4 (though, [77] also suggest clipping γ as a practical matter of algorithm stability).

2.2.2 Main results

Rangan *et al.* [77] show empirical convergence of the vector sequence $\{([\hat{\mathbf{x}}_{1k}]_i, [\mathbf{r}_{1k}]_i, [\mathbf{x}_0]_i)\}_{i=1}^N$ to $(\hat{X}_{1k}, R_{1k}, X_0)$ for the case of the VAMP iterates where $R_{1k} = X_0 + \sqrt{\tau_{1k}}\, Z$ with $\hat{X}_{1k} = g_1(R_{1k}, \overline{\gamma}_{1k})$ and $Z \sim N(0,1)$ independent of $X_0 \sim p(x_0)$. The constants τ_{1k} and $\overline{\gamma}_{1k}$ that describe the limiting variable $\hat{X}_{1k}$ can be characterized exactly by the state evolution equations for VAMP, which we turn our attention to in Section 2.2.3 after we present our main result, Theorem 1, characterizing the asymptotic rate of this empirical convergence for VAMP and GVAMP.

Theorem 1. ***[VAMP and GVAMP Concentration.]*** *Under* ***Assumption 0*** *-* ***Assumption 5*** *given above, for any $\phi \in \mathrm{PL}(2) : \mathbb{R}^{k+1} \to \mathbb{R}$, any $k \geq 0$, and any $\epsilon \in (0,1)$,*

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} \phi([\hat{\mathbf{x}}_{10}]_i, \ldots, [\hat{\mathbf{x}}_{1k}]_i, [\mathbf{x}_0]_i) - \mathbb{E}\{\phi(\hat{X}_{10}, \ldots, \hat{X}_{1k}, X_0)\}\Big| \geq \epsilon\Big) \leq C' C_k e^{-c' c_k N \epsilon^2}, \quad (2.27)$$

where $C_k = C^{2k}(k!)^{16}$ and $c_k = \frac{1}{c^{2k}(k!)^{26}}$ with C, C' and c, c' being universal constants not depending on ϵ or N. Furthermore, for each j, $\hat{X}_{1j} = g_1(R_{1j}, \overline{\gamma}_{1j})$ where $R_{1j} = X_0 + \sigma_{1j} Z_j$ with $Z_j \sim N(0,1)$, and where $(Z_0, \ldots, Z_k)$ is jointly Gaussian and independent of X_0. The constants σ_{1k} and $\overline{\gamma}_{1k}$ are defined in Eq. (2.30) *in Section 2.2.3. In* (2.27)*, the vector $\hat{\mathbf{x}}_{1k}$ is that given in either Algorithm 2 or Algorithm 1.*

Remark 6. *We emphasize that Theorem 1 is true for both GVAMP (Algorithm 2) and VAMP (Algorithm 1), while the assumptions can be slightly simplified for VAMP as discussed in Remark 3.*

Moving from proving VAMP concentration to proving GVAMP concentration only requires the addition of an 'output' channel (whereas as the 'input' channel is analyzed in both cases) and the analysis of the 'output' channel is symmetric to that of the 'input' channel. There is a technical difficulty in that GVAMP requires analyzing denoising functions depending on the vector $\mathbf{z}_0 = \mathbf{A}\mathbf{x}_0$*, whereas a direct analysis of the VAMP algorithm would not consider this, but our theoretical analysis of the GVAMP algorithm below still implies the equivalent result for the (simplified) VAMP algorithm.*

The proof of Theorem 1 is given in Section 2.3.6. The key to proving Theorem 1 comes in Lemma 4, which establishes that the GVAMP iterates are jointly equal in distribution to the iterates of another recursion defined entirely in terms of functions of i.i.d. Gaussian variables. The proof strategy for our main technical concentration lemma (Lemma 6, which implies Theorem 1) then boils down to using the theory concentration for transformations of Gaussians to inductively establish the desired results for this distributionally equivalent recursion.

We finally mention that the proof of Theorem 1 implies a similar concentration result for GAMP as well. This is because, conceptually, we have seen that upgrading the finite sample analysis of an AMP-style algorithm (VAMP) to its generalized version (GVAMP) only requires repeating the same analysis for a symmetric 'output' channel and including a few more assumptions to deal with dependencies between the measurement matrix in the algorithm and the dependence of the denoiser on the output. Hence, GAMP concentration follows from the proof in [81].

To save space and to avoid introducing further notation, we only state an informal theorem of this result here, and we refer to the statement of the GAMP algorithm in [31, Section 4 Equation (55)] and its corresponding state evolution in [31, Section 4 Equation (57)-(58)].

Theorem 2. ***[Informal GAMP Concentration.]*** *Considering the GAMP algorithm in [31, Section 4 Equation (55)] and its state evolution in [31, Section 4 Equation (57)-(58)]. Under assumptions given in Appendix A.5, the asymptotic equivalence of its iterates and its state evolution (see for example for an asymptotic statement in [31, Section 4 Theorem 4.2]) can be upgraded to exponentially fast concentration with rates of* $CC_k e^{-cc_k N\epsilon^2}$ *where* $C_k = C^{2k}(k!)^{\kappa}$ *and* $c_k = \frac{1}{c^{2k}(k!)^{\kappa'}}$ *with*

κ, κ' and c, c' being universal constants not depending on ϵ or N. While we do not specify κ, κ' explicitly in this informal statement, these will be small constant values (like $\kappa = 16$ and $\kappa' = 22$ from Theorem 1).

The assumptions for this result are slightly different than for GVAMP. In particular, all assumptions about the matrix $\mathbf{A}$ above are replaced by the assumption that $\mathbf{A}$ has i.i.d. $N(0,1)$ entries (the same condition under which the large system analysis of GAMP was originally given). Furthermore, the initialization may now be randomly generated without the additional multiplication by $\mathbf{V}$ and $\mathbf{U}$ (as in **Assumption 0**). The remaining assumptions are analogous to those given above, and these represent natural extensions of the assumptions used for original large system analysis of GAMP. In particular, the Lipschitz condition on the denoisers must be extended to the uniformly Lipschitz condition in order to control the rate of change in the τ parameters, the initial data and disturbance vector $\mathbf{w}$ must now be drawn i.i.d. from sub-Gaussian distributions in order to ensure concentration at the appropriate rate, and stronger control is needed over the denoiser derivatives in order to control the probability of algorithm divergence.

2.2.3 State evolution

As mentioned previously, state evolution equations describe the dynamics of AMP-style algorithms. Generally, these are recursive relationships that characterize the effective signal-to-noise ratios at various stages of the algorithm. To specify the GVAMP state evolution, which provides the constants in the expectation in Theorem 1 we must define sensitivity and error functions for each of the approximate models (2.20), (2.21), and (2.22). Following [77], the sensitivity functions are

$$A_{x1}(\gamma_1, \sigma_1^2) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\{g'_{x1}([\mathbf{x}_0]_i + R_1, \gamma_1)\}, \qquad A_{x2}(\gamma_2, \tau_2) = \frac{\gamma_2}{N}\mathrm{Tr}\left[\left(\tau_2 \mathbf{S}^T\mathbf{S} + \gamma_2\right)^{-1}\right], \quad (2.28)$$

$$A_{z1}(\tau_1, \rho_1^2) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\{g'_{z1}(Z_0 + P_1, \tau_1, h\left(Z_0, [\mathbf{w}]_i\right))\}, \qquad A_{z2}(\gamma_2, \tau_2) = \frac{\tau_2}{M}\mathrm{Tr}\left[\mathbf{S}^T\mathbf{S}(\tau_2\mathbf{S}^T\mathbf{S} + \gamma_2)^{-1}\right], \quad (2.29)$$

where $R_1 \sim N(0, \sigma_1^2)$ and $P_1 \sim N(0, \rho_1^2)$ and the derivatives of g_{x1} and g_{z1} are taken with respect to their first input. Next, the error functions are defined as

$$\begin{aligned}
\mathcal{E}_{x1}(\gamma_1, \sigma_1^2) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(g_{x1}([\mathbf{x}_0]_i + R_1, \gamma_1) - [\mathbf{x}_0]_i))^2], \\
\mathcal{E}_{x2}(\gamma_2, \tau_2, \sigma_2^2, \rho_2^2) &= \frac{1}{N}\mathbb{E}[\|g_{x2}(\mathbf{r}_2, \mathbf{p}_2, \gamma_2, \tau_2) - \mathbf{x}_0\|^2], \\
\mathcal{E}_{z1}(\tau_1, \rho_1^2) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(g_{z1}(Z_0 + P_1, \tau_1, h\,(Z_0, [\mathbf{w}]_i)) - Z_0)^2], \\
\mathcal{E}_{z2}(\gamma_2, \tau_2, \sigma_2^2, \rho_2^2) &= \frac{1}{N}\mathbb{E}[\|g_{z2}(\mathbf{r}_2, \mathbf{p}_2, \gamma_2, \tau_2) - \mathbf{z}_0\|^2],
\end{aligned}$$

where $\mathbf{r}_2 \sim N(\mathbf{x}_0, \sigma_2^2\mathbf{I})$ and $\mathbf{p}_2 \sim N(\mathbf{z}_0, \rho_2^2\mathbf{I})$, and where $\mathbf{z}_0 = \mathbf{A}\mathbf{x}_0$. In terms of these functions, we define the scalar evolution equations for GVAMP. This sequence of equations determines the constants used in expectation in Theorem 1. Initializations $\overline{\gamma}_{10}$ and $\overline{\tau}_{10}$ are defined in **Assumption 0** as limits of the precision inputs to the GVAMP algorithm, and we initialize the σ and ρ terms as $\sigma_{10}^2 = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\{(R_{10} - [\mathbf{x}_0]_i)^2\right\}$ and $\rho_{10}^2 = \mathbb{E}\left\{(P_{10} - Z_0)^2\right\}$ where the random variables Z_0, R_{10}, and P_{10} are defined in **Assumption 0**. Then for $k \geq 0$,

$$\begin{aligned}
\overline{\alpha}_{1k} &= A_{x1}(\overline{\gamma}_{1k}, \sigma_{1k}^2), & \overline{\beta}_{1k} &= A_{z1}(\overline{\tau}_{1k}, \sigma_{2k}^2), \\
\sigma_{2k}^2 &= \frac{\mathcal{E}_{x1}(\overline{\gamma}_{1k}, \sigma_{1k}^2) - \overline{\alpha}_{1k}^2\sigma_{1k}^2}{(1-\overline{\alpha}_{1k})^2}, & \rho_{2k}^2 &= \frac{\mathcal{E}_{z1}(\overline{\tau}_{1k}, \rho_{1k}^2) - \overline{\beta}_{1k}^2\rho_{1k}^2}{(1-\overline{\beta}_{1k})^2}, \\
\overline{\gamma}_{2k} &= \overline{\gamma}_{1k}\Big(\frac{1}{\overline{\alpha}_{1k}} - 1\Big), & \overline{\tau}_{2k} &= \overline{\tau}_{1k}\Big(\frac{1}{\overline{\beta}_{1k}} - 1\Big), \\
\overline{\alpha}_{2k} &= A_{x2}(\overline{\gamma}_{2k}, \overline{\tau}_{2k}), & \overline{\beta}_{2k} &= A_{z2}(\overline{\gamma}_{2k}, \overline{\tau}_{2k}), \\
\sigma_{1(k+1)}^2 &= \frac{\mathcal{E}_{x2}(\overline{\gamma}_{2k}, \overline{\tau}_{2k}, \rho_{2k}^2, \sigma_{2k}^2) - \overline{\alpha}_{2k}^2\sigma_{2k}^2}{(1-\overline{\alpha}_{2k})^2}, & \rho_{1(k+1)}^2 &= \frac{\mathcal{E}_{z2}(\overline{\gamma}_{2k}, \overline{\tau}_{2k}, \rho_{2k}^2, \sigma_{2k}^2) - \overline{\beta}_{2k}^2\rho_{2k}^2}{(1-\overline{\beta}_{2k})^2}, \\
\overline{\gamma}_{1(k+1)} &= \overline{\gamma}_{2k}\Big(\frac{1}{\overline{\alpha}_{2k}} - 1\Big), & \overline{\tau}_{1(k+1)} &= \overline{\tau}_{2k}\Big(\frac{1}{\overline{\beta}_{2k}} - 1\Big).
\end{aligned} \tag{2.30}$$

This state evolution closely resembles the state evolution for VAMP derived in [77], with additional terms to account for the added complexity of the output distribution. In short, the VAMP state evolution is structurally the same recursion with only half the terms.

2.3 A general concentration result

Theorem 1 is a consequence of a concentration result for a more general iteration, Algorithm 3, which we introduce now. Let $f_p^{\text{in}}(p_1, p_2, w_p^{(1)}, w_p^{(2)}, \tau, \gamma) : \mathbb{R}^{2d} \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^d$ be a function of $2d+2$ arguments and two parameters, and let f_p^{out}, f_q^{in}, and f_q^{out} be defined similarly. Furthermore let $\mathbf{f}_p^{\text{in}}(\mathbf{p}_1, \mathbf{p}_2, \mathbf{w}_p^{\text{in}}, \tau, \gamma) : \mathbb{R}^{2d\times N} \times \mathbb{R}^{2\times N} \times \mathbb{R}^2 \to \mathbb{R}^{N\times d}$ be given by applying f_p^{in} separably over the N rows of its arguments.

We note that if $M < N$, then $\mathbf{p}_k^{\text{in}}$ and $\mathbf{p}_k^{\text{out}}$ (defined in Algorithm 3 below) have different numbers of columns. When this occurs, we must take care in defining the notion of separability. In general, we will encounter situations in which we want to extend a function $\phi(x, y, w) : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^2 \to \mathbb{R}^d \in PL(2)$ of $2d+2$ arguments x, y and w separably over vectorized inputs $\mathbf{x} \in \mathbb{R}^{d\times N}$, $\mathbf{y} \in \mathbb{R}^{d\times M}$ and $\mathbf{w}^{\text{in}} \in \mathbb{R}^{2\times N}$ or $\mathbf{w}^{\text{out}} \in \mathbb{R}^{2\times M}$ for which $M \le N$. When $\mathbf{w} \in \mathbb{R}^{2\times M}$, define the separable extension by simply truncating $\mathbf{x}$ to an $d \times M$ matrix, and the resulting extended function takes values in $\mathbb{R}^M$. When $\mathbf{w} \in \mathbb{R}^{2\times N}$, it will suffice for our purposes to consider only those ϕ for which $\phi([\mathbf{x}]_i, [\mathbf{y}]_i, [\mathbf{w}]_i)$ is independent of y for all $i > M$. In this case, the separable extension of ϕ over these vectorized arguments clearly makes sense. By convention, unless otherwise noted, we take $[\mathbf{y}]_i = \mathbf{0}$ for $i > M$ so that expressions like $\phi([\mathbf{x}]_i, [\mathbf{y}]_i, [\mathbf{w}]_i)$ are well-defined for all $i \le N$. In particular, we use these conventions to define separable extensions of the functions $\mathbf{f}_p^{\text{in/out}}(\cdot, \cdot, \cdot, \tau, \gamma)$ and $\mathbf{f}_q^{\text{in/out}}(\cdot, \cdot, \cdot, \tau, \gamma)$ over vectorized arguments/parameters in the following sections.

Now, letting $\mathbf{V}$ and $\mathbf{U}$ be orthogonal matrices, we can state Algorithm 3 which defines a recursion in terms of these quantities. Likewise, the VAMP iterates can be generated from a simpler general recursion given in Algorithm 4. The VAMP algorithm is not strictly a special case of GVAMP; similarly, Algorithm 4 is not a special case of Algorithm 3. However, Algorithm 4 can be seen as a subset of Algorithm 3 by discarding all the "output" steps in Algorithm 3 and making

Algorithm 3 General Recursion for GVAMP

Require: Orthogonal matrices $\mathbf{V} \in \mathbb{R}^{N\times N}$ and $\mathbf{U} \in \mathbb{R}^{M\times M}$; separable denoisers $\mathbf{f}_p^{\text{in}}$, $\mathbf{f}_p^{\text{out}}$, $\mathbf{f}_q^{\text{in}}$, and $\mathbf{f}_q^{\text{out}}$ (defined over their argument according to the conventions given above); parameter update functions $\Gamma_{p/q}^{\text{in/out}} : \mathbb{R}^d \to \mathbb{R}^d$; and disturbance vectors $\mathbf{w}_p^{\text{in}}, \mathbf{w}_q^{\text{in}} \in \mathbb{R}^{L\times N}$ and $\mathbf{w}_p^{\text{out}}, \mathbf{w}_q^{\text{out}} \in \mathbb{R}^{2\times M}$.

Initialize $\mathbf{u}_0^{\text{in}} \in \mathbb{R}^{N\times d}$, $\mathbf{u}_0^{\text{out}} \in \mathbb{R}^{M\times d}$, $\gamma_{pk}^{\text{in}} \in \mathbb{R}^d$, and $\gamma_{pk}^{\text{out}} \in \mathbb{R}^d$.

for $k \leftarrow 0, \ldots, K$ **do**

$\mathbf{p}_k^{\text{in}} \leftarrow \mathbf{V}^T\mathbf{u}_k^{\text{in}}$, $\quad \mathbf{p}_k^{\text{out}} \leftarrow \mathbf{U}^T\mathbf{u}_k^{\text{out}}$,

$\alpha_{pk}^{\text{in}} \leftarrow \text{div}\left[\mathbf{f}_p^{\text{in}}\left(\mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}, \mathbf{w}_p^{\text{in}}, \gamma_{pk}^{\text{out}}, \gamma_{pk}^{\text{in}}\right)\right]$, $\quad \alpha_{pk}^{\text{out}} \leftarrow \text{div}\left[\mathbf{f}_p^{\text{out}}\left(\mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}, \mathbf{w}_p^{\text{out}}, \gamma_{pk}^{\text{out}}, \gamma_{pk}\right)\right]$,

$[\mathbf{v}_k^{\text{in}}]_j \leftarrow \frac{1}{1-[\alpha_{pk}^{\text{in}}]_j}\left[\mathbf{f}_p^{\text{in}}\left(\mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}, \mathbf{w}_p^{\text{in}}, \gamma_{pk}^{\text{out}}, \gamma_{pk}^{\text{in}}\right)_j - [\alpha_{pk}^{\text{in}}]_j[\mathbf{p}_k^{\text{in}}]_j\right]$,

$[\mathbf{v}_k^{\text{out}}]_j \leftarrow \frac{1}{1-[\alpha_{pk}^{\text{out}}]_j}\left[\mathbf{f}_p^{\text{out}}\left(\mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}, \mathbf{w}_p^{\text{out}}, \gamma_{pk}^{\text{out}}, \gamma_{pk}^{\text{in}}\right)_j - [\alpha_{pk}^{\text{out}}]_j[\mathbf{p}_k^{\text{out}}]_j\right]$.

$\gamma_{qk}^{\text{in}} \leftarrow \Gamma_q^{\text{in}}(\gamma_{pk}^{\text{in}}, \alpha_{pk}^{\text{in}})$, $\quad \gamma_{qk}^{\text{out}} \leftarrow \Gamma_q^{\text{out}}(\gamma_{pk}^{\text{out}}, \alpha_{pk}^{\text{out}})$,

$\mathbf{q}_k^{\text{in}} \leftarrow \mathbf{V}\mathbf{v}_k^{\text{in}}$, $\quad \mathbf{q}_k^{\text{out}} \leftarrow \mathbf{U}\mathbf{v}_k^{\text{out}}$,

$\alpha_{qk}^{\text{in}} \leftarrow \text{div}\left[\mathbf{f}_q^{\text{in}}\left(\mathbf{q}_k^{\text{in}}, \mathbf{q}_k^{\text{out}}, \mathbf{w}_q^{\text{in}}, \gamma_{qk}^{\text{out}}, \gamma_{qk}^{\text{in}}\right)\right]$, $\quad \alpha_{qk}^{\text{out}} \leftarrow \text{div}\left[\mathbf{f}_q^{\text{out}}\left(\mathbf{q}_k^{\text{in}}, \mathbf{q}_k^{\text{out}}, \mathbf{w}_q^{\text{out}}, \gamma_{qk}^{\text{out}}, \gamma_{qk}^{\text{in}}\right)\right]$,

$[\mathbf{u}_{k+1}^{\text{in}}]_j \leftarrow \frac{1}{1-[\alpha_{qk}^{\text{in}}]_j}\left[\mathbf{f}_q^{\text{in}}\left(\mathbf{q}_k^{\text{in}}, \mathbf{q}_k^{\text{out}}, \mathbf{w}_q^{\text{in}}, \gamma_{qk}^{\text{out}}, \gamma_{qk}^{\text{in}}\right)_j - [\alpha_{qk}^{\text{out}}]_j[\mathbf{q}_k^{\text{in}}]_j\right]$,

$[\mathbf{u}_{k+1}^{\text{out}}]_j \leftarrow \frac{1}{1-[\alpha_{qk}^{\text{out}}]_j}\left[\mathbf{f}_q^{\text{out}}\left(\mathbf{q}_k^{\text{in}}, \mathbf{q}_k^{\text{out}}, \mathbf{w}_q^{\text{out}}, \gamma_{qk}^{\text{out}}, \gamma_{qk}^{\text{in}}\right)_j - [\alpha_{qk}^{\text{out}}]_j[\mathbf{q}_k^{\text{out}}]_j\right]$.

$\gamma_{p(k+1)}^{\text{in}} \leftarrow \Gamma_p^{\text{in}}(\gamma_{qk}^{\text{in}}, \alpha_{qk}^{\text{in}})$, $\quad \gamma_{p(k+1)}^{\text{out}} \leftarrow \Gamma_p^{\text{out}}(\gamma_{qk}^{\text{out}}, \alpha_{qk}^{\text{out}})$,

end for

the denoisers f_p^{in} and f_q^{in} depend only on the input variables. The resulting "input only" algorithm exactly recovers Algorithm 4.

As a result, all of the arguments regarding concentration of the iterates in Algorithm 3 can be specialized to arguments for concentration of the iterates of Algorithm 4. Henceforth, unless specified otherwise, all arguments and discussion will focus on Algorithm 3. To recover the relevant statements for Algorithm 4, it suffices to drop the "in" and "out" superscripts (and, when both occur, eliminate dependence on the "out" variable).

We note that the divergence terms $\boldsymbol{\alpha} \in \mathbb{R}^d$ in Algorithm 3 are computed component-wise, so

Algorithm 4 General Recursion for VAMP

Require: Orthogonal matrix $\mathbf{V} \in \mathbb{R}^{N\times N}$, separable (w.r.t. the first two arguments) denoisers

$$f_p : \mathbb{R}^{d\times N} \times \mathbb{R}^{d\times N} \times \mathbb{R} \to \mathbb{R}^{2\times N} \text{ and } f_q : \mathbb{R}^{d\times N} \times \mathbb{R}^{2\times N} \times \mathbb{R} \to \mathbb{R}^{N},$$

divergence functions $C_i : \mathbb{R} \to \mathbb{R}$ for $i \in \{1, 2\}$, parameter update functions $\Gamma_i : \mathbb{R}^2 \to \mathbb{R}$ for $i \in \{1, 2\}$, and disturbance vectors $\mathbf{w}^p \in \mathbb{R}^N$ and $\mathbf{w}^q \in \mathbb{R}^{2N}$.

Initialize $\mathbf{u}_0 \in \mathbb{R}^N$.
for $k \leftarrow 0, \ldots, K$ **do**
  $\mathbf{p}_k \leftarrow \mathbf{V}^T\mathbf{u}_k$, $\quad \alpha_{1k} \leftarrow \operatorname{div}[f_p(\mathbf{p}_k, \mathbf{w}^p, \gamma_{1k})]$,
  $\mathbf{v}_k \leftarrow C_1(\alpha_{1k}) \left[f_p(\mathbf{p}_k, \mathbf{w}^p, \gamma_{1k}) - \alpha_{1k}\mathbf{p}_k\right]$, $\quad \gamma_{2k} \leftarrow \Gamma_1\left(\gamma_{1k}, \alpha_{1k}\right)$,

  $\mathbf{q}_k \leftarrow \mathbf{V}\mathbf{v}_k$, $\quad \alpha_{2k} \leftarrow \operatorname{div}[f_q(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k})]$,
  $\mathbf{u}_{k+1} \leftarrow C_2(\alpha_{2k}) \left[f_q(\mathbf{q}_k, \mathbf{w}^q, \gamma_{2k}) - \alpha_{2k}\mathbf{q}_k\right]$, $\quad \gamma_{1(k+1)} \leftarrow \Gamma_2\left(\gamma_{2k}, \alpha_{2k}\right)$.
end for

e.g.

$$[\alpha^{\text{in}}_{pk}]_j = \sum_{i=1}^{N} \frac{\partial}{\partial p^{\text{in}}_j} \left[f^{\text{in}}_p\left(p^{\text{in}}_1, \ldots, p^{\text{in}}_d, p^{\text{out}}_1, \ldots, p^{\text{out}}_d, w_p, \gamma^{\text{out}}, \gamma^{\text{in}}\right)\right]_j \Bigg|_{\left([\mathbf{p}^{\text{in}}_k]_i, [\mathbf{p}^{\text{out}}_k]_i, [\mathbf{w}^{\text{in}}_p]_i, \gamma^{\text{in}}_{pk}, \gamma^{\text{out}}_p k\right)},$$

for each $j = 1, \ldots, d$.

We establish a correspondence between the general algorithm, Algorithm 3, and GVAMP in Algorithm 2, through the following translation. We take the orthogonal matrices $\mathbf{U}, \mathbf{V}$ in Algorithm 3 to be the left and right singular vector matrices of $\mathbf{A}$, respectively. Define the following: for $k \geq 0$,

$$\begin{aligned}
\mathbf{u}^{\text{in}}_k &= [\mathbf{r}_{2k} - \mathbf{x}_0,\ \mathbf{x}_0], \quad \mathbf{u}^{\text{out}}_k = [\mathbf{p}_{2k} - \mathbf{z}_0,\ \mathbf{0}], \\
\mathbf{p}^{\text{in}}_k &= \left[\mathbf{V}^T\left(\mathbf{r}_{2k} - \mathbf{x}_0\right),\ \mathbf{V}^T\mathbf{x}_0\right], \quad \mathbf{p}^{\text{out}}_k = \left[\mathbf{U}^T\left(\mathbf{p}_{2k} - \mathbf{z}_0\right),\ \mathbf{0}\right], \\
\mathbf{v}^{\text{in}}_k &= \left[\mathbf{V}^T\left(\mathbf{r}_{1(k+1)} - \mathbf{x}_0\right),\ \mathbf{0}\right], \quad \mathbf{v}^{\text{out}}_k = \left[\mathbf{U}^T\left(\mathbf{p}_{1(k+1)} - \mathbf{z}_0\right),\ \mathbf{U}^T\mathbf{z}_0\right], \\
\mathbf{q}^{\text{in}}_k &= \left[\mathbf{r}_{1(k+1)} - \mathbf{x}_0,\ \mathbf{0}\right], \quad \mathbf{q}^{\text{out}}_k = \left[\mathbf{p}_{1(k+1)} - \mathbf{z}_0,\ \mathbf{z}_0\right].
\end{aligned} \tag{2.31}$$

Furthermore, we make the following function definitions:

$$
\begin{aligned}
f_p^{\text{in}}\left(p^{\text{in}}, p^{\text{out}}, w_p, \tau, \gamma\right) &= \left[\left([\tau]_1[w_p]_1[p^{\text{out}}]_1 + [\gamma]_1[p^{\text{in}}]_1\right) / \left([\tau]_1[w_p]_1^2 + [\gamma]_1\right),\ 0\right], \\
f_p^{\text{out}}\left(p^{\text{in}}, p^{\text{out}}, w_p, \tau, \gamma\right) &= \left[[w_p]_1\left([\tau]_1[w_p]_1[p^{\text{out}}]_1 + [\gamma]_1[p^{\text{in}}]_1\right) / \left([\tau]_1[w_p]_1^2 + [\gamma]_1\right),\ [w_p]_1[p^{\text{in}}]_2\right], \\
f_q^{\text{in}}\left(q^{\text{in}}, q^{\text{out}}, w_q, \tau, \gamma\right) &= \left[g_{x1}\left([q^{\text{in}}]_1 + [w^q]_1, \gamma\right) - [w_q]_1,\ [w_q]_1\right], \\
f_q^{\text{out}}\left(q^{\text{in}}, q^{\text{out}}, w_q, \tau, \gamma\right) &= \left[g_{z1}\left([q^{\text{out}}]_1 + [q^{\text{out}}]_2, \tau, h\left([q^{\text{out}}]_2, [w_q]_1\right)\right) - [q^{\text{out}}]_2,\ 0\right].
\end{aligned}
\tag{2.32}
$$

Finally we define disturbance vectors and constants:

$$
\mathbf{w}_p^{\text{in}} = \mathbf{x}_0, \qquad \mathbf{w}_p^{\text{out}} = \mathbf{w}, \qquad \mathbf{w}_q^{\text{in}} = [(\mathbf{s}\ \mathbf{0}),\ [\mathbf{u}_0^{\text{in}}]_1], \qquad \mathbf{w}_q^{\text{out}} = [\mathbf{s},\ [\mathbf{u}_0^{\text{out}}]_1],
\tag{2.33}
$$

$$
\alpha_{pk}^{\text{in}} = (\alpha_{1k}, 0), \qquad \alpha_{qk}^{\text{in}} = \left(\alpha_{2(k+1)}, 0\right), \qquad \alpha_{pk}^{\text{out}} = (\beta_{1k}, 0), \qquad \alpha_{qk}^{\text{out}} = \left(\beta_{2(k+1)}, 0\right), \tag{2.34}
$$

$$
\gamma_{pk}^{\text{in}} = (\gamma_{1k}, 1), \qquad \gamma_{qk}^{\text{in}} = (\gamma_{2(k+1)}, 1), \qquad \gamma_{pk}^{\text{out}} = (\tau_{1k}, 1), \qquad \gamma_{qk}^{\text{out}} = (\tau_{2(k+1)}, 1). \tag{2.35}
$$

In (2.33), $(\mathbf{s}\ \mathbf{0})$ is the N-vector obtained by padding the positive singular values of $\mathbf{A}$ with 0's, and $[\mathbf{s}\ \ [\mathbf{u}_0^{\text{out}}]_1]$ is the $M \times 2$ matrix with columns $\mathbf{s}$ and $[\mathbf{u}_0^{\text{out}}]_1$. Keeping with our matrix notation, $[\mathbf{w}_p^{\text{out}}]_i$ will denote column i of this matrix, i.e. the vector $([\mathbf{z}_0]_i, [\mathbf{y}]_i)$. In all cases, where we perform truncation of the scalar quantities in GVAMP, we assume the same truncation here. We claim that these quantities satisfy the recursion in Algorithm 3. We start by establishing this for the scalar terms inductively. For the γ terms, if equality has been established for all scalar terms up to the current term, then equality follows for the current term immediately if we define the Γ functions as follows:

$$
\Gamma_p^{\text{in}}(\gamma, \alpha) = \Gamma_p^{\text{out}}(\gamma, \alpha) = \Gamma_q^{\text{in}}(\gamma, \alpha) = \Gamma_q^{\text{out}}(\gamma, \alpha) = \left(\frac{\gamma_1(1-\alpha_1)}{\alpha_1}, 1\right).
$$

For the α terms, we first observe that the equality follows immediately for α_{pk}^{in} and α_{pk}^{out} given the definitions of f_p^{in} and f_p^{out} respectively (assuming equality for all prior scalar terms). It follows directly from the definitions of f_q^{in} and f_q^{out} that

$$[\alpha_{qk}^{\text{in}}]_1 = \frac{1}{N}\sum_{i=1}^{N} \frac{[\gamma_{qk}^{\text{in}}]_1}{[\gamma_{qk}^{\text{out}}]_1[\mathbf{w}_q^{\text{in}}]_{i1}^2 + [\gamma_{qk}^{\text{in}}]_1}, \qquad [\alpha_{qk}^{\text{out}}]_1 = \frac{1}{M}\sum_{i=1}^{M} \frac{[\mathbf{w}_q^{\text{out}}]_{i1}^2[\gamma_{qk}^{\text{out}}]_1}{[\gamma_{qk}^{\text{out}}]_1[\mathbf{w}_q^{\text{out}}]_{i1}^2 + [\gamma_{qk}^{\text{in}}]_1}.$$

Substituting $(\mathbf{s}\ \mathbf{0})$ for $[\mathbf{w}_q^{\text{in}}]_1$ and $\mathbf{s}$ for $[\mathbf{w}_q^{\text{out}}]_1$, it follows that these are exactly the divergences of g_{x2} and g_{z2} derived in [84], and hence are equal to α_{2k} and β_{2k}, respectively.

Observe that lines 3 and 8 in Algorithm 3 follow trivially from the definitions in (2.31) and that

$$\begin{aligned}
[\mathbf{u}_{k+1}^{\text{in}}]_1 = \mathbf{r}_{2(k+1)} - \mathbf{x}_0 &\overset{(a)}{=} \frac{1}{1-\alpha_{1(k+1)}}\left[g_{x1}\left(\mathbf{r}_{1(k+1)}, \gamma_{1(k+1)}\right) - \alpha_{1(k+1)}\mathbf{r}_{1(k+1)}\right] - \mathbf{x}_0 \\
&\overset{(b)}{=} \frac{1}{1-[\alpha_{qk}^{\text{in}}]_1}\left[g_{x1}\left([\mathbf{q}_k^{\text{in}}]_1 + [\mathbf{w}_q^{\text{in}}]_1, [\gamma_{qk}^{\text{in}}]_1\right) - [\alpha_{qk}^{\text{in}}]_1[\mathbf{q}_k^{\text{in}}]_1 - [\alpha_{qk}^{\text{in}}]_1[\mathbf{w}_q^{\text{in}}]_1\right] - [\mathbf{w}_q^{\text{in}}]_1 \\
&= \frac{1}{1-[\alpha_{qk}^{\text{in}}]_1}\left[g_{x1}\left([\mathbf{q}_k^{\text{in}}]_1 + [\mathbf{w}_q^{\text{in}}]_1, [\gamma_{qk}^{\text{in}}]_1\right) - [\mathbf{w}_q^{\text{in}}]_1 - [\alpha_{qk}^{\text{in}}]_1[\mathbf{q}_k^{\text{in}}]_1\right] \\
&\overset{(c)}{=} \frac{1}{1-[\alpha_{qk}^{\text{in}}]_1}\left[\mathbf{f}_q^{\text{in}}\left(\mathbf{q}_k^{\text{in}}, \mathbf{q}_k^{\text{out}}, \mathbf{w}_q^{\text{in}}, \gamma_{qk}^{\text{in}}, \gamma_{qk}^{\text{out}}\right) - [\alpha_{qk}^{\text{in}}]_1\mathbf{q}_k^{\text{in}}\right]_1,
\end{aligned}$$

where (a) follows from the definition of GVAMP in Algorithm 2, and (b) and (c) follow from the definitions (2.31) and (2.32). The proof for $\mathbf{u}_{k+1}^{\text{out}}$ follows in the same way. Next observe that

$$\begin{aligned}
&[\mathbf{v}_k^{\text{in}}]_1 = \mathbf{V}^T(\mathbf{r}_{1k+1} - \mathbf{x}_0) \\
&\overset{(a)}{=} \mathbf{V}^T\left[\frac{1}{1-\alpha_{2k}}\left(\mathbf{V}\mathbf{D}_k\left[\tau_{2k}\mathbf{S}^T\mathbf{U}^T\mathbf{p}_{2k} + \gamma_{2k}\mathbf{V}^T\mathbf{r}_{2k}\right] - \alpha_{2k}\mathbf{r}_{2k}\right)\right] - \mathbf{V}^T\mathbf{x}_0 \\
&\overset{(b)}{=} \frac{1}{1-[\alpha_{pk}^{\text{in}}]_1}\Big(\mathbf{D}_k\left[[\gamma_{pk}^{\text{out}}]_1\mathbf{S}^T([\mathbf{p}_k^{\text{out}}]_1 + \mathbf{U}^T\mathbf{z}_0) + [\gamma_{pk}^{\text{in}}]_1([\mathbf{p}_k^{\text{in}}]_1 + \mathbf{V}^T\mathbf{x}_0)\right] - [\alpha_{pk}^{\text{in}}]_1([\mathbf{p}_k^{\text{in}}]_1 - \mathbf{V}^T\mathbf{x}_0)\Big) - \mathbf{V}^T\mathbf{x}_0 \\
&\overset{(c)}{=} \frac{1}{1-[\alpha_{pk}^{\text{in}}]_1}\Big(\mathbf{D}_k\left[[\gamma_{pk}^{\text{out}}]_1\mathbf{S}^T[\mathbf{p}_k^{\text{out}}]_1 + [\gamma_{pk}^{\text{in}}]_1[\mathbf{p}_k^{\text{in}}]_1 + ([\gamma_{pk}^{\text{out}}]_1\mathbf{S}^T\mathbf{S} + [\gamma_{pk}^{\text{in}}]_1)\mathbf{V}^T[\mathbf{w}_q^{\text{in}}]_1\right] - [\alpha_{pk}^{\text{in}}]_1[\mathbf{p}_k^{\text{in}}]_1 - \mathbf{V}^T[\mathbf{w}_q^{\text{in}}]_1\Big) \\
&\overset{(d)}{=} \frac{1}{1-[\alpha_{pk}^{\text{in}}]_1}\left(\mathbf{D}_k\left[[\gamma_{pk}^{\text{out}}]_1\mathbf{S}^T[\mathbf{p}_k^{\text{out}}]_1 + [\gamma_{pk}^{\text{in}}]_1[\mathbf{p}_k^{\text{in}}]_1\right] - [\alpha_{pk}^{\text{in}}]_1[\mathbf{p}_k^{\text{in}}]_1\right) \\
&\overset{(e)}{=} \frac{1}{1-[\alpha_{pk}^{\text{in}}]_1}\left[\mathbf{f}_p^{\text{in}}\left(\mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}, \mathbf{w}_p^{\text{in}}, \gamma_{pk}^{\text{in}}, \gamma_{pk}^{\text{out}}\right) - \alpha_{pk}^{\text{in}}\mathbf{p}_k^{\text{in}}\right]_1,
\end{aligned}$$

where (a) follows from the definition of GVAMP in Algorithm 2 and the definition (15) in [84], (b) follows from the definitions in (2.31), (c) follows from noting that $\mathbf{U}^T\mathbf{z}_0 = \mathbf{S}\mathbf{V}^T\mathbf{x}_0$, (d) follows from the fact that $\mathbf{D}_k^{-1} = ([\gamma_{pk}^{\text{out}}]_1\mathbf{S}^T\mathbf{S} + [\gamma_{pk}^{\text{in}}]_1)$, and (e) follows from definition (2.32). The proof for $[\mathbf{v}_k^{\text{out}}]_1$ follows in the same way using the fact that $g_{z2}\left(\mathbf{r}_{2k}, \mathbf{p}_{2k}, \tau_{2k}, \gamma_{2k}\right) = \mathbf{U}\mathbf{S}\mathbf{D}_k\left[\tau_{2k}\mathbf{S}^T\mathbf{U}^T\mathbf{p}_{2k} + \gamma_{2k}\mathbf{V}^T\mathbf{r}_{2k}\right]$.

Relative to the large system behavior of the iterates, the primary innovation of (G)VAMP can be understood in terms of the fact that its iterates can be generated by this more general recursion. Utilizing the left and right singular vectors separately allows for the effect of $\mathbf{A}$ to be broken up within the general algorithm, isolating the effects of the (rectangular diagonal) singular values matrix $\mathbf{S} \in \mathbb{R}^{M\times N}$ and of the orthogonal matrices $\mathbf{V} \in \mathbb{R}^{N\times N}$ and $\mathbf{U} \in \mathbb{R}^{M\times M}$ in various stages.

Notice that in Theorem 1, we want to understand the concentration properties of the estimate $\hat{\mathbf{x}}_{1k}$. But this is just a function of $\mathbf{r}_{1k}$ and, thus, by the above identities (2.31) of $\mathbf{p}_k^{\text{in}}$ and $\mathbf{w}_p^{\text{in}} = \mathbf{x}^0$. Therefore, once we have concentration for the iterates of the general recursion, we will be able to easily obtain concentration for $\hat{\mathbf{x}}_{1k}$ in the proof of Theorem 1.

As this translation shows, the effect of the singular values can be entirely subsumed within the denoiser f_q through the disturbance vectors $\mathbf{w}_q^{\text{in/out}}$. Thus, we do not require any distributional assumptions about the singular values beyond empirical convergence to some bounded random variable (**Assumption 0**). In the case of GVAMP, we require only distributional assumptions about $\mathbf{U}$ and $\mathbf{V}$. Moreover, the matrix $\mathbf{U}$ does not occur in Algorithm 4, and it follows that the large system behavior of VAMP can be characterized using only distributional assumptions on $\mathbf{V}$.

2.3.1 Conditions

To prove our general concentration result, we require a number of conditions on the quantities in Algorithm 3. Using the translations above, we show that these conditions are implied by **Assumptions 0**-**5** that we made about the GVAMP quantities in Section 2.2.

Condition 0. The initial sequences $\mathbf{u}_0^{\text{in/out}}$ are independent of each other and the matrices $\mathbf{U}$, $\mathbf{V}$ and concentrate on random variables U_0^{in} and U_0^{out}, respectively. Furthermore, the initial γ_{p0}^{in} and γ_{p0}^{out} converge to $\overline{\gamma}_{p0}^{\text{in}} \geq 0$ and $\overline{\gamma}_{p0}^{\text{out}} \geq 0$, respectively, as $N \to \infty$.

Condition 1. Assume that the components of each of the columns $[\mathbf{w}_p^{\text{in/out}}]_i$ and $[\mathbf{w}_q^{\text{in/out}}]_i$ are either (1) sampled i.i.d. from a sub-Gaussian distribution or (2) are non-random sequences which are bounded independently of N and which converge in $PL(2)$ sense to a limiting random variable with finite second moment. In either case, these columns will converge in $PL(2)$ to limiting variables $W_p^{\text{in/out}}$ and $W_q^{\text{in/out}}$. (This is implied by Lemma 17 if the components are sub-Gaussian.) We further assume that these are independent of $\mathbf{V}$ and $\mathbf{U}$ (thus, independent of the limiting $P_k^{\text{in/out}}$ and $Q_k^{\text{in/out}}$) and independent of each other.

Condition 2. Same as **Assumption 2**.

Condition 3. The denoisers $f_p^{\text{in/out}}$ and $f_q^{\text{in/out}}$ are separable, and both these functions and their derivatives are either

1. uniformly Lipschitz at each $\left(\overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}\right)$ and $\left(\overline{\gamma}_{qk}^{\text{in}}, \overline{\gamma}_{qk}^{\text{out}}\right)$, respectively, for $k \geq 1$, or

2. uniformly bounded conditionally Lipschitz (UBCL) at the same parameters for all $k \geq 1$. This condition requires the following (in the notation for the p case, the q case being symmetric):

   (a) $f_p^{\text{in/out}}$ is continuous in $(p^{\text{in}}, p^{\text{out}}, w_p)$ for all parameters $(\gamma^{\text{in}}, \gamma^{\text{out}})$.

   (b) In a neighborhood of $\left(\overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}\right)$, $f_p^{\text{in/out}}$ is Lipschitz in $(p^{\text{in}}, p^{\text{out}})$ for all w_p, with Lipschitz constant continuous in w_p.

   (c) The domain of w_p is compact.

   (d) The Lipschitz uniformity over the parameters is satisfied for all inputs, i.e.

   $$\left| f_p^{\text{in/out}}\left(p^{\text{in}}, p^{\text{out}}, w_p, \gamma_1\right) - f_p^{\text{in/out}}\left(p^{\text{in}}, p^{\text{out}}, w_p, \gamma_2\right) \right| \leq L\left(1 + \| \left(p^{\text{in}}, p^{\text{out}}, w_p\right) \|\right) |\gamma_1 - \gamma_2|$$

   for all γ_1, γ_2 in a neighborhood of $\overline{\gamma}_{pk}$.

Condition 4. The functions $C_p^{\text{in/out}}, C_q^{\text{in/out}}, \Gamma_p^{\text{in/out}}, \Gamma_q^{\text{in/out}}$ are bounded over their domains and Lipschitz continuous with Lipschitz constant independent of N and k.

Condition 5. Following [81], we define stopping criteria that determine when the algorithm has effectively converged. First, we stop the iteration if $\gamma_{pk}^{\text{out}} < \epsilon_1$ or $\gamma_{qk}^{\text{out}} < \epsilon_2$. In terms Algorithm 1, this is equivalent to stopping the algorithm when the variance of $|R_{1k} - X_0|$ or $|R_{2k} - X_0|$ is sufficiently small. Next, we stop if $\rho_{pk}^{\text{in/out}} < \epsilon_1'$ or $\rho_{qk}^{\text{in/out}} < \epsilon_2'$, where these ρ quantities are defined formally in the next section.

These ρ quantities are equivalent to the limits defined in terms of GVAMP quantities in **Assumption 5**. In terms of the general recursion quantities, we can given these another interpretation. Our analysis will imply that the $\mathbf{p}_k^{\text{in/out}}$ and $\mathbf{q}_k^{\text{in/out}}$ iterates converge empirically to some random variables $P_k^{\text{in/out}}$ and $Q_k^{\text{in/out}}$ respectively. If we regress each $P_k^{\text{in/out}}$ on the previous $P_j^{\text{in/out}}$ for $0 \leq j \leq k-1$, then $\rho_{pk}^{\text{in/out}}$ is just the variance of the (Gaussian) noise in this regression, and similarly for $\rho_{qk}^{\text{in/out}}$. Thus, this stopping criterion amounts to stopping when the limiting amount of excess information gained by the additional iteration $P_k^{\text{in/out}}$ and $Q_k^{\text{in/out}}$ is sufficiently small.

Condition 6. The interval $[t_{\min}, t_{\max}]$ within which the α_1 scalars are truncated includes the $[\overline{\alpha}_{pk}^{\text{in/out}}]_1$ and $[\overline{\alpha}_{qk}^{\text{in/out}}]_1$ (defined in the general recursion state evolution (2.67)) for all iterations k prior to termination.

We now give some remarks on sufficient conditions and special cases of these assumptions.

Remark 7. *In our translation* (2.31)*, we set* $\mathbf{u}_0^{\text{in}} = \mathbf{r}_{20} - \mathbf{x}_0$*. This can be further expanded as*

$$\frac{g_{x1}(\mathbf{r}_{10}, \gamma_{10}) - \alpha_{1k}\mathbf{r}_{10}}{1 - \alpha_{1k}} - \mathbf{x}_0, \tag{2.36}$$

where

$$\alpha_{1k} = \frac{1}{N}\sum_{i=1}^{N} g_{x1}'\left([\mathbf{r}_{10}]_i, \gamma_{10}\right)$$

Our ***Assumption 3*** *ensures that* $g_{x1}'(\cdot, \gamma_{10})$ *is Lipschitz, and thus Lemma 17 gives concentration for* α_{1k}*. Then Lemmas 13, 9, 10, and 17 (using subgaussianity of* $\mathbf{r}_{10}$*) along with* ***Assumption 0*** *give concentration for* (2.36)*. Thus, the elements of* $\mathbf{u}_0^{\text{in}}$ *concentrate as required. Since* g_{x1} *does not depend on the matrices* $\mathbf{U}$ *and* $\mathbf{V}$*, this* $\mathbf{u}_0^{\text{in}}$ *is also independent of* $\mathbf{U}$ *and* $\mathbf{V}$ *as well. An analogous argument shows that our translation for* $\mathbf{u}_0^{\text{out}}$ *and* ***Assumption 0*** *imply* ***Condition 0***

for $\mathbf{u}_0^{\text{out}}$. *Likewise, from* (2.31)*, we can see that the convergence conditions on the* $\gamma_{p0}^{\text{in/out}}$ *are guaranteed by* ***Assumption 0*** *on the corresponding constants in the GVAMP algorithm.*

Remark 8. *Under the translation in* (2.33)*,* ***Condition 1*** *is implied by* ***Assumption 1****.*

Remark 9. *It is easy to see that the* $f_q^{\text{in/out}}$ *in translation* (2.32) *meet the requirements of* ***Condition 3****. They are all clearly separable by definition, and the only functions that are not uniformly Lipschitz are* $[f_q^{\text{in/out}}]_2 = [w_q]^1 q^{\text{in/out}}$*. However, these are UBCL as they are free of* $\gamma_{qk}^{\text{in/out}}$ *and because* $w_q^{\text{in/out}}$ *consists of the singular values, which by* ***Assumption 0*** *lie in a compact interval* $[0, S_{\max}]$*. The* $f_p^{\text{in/out}}$ *are separable and uniformly Lipschitz when the* g_{x1}/g_{z1} *are, which is guaranteed by* ***Assumption 3****.*

Remark 10. *For the functions given in the translation* (2.32)*,* ***Condition 4*** *is satisfied as long as the domains of the functions are compact (and independent of N and k). We note that clipping of the* γ_1 *and* α_1 *scalars is guaranteed by* (2.34) *and* (2.35) *along with* ***Assumption 4****.*

Remark 11. ***Condition 6*** *can be met, for instance, in a log-concave setting satisfying the following assumptions:*

1. g_{x1} *and* g_{z1} *are either the MAP or MMSE denoisers.*
2. ***Assumption 2*** *holds, which stated that the log prior* $\log p(x)$ *and the log likelihood* $\log p(y \mid z)$ *(as a function of z) are concave and* β*-smooth, and that the second derivatives in* (A.125) *are bounded away from zero over any compact subset of their domains.*
3. *The asymptotic variances* $\tau_{pk}^{\text{in/out}}$ *and* $\tau_{qk}^{\text{in/out}}$ *(also defined in* (2.67)*) do not diverge or shrink to* 0 *before termination.*

Under these conditions, the derivatives of the $f_p^{\text{in/out}}$ *and* $f_q^{\text{in/out}}$ *will be bounded away from* 1 *and* 0 *over any compact subset of their domains (see e.g. Lemma 1 in [78]), and the* $[\overline{\alpha}_{pk}^{\text{in/out}}]_1$ *and* $[\overline{\alpha}_{qk}^{\text{in/out}}]_1$ *will likewise we bounded away from* 1 *and* 0 *for all k.*

2.3.2 Notation

In the above translation between the general recursion and GVAMP, we used general iterates of dimension $d = 2$ in order to track the GVAMP iterates, the initial data, and the transformed input $\mathbf{z}_0$ throughout the algorithm. Henceforth, we will analyze the general recursion with $d = 1$ in order to simplify notation and proof ideas. However, these arguments can all be extended to the case of general dimension d (see Appendix A.4 for details on how the following results generalize to $d > 1$). For $k \geq 0$, define matrices $\mathbf{U}_k^{\text{in/out}} \in \mathbb{R}^{N\times(k+1)}$ having columns $\mathbf{u}_i^{\text{in/out}}$ for $0 \leq i \leq k$, and define $N \times (k+1)$ matrices $\mathbf{V}_k^{\text{in/out}}$, $\mathbf{P}_k^{\text{in/out}}$, and $\mathbf{Q}_k^{\text{in/out}}$ analogously.

Now, for $k \geq 1$, define $\mathbf{C}_{pk}^{\text{in/out}}, \mathbf{C}_{uk}^{\text{in/out}} \in \mathbb{R}^{N\times(2k+1)}$ to be the matrices containing, respectively, the $\mathbf{p}^{\text{in/out}}, \mathbf{v}^{\text{in/out}}$ iterates of the algorithm up to and including $\mathbf{p}_k^{\text{in/out}}$ and the $\mathbf{u}^{\text{in/out}}, \mathbf{q}^{\text{in/out}}$ iterates up to and including $\mathbf{u}_k^{\text{in/out}}$. Similarly, let $\mathbf{C}_{vk}^{\text{in/out}}, \mathbf{C}_{qk}^{\text{in/out}} \in \mathbb{R}^{N\times 2k}$ be the matrices containing, respectively, the $\mathbf{p}^{\text{in/out}}, \mathbf{v}^{\text{in/out}}$ iterates up to and including $\mathbf{v}_{k-1}^{\text{in/out}}$ and the $\mathbf{u}^{\text{in/out}}, \mathbf{q}^{\text{in/out}}$ iterates up to and including $\mathbf{q}_{k-1}^{\text{in/out}}$. We use the initial conditions $\mathbf{C}_{p0}^{\text{in/out}} = \mathbf{P}_0^{\text{in/out}} = \mathbf{p}_0^{\text{in/out}}$, $\mathbf{C}_{u0}^{\text{in/out}} = \mathbf{U}_0^{\text{in/out}} = \mathbf{u}_0^{\text{in/out}}$, and $\mathbf{C}_{v0}^{\text{in/out}} = \mathbf{C}_{q0}^{\text{in/out}} = \emptyset$. In other words, these matrices are defined as follows:

$$\begin{aligned}
\mathbf{C}_{pk}^{\text{in/out}} &= [\mathbf{P}_k^{\text{in/out}}\ \mathbf{V}_{k-1}^{\text{in/out}}], \quad \mathbf{C}_{uk}^{\text{in/out}} = [\mathbf{U}_k^{\text{in/out}}\ \mathbf{Q}_{k-1}^{\text{in/out}}], \\
\mathbf{C}_{vk}^{\text{in/out}} &= [\mathbf{P}_{k-1}^{\text{in/out}}\ \mathbf{V}_{k-1}^{\text{in/out}}], \quad \mathbf{C}_{qk}^{\text{in/out}} = [\mathbf{U}_{k-1}^{\text{in/out}}\ \mathbf{Q}_{k-1}^{\text{in/out}}],
\end{aligned} \tag{2.37}$$

where for two matrices $\mathbf{M}_1 \in \mathbb{R}^{N\times p_1}$ and $\mathbf{M}_2 \in \mathbb{R}^{N\times p_2}$, by the notation $\mathbf{M} = [\mathbf{M}_1\ \mathbf{M}_2]$, we mean the $N \times (p_1 + p_2)$ matrix obtained from concatenating the columns of $\mathbf{M}_1$ and $\mathbf{M}_2$.

Next, we define sigma-algebras generated by previous algorithm output at any iteration using the matrices defined in the previous paragraph. We use the initializations $\mathcal{P}_0^{\text{in/out}} = \sigma\{\mathbf{w}_p^{\text{in/out}}, \mathbf{U}_0^{\text{in/out}}\}$ and $\mathcal{Q}_0^{\text{in/out}} = \sigma\{\mathbf{w}_q^{\text{in/out}}, \mathbf{U}_0^{\text{in/out}}, \mathbf{P}_0^{\text{in/out}}, \mathbf{V}_0^{\text{in/out}}\}$, and for $k \geq 1$ define

$$\begin{aligned}
\mathcal{P}_k^{\text{in/out}} &= \sigma\{\mathbf{w}_p^{\text{in/out}}, \mathbf{U}_k^{\text{in/out}}, \mathbf{P}_{k-1}^{\text{in/out}}, \mathbf{V}_{k-1}^{\text{in/out}}, \mathbf{Q}_{k-1}^{\text{in/out}}\}, \\
\mathcal{Q}_k^{\text{in/out}} &= \sigma\{\mathbf{w}_q^{\text{in/out}}, \mathbf{U}_k^{\text{in/out}}, \mathbf{P}_k^{\text{in/out}}, \mathbf{V}_k^{\text{in/out}}, \mathbf{Q}_{k-1}^{\text{in/out}}\}.
\end{aligned} \tag{2.38}$$

Notice that $\mathcal{P}_k^{\text{in/out}}$ includes the iterates of the algorithm up to and including $\mathbf{u}_k^{\text{in/out}}$ (i.e. all output after completing the $(k-1)^{th}$ iteration of the algorithm) and $\mathcal{Q}_k^{\text{in/out}}$ includes the iterates up to and including $\mathbf{q}_k^{\text{in/out}}$ (i.e. all output through step (5) in the k^{th} iteration). We define the combined sigma-algebras $\mathcal{P}_k = \sigma\left\{\mathcal{P}_k^{\text{in}}, \mathcal{P}_k^{\text{out}}\right\}$ and $\mathcal{Q}_k = \sigma\left\{\mathcal{Q}_k^{\text{in}}, \mathcal{Q}_k^{\text{out}}\right\}$.

Finally, for $k \geq 0$, define the matrices $\mathbf{B}_{\mathbf{C}_{qk}^{\text{in/out}}}^{\perp}, \mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}}^{\perp} \in \mathbb{R}^{N\times(N-2k)}$, and $\mathbf{B}_{\mathbf{C}_{uk}^{\text{in/out}}}^{\perp}, \mathbf{B}_{\mathbf{C}_{pk}^{\text{in/out}}}^{\perp} \in \mathbb{R}^{N\times(N-2k-1)}$ where, for example, $\mathbf{B}_{\mathbf{C}_{pk}^{\text{out}}}^{\perp}$ has columns that form an orthonormal basis for $\text{span}(\mathbf{C}_{pk}^{\text{out}})^{\perp}$ where $\mathbf{C}_{pk}^{\text{out}}$ is defined in (2.37), and the other matrices are defined analogously. Moreover, for $k \geq 1$, define the matrix $\mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}} \in \mathbb{R}^{N\times 2k}$ to have as columns an orthonormal basis for $\text{span}(\mathbf{C}_{vk}^{\text{in/out}})$, and for $k \geq 0$, define $\mathbf{B}_{\mathbf{C}_{uk}^{\text{in/out}}} \in \mathbb{R}^{N\times(2k+1)}$ as a matrix having as columns an orthonormal basis for $\text{span}(\mathbf{C}_{uk}^{\text{in/out}})$. Then we let $\mathbf{O}_{p0}^{\text{in/out}} = \mathbf{B}_{\mathbf{C}_{v0}^{\text{in/out}}}^{\perp} \in \mathbb{R}^{n\times n}$ be any (deterministic) orthogonal matrix, and for $k \geq 1$, we define the orthogonal matrices

$$\mathbf{O}_{pk}^{\text{in/out}} = [\mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}}^{\perp} \mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}}] \in \mathbb{R}^{N\times N} \quad \text{and} \quad \mathbf{O}_{q(k-1)}^{\text{in/out}} = [\mathbf{B}_{\mathbf{C}_{u(k-1)}^{\text{in/out}}}^{\perp} \mathbf{B}_{\mathbf{C}_{u(k-1)}^{\text{in/out}}}] \in \mathbb{R}^{N\times N}. \tag{2.39}$$

To wrap up the notation, we state a definition and lemma that will be useful throughout the proof.

Definition 6 (Isotropic Invariance)**.** *A vector* $\mathbf{x} \in \mathbb{R}^N$ *is* ***isotropically invariant*** *(or, spherically symmetric) if* $\mathbf{V}_0\mathbf{x} \overset{d}{=} \mathbf{x}$ *for any orthogonal matrix* $\mathbf{V}_0 \in \mathbb{R}^{N\times N}$*. Notice this implies that* $\mathbf{x}/\|\mathbf{x}\|$ *is uniformly distributed on the unit sphere.*

The following lemma about isotropically invariant vectors will be used throughout the proof.

Lemma 1. $\mathbf{x} \overset{d}{=} \|\mathbf{x}\|(\mathbf{w}_0/\|\mathbf{w}_0\|)$ *where* $\mathbf{w}_0 \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_{N\times N})$ *for any isotropically invariant* $\mathbf{x} \in \mathbb{R}^N$.

Proof. The result follows from two facts. First, $\mathbf{x}/\|\mathbf{x}\| \overset{d}{=} \mathbf{w}_0/\|\mathbf{w}_0\|$ since both $\mathbf{x}/\|\mathbf{x}\|$ and $\mathbf{w}_0/\|\mathbf{w}_0\|$ are uniformly distributed on the unit sphere. Second, $\mathbf{x}/\|\mathbf{x}\|$ and $\|\mathbf{x}\|$ are independent. To see this, notice that $\mathbf{x}/\|\mathbf{x}\|$ is uniformly-distributed on the sphere irrespective of the value of $\|\mathbf{x}\|$; thus, they are independent. Therefore, $\mathbf{x} = \|\mathbf{x}\|(\mathbf{x}/\|\mathbf{x}\|) \overset{d}{=} \|\mathbf{x}\|(\mathbf{w}_0/\|\mathbf{w}_0\|)$, giving the desired result. □

2.3.3 General algorithm concentration

Under the conditions given in Section 2.3.1, we state our general concentration result for Algorithm 3 in Lemmas 3 - 5. To obtain our concentration results, we first introduce deviance terms $\mathbf{\Delta}_{pk}^{\text{in/out}}$ and $\mathbf{\Delta}_{qk}^{\text{in/out}}$ quantifying the discrepancy between the finite sample behavior of the iterates $(\mathbf{p}_1^{\text{in/out}}, ..., \mathbf{p}_k^{\text{in/out}})$ and $(\mathbf{q}_1^{\text{in/out}}, ..., \mathbf{q}_k^{\text{in/out}})$ and their limiting values. This is done in Lemma 3 by studying the distributions of the vectors conditional on the previous output of the algorithm, summarized by the sigma-algebras in (2.38). Before we state and prove Lemma 3, we state a lemma, originally from [77, Lemmas 4 and 5], that characterizes the distribution of random Haar matrices, conditional on linear constraints.

Lemma 2. *[77, Lemmas 4 and 5] Let* $\mathbf{V} \in \mathbb{R}^{N\times N}$ *be a random, Haar-distributed matrix (see Definition 4). Suppose that for deterministic matrices* $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{N\times s}$*, for some* $1 \leq s \leq N$*, we know that* $\mathbf{V}$ *satisfies* $\boldsymbol{M}_1 = \mathbf{V}\boldsymbol{M}_2$*. Then, if* $\boldsymbol{M}_1$ *and* $\boldsymbol{M}_2$ *are full column rank,*

$$\mathbf{V}\Big|_{\{\boldsymbol{M}_1=\mathbf{V}\boldsymbol{M}_2\}} \stackrel{d}{=} \boldsymbol{M}_1(\boldsymbol{M}_1^T\boldsymbol{M}_1)^{-1}\boldsymbol{M}_2^T + \mathbf{B}_{\boldsymbol{M}_1}^{\perp}\widetilde{\mathbf{V}}[\mathbf{B}_{\boldsymbol{M}_2}^{\perp}]^T,$$

where $\widetilde{\mathbf{V}}$ *is Haar distributed, independent of* $\mathbf{V}$*, and* $\mathbf{B}_{\boldsymbol{M}_1}^{\perp}, \mathbf{B}_{\boldsymbol{M}_2}^{\perp} \in \mathbb{R}^{N\times(N-s)}$ *are any matrices whose columns are orthogonal bases for* $Range(\mathbf{M}_1)^{\perp}$ *and* $Range(\mathbf{M}_2)^{\perp}$*, respectively.*

To use the above lemma, we notice that conditioning on sigma-algebras $\mathcal{Q}_k^{\text{in/out}}$ and $\mathcal{P}_k^{\text{in/out}}$, defined in (2.38), is equivalent to conditioning on two linear constraints. As an illustrative example, consider conditioning on the sigma-algebra $\mathcal{Q}_k^{\text{in}} = \sigma\{\mathbf{U}_k^{\text{in}}, \mathbf{P}_k^{\text{in}}, \mathbf{V}_k^{\text{in}}, \mathbf{Q}_{k-1}^{\text{in}}\}$, which contains all of the input iterates up until just before step (8). First, notice that the action of $\mathbf{V}$ in Algorithm 3 is only in step (3), which reads $\mathbf{p}_k^{\text{in}} \leftarrow \mathbf{V}^T\mathbf{u}_k^{\text{in}}$, and step (8), which reads $\mathbf{q}_k^{\text{in}} \leftarrow \mathbf{V}\mathbf{v}_k^{\text{in}}$. Thus, conditioning on $\mathcal{Q}_k^{\text{in}}$ is equivalent to conditioning on the linear constraint $[\mathbf{P}_k^{\text{in}}\ \mathbf{V}_{k-1}^{\text{in}}] = \mathbf{V}^T[\mathbf{U}_k^{\text{in}}\ \mathbf{Q}_{k-1}^{\text{in}}]$ or, equivalently, on $\mathbf{C}_{pk}^{\text{in}} = \mathbf{V}^T\mathbf{C}_{uk}^{\text{in}}$ for $\mathbf{C}_{pk}^{\text{in}}$ and $\mathbf{C}_{uk}^{\text{in}}$ defined in (2.37). By a similar argument, we can see that conditioning on $\mathcal{P}_{k+1}^{\text{in}} = \sigma\{\mathbf{U}_{k+1}^{\text{in}}, \mathbf{P}_k^{\text{in}}, \mathbf{V}_k^{\text{in}}, \mathbf{Q}_k^{\text{in}}\}$ is equivalent to conditioning on $[\mathbf{P}_k^{\text{in}}\ \mathbf{V}_k^{\text{in}}] = \mathbf{V}^T[\mathbf{U}_k^{\text{in}}\ \mathbf{Q}_k^{\text{in}}]$ or on $\mathbf{C}_{v(k+1)}^{\text{in}} = \mathbf{V}^T\mathbf{C}_{q(k+1)}^{\text{in}}$. Continuing this argument, conditioning on $\mathcal{Q}_k^{\text{out}}$ or $\mathcal{P}_{k+1}^{\text{out}}$ is equivalent to conditioning on $\mathbf{C}_{pk}^{\text{out}} = \mathbf{U}^T\mathbf{C}_{uk}^{\text{out}}$ or $\mathbf{C}_{v(k+1)}^{\text{out}} = \mathbf{U}^T\mathbf{C}_{q(k+1)}^{\text{out}}$, respectively.

Now we give the conditional distribution lemma, Lemma 3. In the lemma and in what follows, we use the notation $\mathbf{0}_k$ to denotes a vector of zeros of length k. This lemma and the following, Lemma 4, depend on vectors $\boldsymbol{\beta}_{pk}^{\text{in/out}}, \boldsymbol{\beta}_{qk}^{\text{in/out}} \in \mathbb{R}^k$ and constants $\rho_{pk}^{\text{in/out}}, \rho_{qk}^{\text{in/out}} \in \mathbb{R}_{>0}$, which we leave unspecified for the time being. The statements of these lemmas are valid for any choice of these parameters, and in section 2.3.5 we define them as appropriate limits of GVAMP quantities so as to ensure the desired concentration behavior.

Lemma 3. *If $[\mathbf{C}_{vk}^{\text{in}}]^T \mathbf{C}_{vk}^{\text{in}}$ has full rank for $0 \le k \le K$, then we have for all such k that*

$$\mathbf{p}_0^{\text{in}}|_{\mathcal{P}_0} \stackrel{d}{=} \sqrt{\rho_{p0}^{\text{in}}}\, \mathbf{O}_{p0}^{\text{in}}\, \overline{\mathbf{Z}}_{p0}^{\text{in}} + \boldsymbol{\Delta}_{p0}^{\text{in}}, \qquad \text{and} \qquad \mathbf{p}_k^{\text{in}}|_{\mathcal{P}_k} \stackrel{d}{=} \sum_{\ell=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{in}}]_{\ell+1}\, \mathbf{p}_\ell^{\text{in}} + \sqrt{\rho_{pk}^{\text{in}}}\, \mathbf{O}_{pk}^{\text{in}}\, \overline{\mathbf{Z}}_{pk}^{\text{in}} + \boldsymbol{\Delta}_{pk}^{\text{in}}, \quad (2.40)$$

where $\mathbf{O}_{pk}^{\text{in}}$ is defined in (2.39) *while ρ_{pk}^{in} and $\boldsymbol{\beta}_{pk}^{\text{in}}$ are unspecified for the time being, with appropriate values given in what follows in* (2.72)*, and*

$$\boldsymbol{\Delta}_{p0}^{\text{in}} = \left[\frac{\|\mathbf{u}_0^{\text{in}}\|}{\|\mathbf{Z}_{p0}^{\text{in}}\|} - \sqrt{\rho_{p0}^{\text{in}}} \right] \mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp} \mathbf{Z}_{p0}^{\text{in}}, \qquad \boldsymbol{\Delta}_{pk}^{\text{in}} = \mathbf{C}_{vk}^{\text{in}} \left(([\mathbf{C}_{vk}^{\text{in}}]^T \mathbf{C}_{vk}^{\text{in}})^{-1} [\mathbf{C}_{qk}^{\text{in}}]^T \mathbf{u}_k^{\text{in}} - \begin{bmatrix} \boldsymbol{\beta}_{pk}^{\text{in}} \\ \mathbf{0}_k \end{bmatrix} \right)$$
$$+ \left[\frac{\|[\mathbf{B}_{\mathbf{C}_{qk}^{\text{in}}}^{\perp}]^T \mathbf{u}_k^{\text{in}}\|}{\|\mathbf{Z}_{pk}^{\text{in}}\|} - \sqrt{\rho_{pk}^{\text{in}}} \right] \mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp} \mathbf{Z}_{pk}^{\text{in}} - \sqrt{\rho_{pk}^{\text{in}}}\, \mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}} \check{\mathbf{Z}}_{pk}^{\text{in}}.$$

The matrices $\mathbf{C}_{vk}^{\text{in}}$ and $\mathbf{C}_{qk}^{\text{in}}$ are defined in (2.37)*. In the above, $\overline{\mathbf{Z}}_{pk}^{\text{in}}$ are length-N vectors with independent, standard Gaussian entries that are independent across $0 \le k \le K$ and independent of the corresponding conditioning sigma-algebra $\mathcal{P}_k$. These are decomposed as $\overline{\mathbf{Z}}_{pk}^{\text{in}} = [\mathbf{Z}_{pk}^{\text{in}} \mid \check{\mathbf{Z}}_{pk}^{\text{in}}]$ where $\mathbf{Z}_{pk}^{\text{in}}$ is length $N - 2k$ and $\check{\mathbf{Z}}_{pk}^{\text{in}}$ is length $2k$. In particular, this means $\mathbf{O}_{pk}^{\text{in}} \overline{\mathbf{Z}}_{pk}^{\text{in}} = \mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp} \mathbf{Z}_{pk}^{\text{in}} + \mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}} \check{\mathbf{Z}}_{pk}^{\text{in}}$.*

If $[\mathbf{C}_{uk}^{\text{in}}]^T \mathbf{C}_{uk}^{\text{in}}$ has full rank for $0 \le k \le K$, we have for all such k that

$$\mathbf{q}_0^{\text{in}}|_{\mathcal{Q}_0} \stackrel{d}{=} \sqrt{\rho_{q0}^{\text{in}}}\, \mathbf{O}_{q0}^{\text{in}}\, \overline{\mathbf{Z}}_{q0}^{\text{in}} + \boldsymbol{\Delta}_{q0}^{\text{in}}, \qquad \text{and} \qquad \mathbf{q}_k^{\text{in}}|_{\mathcal{Q}_k} \stackrel{d}{=} \sum_{\ell=0}^{k-1} [\boldsymbol{\beta}_{qk}^{\text{in}}]_{\ell+1}\, \mathbf{q}_\ell^{\text{in}} + \sqrt{\rho_{qk}^{\text{in}}}\, \mathbf{O}_{qk}^{\text{in}}\, \overline{\mathbf{Z}}_{qk}^{\text{in}} + \boldsymbol{\Delta}_{qk}^{\text{in}}, \quad (2.41)$$

where $\mathbf{O}_{qk}^{\text{in}}$ is defined in (2.39) *while ρ_{qk}^{in} and $\boldsymbol{\beta}_{qk}^{\text{in}}$ are unspecified for the time being, with appro-*

priate values given in what follows in (2.72)*, and*

$$\boldsymbol{\Delta}_{q0}^{\text{in}} = \frac{[\mathbf{p}_0^{\text{in}}]^T \mathbf{v}_0^{\text{in}}}{\|\mathbf{u}_0^{\text{in}}\|^2} \mathbf{u}_0^{\text{in}} + \left[\frac{\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{p0}}]^T \mathbf{v}_0^{\text{in}}\|}{\|\mathbf{Z}_{q0}^{\text{in}}\|} - \sqrt{\rho_{q0}^{\text{in}}} \right] \mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{u0}} \mathbf{Z}_{q0}^{\text{in}} - \sqrt{\rho_{q0}^{\text{in}}}\, \mathbf{B}_{\mathbf{C}^{\text{in}}_{u0}} \check{\mathbf{Z}}_{q0}^{\text{in}}, \tag{2.42}$$

$$\begin{aligned} \boldsymbol{\Delta}_{qk}^{\text{in}} = \mathbf{C}_{uk}^{\text{in}} \left(([\mathbf{C}_{uk}^{\text{in}}]^T \mathbf{C}_{uk}^{\text{in}})^{-1} [\mathbf{C}_{pk}^{\text{in}}]^T \mathbf{v}_k^{\text{in}} - \begin{bmatrix} \mathbf{0}_{k+1} \\ \boldsymbol{\beta}_{qk}^{\text{in}} \end{bmatrix} \right) \\ + \left[\frac{\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{pk}}]^T \mathbf{v}_k^{\text{in}}\|}{\|\mathbf{Z}_{qk}^{\text{in}}\|} - \sqrt{\rho_{qk}^{\text{in}}} \right] \mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{uk}} \mathbf{Z}_{qk}^{\text{in}} - \sqrt{\rho_{qk}^{\text{in}}}\, \mathbf{B}_{\mathbf{C}^{\text{in}}_{uk}} \check{\mathbf{Z}}_{qk}^{\text{in}}. \end{aligned} \tag{2.43}$$

The matrices $\mathbf{C}_{uk}^{\text{in}}$ *and* $\mathbf{C}_{pk}^{\text{in}}$ *are defined in* (2.37)*. In the above,* $\overline{\mathbf{Z}}_{qk}^{\text{in}} \in \mathbb{R}^N$ *are length-N vectors with independent, standard Gaussian entries that are independent across* $0 \leq k \leq K$ *and independent of the corresponding conditioning sigma-algebra* $\mathcal{Q}_k$*. As above, these are decomposed as* $\overline{\mathbf{Z}}_{qk}^{\text{in}} = [\mathbf{Z}_{qk}^{\text{in}} \,|\, \check{\mathbf{Z}}_{qk}^{\text{in}}]$ *where* $\mathbf{Z}_{qk}^{\text{in}}$ *is length* $N - 2k - 1$ *and* $\check{\mathbf{Z}}_{qk}^{\text{in}}$ *is length* $2k + 1$*. Again we have that* $\mathbf{O}_{qk}^{\text{in}} \overline{\mathbf{Z}}_{qk}^{\text{in}} = \mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{uk}} \mathbf{Z}_{qk}^{\text{in}} + \mathbf{B}_{\mathbf{C}^{\text{in}}_{uk}} \check{\mathbf{Z}}_{qk}^{\text{in}}$*. The same statement holds with all input variables replaced by their output equivalents.*

In both cases, the output statements for $\mathbf{p}_k^{\text{out}}|_{\mathcal{P}_k}$ *and* $\mathbf{q}_k^{\text{out}}|_{\mathcal{Q}_k}$ *are analogous to those in* (A.118) *and* (2.41) *with* out *replacing* in *everywhere. Furthermore, we have that* $\mathbf{p}_k^{\text{in}}$ *and* $\mathbf{p}_k^{\text{out}}$ *are conditionally independent given* $\mathcal{P}_k$ *and that* $\mathbf{q}_k^{\text{in}}$ *and* $\mathbf{q}_k^{\text{out}}$ *are conditionally independent given* $\mathcal{Q}_k$ *for all* $k \geq 0$*.*

Proof. We prove the lemma for $\mathbf{q}_0^{\text{in}}$ and $\mathbf{q}_k^{\text{in}}$, as the base case for $\mathbf{q}_0^{\text{in}}$ is slightly more involved than that of $\mathbf{p}_0^{\text{in}}$. The proof for $\mathbf{p}_0^{\text{in}}$ and $\mathbf{p}_k^{\text{in}}$ is broadly similar, and the proof for the output variables exactly mirrors the proof for the input variables.

We first prove the statement for $\mathbf{q}_0^{\text{in}}$. Observe from the Algorithm 3 definition, $\mathbf{q}_0^{\text{in}} = \mathbf{V}^T \mathbf{v}_0^{\text{in}}$ and that $\mathbf{v}_0^{\text{in}} \in \mathcal{Q}_0$ since $\mathcal{Q}_0 = \sigma\left\{\mathcal{Q}_0^{\text{in}}, \mathcal{Q}_0^{\text{out}}\right\} = \sigma\{\mathbf{w}_q^{\text{in/out}}, \mathbf{u}_0^{\text{in/out}}, \mathbf{p}_0^{\text{in/out}}, \mathbf{v}_0^{\text{in/out}}\}$, as defined in (2.38).

By the reasoning in the paragraph just before the lemma statement, conditioning on $\mathcal{Q}_k$ is

equivalent to conditioning on the linear constraints $\mathbf{V}\mathbf{C}^{\text{in}}_{pk} = \mathbf{C}^{\text{in}}_{uk}$ and $\mathbf{U}\mathbf{C}^{\text{out}}_{pk} = \mathbf{C}^{\text{out}}_{uk}$. Thus,

$$\mathbf{V}|_{\mathcal{Q}_k} = \mathbf{V}|_{\left\{\mathbf{V}\mathbf{C}^{\text{in}}_{pk}=\mathbf{C}^{\text{in}}_{uk}, \mathbf{U}\mathbf{C}^{\text{out}}_{pk}=\mathbf{C}^{\text{out}}_{uk}\right\}} \overset{(a)}{=} \mathbf{V}|_{\left\{\mathbf{V}\mathbf{C}^{\text{in}}_{pk}=\mathbf{C}^{\text{in}}_{uk}\right\}} = \mathbf{V}|_{\mathcal{Q}^{\text{in}}_k},$$

where (a) follows as $\mathbf{U}$ and $\mathbf{V}$ are assumed to be independent. The right-hand side can be handled by Lemma 2. In particular, the conditional distribution of $\mathbf{V}$ given $\mathcal{Q}^{\text{in}}_k$ can be decomposed as

$$\mathbf{V}|_{\mathcal{Q}^{\text{in}}_k} \overset{d}{=} \mathbf{C}^{\text{in}}_{uk}([\mathbf{C}^{\text{in}}_{uk}]^T\mathbf{C}^{\text{in}}_{uk})^{-1}[\mathbf{C}^{\text{in}}_{pk}]^T + \mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{uk}}\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{pk}}]^T, \tag{2.44}$$

where $\widetilde{\mathbf{V}}$ is a random $(N-(2k+1))\times(N-(2k+1))$ matrix that is independent of $\mathbf{V}$, but taking the same distribution, in that it is Haar (or uniformly) distributed on the group of orthogonal matrices of its dimensions. In what follows, we drop the explicit "in" superscript to save space because throughout we will be working with the input version of the variables.

Thus, for $k = 0$, conditioning on $\mathcal{Q}_0 = \sigma\{\mathbf{w}_q, \mathbf{u}_0, \mathbf{p}_0, \mathbf{v}_0\}$ is equivalent to conditioning on $\mathbf{C}_{p0} = \mathbf{V}\mathbf{C}_{u0}$ (since $\mathbf{V}$ and $\mathbf{w}_q$ are independent), and the result in (2.44) gives

$$\mathbf{V}\mathbf{v}_0|_{\mathcal{Q}_0} \overset{d}{=} \mathbf{C}_{u0}(\mathbf{C}^T_{u0}\mathbf{C}_{u0})^{-1}\mathbf{C}^T_{p0}\mathbf{v}_0 + \mathbf{B}^{\perp}_{\mathbf{C}_{u0}}\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0, \tag{2.45}$$

where we have removed the "in" superscript from the vector $\mathbf{v}_0$ and the matrices $\mathbf{C}_{u0}$ and $\mathbf{C}_{p0}$. Now observe that $\mathbf{C}_{p0} = \mathbf{p}_0$ and $\mathbf{C}_{u0} = \mathbf{u}_0$, by definition in (2.37), and therefore,

$$\mathbf{C}_{u0}(\mathbf{C}^T_{u0}\mathbf{C}_{u0})^{-1}\mathbf{C}^T_{p0}\mathbf{v}_0 = \mathbf{u}_0\|\mathbf{u}_0\|^{-2}\mathbf{p}^T_0\mathbf{v}_0. \tag{2.46}$$

Next, as $\mathbf{p}_0$ and $\mathbf{v}_0$ are both measurable with respect to $\mathcal{Q}_0$, so is the vector $[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0$. We will next argue that the vector $\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0$ is isotropically invariant as defined in Definition 6. To see this, recall that $\widetilde{\mathbf{V}}$ is Haar-distributed (see Definition 4), thus $\mathbf{V}'\widetilde{\mathbf{V}} \overset{d}{=} \widetilde{\mathbf{V}}$ for any other orthogonal matrix $\mathbf{V}' \in \mathbb{R}^{N-1\times N-1}$. Hence, $\mathbf{V}'\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0 \overset{d}{=} \widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0$, meaning $\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0$ is isotropically invariant; therefore, by Lemma 1, its distribution is entirely determined by the distribution of

its magnitude. In other words, since $\|\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0\| = \|[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0\|$, we have

$$\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0 \overset{d}{=} \|[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0\| \left(\mathbf{Z}_{q0}/\|\mathbf{Z}_{q0}\|\right), \tag{2.47}$$

for $\mathbf{Z}_{q0} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_{N-1\times N-1})$, independent of Q_0. This gives us that, conditional on Q_0,

$$\begin{aligned}\mathbf{B}^{\perp}_{\mathbf{C}_{u0}}\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0 &\overset{d}{=} \frac{\|[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0\|}{\|\mathbf{Z}_{q0}\|}\,\mathbf{B}^{\perp}_{\mathbf{C}_{u0}}\mathbf{Z}_{q0} \\ &= \left[\frac{\|[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0\|}{\|\mathbf{Z}_{q0}\|} - \sqrt{\rho_{q0}}\right]\mathbf{B}^{\perp}_{\mathbf{C}_{u0}}\mathbf{Z}_{q0} - \sqrt{\rho_{q0}}\,\mathbf{B}_{\mathbf{C}_{u0}}\check{\mathbf{Z}}_{q0} + \sqrt{\rho_{q0}}\,\mathbf{O}_{q0}\overline{\mathbf{Z}}_{q0}. \end{aligned}\tag{2.48}$$

The second equality in the above uses that $\mathbf{O}_{q0} = [\mathbf{B}^{\perp}_{\mathbf{C}_{u0}}\ \mathbf{B}_{\mathbf{C}_{u0}}] \in \mathbb{R}^{N\times N}$ by (2.39) for $\mathbf{B}_{\mathbf{C}_{u0}} \in \mathbb{R}^{N\times 1}$ and $\mathbf{B}^{\perp}_{\mathbf{C}_{u0}} \in \mathbb{R}^{N\times(N-1)}$; therefore, $\mathbf{B}^{\perp}_{\mathbf{C}_{u0}}\mathbf{Z}_{q0} = \mathbf{O}_{q0}\overline{\mathbf{Z}}_{q0} - \mathbf{B}_{\mathbf{C}_{u0}}\check{\mathbf{Z}}_{q0}$.

Plugging result (2.46) and (2.48) into (2.45), we find the result in (2.41)-(2.42), namely

$$\mathbf{V}^T\mathbf{v}_0|_{Q_0} \overset{d}{=} \left(\frac{\mathbf{p}_0^T\mathbf{v}_0}{\|\mathbf{u}_0\|^2}\right)\mathbf{u}_0 + \left[\frac{\|[\mathbf{B}^{\perp}_{\mathbf{C}_{p0}}]^T\mathbf{v}_0\|}{\|\mathbf{Z}_{q0}\|} - \sqrt{\rho_{q0}}\right]\mathbf{B}^{\perp}_{\mathbf{C}_{u0}}\mathbf{Z}_{q0} - \sqrt{\rho_{q0}}\,\mathbf{B}_{\mathbf{C}_{u0}}\check{\mathbf{Z}}_{q0} + \sqrt{\rho_{q0}}\,\mathbf{O}_{q0}\overline{\mathbf{Z}}_{q0}.$$

To establish the statement for $\mathbf{q}_k^{\text{in}}$, notice that $\mathbf{q}_k^{\text{in}} = \mathbf{V}\mathbf{v}_k^{\text{in}}$ by Algorithm 3, so that $\mathbf{q}_k^{\text{in}}|_{Q_k} = \mathbf{V}\mathbf{v}_k^{\text{in}}|_{Q_k}$ (where $\mathbf{v}_k^{\text{in}} \in Q_k$). Recall from the discussion for the $k = 0$ case above that $\mathbf{V}|_{Q_k} \overset{d}{=} \mathbf{V}|_{Q_k^{\text{in}}}$. Thus, as before, we drop the explicit "in" superscript, and we can use Lemma 2 to write

$$\begin{aligned}\mathbf{q}_k|_{Q_k} &\overset{d}{=} \mathbf{C}_{uk}(\mathbf{C}_{uk}^T\mathbf{C}_{uk})^{-1}\mathbf{C}_{pk}^T\mathbf{v}_k + \mathbf{B}^{\perp}_{\mathbf{C}_{uk}}\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{pk}}]^T\mathbf{v}_k \\ &= \sum_{\ell=0}^{k-1}[\boldsymbol{\beta}_{qk}]_{\ell+1}\mathbf{q}_\ell + \mathbf{C}_{uk}\left((\mathbf{C}_{uk}^T\mathbf{C}_{uk})^{-1}\mathbf{C}_{pk}^T\mathbf{v}_k - \begin{bmatrix}\mathbf{0}_{k+1}\\ \boldsymbol{\beta}_{qk}\end{bmatrix}\right) + \mathbf{B}^{\perp}_{\mathbf{C}_{uk}}\widetilde{\mathbf{V}}[\mathbf{B}^{\perp}_{\mathbf{C}_{pk}}]^T\mathbf{v}_k,\end{aligned}\tag{2.49}$$

where we recall that $\widetilde{\mathbf{V}}$ is Haar-distributed and independent of the conditioning sigma-algebra. The second equality in (2.49) follows from the fact that $\mathbf{C}_{uk} = [\mathbf{U}_k\ \mathbf{Q}_{k-1}]$ from (2.37); thus,

$$\mathbf{C}_{uk}\begin{bmatrix}\mathbf{0}_{k+1}\\ \boldsymbol{\beta}_{qk}\end{bmatrix} = \mathbf{U}_k\mathbf{0}_{k+1} + \mathbf{Q}_{k-1}\boldsymbol{\beta}_{qk} = \sum_{\ell=0}^{k-1}[\boldsymbol{\beta}_{qk}]_{\ell+1}\mathbf{q}_\ell.$$

As above in (2.47)-(2.48), because $\widetilde{\mathbf{V}}$ is Haar distributed and because $\mathbf{B}_{\mathbf{C}_{uk}}^{\perp}\mathbf{Z}_{qk} = \mathbf{O}_{qk}\overline{\mathbf{Z}}_{qk} - \mathbf{B}_{\mathbf{C}_{uk}}\check{\mathbf{Z}}_{qk}$, we find that conditional on $\mathcal{Q}_k$,

$$\begin{aligned}\mathbf{B}_{\mathbf{C}_{uk}}^{\perp}\widetilde{\mathbf{V}}[\mathbf{B}_{\mathbf{C}_{pk}}^{\perp}]^T\mathbf{v}_k &\overset{d}{=} \frac{\|[\mathbf{B}_{\mathbf{C}_{pk}}^{\perp}]^T\mathbf{v}_k\|}{\|\mathbf{Z}_{qk}\|}\,\mathbf{B}_{\mathbf{C}_{uk}}^{\perp}\,\mathbf{Z}_{qk} \\ &= \left[\frac{\|[\mathbf{B}_{\mathbf{C}_{pk}}^{\perp}]^T\mathbf{v}_k\|}{\|\mathbf{Z}_{qk}\|} - \sqrt{\rho_{qk}}\right]\mathbf{B}_{\mathbf{C}_{uk}}^{\perp}\,\mathbf{Z}_{qk} - \sqrt{\rho_{qk}}\,\mathbf{B}_{\mathbf{C}_{uk}}\,\check{\mathbf{Z}}_{qk} + \sqrt{\rho_{qk}}\,\mathbf{O}_{qk}\,\overline{\mathbf{Z}}_{qk}.\end{aligned}$$

Combining with (2.49) and using the definition of $\mathbf{\Delta}_{qk}$ in (2.43), we find the desired result:

$$\begin{aligned}\mathbf{q}_k|_{\mathcal{Q}_k} &\overset{d}{=} \sum_{\ell=0}^{k-1}[\boldsymbol{\beta}_{qk}]_{\ell+1}\mathbf{q}_\ell + \mathbf{C}_{uk}\left((\mathbf{C}_{uk}^T\mathbf{C}_{uk})^{-1}\mathbf{C}_{pk}^T\mathbf{v}_k - \begin{bmatrix}\mathbf{0}_{k+1}\\ \boldsymbol{\beta}_{qk}\end{bmatrix}\right) + \sqrt{\rho_{qk}}\,\mathbf{O}_{qk}\,\overline{\mathbf{Z}}_{qk} \\ &\qquad - \sqrt{\rho_{qk}}\,\mathbf{B}_{\mathbf{C}_{uk}}\,\check{\mathbf{Z}}_{qk} + \left[\frac{\|[\mathbf{B}_{\mathbf{C}_{pk}^{\text{in}}}^{\perp}]^T\mathbf{v}_k^{\text{in}}\|}{\|\mathbf{Z}_{qk}\|} - \sqrt{\rho_{qk}}\right]\mathbf{B}_{\mathbf{C}_{uk}}^{\perp}\mathbf{Z}_{qk} \\ &= \sum_{\ell=0}^{k-1}[\boldsymbol{\beta}_{qk}]_{\ell+1}\mathbf{q}_\ell + \sqrt{\rho_{qk}}\,\mathbf{O}_{qk}\,\overline{\mathbf{Z}}_{qk} + \mathbf{\Delta}_{qk}.\end{aligned}$$

Finally, we establish the conditional independence of $\mathbf{q}_k^{\text{in}}$ and $\mathbf{q}_k^{\text{out}}$ given the σ-algebra $\mathcal{Q}_k$ for all $k \geq 0$. Notice that conditioning on $\mathcal{Q}_k$ is equivalent to conditioning on $\{g(\mathbf{V}) = 0, h(\mathbf{U}) = 0\}$ where g, h are the (deterministic) functions $\mathbf{X} \mapsto \mathbf{C}_{pk}^{\text{in}} - \mathbf{X}\mathbf{C}_{uk}^{\text{in}}$ and $\mathbf{X} \mapsto \mathbf{C}_{pk}^{\text{out}} - \mathbf{X}\mathbf{C}_{uk}^{\text{out}}$, respectively. This is just a reformulation of the linear constraints discussed earlier. Hence,

$$\begin{aligned}\mathbb{P}\left(\mathbf{q}_k^{\text{in}} \in A, \mathbf{q}_k^{\text{out}} \in B \mid \mathcal{Q}_k\right) &= \mathbb{P}\left(\mathbf{q}_k^{\text{in}} \in A, \mathbf{q}_k^{\text{out}} \in B \mid g(\mathbf{V}) = 0, h(\mathbf{U}) = 0\right) \\ &= \mathbb{P}\left(\mathbf{V}^T\mathbf{v}_k^{\text{in}} \in A, \mathbf{U}^T\mathbf{v}_k^{\text{out}} \in B \mid g(\mathbf{V}) = 0, h(\mathbf{U}) = 0\right).\end{aligned} \tag{2.50}$$

In the above, the second equality follows because $\mathbf{q}_k^{\text{in}} = \mathbf{V}^T\mathbf{v}_k^{\text{in}}$ and $\mathbf{q}_k^{\text{out}} = \mathbf{U}^T\mathbf{v}_k^{\text{out}}$ by Algorithm 3

step (8). Now applying the definition of conditional probability,

$$\begin{aligned}\mathbb{P}\left(\mathbf{q}_k^{\text{in}} \in A, \mathbf{q}_k^{\text{out}} \in B \mid \mathcal{Q}_k\right) &= \frac{\mathbb{P}\left(\mathbf{V}^T\mathbf{v}_k^{\text{in}} \in A, \mathbf{U}^T\mathbf{v}_k^{\text{out}} \in B, g(\mathbf{V}) = 0, h(\mathbf{U}) = 0\right)}{\mathbb{P}\left(g(\mathbf{V}) = 0, h(\mathbf{U}) = 0\right)} \\ &= \frac{\mathbb{P}\left(\mathbf{V}^T\mathbf{v}_k^{\text{in}} \in A, g(\mathbf{V}) = 0\right)}{\mathbb{P}\left(g(\mathbf{V}) = 0\right)} \times \frac{\mathbb{P}\left(\mathbf{U}^T\mathbf{v}_k^{\text{out}} \in B, h(\mathbf{U}) = 0\right)}{\mathbb{P}\left(h(\mathbf{U}) = 0\right)} \\ &= \mathbb{P}\left(\mathbf{q}_k^{\text{in}} \in A \mid g(\mathbf{V}) = 0\right)\mathbb{P}\left(\mathbf{q}_k^{\text{out}} \in B \mid h(\mathbf{U}) = 0\right).\end{aligned} \tag{2.51}$$

The second equality of (2.51) follows as $\mathbf{U}$ and $\mathbf{V}$ are independent and $\mathbf{v}_k^{\text{in/out}}$ is measurable with respect to $\mathcal{Q}_k^{\text{in/out}}$ (and because conditioning on $\{g(\mathbf{V}) = 0\}$ or $\mathcal{Q}_k^{\text{in}}$ is equivalent; similarly for $\{h(\mathbf{U}) = 0\}$ and $\mathcal{Q}_k^{\text{out}}$). This establishes conditional independence of $\mathbf{q}_k^{\text{in}}$ and $\mathbf{q}_k^{\text{out}}$ given $\mathcal{Q}_k$ as

$$\begin{aligned}&\mathbb{P}\left(\mathbf{q}_k^{\text{in}} \in A \mid g(\mathbf{V}) = 0\right)\mathbb{P}\left(\mathbf{q}_k^{\text{out}} \in B \mid h(\mathbf{U}) = 0\right) \\ &\quad= \mathbb{P}\left(\mathbf{q}_k^{\text{in}} \in A \mid g(\mathbf{V}) = 0, h(\mathbf{U}) = 0\right)\mathbb{P}\left(\mathbf{q}_k^{\text{out}} \in B \mid g(\mathbf{V}) = 0, h(\mathbf{U}) = 0\right) \\ &\quad= \mathbb{P}\left(\mathbf{q}_k^{\text{in}} \in A \mid \mathcal{Q}_k\right)\mathbb{P}\left(\mathbf{q}_k^{\text{out}} \in B \mid \mathcal{Q}_k\right).\end{aligned}$$

□

Next we extend this lemma to a result characterizing the discrepancy between the finite sample iterates and the limiting behavior of the $\mathbf{p}$ and $\mathbf{q}$ iterates jointly. To do this, we introduce a new recursion in terms of Gaussian vectors, $\overline{\mathbf{Z}}_p^{\text{in/out}}$ and $\overline{\mathbf{Z}}_q^{\text{in/out}}$ defined in Lemma 3. As we see in Lemma 4, the iterates generated by this recursion, referred to as the equivalent Gaussian recursion, are jointly equal in distribution to the iterates in the general recursion, Algorithm 3.

Algorithm 5 Equivalent Gaussian Representation of General Recursion

Require: Separable denoisers $\mathbf{f}_p^{\text{in}}$, $\mathbf{f}_p^{\text{out}}$, $\mathbf{f}_q^{\text{in}}$, and $\mathbf{f}_q^{\text{out}}$; parameter update functions $\Gamma_i : \mathbb{R}^2 \to \mathbb{R}$; initial data $\widetilde{\mathbf{u}}_0^{\text{in}} \in \mathbb{R}^N$ and $\widetilde{\mathbf{u}}_0^{\text{out}} \in \mathbb{R}^M$; and disturbance vectors $\mathbf{w}_p^{\text{in}}, \mathbf{w}_q^{\text{in}} \in \mathbb{R}^N$ and $\mathbf{w}_p^{\text{out}}, \mathbf{w}_q^{\text{out}} \in \mathbb{R}^M$.

1: $\widetilde{\mathbf{p}}_k^{\text{in}} \leftarrow \sum_{r=0}^k [\mathbf{c}_{pk}^{\text{in}}]_r \left(\sqrt{\rho_{pr}^{\text{in}}}\, \widetilde{\mathbf{O}}_{pr}^{\text{in}}\, \overline{\mathbf{Z}}_{pr}^{\text{in}} + \widetilde{\mathbf{\Delta}}_{pr}^{\text{in}} \right)$, $\quad \widetilde{\mathbf{p}}_k^{\text{out}} \leftarrow \sum_{r=0}^k [\mathbf{c}_{pk}^{\text{out}}]_r \left(\sqrt{\rho_{pr}^{\text{out}}}\, \widetilde{\mathbf{O}}_{pr}^{\text{out}}\, \overline{\mathbf{Z}}_{pr}^{\text{out}} + \widetilde{\mathbf{\Delta}}_{pr}^{\text{out}} \right)$,

2: $\widetilde{\mathbf{v}}_k^{\text{in}} \leftarrow \frac{1}{1-\alpha_{pk}^{\text{in}}} \left[\mathbf{f}_p^{\text{in}} \left(\widetilde{\mathbf{p}}_k^{\text{in}}, \widetilde{\mathbf{p}}_k^{\text{out}}, \mathbf{w}_p^{\text{in}}, \gamma_{pk}^{\text{out}}, \gamma_{pk}^{\text{in}} \right) - \alpha_{pk}^{\text{in}} \widetilde{\mathbf{p}}_k^{\text{in}} \right]$,

$\widetilde{\mathbf{v}}_k^{\text{out}} \leftarrow \frac{1}{1-\alpha_{pk}^{\text{out}}} \left[\mathbf{f}_p^{\text{out}} \left(\widetilde{\mathbf{p}}_k^{\text{in}}, \widetilde{\mathbf{p}}_k^{\text{out}}, \mathbf{w}_p^{\text{out}}, \gamma_{pk}^{\text{out}}, \gamma_{pk}^{\text{in}} \right) - \alpha_{pk}^{\text{out}} \widetilde{\mathbf{p}}_k^{\text{out}} \right]$,

3: $\widetilde{\mathbf{q}}_k^{\text{in}} \leftarrow \sum_{r=0}^k [\mathbf{c}_{qk}^{\text{in}}]_r \left(\sqrt{\rho_{qr}^{\text{in}}}\, \widetilde{\mathbf{O}}_{qr}^{\text{in}}\, \overline{\mathbf{Z}}_{qr}^{\text{in}} + \widetilde{\mathbf{\Delta}}_{qr}^{\text{in}} \right)$, $\quad \widetilde{\mathbf{q}}_k^{\text{out}} \leftarrow \sum_{r=0}^k [\mathbf{c}_{qk}^{\text{out}}]_r \left(\sqrt{\rho_{qr}^{\text{out}}}\, \widetilde{\mathbf{O}}_{qr}^{\text{out}}\, \overline{\mathbf{Z}}_{qr}^{\text{out}} + \widetilde{\mathbf{\Delta}}_{qr}^{\text{out}} \right)$,

4: $\widetilde{\mathbf{u}}_{k+1}^{\text{in}} \leftarrow \frac{1}{1-\alpha_{qk}^{\text{in}}} \left[\mathbf{f}_q^{\text{in}} \left(\widetilde{\mathbf{q}}_k^{\text{in}}, \widetilde{\mathbf{q}}_k^{\text{out}}, \mathbf{w}_q^{\text{in}}, \gamma_{qk}^{\text{out}}, \gamma_{qk}^{\text{in}} \right) - \alpha_{qk}^{\text{in}} \widetilde{\mathbf{q}}_k^{\text{in}} \right]$,

$\widetilde{\mathbf{u}}_{k+1}^{\text{out}} \leftarrow \frac{1}{1-\alpha_{qk}^{\text{out}}} \left[\mathbf{f}_q^{\text{out}} \left(\widetilde{\mathbf{q}}_k^{\text{in}}, \widetilde{\mathbf{q}}_k^{\text{out}}, \mathbf{w}_q^{\text{out}}, \gamma_{qk}^{\text{out}}, \gamma_{qk}^{\text{in}} \right) - \alpha_{qk}^{\text{out}} \widetilde{\mathbf{q}}_k^{\text{out}} \right]$.

In the above, all relevant scalar quantities (e.g. $\alpha_{pk}^{\text{in/out}}$) are understood to be defined analogously as in the general recursion, Algorithm 3, but now in terms of the Gaussian equivalent iterates (e.g. $\widetilde{\mathbf{p}}_k^{\text{in}}$ and $\widetilde{\mathbf{p}}_k^{\text{out}}$). Likewise, the discrepancy terms $\widetilde{\mathbf{\Delta}}_{pk}^{\text{in/out}}$ and $\widetilde{\mathbf{\Delta}}_{qk}^{\text{in/out}}$ are defined as those in Lemma 3, but using the Gaussian equivalent iterates. Constant vectors $\mathbf{c}_{pk}^{\text{in/out}}$ and $\mathbf{c}_{qk}^{\text{in/out}}$ are defined explicitly (2.56) below in terms of the $\boldsymbol{\beta}_{pk}^{\text{in/out}}$ and $\boldsymbol{\beta}_{qk}^{\text{in/out}}$ vectors. These vectors, along with $\rho_{pr}^{\text{in/out}}$ and $\rho_{qk}^{\text{in/out}}$ remain unspecified, as the statement of Lemma 4 is valid for any choice of these constants.

The Gaussian equivalent recursion is initialized in the same way as the GVAMP recursion, so in particular $\widetilde{\mathbf{u}}_0^{\text{in/out}} = \mathbf{u}_0^{\text{in/out}}$, and thus $\widetilde{\mathbf{\Delta}}_{p0}^{\text{in/out}} = \mathbf{\Delta}_{p0}^{\text{in/out}}$. Furthermore, $\widetilde{\mathbf{O}}_{p0}^{\text{in/out}}$ and $\mathbf{O}_{p0}^{\text{in/out}}$ are deterministic orthogonal matrices, so we choose them to be equal as well.

In order to state Lemma 4, we first introduce some new notation. For all $j \geq 0$, we will let

$$\mathbf{r}_j = \left(\mathbf{p}_j^{\text{in}}, \mathbf{p}_j^{\text{out}}, \mathbf{q}_j^{\text{in}}, \mathbf{q}_j^{\text{out}} \right), \qquad \widetilde{\mathbf{r}}_j = \left(\widetilde{\mathbf{p}}_j^{\text{in}}, \widetilde{\mathbf{p}}_j^{\text{out}}, \widetilde{\mathbf{q}}_j^{\text{in}}, \widetilde{\mathbf{q}}_j^{\text{out}} \right), \tag{2.52}$$

where the terms in $\mathbf{r}_j$ are those in Algorithm 3, while those in $\widetilde{\mathbf{r}}_j$ are from Algorithm 5. We also define the concatenation of these up through iteration k as

$$\underline{\mathbf{r}}_k = (\mathbf{r}_0, \ldots, \mathbf{r}_k), \qquad \underline{\widetilde{\mathbf{r}}}_k = (\widetilde{\mathbf{r}}_0, \ldots, \widetilde{\mathbf{r}}_k). \tag{2.53}$$

Lemma 4. *Let* $\{\overline{\mathbf{Z}}_{pk}^{\text{in}}\}_{k\geq 0}$ *and* $\{\overline{\mathbf{Z}}_{pk}^{\text{out}}\}_{k\geq 0}$ *be the i.i.d. standard Gaussian vector sequences defined in Lemma 3. Then, with respect to the Gaussian equivalent recursion in Algorithm 5, define*

$$\begin{aligned}
\overset{*}{\mathbf{p}}_0^{\text{in}} &= \sqrt{\rho_{p0}^{\text{in}}}\,\widetilde{\mathbf{O}}_{p0}^{\text{in}}\,\overline{\mathbf{Z}}_{p0}^{\text{in}}, \qquad \textit{and} \qquad \overset{*}{\mathbf{p}}_k^{\text{in}} = \sum_{r=0}^{k-1}[\boldsymbol{\beta}_{pk}^{\text{in}}]_{r+1}\overset{*}{\mathbf{p}}_r^{\text{in}} + \sqrt{\rho_{pk}^{\text{in}}}\,\widetilde{\mathbf{O}}_{pk}^{\text{in}}\,\overline{\mathbf{Z}}_{pk}^{\text{in}}, \\
\overset{*}{\mathbf{p}}_0^{\text{out}} &= \sqrt{\rho_{p0}^{\text{out}}}\,\widetilde{\mathbf{O}}_{p0}^{\text{out}}\,\overline{\mathbf{Z}}_{p0}^{\text{out}}, \qquad \textit{and} \qquad \overset{*}{\mathbf{p}}_k^{\text{out}} = \sum_{r=0}^{k-1}[\boldsymbol{\beta}_{pk}^{\text{out}}]_{r+1}\overset{*}{\mathbf{p}}_r^{\text{out}} + \sqrt{\rho_{pk}^{\text{out}}}\,\widetilde{\mathbf{O}}_{pk}^{\text{out}}\,\overline{\mathbf{Z}}_{pk}^{\text{out}}.
\end{aligned} \tag{2.54}$$

Similarly define $\overset{*}{\mathbf{q}}_k^{\text{in}}$ *and* $\overset{*}{\mathbf{q}}_k^{\text{out}}$*. With these definitions, we have*

1. $$\overset{*}{\mathbf{p}}_k^{\text{in}} = \sum_{r=0}^{k}\sqrt{\rho_{pr}^{\text{in}}}\,[\mathbf{c}_{pk}^{\text{in}}]_r\,\widetilde{\mathbf{O}}_{pr}^{\text{in}}\,\overline{\mathbf{Z}}_{pr}^{\text{in}}, \qquad \overset{*}{\mathbf{p}}_k^{\text{out}} = \sum_{r=0}^{k}\sqrt{\rho_{pr}^{\text{out}}}\,[\mathbf{c}_{pk}^{\text{out}}]_r\,\widetilde{\mathbf{O}}_{pr}^{\text{out}}\,\overline{\mathbf{Z}}_{pr}^{\text{out}}, \tag{2.55}$$

   where for $k \geq 1$ *and* $0 \leq r \leq k-1$, $[\mathbf{c}_{pk}^{\text{in}}]_r$ *and* $[\mathbf{c}_{pk}^{\text{out}}]_r$ *are defined recursively as*

$$[\mathbf{c}_{pk}^{\text{in}}]_r = \sum_{i=r}^{k-1}[\mathbf{c}_{pi}^{\text{in}}]_r[\boldsymbol{\beta}_{pk}^{\text{in}}]_{i+1}, \qquad \textit{and} \qquad [\mathbf{c}_{pk}^{\text{out}}]_r = \sum_{i=r}^{k-1}[\mathbf{c}_{pi}^{\text{out}}]_r[\boldsymbol{\beta}_{pk}^{\text{out}}]_{i+1}, \tag{2.56}$$

   with $[\mathbf{c}_{pk}^{\text{in}}]_k = [\mathbf{c}_{pk}^{\text{out}}]_k = 1$ *for all* $k \geq 0$*.*

2. *Let* $\underline{\overset{*}{\mathbf{r}}}_k = \left(\overset{*}{\mathbf{r}}_0, \ldots, \overset{*}{\mathbf{r}}_k\right)$ *where* $\overset{*}{\mathbf{r}}_j = (\overset{*}{\mathbf{p}}_j^{\text{in}}, \overset{*}{\mathbf{p}}_j^{\text{out}}, \overset{*}{\mathbf{q}}_j^{\text{in}}, \overset{*}{\mathbf{q}}_j^{\text{out}})$ *are defined as above for the ideal variables* $\overset{*}{\mathbf{p}}$ *and* $\overset{*}{\mathbf{q}}$*. Then we have that*

$$\underline{\widetilde{\mathbf{r}}}_k - \underline{\overset{*}{\mathbf{r}}}_k = (\mathbf{d}_0, \ldots, \mathbf{d}_k)\,, \tag{2.57}$$

   where $\underline{\widetilde{\mathbf{r}}}_k$ *is defined in* (2.53) *with*

$$\mathbf{d}_k = \left(\sum_{r=0}^{k}[\mathbf{c}_{pk}^{\text{in}}]_r\widetilde{\boldsymbol{\Delta}}_{pr}^{\text{in}},\ \sum_{r=0}^{k}[\mathbf{c}_{pk}^{\text{out}}]_r\widetilde{\boldsymbol{\Delta}}_{pr}^{\text{out}},\ \sum_{r=0}^{k}[\mathbf{c}_{qk}^{\text{in}}]_r\widetilde{\boldsymbol{\Delta}}_{qr}^{\text{in}},\ \sum_{r=0}^{k}[\mathbf{c}_{qk}^{\text{out}}]_r\widetilde{\boldsymbol{\Delta}}_{qr}^{\text{out}}\right). \tag{2.58}$$

   Furthermore, for all $k \geq 0$*, we have that, conditional on* $\left(\mathbf{w}_p^{\text{in/out}}, \mathbf{w}_q^{\text{in/out}}\right)$*,*

$$\underline{\widetilde{\mathbf{r}}}_k \overset{d}{=} \underline{\mathbf{r}}_k. \tag{2.59}$$

3. *For all $i \geq 1$, we have* $\left([\overset{*}{\mathbf{p}}_0^{\text{in}}]_i, \ldots, [\overset{*}{\mathbf{p}}_k^{\text{in}}]_i, [\overset{*}{\mathbf{p}}_0^{\text{out}}]_i, \ldots, [\overset{*}{\mathbf{p}}_k^{\text{out}}]_i\right) \overset{d}{=} \left(P_0^{\text{in}}, \ldots, P_k^{\text{in}}, P_0^{\text{out}}, \ldots, P_k^{\text{out}}\right)$ *where* $(P_0^{\text{in}}, \ldots, P_k^{\text{in}})$ *and* $(P_0^{\text{out}}, \ldots, P_k^{\text{out}})$ *are independent, zero-mean, jointly Gaussian vectors (and hence the right hand side is itself jointly Gaussian).*

Analogous statements for (1.)-(3.) hold for the $\overset{*}{\mathbf{q}}$ *variables.*

Recall that we introduced the Gaussian equivalent recursion in Algorithm 5 as an idealized 'Gaussian' version of the GVAMP general recursion in Algorithm 3. Lemma 4 makes this precise. Indeed, Lemma 4 introduces a recursion in Algorithm 5 that is shown to be purely Gaussian in point (3.), in that its iterates are elementwise equal in distribution to Gaussian vectors who covariance structure can be given exactly. We see from point (2.) of the lemma, that the difference between the terms in Algorithm 5 (given in $\widetilde{\underline{\mathbf{r}}}_k$) and the purely Gaussian terms from (2.55) (collected in $\overset{*}{\underline{\mathbf{r}}}_k$) is only in the deviation terms (see (2.58)). Moreover, the terms of Algorithm 5 (given in $\widetilde{\underline{\mathbf{r}}}_k$) are equal in distribution to those from Algorithm 3 (given in $\underline{\mathbf{r}}_k$). Taken together, this shows that the difference between the GVAMP general recursion and the purely Gaussian recursion can be summarized via the deviation terms. The major technical part of the proof, given in Lemma 5 is in showing that these deviation terms concentrate exponentially fast to zero, so that the dynamics of the GVAMP general recursion can be characterized by studying the purely Gaussian equivalent.

Before proving Lemma 4, we note that the Gaussian vectors used to characterize the dynamics of the purely Gaussian recursion in point (3.), namely $(P_0^{\text{in}/out}, \ldots, P_k^{\text{in/out}})$ and $(Q_0^{\text{in/out}}, \ldots, Q_k^{\text{in/out}})$, may be explicitly defined in terms of the $\overline{\mathbf{Z}}$ variables defined in Lemma 3 as follows. For $k \geq 1$,

$$P_k^{\text{in/out}} = \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}]_{r+1} P_r^{\text{in/out}} + \sqrt{\rho_{pk}^{\text{in/out}}} [\overline{\mathbf{Z}}_{pk}^{\text{in/out}}]_1, \quad Q_k^{\text{in/out}} = \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{qk}]_{r+1} Q_r^{\text{in/out}} + \sqrt{\rho_{qk}^{\text{in/out}}} [\overline{\mathbf{Z}}_{qk}^{\text{in/out}}]_1, \tag{2.60}$$

with initializations $P_0^{\text{in/out}} = \sqrt{\rho_{p0}^{\text{in/out}}} [\overline{\mathbf{Z}}_{p0}^{\text{in/out}}]_1$ and $Q_0^{\text{in/out}} = \sqrt{\rho_{q0}^{\text{in/out}}} [\overline{\mathbf{Z}}_{q0}^{\text{in/out}}]_1$.

Proof. We prove the results for $\mathbf{p}_k^{\text{in}}$; the arguments for $\mathbf{q}_k^{\text{in}}$ and the output iterates are completely analogous. To establish the first part, namely the result of (2.55), we proceed by induction. First

consider the base case $k = 0$. We have $[c_{p0}^{\text{in}}]_0 = 1$ by definition (2.56), and thus the statement in (2.55) reduces to $\overset{*}{\mathbf{p}}_0^{\text{in}} = \sqrt{\rho_{p0}^{\text{in}}}\, \widetilde{\mathbf{O}}_{p0}^{\text{in}}\, \overline{\mathbf{Z}}_{p0}^{\text{in}}$, which is just the definition of $\overset{*}{\mathbf{p}}_0^{\text{in}}$ in (2.54).

Now assume the statement in (2.55) holds for $\overset{*}{\mathbf{p}}_j^{\text{in}}$ for $0 \leq j \leq k-1$. Then, combining the definition of $\overset{*}{\mathbf{p}}_k^{\text{in}}$ in (2.54) with the induction hypothesis, we find

$$\begin{aligned}
\overset{*}{\mathbf{p}}_k^{\text{in}} &= \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{in}}]_{r+1} \overset{*}{\mathbf{p}}_r^{\text{in}} + \sqrt{\rho_{pk}^{\text{in}}}\, \widetilde{\mathbf{O}}_{pk}^{\text{in}}\, \overline{\mathbf{Z}}_{pk}^{\text{in}} \\
&= \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{in}}]_{r+1} \left[\sum_{j=0}^{r} \sqrt{\rho_{pj}^{\text{in}}}\, [\mathbf{c}_{pr}^{\text{in}}]_j\, \widetilde{\mathbf{O}}_{pj}^{\text{in}}\, \overline{\mathbf{Z}}_{pj}^{\text{in}} \right] + \sqrt{\rho_{pk}^{\text{in}}}\, \widetilde{\mathbf{O}}_{pk}^{\text{in}}\, \overline{\mathbf{Z}}_{pk}^{\text{in}} \\
&= \sum_{j=0}^{k-1} \left[\sum_{r=j}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{in}}]_{r+1}\, [\mathbf{c}_{pr}^{\text{in}}]_j \right] \sqrt{\rho_{pj}^{\text{in}}}\, \widetilde{\mathbf{O}}_{pj}^{\text{in}}\, \overline{\mathbf{Z}}_{pj}^{\text{in}} + \sqrt{\rho_{pk}^{\text{in}}}\, \widetilde{\mathbf{O}}_{pk}^{\text{in}}\, \overline{\mathbf{Z}}_{pk}^{\text{in}} = \sum_{j=0}^{k} [\mathbf{c}_{pk}^{\text{in}}]_j\, \sqrt{\rho_{pj}^{\text{in}}}\, \widetilde{\mathbf{O}}_{pj}^{\text{in}}\, \overline{\mathbf{Z}}_{pj}^{\text{in}},
\end{aligned}$$

as needed, where the final equality uses the definition of $\mathbf{c}_{pk}^{\text{in}}$ in (2.56).

For the second part, we observe that (2.57) becomes trivially true in light of the first result where we showed, for example, that $\overset{*}{\mathbf{p}}_k^{\text{in}} = \sum_{r=0}^{k} \sqrt{\rho_{pr}^{\text{in}}}\, [\mathbf{c}_{pk}^{\text{in}}]_r\, \widetilde{\mathbf{O}}_{pr}^{\text{in}}\, \overline{\mathbf{Z}}_{pr}^{\text{in}}$, and the Gaussian Algorithm 5 step (1) definition, for which $\widetilde{\mathbf{p}}_k^{\text{in}} = \sum_{r=0}^{k} \sqrt{\rho_{pr}^{\text{in}}}\, [\mathbf{c}_{pk}^{\text{in}}]_r\, \widetilde{\mathbf{O}}_{pr}^{\text{in}}\, \overline{\mathbf{Z}}_{pr}^{\text{in}} + \sum_{r=0}^{k} [\mathbf{c}_{pk}^{\text{in}}]_r \widetilde{\boldsymbol{\Delta}}_{pr}^{\text{in}}$.

To establish the distributional equalities in (2.59) of part (3.), we make repeated use of two facts:

Fact 1. *Let* $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathbb{R}^{n_1}$ *and* $\mathbf{y}, \widetilde{\mathbf{y}} \in \mathbb{R}^{n_2}$ *be random vectors. Then we have:*

(A) If $\mathbf{y} \,|_{\mathbf{x}} \overset{d}{=} \widetilde{\mathbf{y}}$*, then* $(\mathbf{x}, \mathbf{y}) \overset{d}{=} (\mathbf{x}, \widetilde{\mathbf{y}})$.

(B) If $\mathbf{y} \overset{d}{=} \widetilde{\mathbf{y}}$ *and* $\mathbf{x}$ *is independent of* $\mathbf{y}$ *and* $\widetilde{\mathbf{y}}$*, then* $f(\mathbf{x}, \mathbf{y}) \overset{d}{=} f(\mathbf{x}, \widetilde{\mathbf{y}})$ *for any measurable* f.

Now we can show $\underline{\mathbf{r}}_k \overset{d}{=} \underline{\widetilde{\mathbf{r}}}_k$ given in (2.59) inductively. For the initialization step, we show $\underline{\mathbf{r}}_0 = \mathbf{r}_0 \overset{d}{=} \widetilde{\mathbf{r}}_0 = \underline{\widetilde{\mathbf{r}}}_0$ where by (2.52) we have that $\mathbf{r}_0 = \left(\mathbf{p}_0^{\text{in}}, \mathbf{p}_0^{\text{out}}, \mathbf{q}_0^{\text{in}}, \mathbf{q}_0^{\text{out}}\right)$ and $\widetilde{\mathbf{r}}_0 = \left(\widetilde{\mathbf{p}}_0^{\text{in}}, \widetilde{\mathbf{p}}_0^{\text{out}}, \widetilde{\mathbf{q}}_0^{\text{in}}, \widetilde{\mathbf{q}}_0^{\text{out}}\right)$. We prove the distributional equality for the $\mathbf{p}^{\text{in}}$ and $\mathbf{p}^{\text{out}}$ terms, while the proof for $\mathbf{q}^{\text{in}}$ and $\mathbf{q}^{\text{out}}$ follows similarly. By **Condition 0** and the definition of $\widetilde{\mathbf{p}}_0^{\text{in/out}}$, we have that

$$\mathbf{p}_0^{\text{in}} |_{\mathcal{P}_0} \overset{d}{=} \sqrt{\rho_{p0}^{\text{in}}}\, \overline{\mathbf{Z}}_{p0}^{\text{in}} + \boldsymbol{\Delta}_{p0}^{\text{in}} \qquad \text{and} \qquad \widetilde{\mathbf{p}}_0^{\text{in}} |_{\mathcal{P}_0} = \sqrt{\rho_{p0}^{\text{in}}}\, [\mathbf{c}_{p0}^{\text{in}}]_0\, \widetilde{\mathbf{O}}_{p0}^{\text{in}}\, \overline{\mathbf{Z}}_{p0}^{\text{in}} + [\mathbf{c}_{p0}^{\text{in}}]_0 \widetilde{\boldsymbol{\Delta}}_{p0}^{\text{in}}.$$

To see the equivalence of the two terms above, notice that $[\mathbf{c}_{p0}^{\text{in}}]_0 = 1$ and $\widetilde{\mathbf{O}}_{p0}^{\text{in/out}} = \mathbf{O}_{p0}^{\text{in/out}} = \mathbf{I}$ by definition, and $\mathbf{\Delta}_{p0}^{\text{in}} = \widetilde{\mathbf{\Delta}}_{p0}^{\text{in}}$ since the GVAMP recursion and general recursion use identical initialization $\mathbf{u}_0^{\text{in}}$. Furthermore, observe that $\mathbf{p}_0^{\text{in}}$ and $\mathbf{p}_0^{\text{out}}$ are independent since $\mathbf{U}$ and $\mathbf{V}$ are independent, and $\widetilde{\mathbf{p}}_0^{\text{in}}$ and $\widetilde{\mathbf{p}}_0^{\text{out}}$ are independent since $\overline{\mathbf{Z}}_{p0}^{\text{in}}$ and $\overline{\mathbf{Z}}_{p0}^{\text{out}}$ are independent (by definition), and $\mathbf{u}_0^{\text{in}}$ and $\mathbf{u}_0^{\text{out}}$ are independent by **Condition 0**.

Next, by definition, conditioning on Q_0 is equivalent to conditioning on $\sigma\left(\mathbf{p}_0^{\text{in}}, \mathbf{p}_0^{\text{out}}\right)$ (since $\mathbf{v}_0$ is a deterministic function of $\mathbf{p}_0$ conditional on the disturbance vectors; see Algorithm 3). So Lemma 3 gives that $\mathbf{q}_0^{\text{in/out}} \mid_{\mathbf{p}_0^{\text{in}}, \mathbf{p}_0^{\text{out}}} = \sqrt{\rho_{q0}^{\text{in/out}}}\mathbf{O}_{q0}^{\text{in/out}}\overline{\mathbf{Z}}_{q0}^{\text{in/out}} + \mathbf{\Delta}_{q0}^{\text{in/out}}$. Furthermore, $\mathbf{q}_0^{\text{in}}$ and $\mathbf{q}_0^{\text{out}}$ are again conditionally independent given Q_0 since $\mathbf{V}$ and $\mathbf{U}$ are independent, so the above holds for $(\mathbf{q}_0^{\text{in}}, \mathbf{q}_0^{\text{out}})$ jointly. Together, this implies

$$\left(\mathbf{p}_0^{\text{in}}, \mathbf{p}_0^{\text{out}}, \mathbf{q}_0^{\text{in}}, \mathbf{q}_0^{\text{out}}\right) \overset{d}{=} \left(\mathbf{p}_0^{\text{in}}, \mathbf{p}_0^{\text{out}}, \sqrt{\rho_{q0}^{\text{in}}}\mathbf{O}_{q0}^{\text{in}}\overline{\mathbf{Z}}_{q0}^{\text{in}} + \mathbf{\Delta}_{q0}^{\text{in}}, \sqrt{\rho_{q0}^{\text{out}}}\mathbf{O}_{q0}^{\text{out}}\overline{\mathbf{Z}}_{q0}^{\text{out}} + \mathbf{\Delta}_{q0}^{\text{out}}\right),$$

by (A) in Fact 1. Now note that $\mathbf{O}_{q0}^{\text{in/out}}$ and $\mathbf{\Delta}_{q0}^{\text{in/out}}$ are deterministic functions of $(\mathbf{p}_0^{\text{in}}, \mathbf{p}_0^{\text{out}}, \overline{\mathbf{Z}}_{q0}^{\text{in}}, \overline{\mathbf{Z}}_{q0}^{\text{out}})$, and thus the entire above vector is also a function of these variables. Furthermore, by Lemma 3, $(\mathbf{p}_0^{\text{in}}, \mathbf{p}_0^{\text{out}})$ and $(\overline{\mathbf{Z}}_{q0}^{\text{in}}, \overline{\mathbf{Z}}_{q0}^{\text{out}})$ are independent. Likewise, $(\widetilde{\mathbf{p}}_0^{\text{in}}, \widetilde{\mathbf{p}}_0^{\text{out}})$ and $(\overline{\mathbf{Z}}_{q0}^{\text{in}}, \overline{\mathbf{Z}}_{q0}^{\text{out}})$ are independent since the former only depends on $\overline{\mathbf{Z}}_{p0}^{\text{in}}$ and $\overline{\mathbf{Z}}_{p0}^{\text{out}}$. Thus, equality in distribution is maintained if we substitute $(\widetilde{\mathbf{p}}_0^{\text{in}}, \widetilde{\mathbf{p}}_0^{\text{out}})$ for $(\mathbf{p}_0^{\text{in}}, \mathbf{p}_0^{\text{out}})$ in the right side of the above by (B) in Fact 1. But this is just $(\widetilde{\mathbf{p}}_0^{\text{in}}, \widetilde{\mathbf{p}}_0^{\text{out}}, \widetilde{\mathbf{q}}_0^{\text{in}}, \widetilde{\mathbf{q}}_0^{\text{out}})$ by the definitions in the Gaussian Algorithm 5 and the fact that $[\mathbf{c}_{q0}^{\text{in/out}}]_0 = 1$. This completes the proof of the base case.

For the inductive step, assume that $\widetilde{\underline{\mathbf{r}}}_{k-1} \overset{d}{=} \underline{\mathbf{r}}_{k-1}$ for some $k \geq 2$. First we will establish that

$\left(\underline{\mathbf{r}}_{k-1}, \mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}\right) \stackrel{d}{=} \left(\widetilde{\underline{\mathbf{r}}}_{k-1}, \widetilde{\mathbf{p}}_k^{\text{in}}, \widetilde{\mathbf{p}}_k^{\text{out}}\right)$. To do this, observe that

$$\begin{aligned}
&\left(\underline{\mathbf{r}}_{k-1}, \mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}\right) \\
&\stackrel{d}{=} \left(\underline{\mathbf{r}}_{k-1}, \sum_{\ell=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{in}}]_{\ell+1}\, \mathbf{p}_\ell^{\text{in}} + \sqrt{\rho_{pk}^{\text{in}}}\, \mathbf{O}_{pk}^{\text{in}}\, \overline{\mathbf{Z}}_{pk}^{\text{in}} + \boldsymbol{\Delta}_{pk}^{\text{in}}, \sum_{\ell=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{out}}]_{\ell+1}\, \mathbf{p}_\ell^{\text{out}} + \sqrt{\rho_{pk}^{\text{out}}}\, \mathbf{O}_{pk}^{\text{out}}\, \overline{\mathbf{Z}}_{pk}^{\text{out}} + \boldsymbol{\Delta}_{pk}^{\text{out}}\right) \qquad (2.61) \\
&\stackrel{d}{=} \left(\widetilde{\underline{\mathbf{r}}}_{k-1}, \sum_{\ell=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{in}}]_{\ell+1}\, \widetilde{\mathbf{p}}_\ell^{\text{in}} + \sqrt{\rho_{pk}^{\text{in}}}\, \widetilde{\mathbf{O}}_{pk}^{\text{in}}\, \overline{\mathbf{Z}}_{pk}^{\text{in}} + \widetilde{\boldsymbol{\Delta}}_{pk}^{\text{in}}, \sum_{\ell=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{out}}]_{\ell+1}\, \widetilde{\mathbf{p}}_\ell^{\text{out}} + \sqrt{\rho_{pk}^{\text{out}}}\, \widetilde{\mathbf{O}}_{pk}^{\text{out}}\, \overline{\mathbf{Z}}_{pk}^{\text{out}} + \widetilde{\boldsymbol{\Delta}}_{pk}^{\text{out}}\right), \qquad (2.62)
\end{aligned}$$

where (2.61) follows from (A) in Fact 1 (using that conditioning on $\underline{\mathbf{r}}_{k-1}$ is equivalent to conditioning on $\mathcal{P}_{k-1}$ because the additional $\mathbf{u}$ and $\mathbf{v}$ terms that generate the latter are deterministic functions of $\underline{\mathbf{r}}_{k-1}$) and the conclusions of Lemma 3 (particularly, that $\mathbf{p}_k^{\text{in}}$ and $\mathbf{p}_k^{\text{out}}$ are independent given $\underline{\mathbf{r}}_{k-1}$), and where (2.62) follows from substituting $(\widetilde{\mathbf{p}}_0^{\text{in/out}}, \ldots, \widetilde{\mathbf{p}}_{k-1}^{\text{in/out}})$ for $(\mathbf{p}_0^{\text{in/out}}, \ldots, \mathbf{p}_{k-1}^{\text{in/out}})$. This substitution is justified by (B) in Fact 1 and the induction hypothesis as $\left(\overline{\mathbf{Z}}_{pk}^{\text{in}}, \overline{\mathbf{Z}}_{pk}^{\text{out}}\right)$ is independent of $(\mathbf{p}_0^{\text{in/out}}, \ldots, \mathbf{p}_{k-1}^{\text{in/out}}) \subset \mathcal{P}_{k-1}$ by construction, and independent of $\widetilde{\underline{\mathbf{r}}}_{k-1}$ as $\widetilde{\underline{\mathbf{r}}}_{k-1}$ only depends on

$$\left(\overline{\mathbf{Z}}_{p0}^{\text{in/out}}, \ldots \overline{\mathbf{Z}}_{p(k-1)}^{\text{in/out}}, \overline{\mathbf{Z}}_{q0}^{\text{in/out}}, \ldots, \overline{\mathbf{Z}}_{q(k-1)}^{\text{in/out}}\right).$$

Next observe that

$$\begin{aligned}
&\sum_{\ell=0}^{k-1}[\boldsymbol{\beta}_{pk}^{\text{in}}]_{\ell+1}\,\widetilde{\mathbf{p}}_{\ell}^{\text{in}}+\sqrt{\rho_{pk}^{\text{in}}}\,\widetilde{\mathbf{O}}_{pk}^{\text{in}}\,\overline{\mathbf{Z}}_{pk}^{\text{in}}+\widetilde{\boldsymbol{\Delta}}_{pk}^{\text{in}}\\
&\quad=\sum_{\ell=0}^{k-1}[\boldsymbol{\beta}_{pk}^{\text{in}}]_{\ell+1}\left[\sum_{j=0}^{\ell}[\mathbf{c}_{p\ell}^{\text{in}}]_j\left(\sqrt{\rho_{pj}^{\text{in}}}\,\widetilde{\mathbf{O}}_{pj}^{\text{in}}\,\overline{\mathbf{Z}}_{pj}^{\text{in}}+\widetilde{\boldsymbol{\Delta}}_{pj}^{\text{in}}\right)\right]+\sqrt{\rho_{pk}^{\text{in}}}\,\widetilde{\mathbf{O}}_{pk}^{\text{in}}\,\overline{\mathbf{Z}}_{pk}^{\text{in}}+\widetilde{\boldsymbol{\Delta}}_{pk}^{\text{in}} \qquad (2.63)\\
&\quad=\sum_{j=0}^{k-1}\left[\sum_{\ell=j}^{k-1}\left[\boldsymbol{\beta}_{pk}^{\text{in}}\right]_{\ell+1}[\mathbf{c}_{p\ell}^{\text{in}}]_j\right]\left(\sqrt{\rho_{pj}^{\text{in}}}\widetilde{\mathbf{O}}_{pj}^{\text{in}}\overline{\mathbf{Z}}_{pj}^{\text{in}}+\widetilde{\boldsymbol{\Delta}}_{pj}^{\text{in}}\right)+\sqrt{\rho_{pk}^{\text{in}}}\,\widetilde{\mathbf{O}}_{pk}^{\text{in}}\,\overline{\mathbf{Z}}_{pk}^{\text{in}}+\widetilde{\boldsymbol{\Delta}}_{pk}^{\text{in}}\\
&\quad=\sum_{j=0}^{k}[\mathbf{c}_{pk}^{\text{in}}]_j\sqrt{\rho_{pj}^{\text{in}}}\widetilde{\mathbf{O}}_{pj}^{\text{in}}\overline{\mathbf{Z}}_{pj}+\sum_{j=0}^{k}[\mathbf{c}_{pk}^{\text{in}}]_j\widetilde{\boldsymbol{\Delta}}_{pj}^{\text{in}} \qquad (2.64)\\
&\quad=\widetilde{\mathbf{p}}_{k}^{\text{in}}, \qquad (2.65)
\end{aligned}$$

where (2.63) follows from the definitions of $\widetilde{\mathbf{p}}_j^{\text{in}}$ in Algorithm 5, (2.64) follows from the definitions of the constants $[\mathbf{c}_{pk}^{\text{in}}]_j$, and (2.65) again follows from the definition of $\widetilde{\mathbf{p}}_k^{\text{in}}$ in Algorithm 5. The same reasoning applies to $\widetilde{\mathbf{p}}_k^{\text{out}}$.

Thus, as claimed, plugging these into the above gives $\left(\underline{\mathbf{r}}_{k-1},\mathbf{p}_k^{\text{in}},\mathbf{p}_k^{\text{out}}\right)\overset{d}{=}\left(\widetilde{\underline{\mathbf{r}}}_{k-1},\widetilde{\mathbf{p}}_k^{\text{in}},\widetilde{\mathbf{p}}_k^{\text{out}}\right)$. Next we need to extend this to the $\left(\mathbf{q}_k^{\text{in}},\mathbf{q}_k^{\text{out}}\right)$ variables. The proof of this works in exactly the same way as above, conditioning now on $(\underline{\mathbf{r}}_{k-1},\mathbf{p}_k^{\text{in}},\mathbf{p}_k^{\text{out}})$ and applying Lemma 2 with (A) and (B) above to get the proper joint distributional equality.

For the third part, we claim $\left(\widetilde{\mathbf{O}}_{p0}^{\text{in}}\overline{\mathbf{Z}}_{p0}^{\text{in}},\widetilde{\mathbf{O}}_{q0}^{\text{in}}\overline{\mathbf{Z}}_{q0}^{\text{in}},\ldots,\widetilde{\mathbf{O}}_{pk}^{\text{in}}\overline{\mathbf{Z}}_{pk}^{\text{in}},\widetilde{\mathbf{O}}_{qk}^{\text{in}}\overline{\mathbf{Z}}_{qk}^{\text{in}}\right)\overset{d}{=}\left(\overline{\mathbf{Z}}_{p0}^{\text{in}},\overline{\mathbf{Z}}_{q0}^{\text{in}},\ldots,\overline{\mathbf{Z}}_{pk}^{\text{in}},\overline{\mathbf{Z}}_{qk}^{\text{in}}\right)$. As a first step, we show that the two above vectors generate the same σ-algebras. For this it suffices to show that the right is a function of the left. (The reverse is immediate since the $\widetilde{\mathbf{O}}$ matrices (which are equal to their $\mathbf{O}$ equivalents defined in (2.39)) are functions of the $\{\widetilde{\mathbf{p}}_i,\widetilde{\mathbf{q}}_i\}_{i=1}^{k-1}$, and these are functions of the right hand side.) We can prove this inductively. For the first component, this is obvious since $\widetilde{\mathbf{O}}_{p0}^{\text{in}}=\mathbf{O}_{p0}^{\text{in}}$ is a deterministic orthogonal matrix. Now suppose the statement is true for all components up to $\widetilde{\mathbf{O}}_{(p/q)r}^{\text{in}}\overline{\mathbf{Z}}_{(p/q)r}^{\text{in}}$. Since $\widetilde{\mathbf{O}}_{(p/q)r}^{\text{in}}$ is a function of the $\overline{\mathbf{Z}}_{(p/q)j}^{\text{in}}$ for $j<r$, it follows by the induction hypothesis that $[\widetilde{\mathbf{O}}_{(p/q)r}^{\text{in}}]^{-1}$ is a function of the prior $\widetilde{\mathbf{O}}_{(p/q)j}^{\text{in}}\overline{\mathbf{Z}}_{(p/q)j}^{\text{in}}$ components. But then $\overline{\mathbf{Z}}_{(p/q)r}^{\text{in}}=[\widetilde{\mathbf{O}}_{(p/q)r}^{\text{in}}]^{-1}\widetilde{\mathbf{O}}_{(p/q)r}^{\text{in}}\overline{\mathbf{Z}}_{(p/q)r}^{\text{in}}$ can be written as a function of the

components of the left hand vector up through $\widetilde{\mathbf{O}}^{\text{in}}_{(p/q)r}\overline{\mathbf{Z}}^{\text{in}}_{(p/q)r}$, as claimed.

Now we can prove the asserted distributional equality inductively using (A) in Fact 1 by establishing distributional equality for each component conditional on the prior components. For the base case, again we know that $\widetilde{\mathbf{O}}^{\text{in}}_{p0}$ is a deterministic orthogonal matrix, so by standard properties of Gaussian random vectors we have that $\widetilde{\mathbf{O}}^{\text{in}}_{p0}\overline{\mathbf{Z}}^{\text{in}}_{p0} \overset{d}{=} \overline{\mathbf{Z}}_{p0}$. Next suppose that the statement is true up through $\widetilde{\mathbf{O}}^{\text{in}}_{(p/q)(r-1)}\overline{\mathbf{Z}}^{\text{in}}_{(p/q)(r-1)}$. Then conditioning on

$$\left(\widetilde{\mathbf{O}}^{\text{in}}_{p0}\overline{\mathbf{Z}}^{\text{in}}_{p0}, \widetilde{\mathbf{O}}^{\text{in}}_{q0}\overline{\mathbf{Z}}^{\text{in}}_{q0}, \ldots, \widetilde{\mathbf{O}}^{\text{in}}_{p(r-1)}\overline{\mathbf{Z}}^{\text{in}}_{p(r-1)}, \widetilde{\mathbf{O}}^{\text{in}}_{q(r-1)}\overline{\mathbf{Z}}^{\text{in}}_{q(r-1)}\right),$$

is equivalent to conditioning on

$$\left(\overline{\mathbf{Z}}^{\text{in}}_{p0}, \overline{\mathbf{Z}}^{\text{in}}_{q0}, \ldots, \overline{\mathbf{Z}}^{\text{in}}_{p(r-1)}, \overline{\mathbf{Z}}^{\text{in}}_{q(r-1)}\right), \tag{2.66}$$

by the induction hypothesis. But, conditional on the above, $\widetilde{\mathbf{O}}^{\text{in}}_{(p/q)r}$ is a deterministic orthogonal matrix (as it depends only on $\{\widetilde{\mathbf{p}}_i, \widetilde{\mathbf{q}}_i\}_{i=1}^{r-1}$, which are functions of (2.66)). Together, this gives that $\widetilde{\mathbf{O}}^{\text{in}}_{pr}\overline{\mathbf{Z}}^{\text{in}}_{pr} \mid_{\{\widetilde{\mathbf{O}}^{\text{in}}_{pi}\overline{\mathbf{Z}}^{\text{in}}_{pi}\}_{i=0}^{r-1}} \overset{d}{=} \widetilde{\mathbf{O}}^{\text{in}}_{pr}\overline{\mathbf{Z}}^{\text{in}}_{pr} \mid_{\{\overline{\mathbf{Z}}^{\text{in}}_{pi}\}_{i=0}^{r-1}} \overset{d}{=} \overline{\mathbf{Z}}^{\text{in}}_{pr}$. Putting this together, we find

$$\left(\widetilde{\mathbf{O}}^{\text{in}}_{p0}\overline{\mathbf{Z}}^{\text{in}}_{p0}, \widetilde{\mathbf{O}}^{\text{in}}_{q0}\overline{\mathbf{Z}}^{\text{in}}_{q0}, \ldots, \widetilde{\mathbf{O}}^{\text{in}}_{pr}\overline{\mathbf{Z}}^{\text{in}}_{pr}\right) \overset{d}{=} \left(\widetilde{\mathbf{O}}^{\text{in}}_{p0}\overline{\mathbf{Z}}^{\text{in}}_{p0}, \widetilde{\mathbf{O}}^{\text{in}}_{q0}\overline{\mathbf{Z}}^{\text{in}}_{q0}, \ldots, \overline{\mathbf{Z}}^{\text{in}}_{pr}\right) \overset{d}{=} \left(\overline{\mathbf{Z}}^{\text{in}}_{p0}, \overline{\mathbf{Z}}^{\text{in}}_{q0}, \ldots, \overline{\mathbf{Z}}^{\text{in}}_{pr}\right),$$

where the first equality follows from the above and (A) in Fact 1 and the second because $\overline{\mathbf{Z}}^{\text{in}}_{pr}$ is independent of $\left(\overline{\mathbf{Z}}^{\text{in}}_{p0}, \overline{\mathbf{Z}}^{\text{in}}_{q0}, \ldots, \overline{\mathbf{Z}}^{\text{in}}_{q(r-1)}, \overline{\mathbf{Z}}^{\text{in}}_{p(r-1)}\right)$ using (B) in Fact 1. The same distributional equality follows in the same way for the vectors of output variables.

Now we can proceed with the proof of the statement in part (iii). For the base case of $k = 0$, we have $\overset{*}{\mathbf{p}}{}^{\text{in}}_0 = \sqrt{\rho^{\text{in}}_{p0}}\,\widetilde{\mathbf{O}}^{\text{in}}_{p0}\,\overline{\mathbf{Z}}^{\text{in}}_{p0}$ and $\overset{*}{\mathbf{p}}{}^{\text{out}}_0 = \sqrt{\rho^{\text{out}}_{p0}}\,\widetilde{\mathbf{O}}^{\text{out}}_{p0}\,\overline{\mathbf{Z}}^{\text{out}}_{p0}$. Because the $\widetilde{\mathbf{O}}^{\text{in/out}}_{p0} = \mathbf{O}^{\text{in/out}}_{p0}$ are deterministic orthogonal matrices, we have that

$$\overset{*}{\mathbf{p}}{}^{\text{in}}_0 = \sqrt{\rho^{\text{in}}_{p0}}\,\widetilde{\mathbf{O}}^{\text{in}}_{p0}\,\overline{\mathbf{Z}}^{\text{in}}_{p0} = \sqrt{\rho^{\text{in}}_{p0}}\,\mathbf{O}^{\text{in}}_{p0}\,\overline{\mathbf{Z}}^{\text{in}}_{p0} \overset{d}{=} \sqrt{\rho^{\text{in}}_{p0}}\,\overline{\mathbf{Z}}^{\text{in}}_{p0} \overset{i.i.d.}{\sim} N(0, \sqrt{\rho^{\text{in}}_{p0}}),$$

and similarly for the output case. The independence of $\overset{*}{\mathbf{p}}_0^{\text{in}}$ and $\overset{*}{\mathbf{p}}_0^{\text{out}}$ follows because $\overline{\mathbf{Z}}_{p0}^{\text{in}}$ and $\overline{\mathbf{Z}}_{p0}^{\text{out}}$ are independent by definition.

To prove the induction step, assume the statement holds for independent Gaussian vectors $(P_0^{\text{in}}, \dots, P_{k-1}^{\text{in}})$ and $(P_0^{\text{out}}, \dots, P_{k-1}^{\text{out}})$ (whose values are given in (2.60)). We first notice that the induction hypothesis along with the definition of $\overset{*}{\mathbf{p}}_k^{\text{in/out}}$ in (2.54) and the above implies that

$$
\begin{aligned}
&\left([\overset{*}{\mathbf{p}}_0^{\text{in}}]_i, [\overset{*}{\mathbf{p}}_0^{\text{out}}]_i, \dots, [\overset{*}{\mathbf{p}}_k^{\text{in}}]_i, [\overset{*}{\mathbf{p}}_k^{\text{out}}]_i \right) \\
&= \left([\overset{*}{\mathbf{p}}_0^{\text{in}}]_i, [\overset{*}{\mathbf{p}}_0^{\text{out}}]_i, \dots, \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{in}}]_r [\overset{*}{\mathbf{p}}_r^{\text{in}}]_i + \sqrt{\rho_{pk}^{\text{in}}}\, [\widetilde{\mathbf{O}}_{pk}^{\text{in}}\, \overline{\mathbf{Z}}_{pk}^{\text{in}}]_i, \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}^{\text{out}}]_r [\overset{*}{\mathbf{p}}_r^{\text{out}}]_i + \sqrt{\rho_{pk}^{\text{out}}}\, [\widetilde{\mathbf{O}}_{pk}^{\text{out}}\, \overline{\mathbf{Z}}_{pk}^{\text{out}}]_i \right) \\
&\overset{d}{=} \left(P_0^{\text{in}}, P_0^{\text{out}}, \dots, \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}]_r P_r^{\text{in}} + \sqrt{\rho_{pk}^{\text{in}}}\, [\widetilde{\mathbf{O}}_{pk}^{\text{in}}\, \overline{\mathbf{Z}}_{pk}^{\text{in}}]_i, \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}]_r P_r^{\text{out}} + \sqrt{\rho_{pk}^{\text{out}}}\, [\widetilde{\mathbf{O}}_{pk}^{\text{out}}\, \overline{\mathbf{Z}}_{pk}^{\text{out}}]_i \right) \\
&\overset{d}{=} \left(P_0^{\text{in}}, P_0^{\text{out}}, \dots, \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}]_r P_r^{\text{in}} + \sqrt{\rho_{pk}^{\text{in}}}\, [\overline{\mathbf{Z}}_{pk}^{\text{in}}]_i, \sum_{r=0}^{k-1} [\boldsymbol{\beta}_{pk}]_r P_r^{\text{out}} + \sqrt{\rho_{pk}^{\text{out}}}\, [\overline{\mathbf{Z}}_{pk}^{\text{out}}]_i \right),
\end{aligned}
$$

where, on the right-hand side, we treat $\overline{\mathbf{Z}}_{pk}^{\text{in/out}}$ as independent of the $P_0^{\text{in/out}}, \dots, P_{k-1}^{\text{in/out}}$ (as the $\overset{*}{\mathbf{p}}_i^{\text{in/out}}$ are independent of $\overline{\mathbf{Z}}_{pk}^{\text{in/out}}$ for $i < k$, which follows directly from the definition of the ideal variables $\overset{*}{\mathbf{p}}_i^{\text{in}}$). In particular, this shows that the last expression is jointly Gaussian (e.g., by Cramer-Wold). The zero-mean property and the independence of the input and output variables follows immediately from the definition of the $\overset{*}{\mathbf{p}}$ variables. Thus this completes the proof of the third part.

□

2.3.4 General state evolution

Next we define a scalar recursion that is the analog of the GVAMP state evolution in (2.30), but for the general algorithm in Algorithm 3. One consequence of our general concentration result, Lemma 5, is that the iterates $\mathbf{p}_k^{\text{in/out}}$ and $\mathbf{q}_k^{\text{in/out}}$ converge empirically to the zero-mean Gaussian variables $P_k^{\text{in/out}}$ and $Q_k^{\text{in/out}}$, respectively, with both defined in (2.60). The variances of these Gaussian variables are completely characterized by the state evolution equations in (2.67). The

general algorithm state evolution is given by the following recursive system for $k \geq 0$:

$$\begin{aligned}
\overline{\alpha}_{pk}^{\text{in/out}} &= \mathbb{E}\left\{\left[f_p^{\text{in/out}}\right]'\left(P_k^{\text{in}}, P_k^{\text{out}}, W^p, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}\right)\right\}, \qquad \overline{\gamma}_{qk}^{\text{in/out}} = \Gamma_{pk}^{\text{in/out}}\left(\overline{\gamma}_{pk}^{\text{in/out}}, \overline{\alpha}_{pk}^{\text{in/out}}\right), \\
\tau_{qk}^{\text{in/out}} &= \left(1 - \overline{\alpha}_{pk}^{\text{in/out}}\right)^{-2}\left[\mathbb{E}\left\{\left[f_p^{\text{in/out}}\right]^2\left(P_k^{\text{in}}, P_k^{\text{out}}, W^p, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}\right) - \left[\overline{\alpha}_{pk}^{\text{in/out}}\right]^2 \tau_{pk}^{\text{in/out}}\right\}\right], \\
\overline{\alpha}_{qk}^{\text{in/out}} &= \mathbb{E}\left\{\left[f_q^{\text{in/out}}\right]'\left(Q_k^{\text{in}}, Q_k^{\text{out}}, W^q, \overline{\gamma}_{qk}^{\text{in}}, \overline{\gamma}_{qk}^{\text{out}}\right)\right\}, \qquad \overline{\gamma}_{p(k+1)}^{\text{in/out}} = \Gamma_{qk}^{\text{in/out}}\left(\overline{\gamma}_{qk}^{\text{in/out}}, \overline{\alpha}_{qk}^{\text{in/out}}\right), \\
\tau_{p(k+1)}^{\text{in/out}} &= \left(1 - \overline{\alpha}_{qk}^{\text{in/out}}\right)^{-2}\left[\mathbb{E}\left\{\left[f_q^{\text{in/out}}\right]^2\left(Q_k^{\text{in}}, Q_k^{\text{out}}, W^q, \overline{\gamma}_{qk}^{\text{in}}, \overline{\gamma}_{qk}^{\text{out}}\right) - \left[\overline{\alpha}_{qk}^{\text{in/out}}\right]^2 \tau_{qk}^{\text{in/out}}\right\}\right].
\end{aligned} \tag{2.67}$$

The initial terms of this recursion $\overline{\gamma}_{p0}^{\text{in/out}}$ are well-defined by **Condition 0**, and the terms $\tau_0^{\text{in/out}}$ are initialized with $\rho_{p0}^{\text{in/out}}$, which we have thought of as a free parameter of Algorithm 5 and which we will define explicitly along with the other limiting quantities in the next section. The (components of the) random vector W^p must be understood relative to the assumptions made in **Condition 0**. Specifically, if $[\mathbf{w}^p]_n$ is assumed to be i.i.d. sub-Gaussian, then W^p is a random variable following this sub-Gaussian distribution. If instead $[\mathbf{w}^p]_i$ is taken to be deterministic, then we define W^p to be distributed as the empirical distribution of the components of $[\mathbf{w}^p]_n$. In this case, the expectation over these components reduce to a simple empirical average For instance, if all components of $\mathbf{w}^p$ are deterministic, then we have

$$\overline{\alpha}_{pk}^{\text{in/out}} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left\{\left[f_p^{\text{in/out}}\right]'\left(P_k^{\text{in}}, P_k^{\text{out}}, \mathbf{w}_n^p, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}\right)\right\},$$

where $\mathbf{w}_n^p$ is the n^{th} row of the matrix $\mathbf{w}^p$, and likewise for the other state evolution quantities. Clearly, this introduces a dependence on N in the state evolution when some components of $\mathbf{w}^p$ are deterministic. Under our assumptions in **Condition 0**, however, this N-dependent state evolution still converges as $N \to \infty$ to a limiting state evolution which is free of N.

In terms of these state evolution quantities, we define two more sequences of random variables.

For $k \geq 0$:

$$\begin{aligned} U_{k+1}^{\text{in/out}} &= \left(1 - \overline{\alpha}_{qk}^{\text{in/out}}\right)^{-1} \left[f_q^{\text{in/out}} \left(Q_k^{\text{in}}, Q_k^{\text{out}}, W^q, \overline{\gamma}_{qk}^{\text{in}}, \overline{\gamma}_{qk}^{\text{out}} \right) - \overline{\alpha}_{qk}^{\text{in/out}} Q_k^{\text{in/out}} \right], \\ V_k^{\text{in/out}} &= \left(1 - \overline{\alpha}_{pk}^{\text{in/out}}\right)^{-1} \left[f_p^{\text{in/out}} \left(P_k^{\text{in}}, P_k^{\text{out}}, W^p, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}} \right) - \overline{\alpha}_{pk}^{\text{in/out}} P_k^{\text{in/out}} \right], \end{aligned} \tag{2.68}$$

with $Q_k^{\text{in/out}}$ and $P_k^{\text{in/out}}$ defined above in (2.60), and where U_0 is the random variable determined by **Condition 0** to which the initial condition $\mathbf{u}_0$ concentrates.

Lemma 5 part (d) (which we prove in what follows) implies that $\|\mathbf{p}_k^{\text{in/out}}\|^2/N$ concentrates on $\mathbb{E}\{[P_k^{\text{in/out}}]^2\}$ and part (e) implies that $\|\mathbf{u}_k^{\text{in/out}}\|^2/N$ concentrates to $\mathbb{E}\{[U_k^{\text{in/out}}]^2\}$.

Following [77], we can use these results to see that

$$\begin{aligned} \mathbb{E}\{[P_k^{\text{in/out}}]^2\} = \lim_{N\to\infty} \frac{\|\mathbf{p}_k^{\text{in/out}}\|^2}{N} &= \lim_{N\to\infty} \frac{\|\mathbf{u}_k^{\text{in/out}}\|^2}{N} \\ &= \frac{\mathbb{E}\left[\left(f_q^{\text{in/out}} \left(Q_k^{\text{in}}, Q_k^{\text{out}}, W^q, \overline{\gamma}_{qk}^{\text{in}}, \overline{\gamma}_{qk}^{\text{out}} \right) - \overline{\alpha}^{\text{in/out}} Q_k^{\text{in/out}} \right)^2 \right]}{(1 - \overline{\alpha}_{qk}^{\text{in/out}})^2}, \end{aligned} \tag{2.69}$$

where the second equality in the above follows because $\mathbf{p}_k^{\text{in}} = \mathbf{V}^T \mathbf{u}_k^{\text{in}}$, by definition in Algorithm 3; hence, $\|\mathbf{p}_k^{\text{in/out}}\|^2 = (\mathbf{u}_k^{\text{in/out}})^T \mathbf{V}\mathbf{V}^T \mathbf{u}_k^{\text{in/out}} = \|\mathbf{u}_k^{\text{in/out}}\|^2$. Now expanding terms and applying Stein's lemma along with the state evolution definitions in (2.67) shows that indeed $\mathbb{E}\{[P_k^{\text{in/out}}]^2\} = \tau_{pk}^{\text{in/out}}$. The same reasoning can be applied to show that $\mathbb{E}\{[Q_k^{\text{in/out}}]^2\} = \tau_{qk}^{\text{in/out}}$.

We can also show, assuming again the result of Lemma 5, that the GVAMP state evolution (2.30) is equivalent to the general state evolution (2.67) under the translation defined in (2.31) - (2.35). The equivalences $\overline{\alpha}_{1k} = \overline{\alpha}_{pk}^{\text{in}}$ and $\overline{\beta}_{1k} = \overline{\alpha}_{pk}^{\text{out}}$ follow directly from the definitions of the sensitivity functions A_{x1} and A_{z1} and the definitions in (2.32). The equivalences $\sigma_{2k}^2 = \tau_{qk}^{\text{in}}$ and $\rho_{2k}^2 = \tau_{qk}^{\text{out}}$ follow similarly. Then $\overline{\gamma}_{2k} = \overline{\gamma}_{qk}^{\text{in}}$ and $\overline{\tau}_{2k} = \overline{\gamma}_{qk}^{\text{out}}$ follows directly from the definition of the $\Gamma_{pk}^{\text{in/out}}$ above.

The equivalences $\overline{\alpha}_{2k} = \overline{\alpha}_{qk}^{\text{in}}$ and $\overline{\beta}_{2k} = \overline{\alpha}_{qk}^{\text{out}}$ follow from as the definitions of the sensitivity functions A_{x2} and A_{z2} are the limits of the averages of the derivatives of f_q^{in} and f_q^{out} respectively

(again by the definitions in (2.32)), and these averages converge to $\overline{\alpha}_{qk}^{\text{in/out}}$ since we have empirical convergence of the arguments. (Indeed, in this case the derivatives do not depend on the $\mathbf{q}_k^{\text{in/out}}$, so we only need empirical convergence of the $\mathbf{w}_q^{\text{in/out}}$, which is guaranteed by **Condition 0**.)

Then, the equivalences $\sigma_{1k}^2 = \tau_{pk}^{\text{in}}$ and $\rho_{1k}^2 = \tau_{pk}^{\text{out}}$ are established by the definitions (2.32), the uniformly Lipschitz property of g_{x2} and g_{z2}, and the fact that the i.i.d. sequences $\mathbf{r}_2$ and $\mathbf{p}_2$ in the definition of the error functions converge empirically to the proper normal variables. Finally, the equivalences $\overline{\gamma}_{qk} = \overline{\gamma}_{pk}^{\text{in}}$ and $\overline{\tau}_{1k} = \overline{\gamma}_{pk}^{\text{out}}$ follow from the definitions of the $\Gamma_{qk}^{\text{in/out}}$ functions above.

2.3.5 Definitions of limiting quantities

The statements of lemmas 3 and 4 are valid for any choice of parameters $\boldsymbol{\beta}_{pk}^{\text{in/out}}$, $\boldsymbol{\beta}_{qk}^{\text{in/out}}$, $\rho_{pk}^{\text{in/out}}$, and $\rho_{qk}^{\text{in/out}}$, and the exact values of these parameters have yet to be specified. To make our analysis work, we choose these parameters to be the limiting values of the quantities that occur in the discrepancy terms $\boldsymbol{\Delta}$ defined in Lemma 3, so that these discrepancy terms concentrate around zero.

To formally define these parameters, we first introduce come intermediate quantities: for $k \geq 0$,

$$\begin{aligned}
\mathbf{b}_{uk}^{\text{in/out}} &= (\mathbb{E}[U_0^{\text{in/out}} U_k^{\text{in/out}}], \ldots, \mathbb{E}[U_{k-1}^{\text{in/out}} U_k^{\text{in/out}}]), \\
\mathbf{b}_{vk}^{\text{in/out}} &= (\mathbb{E}[V_0^{\text{in/out}} V_k^{\text{in/out}}], \ldots, \mathbb{E}[V_{k-1}^{\text{in/out}} V_k^{\text{in/out}}]),
\end{aligned} \tag{2.70}$$

and define matrices $\boldsymbol{\Sigma}_{uk}^{\text{in/out}}, \boldsymbol{\Sigma}_{vk}^{\text{in/out}} \in \mathbb{R}^{k \times k}$ elementwise as

$$\left[\boldsymbol{\Sigma}_{uk}^{\text{in/out}}\right]_{ij} = \mathbb{E}\left[U_i^{\text{in/out}} U_j^{\text{in/out}}\right], \qquad \left[\boldsymbol{\Sigma}_{vk}^{\text{in/out}}\right]_{ij} = \mathbb{E}\left[V_i^{\text{in/out}} V_j^{\text{in/out}}\right], \tag{2.71}$$

where the U and V variables are defined in (2.68). In terms of these quantities, we define our parameters as follows: for $k \geq 1$

$$\begin{aligned}
\boldsymbol{\beta}_{pk}^{\text{in/out}} &= [\boldsymbol{\Sigma}_{u(k-1)}^{\text{in/out}}]^{-1} \mathbf{b}_{uk}^{\text{in/out}}, \qquad & \rho_{pk}^{\text{in/out}} &= \mathbb{E}\{[U_k^{\text{in/out}}]^2\} - (\mathbf{b}_{uk}^{\text{in/out}})^T [\boldsymbol{\Sigma}_{u(k-1)}^{\text{in/out}}]^{-1} \mathbf{b}_{uk}^{\text{in/out}}, \\
\boldsymbol{\beta}_{qk}^{\text{in/out}} &= [\boldsymbol{\Sigma}_{v(k-1)}^{\text{in/out}}]^{-1} \mathbf{b}_{vk}^{\text{in/out}}, \qquad & \rho_{qk}^{\text{in/out}} &= \mathbb{E}\{[V_k^{\text{in/out}}]^2\} - (\mathbf{b}_{vk}^{\text{in/out}})^T [\boldsymbol{\Sigma}_{v(k-1)}^{\text{in/out}}]^{-1} \mathbf{b}_{vk}^{\text{in/out}}.
\end{aligned} \tag{2.72}$$

These parameters are initialized as $\boldsymbol{\beta}_{p0}^{\text{in/out}} = \boldsymbol{\beta}_{q0}^{\text{in/out}} = 0$, $\rho_{p0}^{\text{in/out}}$ given by **Condition 0**, and $\rho_{q0}^{\text{in/out}} = \mathbb{E}\left\{[V_0^{\text{in/out}}]^2\right\}$.

These definitions may appear circular at first, as the definitions of the P and Q variables in Lemma 4 depend on these parameters, which are then defined in terms of moments involving the P and Q. However, it is easy to check that the definitions can be made recursively, avoiding any circular dependence. For instance $U_0^{\text{in/out}}$ is determined by the initial condition $\mathbf{u}_0^{\text{in/out}}$. Since $\boldsymbol{\beta}_{p0}^{\text{in/out}} = 0$ and $\rho_{p0}^{\text{in/out}}$ is also determined by the initial conditions, P_0 is well-defined in terms of U_0. Then it is easy to check from the definition and state evolution equations that V_0 is well-defined in terms of P_0, and Q_0 is well-defined in terms of V_0. Continuing in this fashion allows us to recursively build up the sequences of P, Q, U, and V variables in terms of the previously defined variables.

Finally, we define limiting precision matrices. Let $\boldsymbol{\Pi}_{u0}^{\text{in/out}} = [\boldsymbol{\Sigma}_{u0}^{\text{in/out}}]^{-1}$ and for $k \geq 0$:

$$\boldsymbol{\Pi}_{uk}^{\text{in/out}} = \begin{bmatrix} [\boldsymbol{\Sigma}_{uk}^{\text{in/out}}]^{-1} & 0 \\ 0 & [\boldsymbol{\Sigma}_{v(k-1)}^{\text{in/out}}]^{-1} \end{bmatrix}, \qquad \boldsymbol{\Pi}_{vk}^{\text{in/out}} = \begin{bmatrix} [\boldsymbol{\Sigma}_{uk}^{\text{in/out}}]^{-1} & 0 \\ 0 & [\boldsymbol{\Sigma}_{vk}^{\text{in/out}}]^{-1} \end{bmatrix}, \tag{2.73}$$

Dependence of limiting quantities on N

We also note that, when one or more of the components of $\mathbf{w}^p$ or $\mathbf{w}^q$ is deterministic, the corresponding components of W^p or W^q will follow the corresponding empirical distribution. Consequently, in this case, these state evolution quantities and the limiting quantities defined above will also depend on the system size N. Because these quantities also affect the rate of convergence in our concentration arguments, it is important to note that they can be bounded independently of N, and that they therefore do not alter the overall (exponential) rate of concentration. We sketch here an inductive argument that exploits the recursive definitions (2.60) along with continuity to establish that the existence of an N-independent compact set containing the $\mathbf{w}^{(d)}$ implies that all limiting quantities are also contained within such N-independent compact sets.

For simplicity, we demonstrate these bounds in the case where all components of $W^{p/q}$ follow

an empirical distribution. First note that $\overline{\alpha}_{p0}^{\text{in/out}}$ depends on $\tau_{p0}^{\text{in/out}} = \rho_{p0}^{\text{in/out}}$, $\overline{\gamma}_{pk}^{\text{in/out}}$, and W^p. All but the last of these are independent of N. Since

$$g_{\alpha_p 0}^{\text{in/out}} : w \mapsto \mathbb{E}\left\{\left[f_p^{\text{in/out}}\right]'\left(P_0^{\text{in}}, P_0^{\text{out}}, w, \overline{\gamma}_{p0}^{\text{in}}, \overline{\gamma}_{p0}^{\text{out}}\right)\right\}$$

is continuous in w, and since the support of W^p is contained in a compact set $\mathcal{W}$ that is independent of N, we have that $\mathcal{W}_{\alpha_p 0}^{\text{in/out}} = g_{\alpha_p 0}^{\text{in/out}}(\mathcal{W})$ is compact and independent of N. Thus $\max g_{\alpha_p 0}^{\text{in/out}}(\mathcal{W})$ is an N-independent bound for $\overline{\alpha}_{p0}^{\text{in/out}}$.

For $\overline{\gamma}_{q0}^{\text{in/out}}$, we again have that

$$g_{\gamma_{q0}}^{\text{in/out}} : \alpha \mapsto \Gamma_{p0}^{\text{in/out}}\left(\overline{\gamma}_{p0}^{\text{in/out}}, \alpha\right)$$

is continuous, and since $\overline{\alpha}_{p0}^{\text{in/out}} \in \mathcal{W}_{\alpha_{p0}}^{\text{in/out}}$, we have that $\overline{\gamma}_{q0}^{\text{in/out}} \in g_{\gamma_{q0}}^{\text{in/out}}\left(\mathcal{W}_{\alpha_{p0}}^{\text{in/out}}\right)$, which is again compact and independent of N. Thus we obtain our bound as the maximum of this set.

For $\tau_{q0}^{\text{in/out}}$, we observe that

$$\left(1 - \overline{\alpha}_{pk}^{\text{in/out}}\right)^{-2}\left[\mathbb{E}\left\{\left[f_p^{\text{in/out}}\right]^2\left(P_k^{\text{in}}, P_k^{\text{out}}, W^p, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}\right) - \left[\overline{\alpha}_{pk}^{\text{in/out}}\right]^2 \tau_{pk}^{\text{in/out}}\right\}\right].$$

depends continuously on $\tau_{p0}^{\text{in/out}}$, $\overline{\gamma}_{p0}^{\text{in/out}}$, $\overline{\alpha}_{p0}^{\text{in/out}}$, and W^p. The first two of these are already independent of N, and the latter two are contained in the N-independent compact set $\mathcal{W}_{\alpha_{p0}}^{\text{in/out}} \times \mathcal{W}$. In particular we note that $\mathcal{W}_{\alpha_{p0}}^{\text{in/out}}$ excludes 0 by the clipping assumption (see **Condition 6**), so the above expression is continuous over this compact set. Thus, $\tau_{q0}^{\text{in/out}}$ is contained in a continuous image of this set, $\mathcal{W}_{\tau_{q0}}^{\text{in/out}}$, which is compact and N-independent, and so $\tau_{q0}^{\text{in/out}}$ has an N-independent upper bound.

Next we observe that $\rho_{q0}^{\text{in/out}} = \tau_{q0}^{\text{in/out}}$ and thus is also contained in compact set independent of N. From here, $\overline{\alpha}_{q0}^{\text{in/out}}$, $\overline{\gamma}_{q0}^{\text{in/out}}$, and $\tau_{p1}^{\text{in/out}}$ are analyzed exactly as above. The random variables $P_1^{\text{in/out}}$ are defined in terms of $\boldsymbol{\beta}_{p1}^{\text{in/out}}$ and $\rho_{p1}^{\text{in/out}}$, which are continuous functions of the expectations $\mathbb{E}[U_0^{\text{in/out}}]^2$, $\mathbb{E}[U_1^{\text{in/out}}]^2$, $\mathbb{E}U_0^{\text{in/out}}U_1^{\text{in/out}}$. (In particular, the stopping criteria ensure that the $\boldsymbol{\Sigma}_{pj}^{\text{in/out}}$

are always invertible, so matrix inversion is indeed a continuous operation over the range of the these matrices).

Since U_0 is determined by initial conditions, $\mathbb{E}[U_0^{\text{in/out}}]^2$ is independent of N. Next, $\mathbb{E}[U_1^{\text{in/out}}]^2$ and $\mathbb{E}U_0^{\text{in/out}}U_1^{\text{in/out}}$ can be expanded in terms of the definitions (2.68), which shows that these expectations depend continuously on $\overline{\alpha}_{q0}^{\text{in/out}}$, $\overline{\gamma}_{q0}^{\text{in/out}}$, $\tau_{q0}^{\text{in/out}}$, and $\mathbf{w}^q$. Since these have all been shown to be contained in compact sets independent of N, these expectations are also contained in such sets, and thus so are $\boldsymbol{\beta}_{p1}^{\text{in/out}}$ and $\rho_{p1}^{\text{in/out}}$.

Continuing inductively in this fashion, we can express each of the $V_j^{\text{in/out}}, U_{j+1}^{\text{in/out}}$ in terms of the corresponding $P_j^{\text{in/out}}, Q_j^{\text{in/out}}$ using definitions (2.68), and these can in turn be expanded using definition (2.60) to expressions depending recursively on the previous $\boldsymbol{\beta}$, ρ, α, γ, and τ quantities. Since the dependence remains continuous in every case, this establishes inductively that all of our state evolution and limiting quantities are contained within compact sets independent of N (and, in particular, are bounded independently of N).

Lemma 5. *Throughout the lemma, we use c and C to denote constants that do not depend on the iteration number, but which may vary in their exact value between usages. For $t \geq 0$, let*

$$C_k = C^{2k}(k!)^{16}, \qquad c_k = \frac{1}{c^{2k}(k!)^{26}}, \qquad C'_k = C^k(k+1)^8, \qquad c'_k = \frac{c_k}{c(k+1)^{15}}. \tag{2.74}$$

(a)

$$\mathbb{P}\left(\frac{1}{N}\|\boldsymbol{\Delta}_{qk}^{\text{in/out}}\|^2 \geq \epsilon\right) \leq Ck^2C'_{k-1}e^{-cc'_{k-1}n\epsilon/k^4}, \qquad \mathbb{P}\left(\frac{1}{N}\|\boldsymbol{\Delta}_{pk}^{\text{in/out}}\|^2 \geq \epsilon\right) \leq Ck^2C_{k-1}e^{-cc_{k-1}n\epsilon/k^4}.$$

(b) For $\phi : \mathbb{R}^{k+2} \to \mathbb{R}$ pseudo-Lipschitz, we have

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}\phi\left([\mathbf{q}_0^{\text{in}}]_i, ..., [\mathbf{q}_{k+1}^{\text{in}}]_i, [\mathbf{q}_0^{\text{out}}]_i, ..., [\mathbf{q}_{k+1}^{\text{out}}]_i, [\mathbf{w}_q^{\text{in}}]_i, [\mathbf{w}_q^{\text{out}}]_i\right) - \mathbb{E}\left\{\phi\left(Q_0^{\text{in}}, ..., Q_{k+1}^{\text{in}}, Q_0^{\text{out}}, ..., Q_{k+1}^{\text{out}}, W_q^{\text{in}}, W_q^{\text{out}}\right)\right\}\Big| \geq \epsilon\Big) \leq Ck^4C'_{k-1}e^{-cc'_{k-1}n\epsilon^2/k^7},$$

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}\phi\left([\mathbf{p}_0^{\text{in}}]_i, ..., [\mathbf{p}_{k+1}^{\text{in}}]_i, [\mathbf{p}_0^{\text{out}}]_i, ..., [\mathbf{p}_{k+1}^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i, [\mathbf{w}_p^{\text{out}}]_i\right) - \mathbb{E}\left\{\phi\left(P_0^{\text{in}}, ..., P_{k+1}^{\text{in}}, P_0^{\text{out}}, ..., P_{k+1}^{\text{out}}, W_p^{\text{in}}, W_p^{\text{out}}\right)\right\}\Big| \geq \epsilon\Big) \leq Ck^4C'_{k-1}e^{-cc'_{k-1}n\epsilon^2/k^7}.$$

where the $Q_i^{\text{in/out}}$ are jointly Gaussian with $Q_i^{\text{in/out}} \sim N(0, \tau_{qi}^{\text{in/out}})$ and $\mathbb{E}\{Q_i^{\text{in/out}}Q_j^{\text{in/out}}\} = [\mathbf{\Sigma}_{v,k+1}^{\text{in/out}}]_{ij}$, the $P_i^{\text{in/out}}$ are jointly Gaussian with $P_i^{\text{in/out}} \sim N(0, \tau_{pi}^{\text{in/out}})$ and $\mathbb{E}\{P_i^{\text{in/out}}P_j^{\text{in/out}}\} = [\mathbf{\Sigma}_{uk}^{\text{in/out}}]_{ij}$.

We will use Lemma 5 to prove Theorem 1; details in Section 2.3.6. Because the proof of Lemma 5 is quite technical, in the main body of the text, we give a high level discussion in Section 2.3.7 and relegate the details to Appendix A.1. In particular, Section 2.3.7 discusses the innovation required to prove Lemma 5, and more generally, describes differences between the finite sample analyses for generalized AMP algorithms and for standard AMP algorithms, as in [81, Lemma 5].

2.3.6 Theorem 1 proof

Theorem 1 follows from Lemma 5 part **(b)**. In translating GVAMP (Algorithm 2) to the general iteration (Algorithm 3), recall that we set $\mathbf{p}_k^{\text{in}} = \mathbf{r}_{1k} - \mathbf{x}_0$ and $\gamma_{pk}^{\text{in}} = \gamma_{1k}$ in (2.31). Along with the definition of $\hat{\mathbf{x}}_{1k}$ in Algorithm 2, namely $\hat{\mathbf{x}}_{1k} = g_{x1}(\mathbf{r}_{1k}, \gamma_{1k})$, it follows that

$$\phi([\hat{\mathbf{x}}_{10}]_i, \ldots, [\hat{\mathbf{x}}_{1k}]_i, [\mathbf{x}_0]_i) = \phi(g_{x1}([\mathbf{p}_0^{\text{in}} + \mathbf{x}_0]_i, \gamma_{p0}^{\text{in}}), \ldots, g_{x1}([\mathbf{p}_k^{\text{in}} + \mathbf{x}_0]_i, \gamma_{pk}^{\text{in}}), [\mathbf{x}_0]_i),$$

where now the right-hand side quantities are all from Algorithm 3.

Hence, we can bound the probability in (2.27) by introducing the limiting state evolution quan-

tities $\overline{\gamma}^{\text{in}}_{pk}$ defined in (2.67) as follows:

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}\phi([\hat{\mathbf{x}}_{10}]_i,\dots,[\hat{\mathbf{x}}_{1k}]_i,[\mathbf{x}_0]_i)-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{\phi([\hat{X}_{1k}]_i,[\mathbf{x}_0]_i)\}\Big|\geq\epsilon\Big)\\
&\quad=\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}\phi(g_{x1}([\mathbf{p}_0^{\text{in}}+\mathbf{x}_0]_i,\gamma^{\text{in}}_{p0}),\dots,g_{x1}([\mathbf{p}_k^{\text{in}}+\mathbf{x}_0]_i,\gamma^{\text{in}}_{pk}),[\mathbf{x}_0]_i)-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{\phi([\hat{X}_{1k}]_i,[\mathbf{x}_0]_i)\}\Big|\geq\epsilon\Big)\\
&\quad\leq\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N}\Big|\phi(g_{x1}([\mathbf{p}_0^{\text{in}}+\mathbf{x}_0]_i,\gamma^{\text{in}}_{p0}),\dots,g_{x1}([\mathbf{p}_k^{\text{in}}+\mathbf{x}_0]_i,\gamma^{\text{in}}_{pk}),[\mathbf{x}_0]_i)\\
&\qquad\qquad-\phi(g_{x1}([\mathbf{p}_0^{\text{in}}+\mathbf{x}_0]_i,\overline{\gamma}^{\text{in}}_{p0}),\dots,g_{x1}([\mathbf{p}_k^{\text{in}}+\mathbf{x}_0]_i,\overline{\gamma}^{\text{in}}_{pk}),[\mathbf{x}_0]_i)\Big|\geq\frac{\epsilon}{2}\Big) \quad (2.75)\\
&\qquad+\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}\phi(g_{x1}([\mathbf{p}_0^{\text{in}}+\mathbf{x}_0]_i,\overline{\gamma}^{\text{in}}_{p0}),\dots,g_{x1}([\mathbf{p}_k^{\text{in}}+\mathbf{x}_0]_i,\overline{\gamma}^{\text{in}}_{pk}),[\mathbf{x}_0]_i)-\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\{\phi([\hat{X}_{1k}]_i,[\mathbf{x}_0]_i)\}\Big|\geq \quad (2.76)
\end{aligned}$$

Now we consider the two terms in (2.75) and (2.76) separately and for both we give an upper bound of $CC_k e^{-cc_k N\epsilon^2}$ to establish the desired result. First, we notice that bound on (2.76) follows from Lemma 5 part (b). Indeed, because the $g_{x1}(\cdot,\overline{\gamma}^{\text{in}}_{pj})$ for $0\leq j\leq k$ are Lipschitz, it is easy to verify that the composition $(p_0,\dots,p_k,w^p)\mapsto\phi(g_{x1}(p_0+w^p,\overline{\gamma}^{\text{in}}_{p0}),\dots,g_{x1}(p_k+w^p,\overline{\gamma}^{\text{in}}_{pk}),w^p)$ is pseudo-Lipschitz, and because $[\mathbf{w}^p]_1=\mathbf{x}_0$, we can apply Lemma 5 part (b) to conclude that $\frac{1}{N}\sum_{i=1}^{N}\phi(g_{x1}([\mathbf{p}_0^{\text{in}}+[\mathbf{w}_p^{\text{in}}]_1]_i,,\overline{\gamma}^{\text{in}}_{p0})\dots,g_{x1}([\mathbf{p}_k^{\text{in}}+[\mathbf{w}_p^{\text{in}}]_1]_i,\overline{\gamma}^{\text{in}}_{pk}),[\mathbf{w}_p^{\text{in}}]_1)$ concentrates to

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\phi\left(g_{x1}\left(P_0^{\text{in}}+[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}^{\text{in}}_{p0}\right),\dots,g_{x1}\left(P_k^{\text{in}}+[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}^{\text{in}}_{pk}\right),[\mathbf{w}_p^{\text{in}}]_i\right)\right]\\
&\quad=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\phi\left(g_{x1}\left(\sqrt{\tau^{\text{in}}_{p0}}Z+[\mathbf{x}_0]_i,\overline{\gamma}^{\text{in}}_{p0}\right),\dots,g_{x1}\left(\sqrt{\tau^{\text{in}}_{pk}}Z+[\mathbf{x}_0]_i,\overline{\gamma}^{\text{in}}_{pk}\right),[\mathbf{x}_0]_i\right)\right]\\
&\quad=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\phi\left([\hat{X}_{10}]_i,\dots,[\hat{X}_{1k}]_i,[\mathbf{x}_0]_i\right)\right].
\end{aligned}$$

In the final equality above we have used that for $0\leq j\leq k$, $[\hat{X}_{1j}]_i=g_1([\mathbf{x}_0]_i+R_{1j},\bar{\gamma}_{1j})$ with $R_{1j}=\sigma_{1k}Z$ for $Z\sim\mathcal{N}(0,1)$ independent of $\mathbf{x}_0$, in addition to the facts that $\tau^{\text{in}}_{pi}=\sigma^2_{1k}$, which is discussed when introducing the general state evolution in Section 2.3.4 and that $\gamma^{\text{in}}_{pk}=\gamma_{1k}$; hence, $\overline{\gamma}^{\text{in}}_{pk}=\overline{\gamma}_{1k}$. Now we will upper bound the probability in (2.75). First, notice that by the pseudo-

Lipschitz property of ϕ, we have for $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^d$ that $|\phi(\mathbf{a}_1, b) - \phi(\mathbf{a}_2, b)| \leq L(1 + \|(\mathbf{a}_1, b)\| + \|(\mathbf{a}_2, b)\|)\|\mathbf{a}_1 - \mathbf{a}_2\|$. Moreover, by the Triangle Inequality:

$$\begin{aligned}\|(\mathbf{a}_1, b)\| + \|(\mathbf{a}_2, b)\| = \|(\mathbf{a}_1, b) - (\mathbf{a}_2, b) + (\mathbf{a}_2, b)\| + \|(\mathbf{a}_2, b)\| &\leq \|(\mathbf{a}_1, b) - (\mathbf{a}_2, b)\| + 2\|(\mathbf{a}_2, b)\| \\ &= \|\mathbf{a}_1 - \mathbf{a}_2\| + 2\|(\mathbf{a}_2, b)\| \\ &\leq \sum_{i=1}^{d} |[\mathbf{a}_1]_i - [\mathbf{a}_2]_i| + 2\|(\mathbf{a}_2, b)\|.\end{aligned}$$

Letting $\mathbf{g}_i = \left(g_{x1}\left([\mathbf{p}_0^{\text{in}} + \mathbf{x}_0]_i, \overline{\gamma}_{p0}^{\text{in}}\right), \ldots, g_{x1}\left([\mathbf{p}_k^{\text{in}} + \mathbf{x}_0]_i, \overline{\gamma}_{pk}^{\text{in}}\right)\right)$, we have the following bound.

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\Big|\phi(g_{x1}([\mathbf{p}_0^{\text{in}} + \mathbf{x}_0]_i, \gamma_{p0}^{\text{in}}), \ldots, g_{x1}([\mathbf{p}_k^{\text{in}} + \mathbf{x}_0]_i, \gamma_{pk}^{\text{in}}), [\mathbf{x}_0]_i) \\ &\qquad - \phi(g_{x1}([\mathbf{p}_0^{\text{in}} + \mathbf{x}_0]_i, \overline{\gamma}_{p0}^{\text{in}}), \ldots, g_{x1}([\mathbf{p}_k^{\text{in}} + \mathbf{x}_0]_i, \overline{\gamma}_{pk}^{\text{in}}), [\mathbf{x}_0]_i)\Big| \\ &\leq \frac{L}{N}\sum_{i=1}^{N}\left[1 + \sum_{j=0}^{k}\left|g_{x1}([\mathbf{p}_j^{\text{in}} + \mathbf{x}_0]_i, \gamma_{pj}^{\text{in}}) - g_{x1}([\mathbf{p}_j^{\text{in}} + \mathbf{x}_0]_i, \overline{\gamma}_{pj}^{\text{in}})\right| + 2\left\|(\mathbf{g}_i, [\mathbf{x}_0]_i)\right\|\right] \\ &\qquad \times \sum_{j=0}^{k}\left|g_{x1}\left([\mathbf{p}_j^{\text{in}} + \mathbf{x}_0]_i, \gamma_{pj}^{\text{in}}\right) - g_{x1}\left([\mathbf{p}_j^{\text{in}} + \mathbf{x}_0]_i, \overline{\gamma}_{pj}^{\text{in}}\right)\right| \\ &\leq L(k+3)\sqrt{1 + \frac{1}{N}\sum_{j=0}^{k}\|g_{x1}\left(\mathbf{p}_j^{\text{in}} + \mathbf{x}_0, \gamma_{pj}^{\text{in}}\right) - g_{x1}\left(\mathbf{p}_j^{\text{in}} + \mathbf{x}_0, \overline{\gamma}_{pj}^{\text{in}}\right)\|^2 + \frac{4}{N}\left[\|\mathbf{x}_0\|^2 + \sum_{j=0}^{k}\|[\mathbf{g}]_j\|^2\right]} \\ &\qquad \times \sqrt{\frac{1}{N}\sum_{j=0}^{k}\|g_{x1}\left(\mathbf{p}_j^{\text{in}} + \mathbf{x}_0, \gamma_{pj}^{\text{in}}\right) - g_{x1}\left(\mathbf{p}_j^{\text{in}} + \mathbf{x}_0, \overline{\gamma}_{pj}^{\text{in}}\right)\|^2},\end{aligned} \tag{2.77}$$

where the final inequality follows by Cauchy-Schwarz and Lemma 16. To see that (2.77) concentrates around zero at the correct rate, notice that it suffices to show

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} g_{x1}([\mathbf{p}_j^{\text{in}} + \mathbf{x}_0]_i, \overline{\gamma}_{pj}^{\text{in}})^2 - c_g\right| \geq \epsilon\right) \leq CC_k e^{-cc_k N\epsilon^2}, \tag{2.78}$$

for some universal constant $c_g > 0$, and

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\left|g_{x1}\left([\mathbf{p}_j^{\text{in}}+\mathbf{x}_0]_i,\gamma_{pj}^{\text{in}}\right)-g_{x1}\left([\mathbf{p}_j^{\text{in}}+\mathbf{x}_0]_i,\overline{\gamma}_{pj}^{\text{in}}\right)\right|^2\geq\epsilon^2\right)\leq CC_k e^{-cc_kN\epsilon^2}. \tag{2.79}$$

for all $0 \leq j \leq k$. Indeed if these concentration inequalities hold, then along with either the concentration of $\mathbf{x}_0$ or component-wise boundedness (see **Assumption 0**), we can apply Lemmas 9, 12, and 11 to (2.77) give concentration with the right rates in terms of N and ϵ.

For (2.78), observe that $g_{x1}(\cdot,\overline{\gamma}_{pj}^{\text{in}})$ is uniformly Lipschitz by **Assumption 3**. Thus, the composition, $(p,x)\mapsto\|(g_{x1}(p+x,\overline{\gamma}_{pj}^{\text{in}}),x)\|^2$ is pseudo-Lipschitz, and so with constant $=\mathbb{E}[\|(g_{x1}(P_j^{\text{in}}+X_0,\overline{\gamma}_{pj}^{\text{in}}),X_0)\|]^2$ the result in (2.78) follows by part (b) of Lemma (5). We also use **Assumption 4** that γ_{pj} and, hence, $\bar{\gamma}_{pj}$ is bounded. To demonstrate the result in (2.79), observe that by **Assumption 3**, g_{x1} is uniformly Lipschitz, meaning $|g_{x1}(a,b_1)-g_{x1}(a,b_2)|\leq L(1+|a|)|b_1-b_2|$; hence, $|g_{x1}(a,b_1)-g_{x1}(a,b_2)|^2\leq 2L^2(1+|a|^2)|b_1-b_2|^2$ using Lemma 16. Therefore, applying Lemma 16 again,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\left|g_{x1}\left([\mathbf{p}_j^{\text{in}}+\mathbf{x}_0]_i,\gamma_{pj}^{\text{in}}\right)-g_{x1}\left([\mathbf{p}_j^{\text{in}}+\mathbf{x}_0]_i,\overline{\gamma}_{pj}^{\text{in}}\right)\right|^2 &\leq\left|\gamma_{pj}^{\text{in}}-\overline{\gamma}_{pj}^{\text{in}}\right|^2\frac{2L^2}{N}\sum_{i=1}^{N}\left(1+|[\mathbf{p}_j^{\text{in}}+\mathbf{x}_0]_i|^2\right)\\
&\leq\left|\gamma_{pj}^{\text{in}}-\overline{\gamma}_{pj}^{\text{in}}\right|^2\frac{2L^2}{N}\sum_{i=1}^{N}\left(1+2|[\mathbf{p}_j^{\text{in}}]_i|^2+2|[\mathbf{x}_0]_i|^2\right)\\
&\leq\left|\gamma_{pj}^{\text{in}}-\overline{\gamma}_{pj}^{\text{in}}\right|^2\left(2L^2+\frac{4L^2}{N}\sum_{i=1}^{N}|[\mathbf{p}_j^{\text{in}}]_i|^2+\frac{4L^2}{N}\sum_{i=1}^{N}|[\mathbf{x}_0]_i|^2\right).
\end{aligned}$$

Using the above abound, we can bound the probability in (2.78) as follows:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\left|g_{x1}\left([\mathbf{p}_j^{\text{in}}+\mathbf{x}_0]_i,\gamma_{pj}^{\text{in}}\right)-g_{x1}\left([\mathbf{p}_j^{\text{in}}+\mathbf{x}_0]_i,\overline{\gamma}_{pk}^{\text{in}}\right)\right|^2\geq\epsilon^2\right)\\
&\quad\leq\mathbb{P}\left(\left|\gamma_{pj}^{\text{in}}-\overline{\gamma}_{pj}^{\text{in}}\right|^2\left(1+\frac{2}{N}\sum_{i=1}^{N}|[\mathbf{p}_j^{\text{in}}]_i|^2+\frac{2}{N}\sum_{i=1}^{N}|[\mathbf{x}_0]_i|^2\right)\geq\frac{\epsilon^2}{2L^2}\right)\\
&\quad\leq\mathbb{P}\left(\left|\gamma_{pj}^{\text{in}}-\overline{\gamma}_{pj}^{\text{in}}\right|\geq\frac{\epsilon}{2L}\right)+\mathbb{P}\left(\left|\gamma_{pj}^{\text{in}}-\overline{\gamma}_{pj}^{\text{in}}\right|\sqrt{\frac{1}{N}\|\mathbf{p}_j^{\text{in}}\|^2+\frac{1}{N}\|\mathbf{x}_0\|^2}\geq\frac{\epsilon}{2\sqrt{2}L}\right).
\end{aligned}\tag{2.80}$$

The first term on the right side of (2.80) has the desired bound by Lemma 6 part (f), proved in Appendix A.1.

The second term is bounded as above by using Lemmas 9, 12, and 11 to combine the concentration results for the individual factors and terms that are given by Lemma 6 part (f) (for γ_{pj}^{in}), Lemma 6 part (d) (for $\|\mathbf{p}_j^{\text{in}}\|^2/N$), and the assumed concentration or component-wise boundedness (see **Assumption 0**) of $\mathbf{x}_0$. Specifically, we either have that $|[\mathbf{x}_0]_i| \leq B$ for some universal constant B (in which case $\|\mathbf{x}_0\|^2/N \leq B^2$), or $\mathbb{P}\left(\left|\frac{1}{N}\|\mathbf{x}_0\|^2 - \mathbb{E}[X_0^2]\right| \geq 1\right) \leq C\exp(-cN)$.

2.3.7 Lemma 5 proof discussion

Structurally, Lemma 5 is similar to [81, Lemma 5], which is used to establish the AMP analog of Theorem 1 for the linear model in (2.11). However, there are a number of differences between the algorithms that lead to important differences in the proofs, discussed here.

(1) In the finite sample analysis of AMP, one often studies vectors of the form $\mathbf{Ax}$ where $\mathbf{A}$ has i.i.d. Gaussian entries and $\mathbf{x}$ is deterministic. For a vector $\mathbf{Z}$ having i.i.d. $N(0,1)$ elements, $\mathbf{Ax} \overset{d}{=} \|\mathbf{x}\|\mathbf{Z}$, since $\|\mathbf{x}\|$ is deterministic. Importantly, the elements of $\|\mathbf{x}\|\mathbf{Z}$ are independent.

With GVAMP, the analogous object of interest are vectors of the form $\mathbf{Vx}$ where $\mathbf{V}$ is uniformly distributed on the group of orthogonal matrices. This property of $\mathbf{V}$ implies that $\mathbf{Vx}$ has a rotationally invariant distribution, or mathematically, for a vector $\mathbf{Z}$ having i.i.d. $N(0,1)$ elements, $\mathbf{Vx} \overset{d}{=} (\|\mathbf{x}\|/\|\mathbf{Z}\|)\mathbf{Z}$. Thus, we pay for relaxing the Gaussianity condition by picking up a $\|\mathbf{Z}\|$ factor in the denominator, causing dependencies in the elements of $(\|\mathbf{x}\|/\|\mathbf{Z}\|)\mathbf{Z}$ that complicate the concentration arguments. By observing that the Gaussianity of $\mathbf{Z}$ gives concentration of $\|\mathbf{Z}\|/\sqrt{N}$ around 1, it can be shown that $\mathbf{Vx}$ will concentrate around the same limit as $(\|\mathbf{x}\|/\sqrt{N})\mathbf{Z}$, which essentially returns us to the AMP case.

(2) As in the analysis of GAMP, we need to ensure that the $\mathbf{z}_0$ vector is available at the output denoising stage of the general recursion (as it is an input to the likelihood denoiser). Adding $\mathbf{z}_0$ to the disturbance vectors $\mathbf{w}$ would create dependence between the disturbance vectors and other quantities which would violate our conditions, so instead we must generate it from $\mathbf{x}$ in the general

recursion itself. To achieve this, we consider a version of the general recursion where vector iterates are replaced by matrix iterates in order to track both the GVAMP iterates and to recreate $\mathbf{z}_0$ at each loop through the recursion. This case is not handled explicitly in our proof of Lemma 5, but in Appendix A.4, we demonstrate that when $\mathbf{x}$ is a matrix rather than a vector, we recover a characterization of the form $\mathbf{Vx} \stackrel{d}{=} \mathbf{Z}\mathbf{G}_{\mathbf{Z}}^{-1/2}\mathbf{G}_{\mathbf{x}}^{1/2}$, where $\mathbf{Z}$ is now a matrix of equivalent dimensions with i.i.d. $N(0,1)$ entries and $\mathbf{G}_{\mathbf{M}}$ is the Gram matrix corresponding to matrix $\mathbf{M}$. This further generalizes the expression $(\|\mathbf{x}\|/\|\mathbf{Z}\|)\mathbf{Z}$ above.

(3) When studying AMP, one often works with projection matrices $\mathsf{P}_{\mathbf{M}^\perp}$, which map vectors into the orthogonal complement of the range of a matrix $\mathbf{M}$. On the other hand, when studying GVAMP, we instead work with the transformations of the form $\mathbf{B}_{\mathbf{M}}^{\perp}$ that have columns spanning range$(\mathbf{M})^\perp$. These are related to the corresponding projections by $\mathsf{P}_{\mathbf{M}^\perp} = \mathbf{B}_{\mathbf{M}}^{\perp}[\mathbf{B}_{\mathbf{M}}^{\perp}]^T$. This creates two main differences in the decomposition in Lemma (3).

First, in the deviation terms that describe the distance of the algorithm in finite samples from its idealized (and easily analyzable) counterpart, the transformation $\mathbb{I} - \mathsf{P}_{\mathbf{C}_k}$ that shows up in the AMP case is replaced with the matrices $\mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}} \in \mathbb{R}^{N\times 2k}$, where all we know about this matrix is that it has orthonormal columns that span range$\left(\mathbf{C}_{vk}^{\text{in/out}}\right)$. Notice, on the other hand, the projection matrix $\mathbb{I} - \mathsf{P}_{\mathbf{C}_k}$ is fully specified. However, this knowledge is enough to force images of Gaussian vectors under the transformation given by $\mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}}$ to concentrate around zero, because the images are constrained to live in a $k-$dimensional subspace and $\mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}}$ preserves norms.

The second difference is the presence of the orthogonal matrices $\mathbf{O}_{pk}^{\text{in/out}}$ and $\mathbf{O}_{qk}^{\text{in/out}}$ in the decomposition. These appear because there is no direct way to relate $\mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}}^{\perp}\mathbf{Z}_{pk}^{\text{in/out}}$ and $\mathbf{B}_{\mathbf{C}_{uk}^{\text{in/out}}}^{\perp}\mathbf{Z}_{qk}^{\text{in/out}}$ to $\mathbf{Z}_{pk}^{\text{in/out}}$ and $\mathbf{Z}_{qk}^{\text{in/out}}$. However, these matrices end up having no effect on the resulting joint distribution as they are independent of $\overline{\mathbf{Z}}_{pk}^{\text{in/out}}$ and $\overline{\mathbf{Z}}_{qk}^{\text{in/out}}$, respectively.

(4) Because the parameters γ_{1k}, γ_{2k} and τ_{1k}, τ_{2k} can vary with N, so too can the behavior of the denoisers g_{x1}, g_{x2} and g_{z1}, g_{z2}. This is in contrast to AMP, where the denoisers can vary by iteration but not with N. Furthermore, in the AMP case, the denoisers are Lipschitz in both the iterates and the (analog of) the disturbance vectors $\mathbf{w}$; whereas in the GVAMP case, we must

consider a denoiser of the form $[\mathbf{w}]_i[\mathbf{q}]_i$, which is not Lipschitz in both $[\mathbf{w}]_i$ and $[\mathbf{q}]_i$.

One accounts for these differences by using two extensions of the Lipschitz condition: the *uniformly* Lipschitz and uniformly bounded conditionally Lipschitz conditions. We ultimately view the γ, τ, and disturbance vectors $\mathbf{w}$ as parametrizing families of Lipschitz functions, and the two conditions provide sufficient control to ensure that these Lipschitz functions do not vary too rapidly with the parameters and that the corresponding concentration arguments continue to hold.

(5) The order of denoising and subtracting the Onsager term is interchanged in AMP and GVAMP. This adds an additional layer of complication in obtaining concentration for $\mathbf{v}_k^{\text{in/out}}$: from Algorithm 3,

$$\mathbf{v}_k^{\text{in/out}} = \frac{1}{1-\alpha_{pk}^{\text{in/out}}}\left[f_p^{\text{in/out}}(\mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}, \mathbf{w}_p^{\text{in/out}}, \gamma_{pk}^{\text{in}}, \gamma_{pk}^{\text{out}}) - \alpha_{pk}^{\text{in/out}}\mathbf{p}_k^{\text{in/out}}\right],$$

with $f_p^{\text{in/out}}(\mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}, \mathbf{w}_p^{\text{in/out}}, \gamma_{pk}^{\text{in}}, \gamma_{pk}^{\text{out}}) - \alpha_{pk}^{\text{in/out}}\mathbf{p}_k^{\text{in/out}}$ failing to be Lipschitz (as $\alpha_{pk}^{\text{in/out}}$ also has a dependence on $\mathbf{p}_k^{\text{in/out}}$). This prevents us from simply applying Lemma 5 **(b)** when proving concentration results for $\mathbf{v}_k^{\text{in/out}}$. However, as with the $1/\|\mathbf{Z}\|$ dependence discussed in **(1)** above, we handle this by using concentration of $\alpha_{pk}^{\text{in/out}}$ to its limit $\overline{\alpha}_{pk}^{\text{in/out}}$, allowing one to show that $\mathbf{v}_k^{\text{in/out}}$ will concentrate around the same limit as $\frac{1}{1-\overline{\alpha}_{pk}^{\text{in/out}}}\left[f_p^{\text{in/out}}(\mathbf{p}_k^{\text{in}}, \mathbf{p}_k^{\text{out}}, \mathbf{w}_p^{\text{in/out}}, \gamma_{pk}^{\text{in}}, \gamma_{pk}^{\text{out}}) - \overline{\alpha}_{pk}^{\text{in/out}}\mathbf{p}_k^{\text{in/out}}\right]$. This function is Lipchitz, so (inductively) applying Lemma 5 **(b)** gives the desired concentration.

Chapter 3: Challenges to evidence

Checking the adequacy of a statistical model is an essential step in almost any applied modeling workflow [38, 89, 34, 12]. When a model's assumptions have not been tested against their observable consequences, inferences about unobservable quantities obtained through such models must be interpreted skeptically. However, the process of checking a model is often not straightforward, and it is subject to a number of confusions in the Bayesian setting in particular. For instance, we find substantial disagreement in the literature over questions such as:

1. Is our goal to subject our model to the strongest possible test of its compatibility with (some feature of) the data, in order to have the best possible chance of rejecting the model? Or, is our goal to generate assessments of fitness which can help us to identify particular improvements we can make to the model? We will term the former the *rejection goal*, which is strongly advocated for by [80]. This perspective is explicitly rejected by [35], who advocates instead for more qualitative methods that enable "[understanding] in what ways the fitted model departs from the data." This view closely tracks with our latter goal, which we will refer to as the *discovery goal*.

2. Do we need to know the frequency properties of our model checking procedures in order to interpret their output? Or can we achieve the relevant goals by using "purely" Bayesian calculations? And would using frequency calculations undermine the Bayesian consistency or validity of our analysis? Arguments for the importance of frequency information can found in [80] and [5], whereas arguments in the opposite direction are given in [36] and [35].

In practice, we associate the discovery and rejections goals with what we termed fitness and correctness testing respectively. We note that the distinction between the goals of model rejection

and model discovery can be arbitrarily sharp. An oracle which provides a yes or no answer, for any proposed model, to the question of whether it is a valid description of the true data generating process gives us 100% power against any alternative. But such a binary oracle offers little help in diagnosing the source of the model's inaccuracy or in finding plausible directions for improvement. Combined with Box's famous adage - "all models are wrong, but some are useful" - one might argue on this basis that pursuing the discovery goal is the more practical strategy. In the Bayesian setting, this is most commonly achieved with tests of fitness that compare observed data to simulations from the model's posterior predictive distribution - i.e. the model's sampling distribution $f(\cdot \mid \boldsymbol{\theta})$ averaged over the posterior $q(\boldsymbol{\theta} \mid \mathbf{y})$. In numerical form, this leads to the posterior predictive p-value, but advocates of the posterior predictive check often recommend qualitative visual checks for their higher density of information [34]. In this setting, concerns over frequency properties are either not relevant (in the case of the p-value, which can be interpreted directly as a posterior probability) or not well-defined (in the case of visual assessments, where formal decision processes are rarely defined).

However, when we pursue the rejection goal, frequency evaluations become much more relevant. As noted in chapter 1, [67] showed that the posterior predictive p-value has a frequency distribution which is stochastically less variable than uniform (when data $\mathbf{y}$ are sampled from the model's prior predictive distribution $f(\mathbf{y})$). As a consequence, the frequency of a given posterior predictive p-value is usually less than its nominal value, and sometimes substantially so. If we test the model by comparing the p-value to some threshold, then such tests will be conservative or underpowered compared to a test using the corresponding frequency. Moreover, it has been observed that the size of this power deficit can be quite large in practice [85, 99, 97], limiting its usefulness as a test of model correctness.

We argue that both the discovery and rejection goals are relevant to the applied practice of model checking, and that both goals can be integrated into a unified practice. To see how tests of fitness and correctness may be employed together, we observe that pursuing the discovery goal often entails constructing many models, each designed to improve fitness in response to a check

of a previous model. However, this process of model multiplication must eventually terminate, at least temporarily, due either to the diminishment of identifiable routes for further improvement or the need to use the model for some downstream task. We therefore want to evaluate the risks of stopping at any given time in the model building process. In particular, we may reasonably wish to judge if our current model is acceptable for some task.

In many cases, the acceptability of a model will be directly related to the difficulty of the tests which it has passed. A model which has been checked only superficially is often untrustworthy. Thus, the evaluation of a potential stopping time is connected not just to outcome but also to the power of a test.

Whereas rejection-oriented tools take power-maximization as a first concern, tools oriented towards the discovery goal may be less helpful towards this end, particularly when we consider stopping in response to such discovery-oriented tools becoming less informative. For instance, while checking fitness to individual features of the data is more useful for diagnosing specific sources of misfit, omnibus tests that aim to check the model more holistically can achieve higher power for rejection purposes.

This chapter presents three contributions to the subject of model checking in high dimensions:

1. We present an argument that both traditional fitness and correctness tests tend to become weaker (in their respective senses) as the model dimension grows.

2. We demonstrate that a simple process of disaggregating the posterior predictive p-value (and other posterior predictive assessments) allows for tests to leverage posterior dependence between parameters and fitness, which can reveal directions for model improvement that are masked by corresponding aggregated assessments.

3. We introduce a joint posterior predictive p-value, which leverages the dependence between multiple test statistics under the posterior predictive distribution to provide tests of model correctness that can scale up in power as the dimension of the underlying model grows, compensating for the reduced power of simpler tests in higher dimensions.

3.1 The threat of high dimensionality to testing

To fix ideas, suppose we have selected a test statistic T which we may employ in either fitness or correctness testing. Our main argument in this section proceeds as follows:

1. First, we argue that the reliability of a fitness test using T is threatened by a large posterior variance of the conditional p-value $p_T(\boldsymbol{\theta})$ (defined in chapter 1).

2. Correspondingly, we argue that the power of a correctness test using T is threatened by low prior predictive variance of p_T^{post} around $1/2$.

3. We demonstrate that the posterior variance of $p_T(\boldsymbol{\theta})$ and the prior predictive variance of p_T^{post} are inversely related to one another.

4. Finally, we argue that $\text{Var}_{q(\boldsymbol{\theta}|\mathbf{y})} p_T(\boldsymbol{\theta})$ tends to increase with model expansion.

We conclude from these arguments that traditional means of model checking are likely to become less useful under processes of iterative model building and expansion. This in turn motivates the proposals for new fitness and correctness checking methods that we propose in the next sections.

3.1.1 Reliability and variance in fitness testing

When assessing the fitness of a model, we are typically in interested in whether the model in question can reproduce some qualitative and quantitative feature of the data with sufficiently high frequency. Qualitative assessments are most commonly performed with visual assessments such as replication plots, in which many new data sets are sampled from the posterior predictive distribution $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$ and then visualized along with the observed data in such a way which reveals or emphasizes the feature in question (e.g. histograms to check distribution shape or residual plots to check serial independence of errors).

Quantitative tests usually begin by selecting one or more test statistics $T : \mathcal{Y} \to \mathbb{R}$ which are designed to capture a specific aspect of the data (e.g. a sample mean, quantile, or autocorrelation).

The observed value, $T(\mathbf{y})$, may then be compared to the distribution of $T(\mathbf{y}_{\text{rep}})$ under sampling of replicated data from the posterior predictive distribution $q(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$. In order to make this comparison quantitative, we need a measure of how rare $T(\mathbf{y})$ is relative to the distribution of $T(\mathbf{y}_{\text{rep}})$. Most commonly this comparison is made by computing a tail probability, which results in the posterior predictive p-value p_T^{post} defined in (1.11). The posterior predictive distribution is the modeled distribution of data sets generated from the same process that generated $\mathbf{y}$, aggregating over the modeled uncertainty about that process. Smaller values of p_T^{post} indicate that $T(\mathbf{y})$ is rarer under its posterior predictive distribution, and thus that $\mathbf{y}$ is atypical among datasets generated from (our model of) the same process. This discordance between $\mathbf{y}$ and our model of its data generating process is the characteristic property of a poor fitting model.

Care must be taken in interpreting the posterior predictive p-value, in particular when its value is *not* decisively small. While a value below 0.01, for example, may be seen as proof of poor fit, a larger value like 0.2 cannot be regarded as unambiguous evidence of typicality of the data under the posterior predictive for T. The reason for this can be read off the definition $p_T^{\text{post}} = \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})} p_T(\boldsymbol{\theta})$, i.e. p^{post} is the posterior average of the separate conditional p-values $p_T(\boldsymbol{\theta})$. It can easily be simultaneously true that the posterior mean of $p_T(\boldsymbol{\theta})$ is large (e.g. ≥ 0.2) and that $p_T(\boldsymbol{\theta}) < 0.01$ with substantial probability. In particular, if $\boldsymbol{\theta}^*$ is the (unknown) true value of $\boldsymbol{\theta}$, then larger values of p_T^{post} provide little insurance against $p_T(\boldsymbol{\theta}^*)$ being small, which we claim should be regarded as a clear indication of poor model fitness.

This last point may be succinctly rephrased as the argument that model fitness is most appropriately assessed conditional on $\boldsymbol{\theta}$. Since there are typically infinitely many $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, any single point assessment (such as p^{post}) risks losing useful information about fitness. The degree of this risk in any given circumstance is clearly related to the variability of $p_T(\boldsymbol{\theta})$. In the extreme case that $p_T(\boldsymbol{\theta})$ is constant over $\boldsymbol{\theta}$, then we will have $p_T(\boldsymbol{\theta}) = p_T^{\text{post}}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and the concerns raised above are a nonissue. On the other hand, when the posterior variance of the conditional p-values, i.e. $\text{Var}_{q(\boldsymbol{\theta}|\mathbf{y})} p_T(\boldsymbol{\theta})$, is large, we may expect *most* values of $p_T(\boldsymbol{\theta})$ to be meaningfully far from p_T^{post}.

We also stress that the fundamental problems outlined above are not unique to the posterior

predictive p-value or to quantitative tests of fitness. No matter the means of testing fitness, if the resulting check varies substantially depending on the value of $\boldsymbol{\theta}$ sampled from the posterior, then the result of the check is necessarily more equivocal as to the overall fitness of the model.

3.1.2 Calibration and power in correctness testing

A very common technique for Bayesian model rejection is to compare the posterior predictive p-value (p^{post}) to some threshold. The classic argument against using this procedure for model rejection - that it is overly conservative - can be formalized using the concept of convex order. For distributions p, q, we say that p is less than q in convex order ($p \ll q$) if, for $X \sim p$, $Y \sim q$, and any convex function ψ, we have that

$$\mathbb{E}\psi(X) \leq \mathbb{E}\psi(Y). \tag{3.1}$$

[67] showed that p^{post} is dominated in convex order by a uniform random variable. To demonstrate this, let $p_T(\mathbf{y}, \boldsymbol{\theta})$ be the p-value computed with respect to $f(\mathbf{y} \mid \boldsymbol{\theta})$, i.e.

$$p_T(\mathbf{y}, \boldsymbol{\theta}) = \mathbb{P}_{f(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})}\left[T(\mathbf{y}_{\text{rep}}) \geq T(\mathbf{y})\right]. \tag{3.2}$$

Then we have that

$$\mathbb{E}_{f(\mathbf{y})}\psi\left(p_T^{\text{post}}(\mathbf{y})\right) = \mathbb{E}_{f(\mathbf{y})}\psi\left(\mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})}p_T(\mathbf{y}, \boldsymbol{\theta})\right) \overset{(a)}{\leq} \mathbb{E}_{q(\boldsymbol{\theta})}\mathbb{E}_{f(\mathbf{y}|\boldsymbol{\theta})}\psi\left(p_T(\mathbf{y}, \boldsymbol{\theta})\right) \overset{(b)}{=} \mathbb{E}\psi(U), \tag{3.3}$$

where U is a uniform random variable. Here, (a) follows by Jensen's inequality, and (b) follows from the definition of $p_T(\mathbf{y}, \boldsymbol{\theta})$ and the fact that any p-value has a uniform distribution under its assumed sampling distribution (by the probability integral transform). Roughly, this convex ordering means that p^{post} will tend to have a distribution that is more peaked around 0.5, and thus it will commonly be true that

$$f_T(\alpha) \overset{\text{def}}{=} \mathbb{P}\left(p_T(\mathbf{y}) \leq \alpha\right) < \alpha \tag{3.4}$$

for sufficiently small values of α. We can thus see that for a sufficiently small threshold p^*, when (3.4) holds, the test that rejects when $p_T(\mathbf{y}) < p^*$ is lower power than the test that rejects when $f_T\left(p_T(\mathbf{y})\right) < p^*$.

The Bayesian who does not want to be concerned with frequency calculations may reasonably wonder at this point whether this claimed power deficit will be an issue in practice. Indeed, this argument does not show that p^{post} is useless for model rejection. Meng also showed that

$$\mathbb{P}\left(p_T(\mathbf{y}) \leq \alpha\right) \leq 2\alpha \tag{3.5}$$

for all α. Thus, when $p_T(\mathbf{y})$ is sufficiently small, we will still have sufficient information to reject the model on frequentist grounds without the need to compute or approximate $f_T(p_T(\mathbf{y}))$. Indeed, many examples show that p^{post} can work quite well for this purpose, and one can always choose a more skeptical threshold if power is a substantial concern.

Of course, the viability of this strategy relies entirely on *how* non-uniform p^{post} is in any given case. If p^{post} becomes severely non-uniform and is sharply peaked around 0.5, then the only way to achieve significance levels that aren't extremely conservative may be to place the nominal threshold at levels so large (e.g. > 0.4) that they would never be recommended absent direct evidence of this degree of peakedness (since they would result in unreasonably large significance levels in other cases). Consistent with this, it has been observed that large variation in the conservativity of p^{post} across models and test quantities undermines its consistent interpretation [25]. In short, we can expect p^{post} to give reasonable rejection performance only when it is consistently not-too-severely non-uniform.

3.1.3 Connecting fitness and correctness testing to model expansion

In light of the above arguments, it is clear that we need an understanding of how non-uniform p^{post} may be in practice to adjudicate the relevant concerns. Examining (3.3), we can see that the degree of non-uniformity is entirely controlled by the size of the gap in the inequality (a). It is

well-known that the gap in Jensen's inequality can be bounded above and below as

$$\sigma_{\mathbf{y}}^2 \frac{\inf \psi''}{2} \leq \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})} \psi\left(p_T(\mathbf{y}, \boldsymbol{\theta})\right) - \psi\left(\mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})} p_T(\mathbf{y}, \boldsymbol{\theta})\right) \leq \sigma_{\mathbf{y}}^2 \frac{\sup \psi''}{2}, \tag{3.6}$$

where

$$\sigma_{\mathbf{y}}^2 = \operatorname{Var}\left[p_T(\mathbf{y}, \boldsymbol{\theta}) \mid \mathbf{y}\right]. \tag{3.7}$$

Thus, the non-uniformity of $p_T^{\text{post}}(\mathbf{y})$ is controlled by the average size of $\sigma_{\mathbf{y}}^2$. In particular, taking $\psi(x) = (x - 1/2)^2$, (3.6) shows that larger values of $\sigma_{\mathbf{y}}^2$ imply sharper concentration of the distribution of $p_T^{\text{post}}(\mathbf{y})$ around $1/2$. In light of our argument in Section 3.1.1, this demonstrates that larger values of $\sigma_{\mathbf{y}}^2$ undermine both the reliability of fitness tests and the power of correctness tests.

We next claim that, for at least some T, we should expect $\sigma_{\mathbf{y}}^2$ to increase throughout a discovery-driven modeling workflow. To formalize this claim, we begin with the following assumption.

Workflow Assumption. *In a modeling workflow that emphasizes an open-ended process of model criticism and model improvement, our models will tend to become more complex and require higher-dimensional parameter spaces in order to accommodate those features of the data which are observed empirically but are not accounted for in our existing models.*

This assumption of model improvement as requiring model expansion may not always hold, for instance if we move from a generic initial model to a more specialized model designed with particular domain knowledge. Nevertheless, we believe this assumption is valid in many settings, as model improvement often requires accounting for unanticipated sources of variation (e.g. overdispersion, random effects, nonlinearity), which results in models that are higher-dimensional than their predecessors.

Our workflow assumption can be formalized as the proposition that a discovery-driven modeling workflow will tend to produce models which are expansions of previous models in the sense of Definition 1. Furthermore, when passing from a base model to an expanded model in this way, we

can see by the law of total variance that

$$\begin{aligned}\mathrm{Var}_q\left[p_T(\mathbf{y},(\boldsymbol{\theta},\boldsymbol{\lambda}))\mid\mathbf{y}\right] &= \mathbb{E}\left\{\mathrm{Var}_q\left[p_T(\mathbf{y},(\boldsymbol{\theta},\boldsymbol{\lambda}))\mid\mathbf{y},\boldsymbol{\lambda}\right]\mid\mathbf{y}\right\}+\mathrm{Var}_q\left\{\mathbb{E}\left[p_T(\mathbf{y},(\boldsymbol{\theta},\boldsymbol{\lambda}))\mid\mathbf{y},\boldsymbol{\lambda}\right]\mid\mathbf{y}\right\}\\ &= \mathrm{Var}_{q_{\text{base}}}\left[p_T(\mathbf{y},\boldsymbol{\theta})\right]+\Delta+\mathrm{Var}_q\left\{\mathbb{E}\left[p_T(\mathbf{y},(\boldsymbol{\theta},\boldsymbol{\lambda}))\mid\mathbf{y},\boldsymbol{\lambda}\right]\mid\mathbf{y}\right\},\end{aligned}$$

where we define

$$\Delta=\mathbb{E}\left\{\mathrm{Var}_q\left[p_T(\mathbf{y},(\boldsymbol{\theta},\boldsymbol{\lambda}))\mid\mathbf{y},\boldsymbol{\lambda}\right]\mid\mathbf{y}\right\}-\mathrm{Var}_q\left[p_T(\mathbf{y},(\boldsymbol{\theta},\boldsymbol{\lambda}))\mid\mathbf{y},\boldsymbol{\lambda}=\boldsymbol{\lambda}_0\right].$$

We note that the second equality follows from the fact that $q_{\text{base}}(\boldsymbol{\theta}\mid\mathbf{y})=q(\boldsymbol{\theta}\mid\mathbf{y},\boldsymbol{\lambda}=\boldsymbol{\lambda}_0)$ and $f_{\text{base}}(\mathbf{y}\mid\boldsymbol{\theta})=f(\mathbf{y}\mid\boldsymbol{\theta},\boldsymbol{\lambda}=\boldsymbol{\lambda}_0)$. In any given model expansion, Δ may be positive or negative, as $\mathrm{Var}_q\left[p_T(\mathbf{y},(\boldsymbol{\theta},\boldsymbol{\lambda}))\mid\mathbf{y},\boldsymbol{\lambda}\right]$ may vary arbitrarily over the support of $q(\boldsymbol{\lambda}\mid\mathbf{y})$. On the other hand, we clearly always have $\mathrm{Var}_q\left\{\mathbb{E}\left[p_T(\mathbf{y},(\boldsymbol{\theta},\boldsymbol{\lambda}))\mid\mathbf{y},\boldsymbol{\lambda}\right]\mid\mathbf{y}\right\}\geq 0$. Thus, this identity along with (3.6) strongly suggests that $\sigma^2_{\mathbf{y}}$ - and thus the non-uniformity of p^{post} - tends to increase through the process of model expansion.

We also note that this problem is not exclusive to p^{post}. If the posterior predictive p-value is highly non-uniform, then we should expect similar posterior predictive checks, such as replication plots, to be problematic for purposes of fitness and correctness testing as well. For instance, a check which produces replications that appear visually similar to the observed data 20% of the time would usually be considered a positive result for the proposed model. From a correctness perspective, it may just as easily be true that if the model were correct, such a visual check would produce sampled data similar to the observed data in a much higher proportion of replications. In short, visual checks can be conservative in the same way as numerical p-values. From a fitness perspective, a test that inspects many simulations from the posterior predictive distribution (such as a replication plot) can at least make us aware of greater variance in fitness. But such checks do not typically allow us to distinguish between variance driven by the sampling distribution of the statistic (conditional on $\boldsymbol{\theta}$) versus variance driven by differential fitness across distinct values of $\boldsymbol{\theta}$. As we will see in the next section, disentangling these sources of variance can reveal particular

directions for model improvement.

The above shows that when our tools for model discovery lead us to larger models, they also tend to lead toward models that are harder to reject or improve using observable data. Specifically, these calculations show that existing and common model checking tools such as p^{post} will tend to perform worse for both purposes of fitness and correctness testing as the model dimension grows within an expansion-based workflow. We therefore believe there is a practical need for correctness and fitness checking tools for which the difficulty and richness of the test can be scaled to match the complexity of the model appropriately.

3.2 Extracting useful explanations of fitness variance from the posterior

As discussed in the previous section, when the variance of the conditional p-value $p_T(\boldsymbol{\theta})$ is large, the posterior predictive p-value becomes a worse summary of the (posterior) distribution of $p_T(\boldsymbol{\theta})$. In this case, accounting for the uncertainty the posterior places on the true data generating process $f(\cdot \mid \boldsymbol{\theta})$, there is no unequivocal and unambiguous answer to the question "does the model fit the data?". This creates a challenge for efforts to improve the model insofar as the ambiguity of our model test is inversely related to the ease with which we can identify avenues for model improvement. As an example, consider a case where we model an error term as i.i.d. and wish to test whether the model can explain the observed autocorrelation in the observed errors. If this results in a vanishing posterior predictive p-value, then we can relatively easily conclude that the true error process involves serial correlation which we could then model explicitly to improve fitness.

To take a more complex and subtle case, consider modeling two-dimensional data of the form $x = \{x_{1i}, x_{2i}\}_{i=1}^{n}$. Suppose that we model these two-dimensional vectors as i.i.d. from a joint distribution $p(X_1, X_2)$, and suppose that we wish to test the sample variance of the sum $x_{1i} + x_{2i}$. Note that the sample variance of the sum depends on both the marginal variances $\text{Var}(X_1)$ and $\text{Var}(X_2)$ as well as the covariance $\text{Cov}(X_1, X_2)$. If the data were generated from a process with smaller marginal variances for X_1 and X_2 than are possible in our model, then we might expect the

modeled variance of the sum $X_1 + X_2$ to exceed the sample variance and to conclude that our model does not fit the data. However, if there is sufficient posterior uncertainty about the correlation ρ between X_1 and X_2, then more negative values of this correlation may be able to mask this problem by compensating the too-large variance with a sufficiently negative covariance. In this case, we might also expect to see greater variance in the conditional p-values $p_T(\boldsymbol{\theta})$, since the model's fit to the sum is highly associated with the modeled correlation between X_1 and X_2.

However, this example also points to a way out of this confusion. If we can discover that $p_T(\boldsymbol{\theta})$ and ρ are negatively associated under the posterior distribution, then we both be able to

1. interpret the otherwise-ambiguous posterior predictive p-value by explaining the pattern of poor fitness, and

2. identify possible directions for model improvement.

This latter point is worth elaborating on. If there was no identifiable relation between $p_T(\boldsymbol{\theta})$ and the model parameters, then it may be unclear how changes to the model might improve fitness. With the benefit of the observed negative association, we can conclude that we may either improve model fitness by modifying the model to favor more negative correlations, or by increasing the variance of the sum when the correlation is closer to 0 (i.e. by increasing the marginal variances). Which of these directions is more reasonable will depend on our understanding of the specific domain, but this conclusion still serves to organize our thinking and suggest directions for improvement. In this simple example, this conclusion is obvious. We know that the variance of the sum is decomposed into these marginal variances and covariance, and thus one of these terms must shrink in order to shrink the modeled variance of the sum. But such constraints may not be obvious in more complex examples, and, critically, we did not need prior knowledge of this decomposition for the association between $p_T(\boldsymbol{\theta})$ and ρ to be interpretable in this way. In the rest of this section, we will walk through a real-data example in which avenues for model improvement are less obvious, and we will see how exploring the associations between conditional p-values and parameters can substantially aid our search.

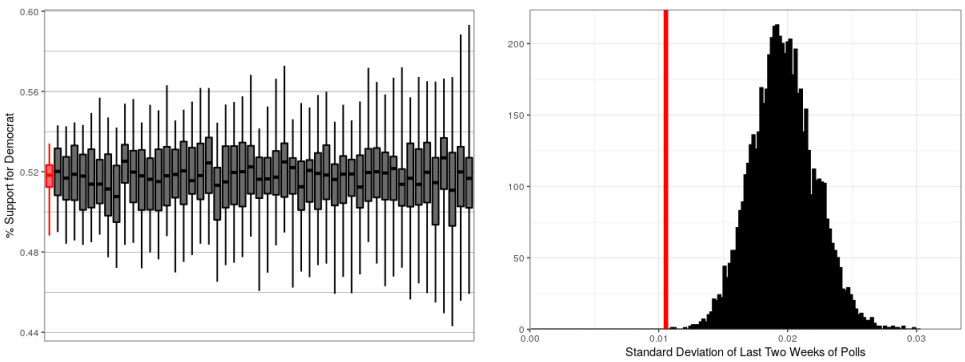

Figure 3.1: The observed data is highlighted in red. Left: box plots of the last ten days of polls for the observed data and fifty posterior predictive replications, ordered by range. Right: histogram of standard deviations of the last ten days of polls for the observed data and 6000 posterior predictive replications.

3.2.1 Discovering directions for model improvement in a model of presidential polls

We take as our base model the U.S. presidential election forecasting model of [43]. We pay particular attention to herding, a hypothesized process whereby pollsters augment their raw data in such a way that their published poll numbers are closer to the existing consensus of recent polls than they would otherwise be. That the variance among presidential election polls has fallen well below the minimal expected variance (assuming independence between polls) just before recent elections has been presented as evidence for the existence of herding [75].

The base forecasting model and a model of herding

In this model, the result of state or national poll i with n_i respondents is quantified by the number of respondents y_i who support the Democratic candidate. These y_i are modeled as binomial with parameters $\theta_i = \mathsf{logit}^{-1}(t_i)$, where t_i contains terms including the true level of support as well as sources of polling bias. See the appendix for model specification details.

In any close race, we have $y_i/n_i \in (0.4, 0.6)$ for all national polls, and the binomial model has standard deviation at least $\approx 0.49/\sqrt{n_i}$. Sufficiently low poll variance over any interval will violate this model. Indeed, The left panel of Figure 3.2 shows that the observed polling variance in the ten days before the election is substantially smaller than in posterior model simulations.

By allowing dependence among polls, adding a herding mechanism can resolve this misfit. First let θ_j^{nat} be the binomial parameter for national poll j. This represents the expected average support for the Democratic candidate in random samples drawn from the sampling frame of the corresponding poll. We model the sampling process of national poll j as follows.

1. The pollster samples their sampling frame, observing unherded result:

$$y_j^{\text{nat}} \mid \theta_j^{\text{nat}} \sim \text{binomial}\left(\theta_j^{\text{nat}}, n_j^{\text{nat}}\right).$$

2. The pollster calculates a herding target μ_j^{herd} representing their quantification of the consensus of polls to which they compare their $y_j^{\text{nat}}/n_j^{\text{nat}}$.

3. The pollster herds $y_j^{\text{nat}}/n_j^{\text{nat}}$ by some fraction $\lambda_j^{\text{herd}} \in (0, 1)$ towards μ_j^{herd}. They publish this herded result:

$$p_j^{\text{nat}} = \left(1 - \lambda_j^{\text{herd}}\right) \frac{y_j^{\text{nat}}}{n_j^{\text{nat}}} + \lambda_j^{\text{nat}} \mu_j^{\text{herd}}.$$

To simplify implementation, we use the normal approximation to the binomial, getting

$$p_j^{\text{nat}} \mid \theta_j^{\text{nat}}, \mu_j^{\text{herd}}, \lambda_j^{\text{herd}} \sim \text{normal}\Bigg[\left(1 - \lambda_j^{\text{herd}}\right) \theta_j^{\text{nat}} + \lambda_j^{\text{nat}} \mu_j^{\text{herd}}, \tag{3.8}$$

$$\left(1 - \lambda_j^{\text{herd}}\right) \sqrt{\theta_j^{\text{nat}} \left(1 - \theta_j^{\text{nat}}\right) \Big/ n_j^{\text{nat}}}\Bigg]. \tag{3.9}$$

The herding model for the state pols uses analogous targets $\mu_{s_j,j}^{\text{herd}}$ and percentages $\lambda_{s_j,j}^{\text{herd}}$, where s_j is the state in which poll j was conducted. Now let m^{nat}, s^{nat}, m_s^{state} and s_s^{state} be the sample averages and standard deviations of the national polls and polls of state $1 \leq s \leq 51$ respectively. We set the following priors for these parameters.

$$\mu_i^{\text{herd}} \overset{iid}{\sim} \text{normal}\left(m^{\text{nat}}, \frac{2}{3} s^{\text{nat}}\right), \quad \mu_{s_j,j}^{\text{herd}} \overset{iid}{\sim} \text{normal}\left(m_{s_j}^{\text{state}}, \frac{2}{3} s_{s_j}^{\text{state}}\right). \tag{3.10}$$

These priors essentially constrain the targets to lie within ±2 standard deviations of the overall

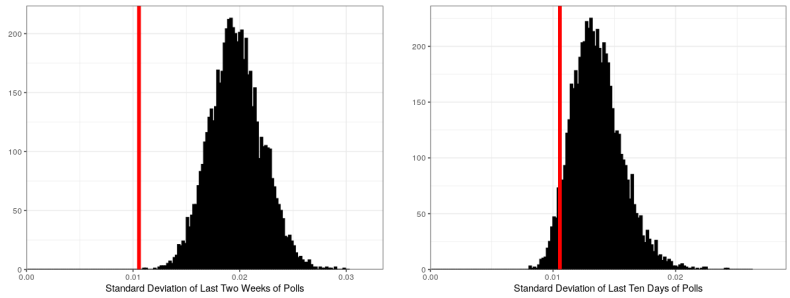

Figure 3.2: (Observed data shown in red). Histogram of the s.d. of the last ten days of polls for the observed data and six thousand posterior replications. Left: base model. Right: first expanded model.

polling means, away from the extremes of our data. Next let $N^{\text{nat}}, N^{\text{state}}$ be the number of national and state polls, and let $C = \{1 \leq i \leq N^{\text{nat}} \mid t_i \geq T - 10\}$ be the indices of national polls conducted in the last ten days prior to the election. We set priors:

$$\begin{aligned} \lambda_i^{\text{nat}}, \lambda_{j,s_j}^{\text{state}} &\overset{iid}{\sim} \text{Beta}\,(1,9)\,, \text{ for all } 1 \leq j \leq N^{\text{state}} \text{ and } i \in C^c, \\ \lambda_i^{\text{nat}} \Big| \mu_\lambda^{\text{last}}, k_\lambda^{\text{last}} &\overset{iid}{\sim} \text{Beta}\left(\mu_\lambda^{\text{last}} k_\lambda^{\text{last}}, \left(1 - \mu_\lambda^{\text{last}}\right) k_\lambda^{\text{last}}\right), \text{ for all } i \in C, \end{aligned} \tag{3.11}$$

with hyperpriors $\mu_\lambda^{\text{last}} \sim \text{uniform}\,[0,1]$ and $k_\lambda^{\text{last}} \sim \text{normal}\,(200, 120)$. We also constrain the λ explicitly to $[0, 0.9]$ since values too close to 1 create adverse geometry that frustrates the sampler, and since our prior beliefs rule out such large herding percentages. The hierarchical prior on the last ten days of national herding parameters serves two purposes. Since the polls in this period are our primary evidence for herding, inference for $\mu_\lambda^{\text{last}}$ and k_λ^{last} is of substantive interest. Also, since data is especially dense in this period, we can better estimate the herding parameters for these polls (and can thus afford the weaker marginal priors implied by the hierarchical structure).

The right panel of Figure 3.2 displays the posterior predictive check of polling variation in the last ten days for the expanded model. While the observed variation is still relatively small, it is no longer implausible. We may now be tempted to stop and declare our model good enough, as it is not obvious what further improvements to make. We can, however, extract more information than is revealed by this (marginal) posterior predictive check.

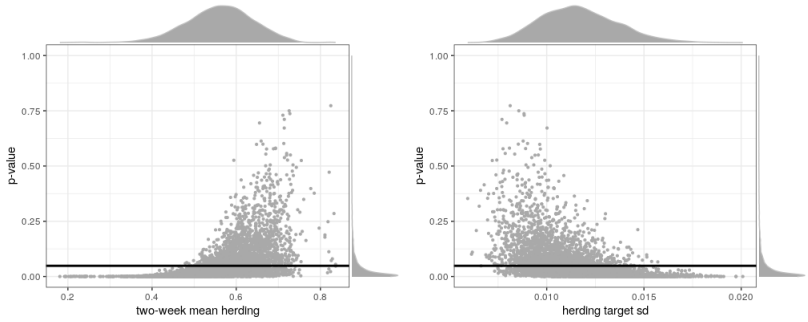

Figure 3.3: Cppp-vs for the s.d. of the last ten days of national polls in the expanded model plotted against the population average herding percentage over that time frame (left) and the standard deviation of the herding targets over that time frame (right). Black lines indicate the marginal ppp-v. Estimated marginal distributions are displayed on the margins.

A more informative conditional model check

Recall that the conditional posterior predictive p-values (cppp-vs) are defined for test statistic T as $p_T(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})}\mathbb{1}\left\{T(\mathbf{y}_{\text{rep}}) \geq T(\mathbf{y})\right\}$. The left panel of Figure 3.3 plots the cppp-vs with T equal to the standard deviation of the last ten days of national polls against posterior samples of $\mu_\lambda^{\text{last}}$, the average herding percentage in the last ten days of polling. The marginal distributions displayed on the axes show that the distribution of cppp-vs has a heavy right tail with the preponderance of sampled p-values less than the average of 0.05. Furthermore, the fit of the model appears much better for larger values of $\mu_\lambda^{\text{last}}$, suggesting that model fitness may improve if the posterior favored more herding. But it is unclear *how* we should achieve this since using a more informative prior to favor higher values of $\mu_\lambda^{\text{last}}$ would be inconsistent with our (lack of) prior knowledge.

Another comparison tells us more. The right panel of Figure 3.3 plots the same cppp-vs against the standard deviation of herding targets μ_i^{herd} in the last ten days of polls. Less variation in the targets is associated with better fit, indicating another avenue for improvement. The prior on μ_i^{herd} only enforces that μ_i^{herd} not be extreme compared to all polls in the series. But clearly μ_i^{herd} should also be non-extreme compared to polls around the *specific* time that poll i was conducted. Including this prior information should reduce the posterior standard deviation of herding targets and thus hopefully improve the fit.

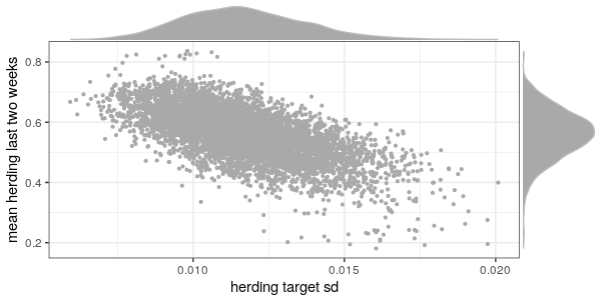

Figure 3.4: Posterior samples from the first expanded model of the average herding percentage for national polls against the standard deviation of herding targets for national polls both over the last ten days of polls. Estimated marginals are given on the margins.

Using the conditional check to improve the model

To construct a new prior on the targets μ_i^{herd}, we calculate the trailing averages

$$c_d^i = \frac{1}{|C_{t_i}^d|} \sum_{k \in C_d^i} \frac{y_k^{\text{nat}}}{n_k^{\text{nat}}} \tag{3.12}$$

for $d = 2, \ldots, 7$ days before each national poll i, where $C_t^d = \left\{1 \leq j \leq N^{\text{nat}} \mid t - d \leq t_j \leq t\right\}$.

We then define the average and spread of these:

$$m_c^i = \frac{1}{6} \sum_{d=2}^{7} c_d^i, \text{ and } s_c^i = \sqrt{\frac{1}{6} \sum_{d=2}^{7} \left(c_d^i - m_c^i\right)^2} + 0.0002. \tag{3.13}$$

The addition of a small constant 2^{-4} ensures the spread cannot shrink to 0 when data is sparse. We then define the improved prior $\mu_i^{\text{herd}} \sim \text{normal}\left(m_c^i, s_c^i\right)$, with priors for the state herding targets constructed similarly from the state-level time series. This prior now encodes the idea that μ_i^{herd} should look like a consensus of *recent* polls.

Figure 3.5 displays the same pair of conditional posterior predictive checks using this more informative prior. The (marginal) ppp-v is nearly 0.3, and the unpleasant 'spike and slab' shape of the cppp-vs that appeared in Figure 3.3 has been attenuated. Furthermore, we no longer find intervals of large posterior probability where the cppp-vs are vanishingly small in either plot. Thus, the marginal check looks good, and the conditional check no longer indicates obvious directions

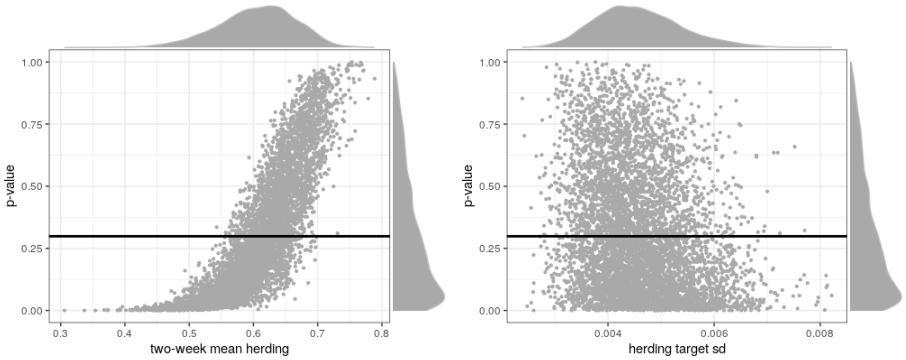

Figure 3.5: cppp-vs for the standard deviation of national polls under the second expanded model, plotted against the population average herding percentage (left) and the standard deviation of the herding targets (right), all computed over the last ten days of polls. Black lines indicate marginal ppp-vs. Estimated marginals are given on the margins.

for improvement.

The improvement in cppp-vs versus the standard deviation of μ_i^{herd} is expected, but the improvement in the comparison with the mean herding percentage (achieved by higher mean herding in the posterior) may be somewhat surprising. The source of the improvement is revealed in Figure 3.4, which displays a strong negative association between the standard deviations of the μ_i^{herd} and the $\mu_\lambda^{\text{last}}$ in the last ten days for the first expanded model. Thus, solving the problem for the herding targets also solved the herding percentage problem for free.

This example shows that marginalized model checks like the ppp-v can obscure information useful for assessing model fitness in practice. More positively, it also clearly demonstrates how the cppp-v can be a powerful tool for motivating specific model improvements, especially when the marginal ppp-v is neither implausibly small nor reassuringly large.

3.3 Increasing power using the joint posterior predictive distribution of multiple statistics

We may try to increase the difficulty of correctness tests by modifying our choice of test quantity. One way to achieve this is with ancillary statistics, which yield posterior predictive p-values that are exactly uniform [36]. However, discovering ancillary (or approximately ancillary) statistics is often difficult. And if our workflow assumption above holds, then we expect the discovery of

ancillary statistics to become more difficult as our model size increases and our sampling distributions accommodate a greater variety of data behaviors. Since our primary concern is constructing tests for model rejection which can scale with model complexity, this is particularly worrying.

Similarly, the use of pivotal discrepancy measures (which may depend on parameters $\boldsymbol{\theta}$ as well as data $\mathbf{y}$) has been proposed since calibrating the corresponding p-values is easier [37, 54, 98]. But pivotal quantities may not exist when the observed data are not independent given the parameters, and there are no guarantees that pivotal quantities exist which quantify any particular feature of interest even when this assumption holds.

Another method for increasing the difficulty of model checks is to hold out some portion of the data with which the test quantity is computed and then compare this quantity to the model fit to the remainder of the data [91, 86]. Like the exactly calibrated p-value, this approach requires repeated sampling from the posterior distribution for different sets of observed data, which is often prohibitively computationally expensive. Faster approximate procedures have been proposed, but none of these can be applied across all types of models reliably [61, 60, 66].

A more general and easily-applied approach for generating harder tests of our models is obtained by using many test statistics at once. If $\mathcal{T} = \{T_s\}_{s=1}^{d}$ is a collection of test statistics, then the corresponding joint posterior predictive p-value is

$$\mathsf{joint}\text{-}p_{\mathcal{T}}(\mathbf{y}) = \mathbb{P}_{f(\mathbf{y}_{\text{rep}}|\mathbf{y})}\left(T_1(\mathbf{y}_{\text{rep}}) > T_1(\mathbf{y}) \text{ and } T_2(\mathbf{y}_{\text{rep}}) > T_2(\mathbf{y}) \text{ and } \cdots\ T_d(\mathbf{y}_{\text{rep}}) > T_d(\mathbf{y})\right). \tag{3.14}$$

3.3.1 Extracting usable information from the joint p-value

The first obstacle to using a joint p-value in practice is that we expect its observed value to shrink towards 0 as d increases even if the proposed model is correct. Furthermore, the joint p-value no longer satisfies Meng's bound (3.5). The first step towards making $\mathsf{joint}\text{-}p$ useful is thus a simple generalization of Meng's bound which applies to multiple test statistics.

Theorem 3 (Frequency Bound for joint$-p$)**.** *For any level $\alpha \in [0, 1]$, we have that*

$$\mathbb{P}_{f(\mathbf{y})}\left(\text{joint}-p_{\mathcal{T}}(\mathbf{y}) \leq \alpha\right) \leq \inf_{s \in [\alpha, 1]} \frac{\int_0^s F(t)dt}{s - \alpha}, \tag{3.15}$$

where F is the cumulative distribution function of the random variable

$$\mathbb{P}_{f(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})}\left[T_1(\mathbf{y}_{\text{rep}}) > T_1(\mathbf{y}) \textit{ and } T_2(\mathbf{y}_{\text{rep}}) > T_2(\mathbf{y}) \textit{ and } \cdots\ T_d(\mathbf{y}_{\text{rep}}) > T_d(\mathbf{y})\right]. \tag{3.16}$$

Proof. See Appendix B.1 □

To directly estimate the cumulative distribution function of joint$-p$ or p^{post} requires repeated simulation of the posterior predictive distribution $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$ for each draw of $\mathbf{y}$ from the model. In all but the simplest models, this requires sampling from $q(\boldsymbol{\theta} \mid \mathbf{y})$ for each such $\mathbf{y}$, which will often be prohibitively expensive. Theorem 3 shows that we can bound the cumulative distribution function of joint$-p$ by an optimum involving the cumulative distribution function of (3.16), which can be simulated directly and efficiently with draws from $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$.

Thus, Theorem 3 establishes that joint$-p$ can be interpreted for purposes of model rejection by computing a bound on its frequency. But we have yet to establish that joint$-p$ improves on p^{post} for rejection purposes in general. Since we expect F to increase more sharply at 0 as d increases, the bound (3.15) will generally get worse with increasing d for a fixed level of the joint p-value. Nevertheless, this bound can still provide value over p^{post} for rejection purposes if the nominal p-value (3.14) compensates by falling fast enough with d.

This can occur, for instance, when the values of $\boldsymbol{\theta}$ for which $f(\mathbf{y} \mid \boldsymbol{\theta})$ best fits each $T_s \in \mathcal{T}$ lie in mostly distinct subsets of the parameter space. In particular, define

$$\boldsymbol{\Theta}_{s,\alpha} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid \mathbb{P}_{f(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})}(T_s(\mathbf{y}_{\text{rep}}) \geq T_s(\mathbf{y})) \geq \alpha\}. \tag{3.17}$$

Each $\boldsymbol{\Theta}_{s,\alpha}$ can be thought of as the subspace corresponding to data generating processes for which the observed T_s is not atypical. If the $\boldsymbol{\Theta}_{s,\alpha}$ each have sufficient posterior probability for moderate

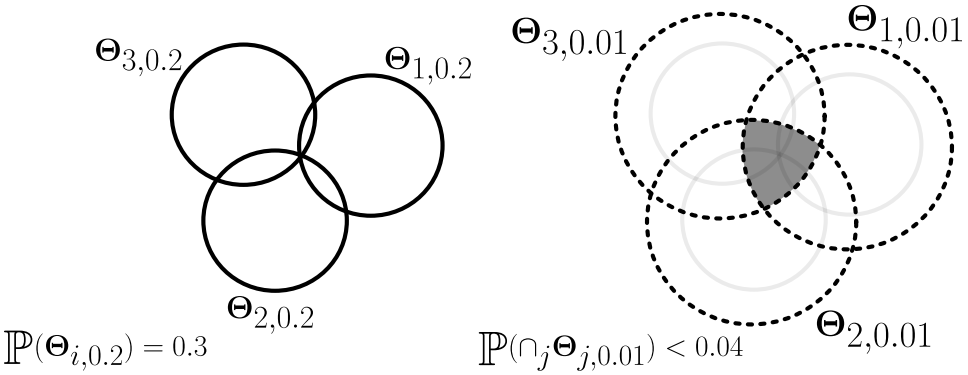

Figure 3.6: A schematic representation of how the marginal posterior predictive p-values can be relatively large while the joint p-value is small. In the left panel, because the $\boldsymbol{\Theta}_{s,0.2}$ have posterior probability 0.3, $p^{\text{post}}_{T_s}$ is bounded below by $0.3 \times 0.2 = 0.06$. In the right panel, because the intersection of the $\boldsymbol{\Theta}_{j,0.01}$ has posterior probability less than 0.04, joint$-p$ is bounded above by $0.01 \times 0.96 + 1 \times 0.04 < 0.05$.

α, then the corresponding p^{post} values will be too large to reject. Nevertheless, if the $\boldsymbol{\Theta}_{s,\alpha}$ also have small overlap, then the nominal value of joint$-p$ can be vanishingly small. In such a case, the bound (3.15) may still be sufficient to reveal the lack of fit. This situation is illustrated in Figure 3.6.

The second obstacle to using joint$-p$ comes in computing the quantities (3.14) and (3.15). The nominal value of joint$-p$ can be estimated for any $\mathcal{T}$ in the same manner as p^{post} is estimated for a single statistic. Because joint$-p$ will concentrate near 0 as d increases, the estimation of joint$-p$ may require a greater number of simulations from $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$ in order to resolve the order of magnitude to acceptable accuracy. However, in practice it often suffices to retain a fixed number of posterior draws $\boldsymbol{\theta}_i$ and take multiple draws from $f(\mathbf{y} \mid \boldsymbol{\theta}_i)$ for each i. Thus, the increase in computational overhead from this step is usually modest.

The estimation of the corresponding frequency bound (3.15) is more taxing, as we will usually not know the cumulative distribution function F in closed form. However, this function can be estimated with inexpensive Monte Carlo simulations of the joint model. Algorithm 6 describes the procedure, repeatedly estimating the empirical CDF of the random variable (3.16) conditional on $\boldsymbol{\theta}$ and then aggregating the results. In particular, we only sample from the prior and sampling distributions, never from the posterior, significantly speeding up the Monte Carlo operations com-

Algorithm 6 Joint p-value empirical CDF estimator

Require: Observed data $\mathbf{y}$, test statistics $\{T_s\}_{s=1}^{S}$, # of prior samples N_{prior}, # of sampling distribution samples M_{sampling}, # of samples at which to estimate CDF L_{estimate}.

for $n \leftarrow 1, \ldots, N_{\text{prior}}$ **do**
  Sample $\{\boldsymbol{\theta}^{(n)}\} \sim q\,(\boldsymbol{\theta})$.
  Sample $\left\{\mathbf{y}_{\text{rep}}^{(m)}\right\}_{m=1}^{M_{\text{sampling}}} \overset{iid}{\sim} f\left(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta}^{(n)}\right)$.
  for $l \leftarrow 1, \ldots L_{\text{estimate}}$ **do**
    Compute $\hat{p}^{(n,l)} = \frac{1}{M_{\text{sampling}}} \sum_{m=1}^{M_{\text{sampling}}} \mathbb{1}\left\{T_s\left(\mathbf{y}_{\text{rep}}^{(m)}\right) \geq T_s\left(\mathbf{y}_{\text{rep}}^{(l)}\right) \text{ for all } 1 \leq s \leq S\right\}$
  end for
  Compute $\hat{F}^{(n)} = q \mapsto \frac{1}{L_{\text{estimate}}} \sum_{l=1}^{L_{\text{estimate}}} \mathbb{1}\left\{\hat{p}^{(n,l)} \leq q\right\}$.
end for
Compute $\hat{F} = q \mapsto \frac{1}{N_{\text{prior}}} \sum_{n=1}^{N_{\text{prior}}} \hat{F}^{(n)}(q)$
return $\hat{F}$.

pared to exact calibration. Once we have our estimate $\hat{F}$, we can estimate the bound (3.15) using quadrature and a grid search for the optimum of $\int_0^s \hat{F}(t)dt/(s-\alpha)$. Since $\hat{F}$ is a one-dimensional function and $[\alpha, 1]$ is compact, this last step can generally be performed very quickly.

When the nominal observed value of $\mathsf{joint}{-}p$ is small, we will need a high resolution estimate of F near 0 in order to accurately estimate the optimum, and this can require large values of M_{sampling} and L_{estimate}. Because the complexity of the algorithm scales as

$$N_{\text{prior}} \times M_{\text{sampling}} \times L_{\text{estimate}},$$

the cost of estimating F will almost always dominate the computation. In practice, this cost can be substantially reduced by taking advantage of the fact that the computation can be carried out in parallel over the samples $\{\boldsymbol{\theta}^{(n)}\}$.

The reader may also notice that the $\hat{p}^{(n,l)}$ in Algorithm 6 are not independent across $1 \leq l \leq L_{\text{estimate}}$. While this does introduce correlation in the errors, the estimator remains asymptotically unbiased. In particular, [2] showed under weak regularity conditions that the $\sqrt{L_{\text{estimate}}}\left(\hat{F}^{(n)} - F^{(n)}\right)$ converge in distribution to centered Gaussian processes, where $F^{(n)}$ is the CDF of (3.16) conditional on $\boldsymbol{\theta}_n$.

We briefly mention a simply extension of the joint p-value which slightly simplifies computation, and which can provide further gains to power. This is achieved by computing the joint p-value with respect to a particular sampling distribution $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} \sim q(\cdot \mid \mathbf{y})$, rather than the posterior predictive distribution. This yields the sampled joint p-value $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ defined in (3.18).

$$\mathsf{joint}{-}p_{\mathcal{T}}(\mathbf{y}, \boldsymbol{\theta}) = \mathbb{P}_{f(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})}\left[T_1(\mathbf{y}_{\text{rep}}) > T_1(\mathbf{y}) \text{ and } T_2(\mathbf{y}_{\text{rep}}) > T_2(\mathbf{y}) \text{ and } \cdots\ T_d(\mathbf{y}_{\text{rep}}) > T_d(\mathbf{y})\right], \tag{3.18}$$

We emphasize that the resulting p-value is random, since $\boldsymbol{\theta}$ must be sampled from the posterior before (3.18) may be computed. When we test only a single statistic T this reduces to the sampled posterior p-value [39], another alternative to p^{post} designed to remedy its conservativity, to which we will compare the joint and sampled joint p-values in Section 3.3.3.

When $\mathbf{y}$ is drawn from the prior predictive distribution $f(\mathbf{y})$, the cumulative distribution function of the sampled joint p-value is exactly the $F(t)$ in the bound (3.15). Since estimating F represents the majority of the computational burden in calculating this bound, the sampled joint p-value is generally only slightly faster to compute than the bound on the ordinary joint p-value. However, by exploiting the skew of the distribution of (3.16) and obviating the need to optimize the integral in (3.15), the sampled joint p-value can have substantially higher power than the joint p-value alone.

3.3.2 Validating $\mathsf{joint}{-}p$ with non-positively associated extremes

We now examine the behavior of $\mathsf{joint}{-}p$ under different assumptions about how the exceedance events $\{T_s(\mathbf{y}_{\text{rep}}) \geq T_s(\mathbf{y})\}$ are associated under the posterior predictive distribution $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$ and the sampling distributions $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$. We test the behavior of our bound (3.15) under two conditions: an easier condition in which our test statistics are non-negatively associated under the sampling distributions $f(\mathbf{y} \mid \boldsymbol{\theta})$, which we can study with exact computations, and a

harder condition in which the test statistics can be non-positively associated, which we study with simulation experiments using a parametric model for the copula of the statistics.

Our main purpose is to establish quantitative evidence under reasonable assumptions for the intuition given after Theorem 3, viz., that our frequency bound (3.15) will in fact shrink to 0 as $d \to \infty$ when our extreme exceedances are not positively associated under the posterior predictive and have corresponding marginal p-values which are not too large. First, we must establish precisely what we mean by non-positively and non-negatively associated test statistics. To do this, we first generalize our definition of F from Theorem 3 to arbitrary random variables. If $(Y_1, \ldots, Y_d)$ are random variables with joint cumulative distribution function $\Phi(y_1, \ldots, y_d)$, then the cumulative distribution function $F_\Phi(t)$ of the random variable $\Phi(Y_1, \ldots, Y_d)$ is the Kendall function associated to the distribution Φ.

If we denote the joint CDFs associated to $(-T_1, \ldots, -T_d)$ under $f(\mathbf{y} \mid \boldsymbol{\theta})$ by $\Phi_{\boldsymbol{\theta}}$, then $F(t) = \mathbb{E}_{q(\boldsymbol{\theta})} F_{\Phi_{\boldsymbol{\theta}}}(t)$, and we can study the behavior of our frequency bound (3.15) by studying the Kendall functions $F_{\Phi_{\boldsymbol{\theta}}}(t)$. (We negate the test statistics in constructing the Kendall function simply to keep the inequality direction consistent with (3.15), but nothing of importance is changed since this direction is arbitrary.) Furthermore, if F_{Φ_1}, F_{Φ_2} are two Kendall functions, then Φ_2 is larger than Φ_1 in positive K-dependence order ($\Phi_1 \prec_{\text{PKD}} \Phi_2$) if $F_{\Phi_1}(t) \geq F_{\Phi_2}(t)$ for all $t \in [0, 1]$ [23].

To see that this ordering is related to the dependence structure of the distributions Φ, note that if the corresponding random variables $\{Y_i\}_{i=1}^d$ have medians $m_1, \ldots, m_d$ and are strongly positively associated, then there is a stronger probability of all the Y_i lying on the same side of their medians, and $\Phi(m_1, \ldots, m_d)$ can be relatively large. By contrast, if the $\{Y_i\}_{i=1}^d$ are strongly negatively associated, then we would expect some Y_i to be larger than their medians m_i and some to be smaller, resulting in much smaller $\Phi(m_1, \ldots, m_d)$. Combining these ideas, we expect $F_\Phi(t)$ to increase more rapidly when the $\{Y_i\}_{i=1}^d$ are negatively associated and more slowly when they are positively associated.

This idea can be formalized somewhat by noting that $\Phi_1 \prec_{\text{PKD}} \Phi_2$ implies that $\tau_1 \leq \tau_2$, where τ_i is the value of Kendall's tau associated to Φ_i. In two dimensions, Kendall's τ is a form of rank

correlation and is given by the formula

$$\tau = \mathbb{E}\text{sign}\left[\left(Y_1 - Y_1'\right)\left(Y_2 - Y_2'\right)\right],$$

where $(Y_1, Y_2), (Y_1', Y_2') \overset{iid}{\sim} \Phi$. This definition can be generalized to higher dimensions, where it measures an overall level and direction of association between the random variables $\{Y_i\}_{i=1}^d$ [52].

We will say simply that $(Y_1, \ldots, Y_d) \sim \Phi$ are positive K-dependent if $F_\Phi \prec_{\text{PKD}} F_\Psi$, where Ψ is the joint CDF corresponding to independent random variables $U_1, \ldots, U_d$. Because the Kendall function F_Ψ is independent of the marginal distributions of the U_i, we may take them to be uniform on $[0, 1]$. This simplification allows for direct calculation of F_Ψ, which is shown in [2] to be given by the formula

$$F_\Psi(t) = t\left[1 + \sum_{i=1}^{d-1} \frac{\log(1/t)^i}{i!}\right]. \tag{3.19}$$

With these facts established, we can now see that if our test statistics are positively associated in the sense of being positive K-dependent under $f(\mathbf{y} \mid \boldsymbol{\theta})$ for each $\boldsymbol{\theta}$, then our frequency bound (3.15) is further upper bounded by

$$\inf_{s \in [\alpha, 1]} \frac{\int_0^s F_\Psi(t)dt}{s - \alpha}. \tag{3.20}$$

We study the behavior of this bound under the following assumptions on our posterior predictive p-values:

1. The exceedances for our test statistics T_s are non-positively associated:

$$\begin{aligned}\mathbb{P}\left(T_s(\mathbf{y}_{\text{rep}}) \geq T_s(\mathbf{y}) \mid T_1(\mathbf{y}_{\text{rep}}) \geq T_1(\mathbf{y}), \ldots, T_{s-1}(\mathbf{y}_{\text{rep}}) \geq T_{s-1}(\mathbf{y})\right) \\ \leq \mathbb{P}\left(T_s(\mathbf{y}_{\text{rep}}) \geq T_s(\mathbf{y})\right),\end{aligned} \tag{3.21}$$

2. The posterior predictive p-values are upper bounded by some number $p \in (0, 1)$:

$$p_{T_s}^{\text{post}}(\mathbf{y}) \leq p \text{ for all } 1 \leq s \leq d. \tag{3.22}$$

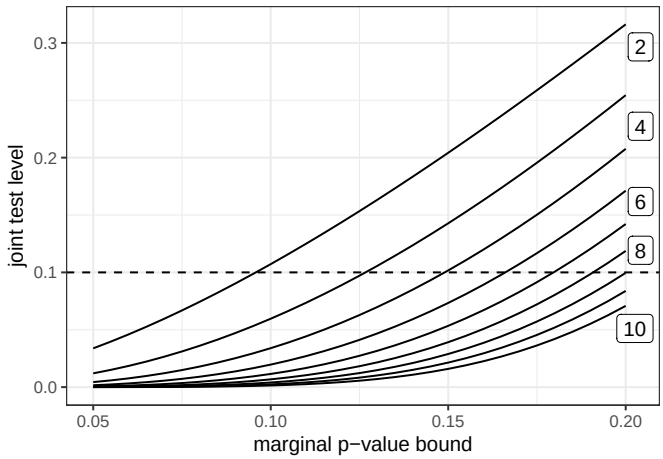

Figure 3.7: Upper bounds on the frequency of a nominal joint p-value p^d for non-negatively associated exceedances versus the bound on the posterior predictive p-values of the test statistics separately for various numbers d of test statistics. These curves can be interpreted as lower bounds on the level of a test that would always reject under our assumptions. This level decreases with the number of test statistics and with the bound on the marginal p-values.

Under these conditions, we clearly have that $\mathsf{joint}{-}p(\mathbf{y}) \leq p^d$. In order to assess the performance of $\mathsf{joint}{-}p$ under these assumptions, we plot the relation between (3.20) and the marginal bound p for dimensions $d = 2, \ldots, 10$ in Figure 3.7. For each value of the bound p, the level monotonically decreases as the number of test statistics increases. This suggests that, in this setting, the joint p-value is asymptotically successful in the sense that we will eventually reject the model if we have enough test statistics with non-positively associated observed extremal exceedances.

It is also apparent that the efficiency of this procedure depends strongly on the marginal p-value bound p. For smaller values of p, passing from two to three test statistics is sufficient to halve the resulting level of the test, but for larger values, the drop is less than a fifth. This may be particularly troubling since those cases where p is larger are exactly those in which p^{post} is least capable of model rejection. We note, however, that this represents a worst-case scenario in the sense that we make no assumptions about the gap in the inequalities (3.21). In practice, when these exceedance events are non-positively associated, we usually observe nominal values of $\mathsf{joint}{-}p$ smaller than p^d, and thus we can often obtain bounds that fall below the corresponding curve in Figure 3.7.

We also note that these conclusions continue to hold if we relax the assumption of positive

K-dependence for all $\boldsymbol{\theta}$ to the assumption that

$$\mathbb{E}_{q(\boldsymbol{\theta})} F_{\Phi_{\boldsymbol{\theta}}}(t) \leq F_{\Psi}(t) \text{ for } t < \epsilon$$

for some $\epsilon > 0$.

The results of this experiment suggest that our proposed joint p-value works when our test statistics are non-negatively associated under the model in the sense that, when our exceedance events $\{T_s(\mathbf{y}_{\text{rep}}) \geq T_s(\mathbf{y})\}$ are non-positively associated, the resulting frequency bound for $\mathsf{joint}{-}p$ shrinks to 0 as d grows. The situation for $\mathsf{joint}{-}p$ is harder, however, in situations when the test statistics are non-positively associated under the proposed model. In this case, smaller joint p-values are more common under the proposed model, so rejecting the model becomes harder.

Furthermore, investigating the behavior of $\mathsf{joint}{-}p$ in this setting is more challenging since there is no general upper bound available for the average Kendall function F. Instead, we use a parametric model of the test statistics $\{T_s\}_{s=1}^{d}$ to investigate the performance of $\mathsf{joint}{-}p$. In particular, we use a copula model, which specifies only the dependence structure of the statistics while leaving the marginal distributions arbitrary. As noted above, modeling only the dependence structure is possible since the Kendall function is independent of the marginal distributions.

We need our copula model to be defined in arbitrarily many dimensions and to allow for modeling negative association between the test statistics, particularly in the tails of their distribution. The Gaussian copula is a natural choice in this setting since it is definable in any dimension and can be parametrized to represent negative association between each component variable. It is also critical that the Gaussian copula has zero tail dependence - a measure of the positive association between component variables when one takes an extreme value - since the behavior of the Kendall function near 0 is particularly sensitive to these extremes. The t copula, for instance, can be parametrized to represent negative overall association, but the resulting copulas always exhibit positive dependence in the tails [51]. As a result, the corresponding Kendall functions are dominated by F_{Ψ} near zero.

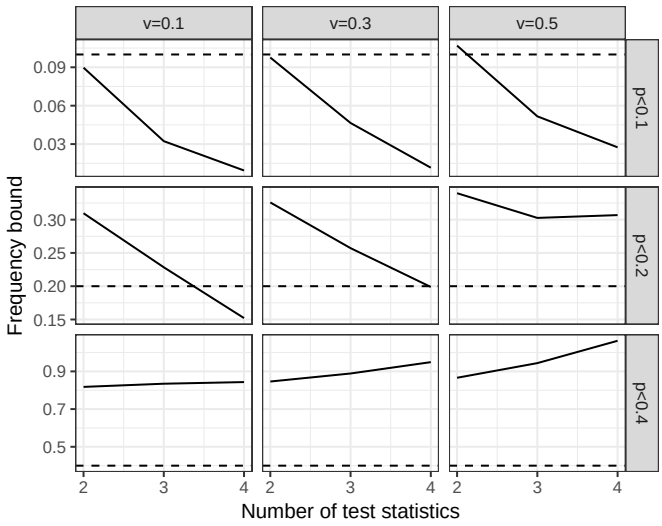

Figure 3.8: Upper bounds on frequency of joint p-value against dimension for negatively associated test statistics and for varying levels of negative dependence (columns) and varying bounds on the marginal posterior predictive p-values (rows).

The Gaussian copula Φ^G is parametrized by a correlation matrix R, which we define as

$$R_{ij} = \begin{cases} 1, & i = j \\ -v/(d-1), & i \neq j, \end{cases} \tag{3.23}$$

where d is the number of test statistics. When $v = 1$, this is the minimum value that results in a valid correlation matrix R when the off-diagonal entries are constant. In light of the above, we view this as a reasonably hard case for joint$-p$, since we assume that every pair of test statistics is negatively associated in the proposed model and we thus have that $\Phi^G \prec_{\text{PKD}} \Psi$.

There is no known analytic expression for the Kendall function of the Gaussian copula, so we estimate it using the empirical CDF of $\Phi^G\left(U_1^{(i)}, \ldots, U_d^{(i)}\right)$, where $(U_1^{(i)}, \ldots, U_d^{(i)})$ are Monte Carlo samples from Φ^G. Plugging this estimate in for F in (3.15), we can compute frequency bounds with $\alpha = p^d$ while varying the p^{post} bound p and the level of negative association v. Figure 3.8 plots the resulting bounds against dimension $d = 2, 3, 4$ for $p = 0.1, 0.2, 0.4$ and $v = 0.1, 0.3, 0.5$.

The effectiveness of joint$-p$ in this setting is now contingent on the combination of p and v. In the first row, for $p = 0.1$, the joint p-value continues to work well in the sense that the

frequency bound decreases with d fast enough to be significant at the 0.1 level for just three test statistics and for all tested values of v. In the second row, for $p = 0.2$, our bound falls with d and is below the corresponding bound for p^{post} for $v = 0.1$ and $v = 0.3$, indicating that the joint p-value is improving on our ability to reject the model. But for $v = 0.5$ the bound actually increases with d, indicating that our proposed procedure is no longer able to use the negative association in the observed extremal exceedances to reject the model. When p increases to 0.4, we find that the frequency bound increases with d regardless of the value of v. We conclude that $\mathsf{joint}{-}p$ can still provide value for model rejection when our test statistics are negatively associated under the proposed model, but, for this to be possible and efficient, we need one of the following to hold:

1. The corresponding marginal p-values are not too large.

2. The test statistics are not too negatively associated on average under $f(\mathbf{y} \mid \boldsymbol{\theta})$.

3. The exceedance events $\{T_s(\mathbf{y}_{\text{rep}}) \geq T_s(\mathbf{y})\}$ are sufficiently negatively associated under $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$ (i.e. the gap in the inequalities (3.21) must be sufficiently large).

The broad conclusions from these copula-based experiments continue to hold when the joint p-value is replaced by the sampled joint p-value. Since $\mathsf{joint}{-}p(\mathbf{y})$ is just the expected value of the sampled joint p-value $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ under $q(\boldsymbol{\theta} \mid \mathbf{y})$, the upper bound p^d for $\mathsf{joint}{-}p$ can be interpreted as an upper bound on the average of $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$. We may still hope that, due to the skewness of the distribution of $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ under $\boldsymbol{\theta} \sim q(\boldsymbol{\theta} \mid \mathbf{y})$, the median of $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ over $q(\boldsymbol{\theta} \mid \mathbf{y})$ could tend to zero faster than the mean $\mathsf{joint}{-}p(\mathbf{y})$, and could thus achieve better performance after calibration. However, this is only possible if the distribution of $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ has a sufficiently heavy right tail under $q(\boldsymbol{\theta} \mid \mathbf{y})$, as measured by the coefficient of variation $\mathsf{cv} = \sigma/\mu$, where μ and σ are the mean and standard deviation of this distribution. If $\mathsf{cv} < C$ for some fixed, finite $C > 0$ for all choices of test statistics, then the Paley-Zygmund inequality implies a nonvanishing probability of obtaining sampled joint p-values at least a constant fraction of the

mean. In particular, we have for $\delta \in (0, 1)$ that

$$\mathbb{P}_{q(\boldsymbol{\theta}|\mathbf{y})}\left[\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta}) \geq (1-\delta)p^{\text{post}}(\mathbf{y})\right] \geq \frac{1-\delta}{C+1-\delta}. \tag{3.24}$$

Thus, we do not generally expect typical sampled joint p-values to tend to zero at a faster exponential rate than the joint p-value. To analyze the *calibrated* sampled joint p-values, the bounds (3.15) would be replaced by the evaluated cumulative distribution functions $F(\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta}))$ in Figures 3.7 and 3.8. These are smaller than the corresponding bounds (since they avoid the optimization step in (3.15)), but generally only by a factor of about 2. Thus, while $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ is usually smaller than $\mathsf{joint}{-}p(\mathbf{y})$, the trends established in this section - particularly the relationship between p-value performance and dependence among test statistics - generally apply to $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ as well.

3.3.3 Comparison to other Bayesian p-values

[44] propose to overcome the conservativity of p^{post} by plugging it into (an estimate of) its distribution function, which will result in a uniformly distributed quantity when the model is correctly specified. In particular, if H is the distribution function of $p_{\mathbf{y}}^{\text{post}}$ when $\mathbf{y}$ is drawn from the prior predictive distribution, then we can estimate H by the empirical distribution function $\hat{H}(p) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\left\{p_T^{\text{post}}(\mathbf{y}_{\text{rep}}^{(s)}) \leq p\right\}$, where $\left\{\mathbf{y}_{\text{rep}}^{(s)}\right\}_{s=1}^{S} \overset{iid}{\sim} p(\mathbf{y})$ is a sample from the prior predictive distribution. The calibrated posterior predictive p-value is then

$$\mathsf{cal}-p_{\mathbf{y}} = \hat{H}\left(p_T^{\text{post}}(\mathbf{y})\right). \tag{3.25}$$

This calibration step fully resolves the conservativity problem when H is well-estimated by $\hat{H}$. However, the computation of $\hat{H}$ generally requires sampling from $q(\boldsymbol{\theta} \mid \mathbf{y}_{\text{rep}}^{(s)})$ separately for each $s = 1, \ldots, S$. This can quickly become computationally infeasible for moderate S if the model is sufficiently complex.

The sampled posterior predictive p-value retains exact uniformity for continuously distributed

test statistics T, but it is almost universally easy to compute [39, 55, 53]. Unlike the other approaches we consider, this method generates a *random* p-value by first drawing a sample $\boldsymbol{\theta}$ from the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$ and then computing a p-value with respect to $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$. In symbols:

$$\mathsf{sampled}{-}p_T(\mathbf{y}, \boldsymbol{\theta}) = \mathbb{P}_{f(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})}\left(T(\mathbf{y}_{\text{rep}}) \geq T(\mathbf{y})\right), \quad \boldsymbol{\theta} \sim q(\boldsymbol{\theta} \mid \mathbf{y}). \tag{3.26}$$

The posterior predictive p-value is just the expected value of $\mathsf{sampled}{-}p_T$ over the posterior distribution. Estimating (3.26) by a Monte Carlo average is generally extremely fast since sampling $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$ is trivial for most models. The uniformity of $\mathsf{sampled}{-}p_T$ is achieved by not aggregating over the posterior distribution. However, this calibration comes at the cost of increased variance due to the additional randomization of $\boldsymbol{\theta}$. As a result, the sampled p-value can exhibit lower power when this variance is large.

In particular, this property means that the sampled p-value, like the posterior predictive p-value, generally scales poorly with model expansion. To see this quantitatively, recall that $p_T^{\text{post}}(\mathbf{y})$ tends to become more concentrated around $1/2$ with model expansion. Defining $Q(p) = \mathbb{P}_{f(\mathbf{y})}\left(p_T^{\text{post}}(\mathbf{y}) \leq p\right)$, we think of the calibrated p-value $Q(p_T^{\text{post}}(\mathbf{y}))$ as representing the true level of fitness of the model to T. The contraction of the distribution of $p_T^{\text{post}}(\mathbf{y})$ then implies that the nominal value $p_T^{\text{post}}(\mathbf{y})$ must increase with model expansion in order to maintain a fixed true level of fitness $Q(p^{\text{post}}(\mathbf{y}))$. Then, for any $\delta < p^{\text{post}}(\mathbf{y})$, the Paley-Zygmund inequality implies that

$$\mathbb{P}\left(\mathsf{sampled}{-}p_T(\mathbf{y}, \boldsymbol{\theta}) > \delta\right) \geq \phi\left(4\left[p_T^{\text{post}}(\mathbf{y}) - \delta\right]\right), \tag{3.27}$$

where $\phi = \frac{x^2}{1+x^2}$. Because ϕ is an increasing function, this implies that our lower bound (3.27) - and thus the risk of observing large sampled p-values - will also tend to increase with model expansion if the true level of fitness is nondecreasing.

The broad conclusions from these copula-based experiments continue to hold when the joint p-value is replaced by the sampled joint p-value. Since $\mathsf{joint}{-}p(\mathbf{y})$ is just the expected value of the sampled joint p-value $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ under $q(\boldsymbol{\theta} \mid \mathbf{y})$, the upper bound p^d for $\mathsf{joint}{-}p$ can be

interpreted as an upper bound on the average of $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$. We may still hope that, due to the skewness of the distribution of $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ under $\boldsymbol{\theta} \sim q(\boldsymbol{\theta} \mid \mathbf{y})$, the median of $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ over $p(\boldsymbol{\theta} \mid \mathbf{y})$ could tend to zero faster than the mean $\mathsf{joint}{-}p(\mathbf{y})$, and could thus achieve better performance after calibration. However, this is only possible if the distribution of $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ has a sufficiently heavy right tail under $q(\boldsymbol{\theta} \mid \mathbf{y})$, as measured by the coefficient of variation $\mathsf{cv} = \sigma/\mu$, where μ and σ are the mean and standard deviation of this distribution. If $\mathsf{cv} < C$ for some fixed, finite $C > 0$ for all choices of test statistics, then the Paley-Zygmund inequality implies a nonvanishing probability of obtaining sampled joint p-values at least a constant fraction of the mean. In particular, we have for $\delta \in (0, 1)$ that

$$\mathbb{P}_{q(\boldsymbol{\theta}|\mathbf{y})}\left[\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta}) \geq (1-\delta)p^{\text{post}}(\mathbf{y})\right] \geq \frac{1-\delta}{C+1-\delta}. \tag{3.28}$$

Thus, we do not generally expect typical sampled joint p-values to tend to zero at a faster exponential rate than the joint p-value. To analyze the *calibrated* sampled joint p-values, the bounds (3.15) would be replaced by the evaluated cumulative distribution functions $F(\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta}))$ in Figures 3.7 and 3.8. These are smaller than the corresponding bounds (since they avoid the optimization step in (3.15)), but generally only by a factor of about 2. Thus, while $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ is usually smaller than $\mathsf{joint}{-}p(\mathbf{y})$, the trends established in this section - particularly the relationship between p-value performance and dependence among test statistics - generally apply to $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\theta})$ as well.

Other methods have been proposed that trade exact calibration for approximate calibration and improved computability (compared to $\mathsf{cal} - p$) or power (compared to $\mathsf{sampled}{-}p$). [6] propose p-values which are Bayesian in the sense that they account for posterior uncertainty but which enjoy reduced conservativity relative to p^{post} by having a uniform frequency distribution in appropriate asymptotics. The key idea for achieving asymptotic uniformity comes from the observation that p^{post} involves a double use of the data whereby the posterior "sees" the statistic T against which it will subsequently be tested. This artificially reduces the difficulty of the test,

leading to conservativity.

This diagnosis is partly justified by considering tests with ancillary statistics T. Since these have distributions which are independent of $\boldsymbol{\theta}$, the posterior contains no information about T, and p^{post} becomes exactly uniform for such T. The proposed p-values attempt to formalize the idea of "removing" the information in T from the posterior before testing. The first of these is the conditional predictive p-value, defined for a test statistic T as

$$\text{cond}-p_T(\mathbf{y}) = \mathbb{P}_{f\left(\mathbf{y}_{\text{rep}} \mid \hat{\boldsymbol{\theta}}_T\right)}\left(T(\mathbf{y}_{\text{rep}}) \geq T(\mathbf{y})\right), \tag{3.29}$$

where we define $\hat{\boldsymbol{\theta}}_T = \arg\max f\left(\mathbf{y} \mid \boldsymbol{\theta}, T(\mathbf{y})\right)$ as the T-conditional maximum likelihood estimate of $\boldsymbol{\theta}$, and

$$f\left(\mathbf{y}_{\text{rep}} \mid \hat{\boldsymbol{\theta}}_T\right) = \int f\left(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta}, T(\mathbf{y})\right) q\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_T\right) d\boldsymbol{\theta}. \tag{3.30}$$

The key idea in this definition is that $\hat{\boldsymbol{\theta}}_T$ should capture as much of the information about $\boldsymbol{\theta}$ contained in the data as possible while excluding the information in T. When $\hat{\boldsymbol{\theta}}_T$ is sufficient for $\boldsymbol{\theta}$, cond$-p_T$ is exactly uniform. However, forming and conditioning on the conditional MLE is often computationally difficult. We can instead try to remove the information contained in T from the posterior directly by conditioning T out of the likelihood. This results in Bayarri and Berger's partial predictive p-value:

$$\text{part}-p_T\left(\mathbf{y}\right) = \mathbb{P}_{f\left(\mathbf{y}_{\text{rep}} \mid \mathbf{y} \setminus T(\mathbf{y})\right)}\left(T(\mathbf{y}_{\text{rep}}) > T(\mathbf{y})\right), \tag{3.31}$$

where we define the partial posterior and posterior predictive distributions as

$$q(\boldsymbol{\theta} \mid \mathbf{y} \setminus T(\mathbf{y})) \propto f(\mathbf{y} \mid \boldsymbol{\theta}, T(\mathbf{y}))q(\boldsymbol{\theta}), \quad f(\mathbf{y}_{\text{rep}} \mid \mathbf{y} \setminus T(\mathbf{y})) = \int f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})q(\boldsymbol{\theta} \mid \mathbf{y} \setminus T(\mathbf{y}))d\boldsymbol{\theta}. \tag{3.32}$$

Since $T(\mathbf{y})$ is determined exactly by $\mathbf{y}$, the partial posterior differs from the posterior by a factor proportional to $f(T(\mathbf{y}) \mid \boldsymbol{\theta})^{-1}$.

That these p-values approximately succeed in removing the conservativity problem is justified

by Theorem 2 of [80], which implies that cond$-p$ and part$-p$ both have asymptotically uniform frequency distributions under sampling models of the form

$$f\left(\mathbf{y} \mid \boldsymbol{\theta}, \psi_n\right) = \prod_{i=1}^{n} f_i\left(\mathbf{y}_i \mid \boldsymbol{\theta}, \psi_n\right), \tag{3.33}$$

where $\psi_n \in \mathbb{R}$ is a one-dimensional nuisance parameter. Robins et al. also propose a number of other methods for deriving approximately calibrated p-values which depend on either modifications of the test statistic T or on approximate recalibrations of simpler p-values such as p^{post}. We do not treat these approaches in detail here since any generally available computational speedups relative to part$-p$ and cond$-p$ are usually achieved by exploiting some aspect of the asymptotics of (3.33), which we argue in the next section is an overly limiting model in many cases. The interested reader can consult [80] for details. Because part$-p$ and cond$-p$ have identical asymptotic performance under (3.33) and part$-p$ is generally easier to compute, we will focus all subsequent comparisons on part$-p$.

We note that part$-p$ can still suffer from substantial computational costs when $f(T(\mathbf{y}) \mid \boldsymbol{\theta})^{-1}$ is not analytically available, which is usually the case when the model is sufficiently complex. We are unaware of a scheme for estimating this quantity in general other than estimating $f(T(\mathbf{y}) \mid \boldsymbol{\theta})$ with a kernel density estimator and inverting the result (which is the recommended strategy in [5]). Such kernel density estimates can be highly inefficient in the tails of the density, leading to explosive errors in the inverse.

We now turn to a comparison between joint$-p$, part$-p$, sampled$-p$, and cal $- p$ in two simple simulation examples. In the first example, we examine a case where the joint p-value is more powerful than less computationally expensive alternatives while getting close to the power of approaches which are much more computationally costly. In the second example, we present a case that is challenging for the joint p-value and demonstrate that the sampled joint p-value can consistently dominate both joint$-p(\mathbf{y})$ and sampled$-p(\mathbf{y}, \boldsymbol{\theta})$.

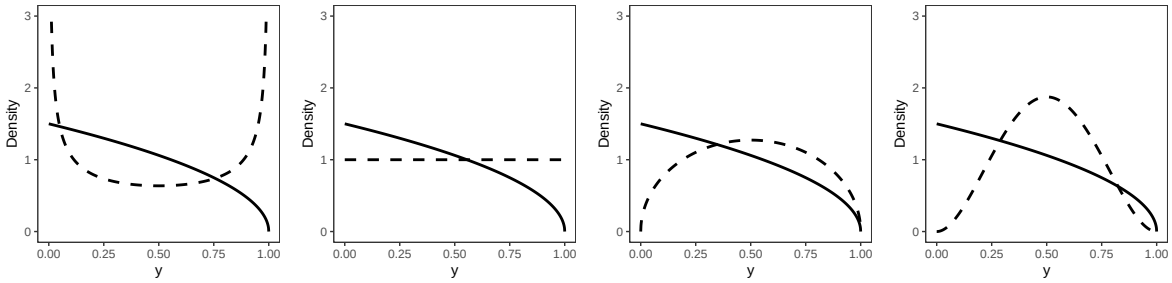

Figure 3.9: Solid: the beta$(1, 1.5)$ distribution from which our observed data was generated. Dashed: our assumed beta(θ, θ) sampling distribution for $\theta = 0.5, 1, 1.5, 3$.

Testing the quantiles of a beta population

We take our observed data to be a random sample of size $N = 100$ from a beta $(1, 1.5)$ distribution. We will assume a misspecified model with a beta (θ, θ) sampling distribution and a uniform prior over $[0, 3]$. While our assumed sampling distribution is symmetric for all values of θ, the true data generating distribution is substantially skewed to the right, as shown in Figure 3.9. We test our assumed model with statistics T_1 and T_2 taken to be the 0.05 and 0.95 sample quantiles respectively.

Qualitatively, values of θ closer to zero yield sampling distributions that better match the observed lower quantile but overshoot the observed upper quantile. Similarly, larger values of θ better match the observed upper quantile but now overshoot the observed lower quantile. If the posterior splits the difference between these regions of parameter space, then we should expect that both observed quantiles will be lower than what is typical in posterior predictive replications from the assumed model. Indeed, this is precisely what we see when we compute the probabilities of $T_s(\mathbf{y}_{\text{rep}}) \leq T_s(\mathbf{y})$ for $s = 1, 2$, yielding posterior predictive p-values of ≈ 0.07 for both test statistics.

Computing our bound for joint$-p$ and cal $- p$ requires estimating distribution functions of certain exceedance probabilities, and computing part$-p$ requires estimating the partial posteriors (3.32) for T_1 and T_2. Figure 3.10 displays the estimated distribution function $\hat{F}$ around the nominal joint p-value $\alpha = 0.0028$ along with the optimization objective $\int_\alpha^s \hat{F}(s)/(s - \alpha)$. This shows in

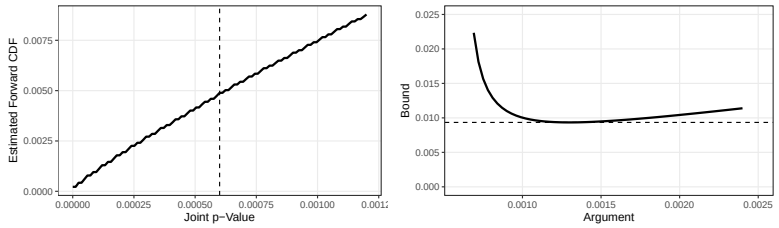

Figure 3.10: Left: the estimated distribution function of joint extremal exceedances (3.16) with vertical line indicating the nominal joint p-value. Right: the optimization objective on the right-hand side of (3.15) for a range of s with horizontal line at the optimum.

$p^{\text{post}}_{T_s}$ (bound)	joint$-p$ (bound)	part$-p_{T_s}$	median sampled$-p_{T_s}$	cal $- p_{T_s}$
0.067 (0.135), 0.027 (0.055)	0.0006 (0.009)	0.021, 0.009	0.039, 0.018	0.025, 0.012

Table 3.1: Candidate p-values and corresponding bounds (in parentheses, where applicable) for T_1 and T_2 equal to the 0.05 and 0.95 sample quantiles. The partial and calibrated p-value give strongest evidence against the model, followed by the joint p-value, median sampled p-value, and posterior predictive p-value.

particular that we have estimated the distribution F to sufficient resolution around α to trust our estimated bound. Plots of estimated intermediate quantities for part$-p$ and cal $- p$ are given in the appendix.

Table 3.1 displays estimates of the various candidate p-values for this problem. Because the nominal joint p-value is two orders of magnitude smaller than either of the p^{post} values, we achieve a frequency upper bound which matches the smaller of the two partial p-value and calibrated p-values. Furthermore, while joint$-p$ automatically controls for multiple testing by accounting for dependence, the other results have somewhat artificially higher power due to a lack of multiple testing adjustment.

Compared to the sampled p-value, our frequency bound is less than median sampled p-value for either statistic. More importantly, the sampled p-value is a random quantity and is larger than our frequency bound in more than 66% of samples for either statistic. Figure 3.11 displays the survival function of the sampled p-values for T_1 and T_2 along with various other p-values, showing that the sampled p-value is only less conservative on average compared to the posterior predictive p-value. Thus, while the sampled p-value is much easier to compute than our bound, it is more

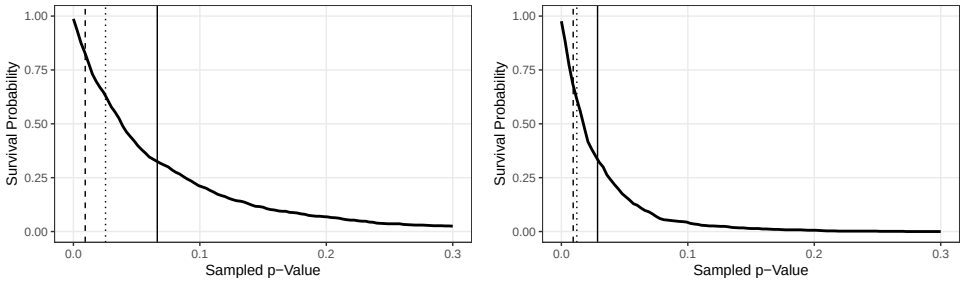

Figure 3.11: Survival functions for the sampled p-value computed for test statistics T_1 and T_2 equal to the 0.05 and 0.95 sample quantiles respectively. The solid, dashed and dotted lines represent the corresponding posterior predictive, joint, and calibrated p-values respectively. The sampled p-value is less conservative than p^{post} on average in this problem, but more conservative than $\mathsf{joint}-p$ and $\mathsf{cal}-p$ on average.

conservative in this problem.

Unsurprisingly, at about a sixth the magnitude, the joint p-value bound is a substantial improvement over the corresponding frequency bounds for the individual posterior predictive p-values. Overall, $\mathsf{joint}-p$ displays a useful balance of trade-offs in this problem. It is substantially less conservative than the alternatives which are easier to compute, and it comes within a factor of two of the more powerful alternatives while being easier to compute and offering preasymptotic guarantees.

Testing for large effects in a linear regression model

We now consider a conjugate Bayesian linear regression model of the following form.

$$\mathbf{y} \mid \boldsymbol{\beta} \sim \mathsf{normal}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right), \quad \boldsymbol{\beta} \sim \mathsf{normal}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \tag{3.34}$$

where σ^2 and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ are taken to be fixed hyperparameters for simplicity, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ is known matrix of covariates. We construct the prior covariance as

$$\boldsymbol{\Sigma}_{ij} = \begin{cases} 1, & i = j \\ \rho, & \{i, j\} = \{1, 2\} \\ 0, & \text{otherwise} \end{cases}. \tag{3.35}$$

In other words, $\boldsymbol{\Sigma}$ is the identity matrix except for the $(1, 2)$ and $(2, 1)$ entries, which encode a possibly nonzero prior correlation between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. We test the fit of this model to data generated from the following process.

$$\mathbf{y} \mid \boldsymbol{\beta}^* \sim \mathsf{normal}\left(\mathbf{X}\boldsymbol{\beta}^*, \sigma^2 \mathbf{I}\right), \quad \{\mathbf{X}_{ij}\} \overset{iid}{\sim} \mathsf{normal}(0, 1), \tag{3.36}$$

where $[\boldsymbol{\beta}^*]_j = 4$ if $j = 1, 2$ and $[\boldsymbol{\beta}^*]_j = 1$ if $j \geq 3$. The model (3.34) is misspecified for this data generating process in the sense that values of $\boldsymbol{\beta}_j$ as large as 4 are extremely rare under the prior covariance (3.35).

In our experiment, we generate $n = 200$ observations from $d = 100$ covariates. To test the fit of the model (3.34) to datasets generated from the process (3.36), we employ test statistics $T_1(\mathbf{y}) = \mathbf{X}_1^T \mathbf{y}$ and $T_2(\mathbf{y}) = \mathbf{X}_2^T \mathbf{y}$, where $\mathbf{X}_j$ is column j of $\mathbf{X}$. In particular, we compare the sampled p-value with respect to T_1, the joint p-value with respect to $\{T_1, T_2\}$, and the sampled joint p-value with respect to the same statistics. We are particularly interested in the relative performance of these p-values as we vary the correlation hyperparameter ρ over the interval $[0, -0.8]$.

The law of total covariance implies that

$$\begin{aligned} \mathrm{Cov}\left(T_1(\mathbf{y}_{\text{rep}}), T_2(\mathbf{y}_{\text{rep}}) \mid \mathbf{y}\right) = \mathbb{E}\left[\mathrm{Cov}\left(T_1(\mathbf{y}_{\text{rep}}), T_2(\mathbf{y}_{\text{rep}}) \mid \boldsymbol{\beta}\right) \mid \mathbf{y}\right] \\ + \mathrm{Cov}\left(\mathbb{E}\left[T_1(\mathbf{y}_{\text{rep}}) \mid \boldsymbol{\beta}\right], \mathbb{E}\left[T_2(\mathbf{y}_{\text{rep}}) \mid \boldsymbol{\beta}\right] \mid \mathbf{y}\right) \end{aligned} \tag{3.37}$$

The first term on the right-hand side affects the slope of the distribution function $F(t)$ in the

bound (3.15) near 0 (for the reasons discussed in Section 3.3.2), whereas the second term affects the magnitude of the ratio of joint$-p_{\{T_1,T_2\}}$ to the $p_{T_i}^{\text{post}}$ (since, e.g., we expect a negative covariance between the conditional expectations to increase the gap in (3.21)). As the computations in Section 3.3.2 demonstrated, the joint p-value bound is most powerful when it can exploit a combination of relatively slowly increasing $F(t)$ and a very small value of joint$-p$, as is most likely to occur when the first covariance on the right-hand side of (3.37) is positive and the second is negative.

With our model (3.34) and our chosen test statistics, these two covariances are given by $\sigma^2[\mathbf{X}^T\mathbf{X}]_{12}$ and

$$\sigma^2[\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\mathbf{X}^T\mathbf{X}]_{12}, \tag{3.38}$$

respectively. When the influence of the prior is not too strong, the latter approaches the former. And in our simulation experiments, these covariances always had the same sign when $\rho = 0$, leading to either relatively large values of joint$-p$ or relatively steep distribution functions $F(t)$. This is thus a particularly challenging problem for joint$-p$. Decreasing ρ introduces additional negative correlation between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in the posterior, which allows for cases in which $[\mathbf{X}^T\mathbf{X}]_{12}$ is positive but (3.38) is negative and creates more favorable conditions for joint$-p$.

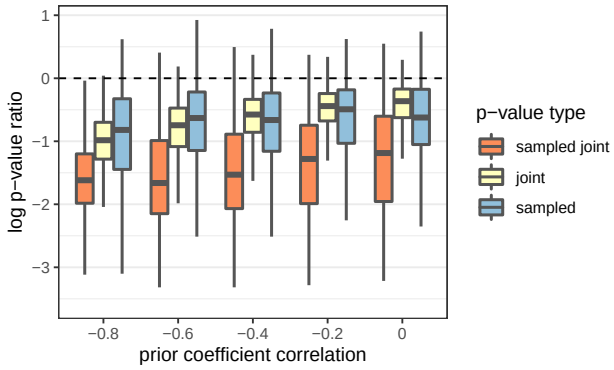

Figure 3.12: Box plots of the log ratio of candidate p-values to twice the posterior predictive p-value under repeated sampling of the regression model (3.36).

Figure 3.12 displays box plots for the log ratio of the joint, sampled, and sampled joint p-values to twice the posterior predictive p-value as a function of $\rho \in [0, -0.8]$. Normalizing by $2p^{\text{post}}$ in this way adjusts for random variation in the fitness of the model (3.34) to test statistics

sampled from (3.36), allowing for a direct comparison of relative performance. Values below the dashed line at 0 indicate p-values smaller than $2p^{\text{post}}$. For ρ near 0, $\mathsf{joint}{-}p(\mathbf{y})$ tends to perform worse than $\mathsf{sampled}{-}p(\mathbf{y}, \boldsymbol{\beta})$ while exhibiting substantially lower variance, reflecting the relative lack of joint posterior information for $\mathsf{joint}{-}p(\mathbf{y})$ to exploit. However, as we expect from the above discussion, the performance of $\mathsf{joint}{-}p$ is substantially more sensitive to decreasing ρ than the performance of $\mathsf{sampled}{-}p$. Consequently, for $\rho \leq -0.6$, the median $\mathsf{joint}{-}p$ is eventually smaller than the median $\mathsf{sampled}{-}p$.

On the other hand, we see that the sampled joint p-value $\mathsf{joint}{-}p(\mathbf{y}, \boldsymbol{\beta})$ is able to dominate both $\mathsf{joint}{-}p(\mathbf{y})$ and $\mathsf{sampled}{-}p(\mathbf{y}, \boldsymbol{\beta})$ across all values of ρ, with a median value at least half an order of magnitude smaller than the median sampled or joint p-values in each case. This makes sense, since combining these strategies can attenuate their individual weaknesses. Compared to $\mathsf{joint}{-}p(\mathbf{y})$, the sampled joint p-value is not as sensitive to skewness in the distribution of (3.16) and can be calibrated exactly rather than relying on bounds. Compared to $\mathsf{sampled}{-}p(\mathbf{y}, \boldsymbol{\beta})$, the sampled joint p-value can exploit negative posterior predictive dependence between test statistics to obtain more consistently small p-values across all values of $\boldsymbol{\beta}$, resulting in reduced variance compared to $\mathsf{sampled}{-}p$.

Chapter 4: Challenges to inference

In the last section, we noted that the posterior variance of the conditional p-values $p_T(\boldsymbol{\theta})$ tends to increase under the process of model expansion for any test statistic T. This tendency arose from a nonnegative term appearing in a decomposition of this variance which encoded variation in the distribution of T driven by the new parameters $\boldsymbol{\lambda}$ introduced in the expanded model. As we show in Section 4.3, a similar tendency continues to hold not just for the conditional p-values $p_T(\boldsymbol{\theta})$, but for the data distribution $f(\cdot \mid \boldsymbol{\theta})$ itself. While the former is a scalar quantity, the latter is a distribution, and thus we cannot speak about the variance of $f(\cdot \mid \boldsymbol{\theta})$ under sampling $\boldsymbol{\theta} \sim q(\boldsymbol{\theta} \mid \mathbf{y})$ in the same way. Nevertheless, we will define an appropriate analog of the posterior variance for these distributions and show a similar tendency towards greater complexity under conditions of model expansion. In particular, if we think of the posterior predictive distribution

$$f(\mathbf{y}_{\text{rep}} \mid \mathbf{y}) = \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})} f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta}) \tag{4.1}$$

as the average of these data distributions, then our decomposition will show that, for $\boldsymbol{\theta}$ randomly sampled from the posterior, $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$ will tend to be further from its average $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$ after model expansion.

Yet, as in the case of the conditional p-value, this tendency is not fate. This tendency is represented through the addition of a nonnegative term analogous to the variance term

$$\text{Var}_q \left\{ \mathbb{E} \left[p_T(\mathbf{y}, (\boldsymbol{\theta}, \boldsymbol{\lambda})) \mid \mathbf{y}, \boldsymbol{\lambda} \right] \mid \mathbf{y} \right\} \geq 0 \tag{4.2}$$

in the decomposition of the posterior p-value variance $\text{Var}_{q(\boldsymbol{\theta}|\mathbf{y})}(p_T(\boldsymbol{\theta}))$ under model expansion given in chapter 3. In any given case, the additional nonnegative term could be offset by a cor-

responding decrease in another term. This leaves open the question of whether there are generic strategies for model expansion that take advantage of this possibility and attenuate the challenges to model checking posed by an increasing parameter dimension.

Understanding whether such strategies are available is of substantial practical importance. As we demonstrated in the last chapter, while it may be possible to overcome the challenges of high dimensions to model checking with more sophisticated tools, these tools come with an increased computational and interpretational burden, and thus the applied modeler should always prefer to avoid, rather than compensate for, these potential challenges whenever possible. On the other hand, if these challenges are *not* generally avoidable through a more careful choice of model, then the impetus for developing, maintaining, and disseminating more sophisticated tools is that much greater.

In order to make a first attempt at answering this question, we relate it to another challenge long apparent in the high-dimensional context - that of parameter identification. In Section 4.4, we formalize the notion of parameter identification for Bayesian models and present a decomposition which shows a tendency for identification to weaken after model expansion. As in the decomposition of the conditional p-value variance, our decomposition also shows that this tendency towards weakening identification can be offset in particular cases by other terms. Thus, a parallel question arises: can weakening identification be generally avoided through careful choice of expansion?

In this chapter, we develop an argument to the effect that, absent the careful introduction of addition constraints through the prior distribution, we should not expect to be able to avoid the challenges of dimensionality to model checking through our choice of model without exacerbating the challenges of dimensionality with respect to parameter identification. In cases where one or the other of these problems is not a concern (e.g. in pure prediction problems where inference for unobserved parameters is not of interest), this trade-off may be acceptable. In general, however, we view this as a fundamental limit on the extent to which the challenges of high dimensional modeling, taken as a whole, can be circumvented by the modeler.

In Section 4.1, we present a simplified regression model to develop intuition for the relevant

trade-offs. Then in Section 4.2 we review related ideas from the existing literature. Sections 4.3 and 4.4 develop the necessary theoretical constructions to study model checking and parameter identification formally. We then present our main result on the trade-off between these challenges in Section 4.5 before presenting our main conclusions in Section 4.6.

4.1 A simple regression example

Suppose we have two observations (y_1, y_2) and known residual variance $\sigma^2 = 1$. In our first model, we have one predictor $\mathbf{x}_1 = (0, 1)$ with coefficient β_1. Assigning a normal prior, the model is

$$y_j \mid \beta_1 \sim \text{normal}\left(\beta_1 x_{1j}, 1\right) \text{ for } j = 1, 2, \quad \beta_1 \sim \text{normal}\left(0, \sigma_b\right), \tag{4.3}$$

where the hyperparameter σ_b is taken large so that the prior is weakly informative. We expand this model by adding a second predictor $\mathbf{x}_2$ with $\|\mathbf{x}_2\| = 1$ and with coefficient β_2. Assuming β_2 is a priori independent of β_1 and assigning an identical marginal prior, we get

$$y_j \mid \beta_1, \beta_2 \sim \text{normal}\left(\beta_1 x_{1j} + \beta_2 x_{2j}, 1\right) \text{ for } j = 1, 2, \quad \beta_1, \beta_2 \overset{iid}{\sim} \text{normal}\left(0, \sigma_b\right). \tag{4.4}$$

We consider five levels of correlation between the predictors $\mathbf{x}_1$ and $\mathbf{x}_2$, which are plotted in the top row of Figure 4.1. To assess the effect of adding a predictor on identifiability, we compare the marginal posterior and prior of β_1, plotted in the second row of Figure 4.1. As we expect, as predictor correlation increases, the identification of the coefficient β_1 decreases.

It is less obvious how we should assess the falsifiability of the model. We argue in Section 4.3 that falsifiability is connected to a measure of posterior confidence about the true data generating process, expressed as a distribution over an independent draw or replication of the data, $\mathbf{y}_{\text{rep}}$. In particular, we will argue that falsifiability tends to decrease as the sampling distributions $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\beta})$ and the posterior predictive distribution $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y}) = \int f\left(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\beta}\right) q\left(\boldsymbol{\beta} \mid \mathbf{y}\right) d\boldsymbol{\theta}$ become more dissimilar. The third row of Figure 4.1 visualizes this by plotting the sampling distributions for a replicated first observation y_1^{rep} at the posterior mean $\overline{\boldsymbol{\beta}}$ with the posterior predictive distributions.

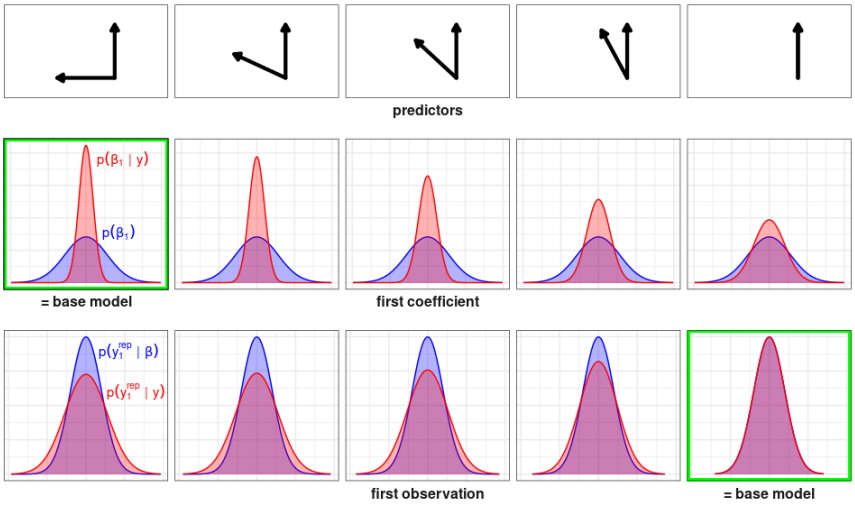

Figure 4.1: First row: $\mathbf{x}_1$ and $\mathbf{x}_2$, in order of increasing absolute correlation. Second row: the priors $q(\beta_1)$ (blue) and the posteriors $q(\beta_1 \mid \mathbf{y})$ (red), both centered to allow for an easier comparison of scales. Narrower posteriors relative to the prior indicate better identification. Third row: the posterior predictive $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$ (red) and the sampling distributions $f(\mathbf{y} \mid \overline{\boldsymbol{\beta}})$ (blue). More dissimilarity is connected to worse falsifiability on average.

As the predictor correlation decreases, the distributions $f(y_1^{\text{rep}} \mid \overline{\boldsymbol{\beta}})$ and $f(y_1^{\text{rep}} \mid \mathbf{y})$ become less similar.

The highlighted panels also show the relationships between the corresponding distributions that occur in the single-predictor model (4.3). These highlighted panels represent the best-case behavior, showing that the expanded model can only perform worse than the base model on either metric. Furthermore, these behaviors are inversely related among the expanded models, i.e. the most precise marginal inference occurs when the sampling and posterior predictive distributions are most dissimilar and vice versa.

4.2 Related work

Recently, statistical workflow has enjoyed increased attention as a discrete topic in statistics. This literature has sought to provide a consistent framework and practical advice for each step of a statistical analysis, including model expansion (see, e.g. [38, 89, 34]). We seek to complement this perspective by studying model expansion as a distinct regime. To this end, our main result

provides interpretable bounds on the mutual and conditional mutual information.

In [40], the authors study a formalization of model expansion which is similar to our definition of model expansion (given in chapter 3) and considers the consequences of nonidentification of parameters in the expanded model. Whereas Greenland assumes that the expanded model is (strictly) nonidentified, this work connects model expansion with nonidentification generally. We believe our results thus give additional weight to Greenland's conclusion that, when a model suffers from weak identification, "any analysis should thus be viewed as a part of a sensitivity analysis which depends on external plausibility considerations to reach conclusions."

Outside of the context of model expansion, the problem of weak/non-identification has been extensively studied. Gustafson considers the asymptotic properties of posterior distributions for strictly nonidentified likelihood models, showing that the corresponding infinite-data posteriors may still identify parameters substantially better than the prior [42, 41]. In particular, he shows this can be achieved in special cases with priors that encode dependence between identified and non-identified parameters. Our main result echoes this observation, but establishes it in a preasymptotic setting and under a much more general class of models. Other methods for detecting and dealing with identification problems have been studied in, e.g. [95, 57].

As we will argue in Section 4.3, problems of falsifiability are directly connected to debates over the conservativity of the posterior predictive p-value. Various forms of this problem have been described, and possible solutions have been proposed in [6, 5, 80, 97, 39, 53, 55, 44]. We propose another possible solution - conditional p-values - which we argue are better suited to the task of revealing specific deficiencies with a model and uncovering possible solutions compared to these alternatives while also being computationally inexpensive.

Our approach to studying the problems of identifiability and falsifiability follows many previous successes in using information-theoretic tools to understand and quantify model behaviors in substantial generality. We enumerate a few connections of particular note:

1. We study identifiability by thinking of information entropy as quantifying uncertainty. This representation of uncertainty as entropy can be traced back to Jaynes, who used it to justify

the application of maximum entropy prior distributions [50].

2. Many popular predictive performance metrics for models are based on information-theoretic quantities [94]. In the Bayesian context, the expected log predictive density (ELPD) has been used as a flexible objective for model evaluation and comparison [90]. When our data consists of a scalar quantity y, and the model is correctly specified, the ELPD can be given as

$$\mathrm{D}\left(f(y \mid \boldsymbol{\theta}^*) \,||\, f(y \mid y^{\text{rep}})\right) + C,$$

where $f(y \mid \boldsymbol{\theta}^*)$ is the true data generating process, C is a constant depending only on this true distribution, and $y^{\text{rep}} \sim f(y \mid \boldsymbol{\theta})$ is an independent replication of the data. If we substitute the true value $\boldsymbol{\theta}^*$ with an average over the posterior $q(\boldsymbol{\theta} \mid \mathbf{y})$, then the first term recovers the conditional mutual information $\mathbf{I}\left(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}\right)$.

3. The Rashomon effect, first described by Breiman in [15], occurs when many models which achieve similar overall loss provide very different point predictions. Our concept of falsifiability is also threatened by the multiplicity of predictive distributions. And our conditional mutual information rests on a similar KL divergence as a recently proposed metric for quantifying the Rashomon effect, the Rashomon capacity [46].

4. Mutual information-based quantities have also been deployed to bound measures of other adverse model behaviors, including bias and generalization error [83, 96].

4.3 Weak falsifiability and model expansion

We now turn to the behavior of the posterior predictive distribution (ppd) $f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})$ under model expansion. For joint model $q(\mathbf{y}, \boldsymbol{\theta})$, the ppd is

$$f(\mathbf{y}_{\text{rep}} \mid \mathbf{y}) = \int f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta}) q(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}. \tag{4.5}$$

Comparisons between posterior predictive samples and observed data are commonly used to

check Bayesian models. It is often convenient to formalize these checks as posterior predictive p-values (ppp-vs).

Definition 2 (Posterior Predictive p-Value)**.** *For observed data* $\mathbf{y}$*, joint model* $q(\mathbf{y}, \boldsymbol{\theta})$*, and real-valued test statistic* T*, the right-tailed ppp-v for* T *is*

$$p_T = \int_{\{T(\mathbf{y}_{\text{rep}}) \geq T(\mathbf{y})\}} f(\mathbf{y}_{\text{rep}} \mid \mathbf{y}) d\mathbf{y}_{\text{rep}}. \tag{4.6}$$

The left-tailed and two-tailed p-values are defined analogously.

It will be useful to set ppp-vs in a general framework of model evaluations.

Definition 3 (Data Distribution Evaluation)**.** *Given* $q(\mathbf{y}, \boldsymbol{\theta})$*, let* $\mathcal{Y}$ *be the (common) support of the densities* $f(\mathbf{y} \mid \boldsymbol{\theta})$*,* $P(\mathcal{Y})$ *be the space of all densities on* $\mathcal{Y}$*, and* $\mathcal{E} \subset \mathbb{R}$*. Then for any data* $\mathbf{y}$*, an evaluation is a (measurable) map*

$$e_{\mathbf{y}} : P(\mathcal{Y}) \to \mathcal{E}$$

For posterior sample $\{\boldsymbol{\theta}_{(s)}\}_{s=1}^{S}$ and evaluation $e_{\mathbf{y}}$, the modeler has data

$$\left\{\boldsymbol{\theta}_{(s)}, e_{\mathbf{y}}\left(q(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta}_{(s)})\right)\right\}_{s=1}^{S} \tag{4.7}$$

with which to evaluate the model. Since (4.7) can be complex and high-dimensional, it may not provide easy conclusions about overall model fitness. Posterior predictive checks solve this by providing simple summaries of (4.7). For statistic T, we define conditional ppp-vs $p_T(\boldsymbol{\theta})$ as the evaluations $e_{\mathbf{y}}\left(f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})\right)$ for the map

$$e_{\mathbf{y}} : f(\cdot) \to \int_{\{T(\mathbf{x}) \geq T(\mathbf{y})\}} f(\mathbf{x}) d\mathbf{x} \tag{4.8}$$

The usual ppp-v is then just the average:

$$p_T = \int e_{\mathbf{y}}\left(f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})\right) q(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}, \tag{4.9}$$

which is naturally estimated by $\frac{1}{S}\sum_{s=1}^{S} e_{\mathbf{y}}\left(f(\cdot \mid \boldsymbol{\theta}_{(s)})\right)$.

These conditional p-values are similar to the sampled p-value developed in [39, 53, 55]. These sampled p-values similarly result from first drawing a parameter $\boldsymbol{\theta}$ from the posterior distribution and evaluating a p-value for T with respect to the corresponding $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$. The difference between our procedures is entirely in how they are used. The sampled p-value is meant to be computed once and compared to a threshold for rejection. The benefit is that the resulting p-value has an asymptotically uniform distribution under the assumed model.

Because we are primarily interested in model improvement over model rejection, and because we are focused on retaining as much information for this purpose as possible, we instead construct the conditional p-values as a distribution of model diagnostic information. Rather than compare to any threshold, we exploit the joint structure between the random conditional p-values and the other model quantities under the posterior to help us determine fruitful directions for model improvement (as in the following example).

In this context, it becomes clear that the ppp-v may be limited if relevant information in (4.7) is lost in the one-dimensional reduction (4.9). If $f(\mathbf{y} \mid \boldsymbol{\theta}_{(s)}) = f(\mathbf{y} \mid \boldsymbol{\theta}_{(t)})$ for all $\mathbf{y}$ and $1 \leq s, t \leq S$, the ppd reduces to this single distribution, and no information is lost in (4.9). But in general the ppd will not capture all the sampling distributions which are plausible under the posterior. We quantify this loss of information with the Kullback-Leibler (KL) divergence.

Definition 4 (Posterior Sampling Divergence)**.** *For data* $\mathbf{y}$ *and model* $q(\boldsymbol{\theta}, \mathbf{y})$*, the posterior sampling divergence is*

$$\text{psd}\,(\mathbf{y}) = \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})} D\left(f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta}) || f(\mathbf{y}_{\text{rep}} \mid \mathbf{y})\right). \tag{4.10}$$

Using the Donsker-Varadhan representation and Jensen's inequality, we get

$$\text{psd}(\mathbf{y}) \leq \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{y})} \left\{ \sup_{T:\mathcal{Y}\to\mathbb{R}} \left| \mathbb{E}_{f(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})} T\left(\mathbf{y}_{\text{rep}}\right) - \mathbb{E}_{f(\mathbf{y}_{\text{rep}}|\mathbf{y})} T\left(\mathbf{y}_{\text{rep}}\right) \right| \right\}$$

The psd lower bounds the degree to which typical sampling distributions $f(\mathbf{y}_{\text{rep}} \mid \boldsymbol{\theta})$ (under the posterior) and the ppd can be distinguished by a statistic T. A large psd thus indicates increased

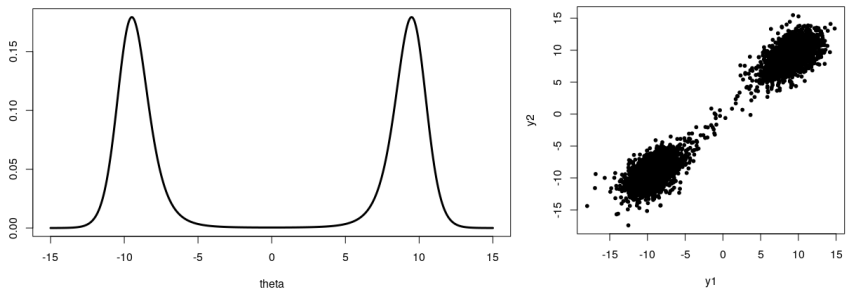

Figure 4.2: Left: The posterior of θ. Right: The posterior predictive distribution of (y_1, y_2).

risk of information loss when using a ppp-v compared to (4.7). An example shows how this information loss can make falsification with a ppp-v difficult in practice.

Let $\mathbf{y} = (-10, 10)$, with model

$$y_1, y_2 \overset{iid}{\sim} \text{student-t}\,(\theta, 1, 10)\,, \qquad \theta \sim \text{uniform}\,(-15, 15)\,, \tag{4.11}$$

where student-t(μ, σ, d) is the t distribution with location μ, scale σ, and d degrees of freedom. This results in a multimodal posterior, plotted in the left panel of Figure 4.2. The right panel plots joint samples from the ppd, which is also bimodal.

Consider test statistics $T_1(y_1, y_2) = -y_1$ and $T_2(y_1, y_2) = y_2$, and let $p_{T_1}(\theta)$ and $p_{T_2}(\theta)$ be the corresponding conditional ppp-vs for $f(y_1, y_2 \mid \theta)$. Figure 4.3 plots these against θ. The ppp-vs p_{T_1} and p_{T_2} are ≈ 0.165, above usual thresholds for rejection. However, the $c$ppp-vs are vanishingly small over the bulk of the posterior support, suggesting that the model may be improved by introducing a scale parameter, or allowing the means to differ for the two observations, for example.

This example points towards a notion of power which does not make reference to the frequency properties of the ppp-v. Specifically, we will consider a model assessment underpowered if there is additional data (e.g. that contained in (4.7)) which would cause us to consider the relevant aspect of model fitness deficient despite the particular model assessment passing (e.g. a ppp-v not falling below a rejection threshold). Therefore, we view an increasing psd as increasing the risk of our chosen model assessments suffering such power deficits. In these cases, we have to work harder to

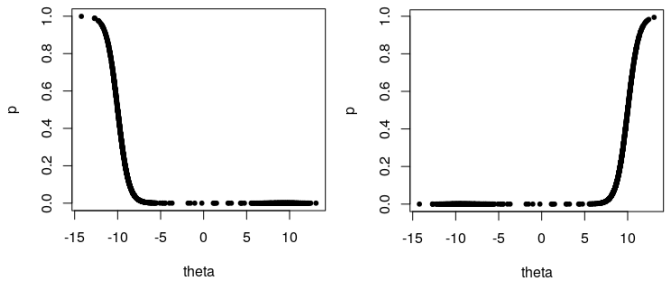

Figure 4.3: Conditional posterior predictive p-values for T_1 (left panel) and T_2 (right panel), evaluated at and plotted against posterior draws of θ.

find strong evidence for the falsity of the model (e.g. by examining the conditional p-value plots in Figure 4.3), and in this sense the model exhibits weaker falsifiability.

4.3.1 Posterior sampling divergence and model expansion

Since $\mathbf{y}$ and $\mathbf{y}_{\text{rep}}$ are conditionally independent given $\boldsymbol{\theta}$, it follows that the average psd reduces to the conditional mutual information (cmi): $\mathbb{E}_{f(\mathbf{y})}\mathsf{psd}\,(\mathbf{y}) = \mathbf{I}_{q(\boldsymbol{\theta},\mathbf{y},\mathbf{y}_{\text{rep}})}\left(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}\right)$. For base model $q_{\text{base}}(\boldsymbol{\theta}, \mathbf{y})$ and expansion $q(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{y})$, we define

$$\Delta^{\text{pred}} = \mathbb{E}_{q(\lambda)}\left[\mathbf{I}_{q(\boldsymbol{\theta},\mathbf{y}_{\text{rep}},\mathbf{y}|\boldsymbol{\lambda})}(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}) - \mathbf{I}_{q(\boldsymbol{\theta},\mathbf{y}_{\text{rep}},\mathbf{y}|\boldsymbol{\lambda}_0)}(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y})\right].$$

Using the chain rule for cmi, we can then establish the following cmi decomposition:

$$\mathbf{I}_q\left((\boldsymbol{\theta}, \boldsymbol{\lambda}), \mathbf{y}_{\text{rep}} \mid \mathbf{y}\right) = \mathbf{I}_{q_{\text{base}}}\left(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}\right) + \Delta^{\text{pred}} + \mathbf{I}_q\left(\boldsymbol{\lambda}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}\right). \tag{4.12}$$

As in our decomposition for the mi (4.15), the nonnegative term $\mathbf{I}_q\left(\boldsymbol{\lambda}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}\right)$ creates an upward bias for the cmi. In some simple examples, the cmi can be computed exactly:

1. For base model $y \mid \theta \sim \text{normal}\,(\theta, 1)$ and $\theta \sim \text{normal}\,(0, 1)$, we get $\mathbf{I}(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}) = \log(3/2)/2$. Adding a redundant location parameter, we get

$$y \mid \boldsymbol{\theta} \sim \text{normal}\left((\theta_1 + \theta_2)/\sqrt{2}, 1\right), \qquad \theta_1, \theta_2 \overset{iid}{\sim} \text{normal}\,(0, 1).$$

After an invertible reparametrization $(\mu_1, \mu_2) = \phi(\theta_1, \theta_2)$, this model is just $y \mid \boldsymbol{\theta} \sim$ normal $(\mu_1, 1)$ and $\mu_1, \mu_2 \overset{iid}{\sim}$ normal $(0, 1)$. By invariance of the cmi under invertible reparametrization, we have:

$$\mathbf{I}(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}) = \mathbf{I}(\boldsymbol{\mu}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}) = h\left(\mu_1 \mid \mathbf{y}\right) - h\left(\mu_1 \mid \mathbf{y}, \mathbf{y}_{\text{rep}}\right) = \log\left(3/2\right)/2.$$

2. Now consider a normal location model with data $\mathbf{y} \in \mathbb{R}^{2n}$ for $n \geq 1$:

$$\mathbf{y}_i \overset{iid}{\sim} \text{normal}\left(\theta, 1\right) \text{ for } 1 \leq i \leq 2n, \qquad \theta \sim \text{normal}\left(0, 1\right).$$

Dividing $\mathbf{y}$ as $\mathbf{y} = \left(\mathbf{y}^1, \mathbf{y}^2\right) \in \mathbb{R}^{n \times 2}$ and introducing separate means θ_1, θ_2, we get:

$$\mathbf{y}_i^j \overset{iid}{\sim} \text{normal}\left(\theta_j, 1\right) \text{ for } 1 \leq i \leq n \text{ and } j = 1, 2, \qquad \theta_1, \theta_2 \overset{iid}{\sim} \text{normal}\left(0, 1\right).$$

Now the cmi of the base model is $\text{CMI}_{\text{base}}(n) = \frac{1}{2}\log\left(\frac{4n+1}{2n+1}\right)$, whereas the cmi of the expanded model is $\text{CMI}_{\text{exp}}(n) = \log\left(\frac{2n+1}{n+1}\right)$. We have $\text{CMI}_{\text{exp}}(n) > \text{CMI}_{\text{base}}(n)$ for all n with $\text{CMI}_{\text{exp}}(n) \to 2\text{CMI}_{\text{base}}(n)$ as $n \to \infty$.

The change in cmi can be separated into two pieces. First, by splitting $\mathbf{y}$, we reduce the data we have to estimate each of the means θ_1 and θ_2. This is reflected in the inequality $\log\left(\frac{4n+1}{2n+1}\right) > \log\left(\frac{2n+1}{n+1}\right)$. But parametrizing with two independent means adds a degree of freedom in the sampling distribution of the expanded model, doubling the constant factor, which dominates the comparison. However, the latter effect will not always determine the change in cmi between models, as the next example shows.

3. We take the normal-normal base model from the last example with $\mathbf{y} \in \mathbb{R}^n$ and expand it by adding a precision parameter and using a normal-gamma prior:

$$\mathbf{y}_i \overset{iid}{\sim} \text{normal}\left(\theta_1, \theta_2^{-1/2}\right) \text{ for } 1 \leq i \leq n, \qquad (\theta_1, \theta_2) \sim \text{NG}\left(0, \mu_{\theta_2^{-1}}, 2, \mu_{\theta_2^{-1}}\right)$$

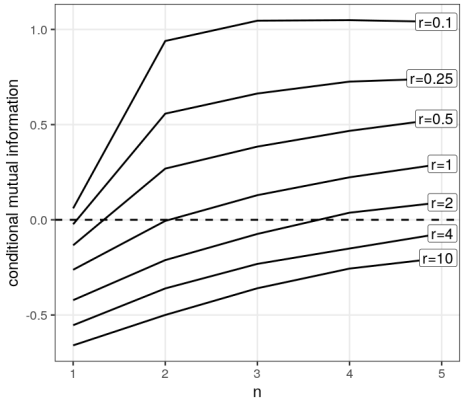

Figure 4.4: Percentage change in cmi from the base model to the expanded model against data size n for a range of noise levels $r = \mu_{\theta_2^{-1}}/n$.

Here, $\mu_{\theta_2^{-1}} > 0$ is the prior mean of the variance θ_2^{-1}. The marginal prior on θ_1 is normal $(0, 1)$, matching the prior in the base model. Figure 4.4 shows estimated percentage changes in cmi from base to expanded model against n for a range of noise levels $r = \mu_{\theta_2^{-1}}/n$.

As before, increasing n makes it easier to distinguish sampling distributions, increasing the cmi. Likewise, the added degree of freedom introduced by the precision parameter increases the cmi. Hence, most points in Figure 4.4 lie above 0. However, unlike the last example, the cmi can decrease in the expanded model if r is sufficiently large. This is because, for larger r, the prior favors sampling distributions with larger scales that are harder to distinguish. Nevertheless, the effect of the added degree of freedom usually dominates this comparison. For example, for $n = 2$, the noise level in the base model is $r = 0.5$. Unless the prior mean noise level in the expanded model is more than double that of the base model (i.e. $r = 1$), the cmi will increase.

4.4 Formalizing weak identification

We now formalize our notion of identification using the information entropy. First we establish some notation. For a joint model $q(\boldsymbol{\theta}, \mathbf{y})$, the (differential) entropy of $q(\boldsymbol{\theta})$ is denoted $h_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})$, and the conditional entropy of $\boldsymbol{\theta}$ given $\mathbf{y}$ is $h_{q(\boldsymbol{\theta},\mathbf{y})}(\boldsymbol{\theta} \mid \mathbf{y})$. The mutual information (mi) is denoted $\mathbf{I}_q(\boldsymbol{\theta}, \mathbf{y})$, which will at times be extended to a conditional mutual information (cmi), denoted by

$\mathbf{I}_q(\boldsymbol{\theta}, \mathbf{y} \mid \mathbf{x})$, when the joint model extends over an additional quantity $\mathbf{x}$. When distributions are clear from context, we may drop subscripts from entropies and mutual informations, writing e.g. $h(\boldsymbol{\theta})$ and $\mathbf{I}(\boldsymbol{\theta}, \mathbf{y})$. The reader who is unfamiliar with information theory may consult the appendix for definitions of these quantities and statements of the basic results that we use. With these definitions, we can now give quantitative operational definitions of our notion of weak marginal identification for arbitrary subsets of $\boldsymbol{\theta}$.

Definition 5 (ϵ-Weak Identification)**.** *Let $I \subset [d]$. We say for any $\epsilon > 0$ that $\boldsymbol{\theta}_I = (\theta_i)_{i \in I}$ is ϵ-weakly identified for data $\mathbf{y}$ if*

$$h_{q(\boldsymbol{\theta}_I \mid \mathbf{y})}(\boldsymbol{\theta}_I) > h_{q(\boldsymbol{\theta}_I)}(\boldsymbol{\theta}_I) - \epsilon. \tag{4.13}$$

We also define weak identification for entire models (regardless of data $\mathbf{y}$) by averaging over the prior predictive distribution.

Definition 6 (ϵ-Weakly Identifiable Model)**.** *Let $I \subset [d]$. We say $\boldsymbol{\theta}_I$ is ϵ-weakly identifiable in $p(\boldsymbol{\theta}, \mathbf{y})$ if*

$$h(\boldsymbol{\theta}_I \mid \mathbf{y}) > h(\boldsymbol{\theta}_I) - \epsilon, \tag{4.14}$$

or, equivalently, if $\mathbf{I}(\boldsymbol{\theta}_I, \mathbf{y}) < \epsilon$.

Henceforth, we will leave the ϵ-dependence of this definition implicit and simply say that a parameter is weakly identified if it is ϵ-weakly identified for an appropriate value of ϵ (which will usually be given by domain understanding).

This operational definition of weak identifiability can only be interpreted relative to the prior. Often, this is a natural quantity to focus on (e.g. when we are concerned with the contribution of a research finding to existing prior knowledge). However, if we expand a model in such a way that adds prior information about $\boldsymbol{\theta}$, then it is possible for both the posterior entropy of $\boldsymbol{\theta}$ and the mutual information to decrease. In other words, identification relative to the prior may decrease while the posterior becomes more concentrated relative to the base model. This divergence between absolute

and relative notions of identification can be avoided if we exclude from consideration expansions which decrease $h(\boldsymbol{\theta})$. This will hold for most interesting cases of model expansion.

We can now show that certain model expansions tend to weaken identification in the above sense. If $q(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{y})$ is an expansion of $q_{\text{base}}(\boldsymbol{\theta}, \mathbf{y})$, then we have the following decomposition of the mutual information:

$$\mathbf{I}_{q(\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{y})}(\boldsymbol{\theta}, \mathbf{y}) = \mathbf{I}_{q_{\text{base}}(\boldsymbol{\theta},\mathbf{y})}(\boldsymbol{\theta}, \mathbf{y}) + \Delta_I^{\text{exp}} + \Delta_I^{\text{post}}, \tag{4.15}$$

where we define

$$\begin{aligned}
\Delta_I^{\text{exp}} &= \mathbf{I}_{q(\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{y})}(\boldsymbol{\theta}, \mathbf{y} \mid \boldsymbol{\lambda}) - \mathbf{I}_{q(\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{y})}(\boldsymbol{\theta}, \mathbf{y} \mid \boldsymbol{\lambda}_0) \\
\Delta_I^{\text{post}} &= \mathbf{I}_{q(\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{y})}(\boldsymbol{\lambda}, \boldsymbol{\theta}) - \mathbf{I}_{q(\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{y})}(\boldsymbol{\lambda}, \boldsymbol{\theta} \mid \mathbf{y}).
\end{aligned}$$

The Δ_I^{exp} term is the difference in the amount of information $\mathbf{y}$ provides about $\boldsymbol{\theta}$ given $\boldsymbol{\lambda}$ and given $\boldsymbol{\lambda}_0$, averaging $\boldsymbol{\lambda}$ over the expanded model. The Δ_I^{post} term is the difference in the amount of information $\boldsymbol{\lambda}$ provides about $\boldsymbol{\theta}$ before and after observing the data $\mathbf{y}$. We regard this as a measure of the a priori informativeness of $\boldsymbol{\lambda}$ about $\boldsymbol{\theta}$, which is justified by the fact that we have, $\Delta_I^{\text{post}} \geq -\mathbf{I}(\boldsymbol{\lambda}, \boldsymbol{\theta} \mid \mathbf{y})$ with equality if and only if $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are independent in the expanded prior (i.e. if $q(\boldsymbol{\theta}, \boldsymbol{\lambda}) = q(\boldsymbol{\theta})q(\boldsymbol{\lambda})$). When $\Delta_I^{\text{post}} < 0$, (4.15) shows the expanded model is biased towards weaker identification of $\boldsymbol{\theta}$ compared to the base model.

4.5 Marginal entropy and sampling divergence

Let $q_{\text{base}}(\mathbf{y}, \boldsymbol{\theta})$ and $q(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ be a base model and expansion. Consider the following two extreme scenarios:

- Let $f_{\text{base}}(\mathbf{y} \mid \boldsymbol{\theta}) = f(\mathbf{y})$ for a density f. Then the likelihood is constant, and $h_{q_{\text{base}}}(\boldsymbol{\theta}) = h_{q_{\text{base}}}(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{y}_{\text{rep}})$ for all $\mathbf{y}$ and $\mathbf{y}_{\text{rep}}$. It follows immediately that $\mathbf{I}(\boldsymbol{\theta}, \mathbf{y}) = \mathbf{I}(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}) = 0$, and so any nontrivial expansion of q_{base} with $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ independent must decrease the marginal posterior entropy of $\boldsymbol{\theta}$ and increase the cmi.

- Let $pq\text{base}(\mathbf{y} \mid \theta) = \text{normal}(\mathbf{y} \mid \theta, 1)$, and then take the expansion $f(\mathbf{y} \mid \theta, \lambda) = f_{\text{base}}(\mathbf{y} \mid \theta + \lambda)$ with priors $\theta \sim \text{normal}(0, \sigma_\theta^2)$ and $\lambda \sim \text{normal}\left(0, \sigma_\lambda^2\right)$. We then have for the expanded model that

$$\mathbf{I}(\boldsymbol{\theta}, \mathbf{y}) = \frac{1}{2} \log\left(1 + \frac{\sigma_\theta^2}{1 + \sigma_\lambda^2}\right), \quad \mathbf{I}\left((\boldsymbol{\theta}, \boldsymbol{\lambda}), \mathbf{y}_{\text{rep}} \mid \mathbf{y}\right) = \frac{1}{2} \log\left(\frac{1 + 2(\sigma_\theta^2 + \sigma_\lambda^2)}{1 + \sigma_\theta^2 + \sigma_\lambda^2}\right).$$

  Similarly, the cmi in the base model is

$$\mathbf{I}\left(\boldsymbol{\theta}, \mathbf{y}_{\text{rep}} \mid \mathbf{y}\right) = \frac{1}{2} \log\left(\frac{1 + 2\sigma_\theta^2}{1 + \sigma_\theta^2}\right).$$

  Both cmi expressions are bounded above by $\frac{1}{2}\log(2)$. Furthermore, by taking σ_θ^2 large, we can ensure that the cmi for the base model is arbitrarily close to this limit. Then the cmi can increase only negligibly in the expanded model regardless of σ_λ^2. However, fixing σ_θ^2, the mi in the expanded model tends to 0 as $\sigma_\lambda^2 \to \infty$.

In each case, one of the models has singular Fisher information, and either the mi or cmi degrades (i.e. decreases or increases respectively) while the other changes at most negligibly. We can show in certain cases that a similar phenomenon occurs in between these extremes (i.e. when the Fisher information is nonsingular and the (conditional) mutual information is not (maximized) minimized). In particular, the following result demonstrates a trade-off asymptotically between worsening (i.e. decreasing) mi and worsening (i.e. increasing) cmi.

We can now state our main result which connects the patterns observed in the last two sections and demonstrates a trade-off between mitigating the downward bias on identifiability and mitigating the upward bias on falsifiability in any particular instance of model expansion.

Theorem 4. *Let $q(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{y})$ be an expansion of $q_{\text{base}}(\boldsymbol{\theta}, \mathbf{y})$, and let $\{\iota_i\}_{i=1}^d$ and $\{\iota_i^{\text{cond}}\}_{i=1}^d$ be the eigenvalues of $\mathbb{E}\boldsymbol{\mathcal{I}}_{\text{base}}(\boldsymbol{\theta})$ and $\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ respectively, where $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ is the Fisher information of the expanded model with respect to the components of $\boldsymbol{\theta}$ alone. Assume that:*

1. *the marginal priors $q_{\text{base}}(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ are log concave, and their covariance matrices have*

unit spectral norm. (This norm assumption can always be satisfied as long as $\boldsymbol{\lambda}$ contains a hyperparameter governing the scale of $\boldsymbol{\theta}$; see appendix for details.)

2. *the posteriors $q_{\text{base}}(\boldsymbol{\theta} \mid \mathbf{y}_M)$ and $q(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \mathbf{y}_M)$ are normal for all $\mathbf{y}_M$ for M large, where $\mathbf{y}_M$ is a vector of M i.i.d. draws from the models $f_{\text{base}}(\mathbf{y} \mid \cdot)$ and $f(\mathbf{y} \mid \cdot)$ respectively,*

3. *the matrix norm moments $\mathbb{E}\|\boldsymbol{\mathcal{I}}_{\text{base}}(\boldsymbol{\theta})\|^2$, $\mathbb{E}\|\boldsymbol{\mathcal{I}}_{\text{base}}(\boldsymbol{\theta})^{-1}\|^2$, $\mathbb{E}\|\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}, \boldsymbol{\lambda})\|^2$, $\mathbb{E}\|\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}, \boldsymbol{\lambda})^{-1}\|^2$ are bounded above by universal constants, and similarly for squared norms of the posterior covariance/precision matrices,*

4. *and a few other regularity conditions on the smoothness of the relevant densities and tails of the spectrum of the observed information. (See appendix for details.)*

Additionally, assume the distributions of the Fisher information and posterior covariance spectra are not too skewed in the sense of Lemma 9 in the appendix. This ensures that these (random) matrices are sufficiently well-summarized by their means. Under these conditions, for increasing functions ψ_i and ψ_f, we have the following inequalities:

$$\mathbf{I}_{\text{base}}(\mathbf{y}, \boldsymbol{\theta}) \leq \psi_i\left(\sum_{j=1}^{d} \iota_j\right), \qquad \mathbf{I}_{\text{base}}\left(\mathbf{y}_{\text{rep}}, \boldsymbol{\theta} \mid \mathbf{y}\right) \geq \frac{1}{\log d} \sum_{j=1}^{d} \psi_f\left(\iota_j\right) \tag{4.16}$$

$$\begin{aligned} \mathbf{I}(\mathbf{y}, \boldsymbol{\theta}) &\leq \psi_i\left(\sum_{j=1}^{d}\left[\iota_j^{\text{cond}} - [\boldsymbol{\Delta}^i]_j\right]\right), \\ \mathbf{I}\left(\mathbf{y}_{\text{rep}}, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \mid \mathbf{y}\right) &\geq \frac{1}{\log d^{\text{exp}}}\left[\sum_{j=1}^{d} \psi_f\left(\iota_j^{\text{cond}}\right) + \Delta^f\right], \end{aligned} \tag{4.17}$$

where $\Delta^f \geq 0$. We also have $[\boldsymbol{\Delta}^i]_j > 0$ for $j = 1, \ldots, d$ when $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are independent under the expanded prior and in general whenever $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are sufficiently more dependent in the posterior than the prior. Expressions for the Δ terms and ψ functions can be found in the appendix.

Further, suppose that $\boldsymbol{\lambda}$ tends to reduce the information gained about $\boldsymbol{\theta}$ in the sense that

$$\mathbb{E}_{q_{\text{base}}(\boldsymbol{\theta})} \boldsymbol{\mathcal{I}}_{\text{base}}(\boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{\lambda})} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) > 0, \tag{4.18}$$

then we have

$$\psi_i\left(\sum_{j=1}^{d} \iota_j^{\text{cond}}\right) \leq \psi_i\left(\sum_{j=1}^{d} \iota_j\right). \tag{4.19}$$

If on the other hand (4.18) *is reversed, we get*

$$\sum_{j=1}^{d} \psi_f\left(\iota_j^{\text{cond}}\right) \geq \sum_{j=1}^{d} \psi_f\left(\iota_j\right). \tag{4.20}$$

This formalizes the trade-off between identifiability and falsifiability that we noted in the previous sections.

Proof. See appendix. □

We remark that the $\boldsymbol{\Delta}^i$ and Δ^f terms above are analogous to the bias terms in the identities (4.15) and (4.12), pushing the bounds (4) in the adverse directions. These terms also tend to increase in magnitude with the expanded model dimension. Along with the observed trade-off behavior, this substantially generalizes the pattern we observed in our introductory regression example, where predictor correlation structures that offered better identification created greater posterior uncertainty about the sampling distribution and vice versa.

We can interpret this theorem in a positive and negative light. Negatively, when our prior information is relatively unstructured, the process of model expansion may force us to confront either weakly identified marginal inferences or a large posterior sampling divergence. Positively, a prior with sufficient dependence between the parameters may allow us to avoid these difficulties, even if this prior is weak in the sense of carrying relatively little marginal information about any particular parameter. In the next sections, we explore examples where these two difficulties occur and demonstrate how they can be partially overcome with sufficiently rich posterior summaries.

4.5.1 Example: inference under poor identification

The next two sections explore methodological implications of the above results with concrete examples. Here we consider an example in which we get 'stuck' between an implausibly simple base model and an expanded model where the parameter of interest is too weakly identified to allow strong conclusions. We then show how the expanded model can still support nontrivial inferences which, in this case, inform how we should collect future data. In so doing, we hope to demonstrate (i) how the concerns raised in Sections 4.4 and 4.5 motivate looking beyond standard marginal posterior summaries, and (ii) that weak identification need not be an inferential dead end for a statistical analysis.

Two models for grouped data

We first define three simulated data sets, each generated by drawing random samples of M measurements from each of L subpopulations. We take $M = 2$ and $L = 20$ for all data sets, but vary the ratio of within- and between-subpopulation variances between them. Section 2 of the Supplementary Materials provides a complete description of the data generating process.

Figure 4.5 plots the data, with a row for each subpopulation, a column for each data set, and dots for the individual measurements.

The unobserved grand (i.e. superpopulation-level) mean will be our quantity of interest. We also assume that more positive values of the grand mean represent "better" outcomes (i.e. more desirable from the researcher's standpoint). We will think about identification of this mean in terms of both the change in entropy from prior to posterior and the degree to which the posterior can rule out "practically insignificant" values, which we define as those with magnitude less than 1.

In our base model, we assume the subpopulation-level distributions are identical. Letting $\mathbf{y}_{ml}$ be the data with $1 \leq m \leq M$ indexing measurements and $1 \leq l \leq L$ indexing subpopulations, the sampling distribution is $\mathbf{y}_{ml} \overset{iid}{\sim} \text{normal}(\mu, \sigma_*)$. We assign the grand mean μ a normal $(0, |\mu_0|)$ prior, with μ_0 the "true" value used in simulating the data sets. We treat σ_* as a hyperparameter which we set to a prior guess.

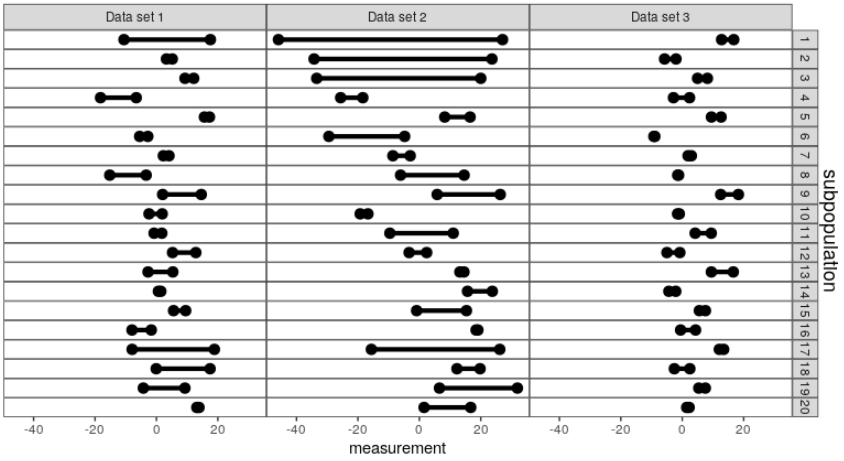

Figure 4.5: Columns: the three data sets. Rows: the 20 subpopulations. Cells: the two data points drawn from each subpopulation, connected with a horizontal line to show their range.

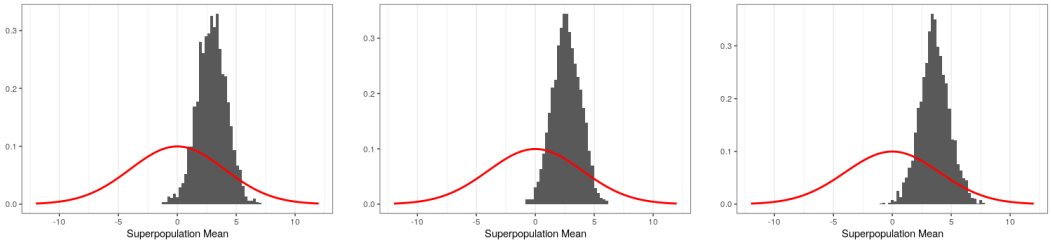

Figure 4.6: Histograms of samples from the posterior distributions of μ under the base model fit to each data set. The red curve shows the density of the prior on μ.

Figure 4.6 shows the histograms of draws from the posteriors with the prior density. For each data set, the posterior standard deviation is $\approx 1/4$ the prior standard deviation, and the probabilities of a practically insignificant effect are between 2% and 10%. Thus, we have improved on our prior knowledge, but the posterior cannot rule out an insignificant effect.

Two features of the base model stand out for criticism:

1. We usually cannot confidently guess the scale σ_* with just prior information.

2. The subpopulation distributions could be distinct, so we should not assume otherwise.

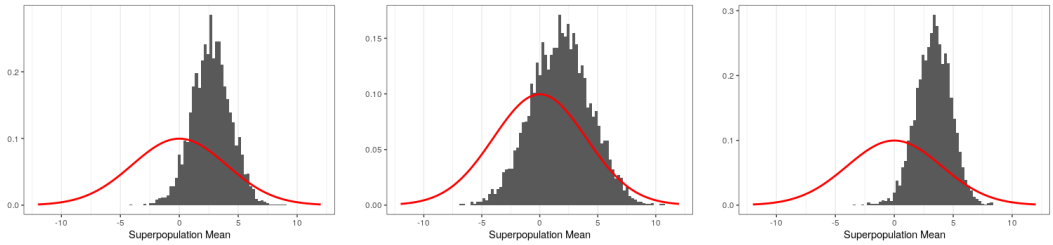

Figure 4.7: Histograms of samples from the marginal posteriors of μ under the expanded model fit to each data set. The red curve shows the density of the prior on μ.

Our expanded model makes the scale unknown and allows for different subpopulations:

$$\begin{gathered} \mathbf{y}_{ml} \overset{iid}{\sim} \text{normal}\left(\theta_l, \sigma\right), \qquad \theta_l \overset{iid}{\sim} \text{normal}\left(\mu, \tau\right), \\ \mu \sim \text{normal}\left(0, |\mu_0|\right), \qquad \sigma \sim \text{gamma}\left(a_\sigma, b_\sigma\right), \qquad \tau \sim \text{gamma}\left(a_\tau, b_\tau\right). \end{gathered} \tag{4.21}$$

Now, $a_\sigma, b_\sigma, a_\tau, b_\tau$ are hyperparameters of the model. The a_σ and b_σ are taken so that the prior mode for σ equals our guess σ_*. The prior $p(\mu)$ and μ_0 are unchanged from the base model. Setting $\sigma = \sigma_*$ and $\tau = 0$ recovers the base model as a special case.

Figure 4.7 shows the posterior inferences for μ in the expanded model. In the first and third data sets, the posterior standard deviation is $\approx 1/3$ larger than in the base model. For the second data set, it is more than twice as large as in the base model. In all cases, the standard deviation is at least a third that of the prior. Similarly, the probabilities of a practically insignificant effect are all at least three times larger than in the base model, ranging from $\approx 6\%$ in the best case to over 30%.

We might now return to the base model or focus on subpopulation means to improve identification. However, the former is inadvisable since the base model was implausible, and the latter will likely fail since the subpopulation sample sizes are tiny. A less convenient but much safer approach is to gather more data. In fact, inferences from the expanded model can yield relevant information for future data collection despite the weak identification.

A bootstrap comparison of two sampling schemes

It has been recognized that bootstrapping new data sets, fitting a Bayesian model to each, and then aggregating the results can provide more information than the fit to the observed data alone [17]. This idea has been used, e.g., for model criticism and selection tasks [48, 47]. Here, we demonstrate that a similar procedure with a Bayesian parametric bootstrap of future data can help us learn about how to sample future data despite the underwhelming marginal identification of the posterior.

First, we must identify the candidate sampling schemes. We may sample the same subpopulations that we sampled originally, or we may sample from new subpopulations, or some combination of these. If the within-subpopulation variance is low, then sampling the same subpopulations may yield little improvement for identification. If instead the between-subpopulation variance is low, sampling the same subpopulations may provide enough information about the individual θ_l to identify μ well. The issue may be further complicated by cost differences. For our purposes, we assume that a sample from a new subpopulation is four times the cost of a sample from an existing subpopulation.

To simplify our comparison, we focus on whether it is more efficient to sample *exclusively* from existing subpopulations or from new subpopulations. To answer this, we use the joint posterior over parameters and replicated data to simulate sampling new data with these two schemes. We then compare posteriors on μ after refitting the model to these enlarged data sets. Replicating many times then allows us to assess the relative risks and benefits of each approach. Figure 4.8 diagrams these resampling schemes with full pseudocode in appendix. We take $M^{\text{new}} = 8$ additional samples from each of the $L = 20$ existing subpopulations in the first scheme and $M = 2$ samples from $L^{\text{new}} = 20$ new subpopulations in the second scheme. These are performed at equivalent overall cost, hence the difference in total sample size. We replicate each scheme $R = 500$ times. We also repeat this process with the prior replacing the posterior at each step in order to demonstrate that the posterior inferences differ in relevant aspects from prior inferences despite the marginal weak identification.

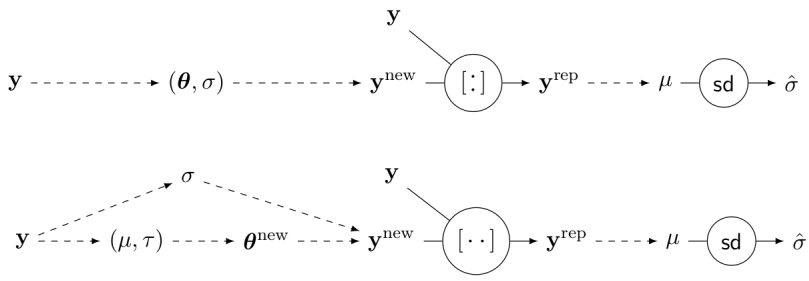

Figure 4.8: Schematic representation of our two resampling schemes. Dashed arrows represent sampling the output variable conditional on the input. Solid arrows represent that the output variable is formed by evaluating the circled function on the input variables.

For each of the $r = 1, \ldots, R$ replications and each method, define the ratio $\rho_r = \sigma^{(r)}/\sigma_{\text{obs}}$, where $\sigma^{(r)}$ is the posterior standard deviation of μ for the r^{th} expanded data set simulated from the given method, and where σ_{obs} is the posterior standard deviation of μ given just the observed data set. Table 4.1 presents the averages of these ratios over all replications $\overline{\rho} = \frac{1}{R}\sum_{r=1}^{R} \rho_r$ for each method and each of our observed data sets.

Despite weak identification, the posterior and prior columns differ, indicating that our simulated data reflects information learned from the observed data. Furthermore, the average improvement to identification for each scheme depends substantially on the data set, with each scheme winning in one data set and tying in the third. This clearly demonstrates that the posterior contains useful information, and that the weak identification of our parameter of interest is not a dead end for our statistical analysis using this model.

Figure 4.9 gives a finer-grained comparison, plotting histograms of the ρ_r for the two schemes. While the first data set yields a tie in the comparison of averages, the distribution of ρ_r is wider when sampling the same subpopulations than when sampling new subpopulations. This suggests a risk-reward trade-off: the latter scheme gives a more predictable reduction in uncertainty while the former carries the possibility of a greater reduction. Together this demonstrates that we can get nontrivial inferences which depend strongly on the particular data observed despite weak identification.

	Same sub-pops	New sub-pops	Same sub-pops (prior)	New sub-pops (prior)
Data set #1	0.74	0.74	1.18	1.04
Data set #2	0.69	0.8	0.74	0.77
Data set #3	0.94	0.71	1.16	1.03

Table 4.1: Values of $\overline{\rho}$ computed with 500 parametric bootstrap resamples separately for each (i) data set, (ii) future data sampling scheme, and (iii) choice of posterior or prior.

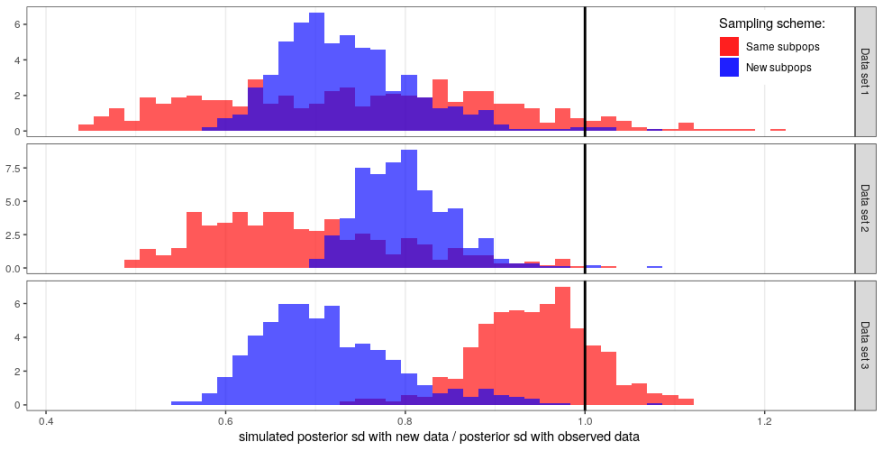

Figure 4.9: Histograms of estimated posterior standard deviations of μ for 500 simulations of future data under each sampling scheme and each data set.

4.6 Conclusions

When constructing a model for a given data analysis, a statistician should balance various desiderata, including:

- model predictions compatible with what is known about the world;
- inferences sufficiently well identified to support nontrivial conclusions;
- model checks powerful enough to reveal frictions between model and data.

When model checks reveal deficiencies, the first item is no longer satisfied, and a better model should be sought. In practice, this is often an expansion of the previous model. If such expansions are not accompanied by sufficiently strong prior information (in the form of prior dependence of parameters, not the marginal scales), then our results demonstrate that a tension may easily arise

within these three goals as the model dimension grows.

Insofar as the first desideratum is most essential, this motivates methods that can extract useful information even when one of the latter two desiderata are not satisfied. One avenue is to pursue richer inferential summaries. As demonstrated by the last two examples, the full posterior often contains such rich inferential data. This data is both accessible (by looking beyond e.g. marginal means, standard deviations, and p-values) and capable of supporting nontrivial conclusions (which are concealed by the common marginal summaries).

References

[1] E. Aras, K. Lee, A. Pananjady, and T. A. Courtade, "A family of bayesian cramér-rao bounds, and consequences for log-concave priors," *CoRR*, vol. abs/1902.08582, 2019. arXiv: 1902.08582.

[2] P. Barbe, C. Genest, K. Ghoudi, and B. Rémillard, "On kendall's process," *Journal of Multivariate Analysis*, vol. 58, no. 2, pp. 197–229, 1996.

[3] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, "Optimal errors and phase transitions in high-dimensional generalized linear models," *Proceedings of the National Academy of Sciences*, vol. 116, no. 12, pp. 5451–5460, 2019.

[4] J. Barbier, N. Macris, and C. Rush, "All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 915–14 926, 2020.

[5] M. J. Bayarri and J. O. Berger, "P values for composite null models," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1127–1142, 2000.

[6] M. Bayarri and J. Berger, "Quantifying surprise in data and model verification," in *Bayesian Statistics 6*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Eds., Oxford University Press, 1999, pp. 53–82.

[7] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *Ann. Appl. Probability*, vol. 25, no. 2, pp. 753–822, 2015.

[8] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *The Annals of Applied Probability*, vol. 25, no. 2, pp. 753–822, 2015.

[9] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, pp. 764–785, 2011.

[10] R. Berthier, A. Montanari, and P.-M. Nguyen, "State evolution for approximate message passing with non-separable functions," *Information and Inference: A Journal of the IMA*, vol. 9, no. 1, pp. 33–79, 2020.

[11] E. Biyik, J. Barbier, and M. Dia, "Generalized approximate message-passing decoder for universal sparse superposition codes," in *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 1593–1597.

[12] D. M. Blei, "Build, compute, critique, repeat: Data analysis with latent variable models," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 203–232, 2014. eprint: https://doi.org/10.1146/annurev-statistics-022513-115657.

[13] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, 859–877, Apr. 2017.

[14] E. Bolthausen, "An iterative construction of solutions of the TAP equations for the Sherrington-Kirkpatrick model," *arXiv:1201.2891*, 2012.

[15] L. Breiman, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," *Statistical Science*, vol. 16, no. 3, pp. 199 –231, 2001.

[16] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, May 2011, ISBN: 9780429138508.

[17] P. Bühlmann, "Discussion of big bayes stories and BayesBag," *Statistical Science*, vol. 29, no. 1, 2014.

[18] C. Cademartori, *Identifiability and falsifiability: Two challenges for bayesian model expansion*, 2023. arXiv: 2307.14545 [math.ST].

[19] C. Cademartori, *Joint p-values for higher-powered bayesian model checking with frequentist guarantees*, 2023. arXiv: 2309.13001 [stat.ME].

[20] C. Cademartori and C. Rush, "A non-asymptotic analysis of generalized vector approximate message passing algorithms with rotationally invariant designs," *IEEE Transactions on Information Theory*,

[21] C. Cademartori and C. Rush, "Exponentially fast concentration of vector approximate message passing to its state evolution," in *2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2020, pp. 2670–2675.

[22] E. J. Candès and P. Sur, "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression," *Ann. Statist.*, vol. 48, no. 1, pp. 27–42, 2020.

[23] P. Capéraà, A.-L. Fougères, and C. Genest, "A stochastic ordering based on a decomposition of kendall's tau," in *Distributions with given Marginals and Moment Problems*, V. Beneš and J. Štěpán, Eds. Dordrecht: Springer Netherlands, 1997, pp. 81–86.

[24] W.-K. Chen and W.-K. Lam, "Universality of approximate message passing algorithms," *Electronic Journal of Probability*, vol. 26, pp. 1–44, 2021.

[25] P. B. Conn, D. S. Johnson, P. J. Williams, S. R. Melin, and M. B. Hooten, "A guide to bayesian model checking for ecologists," *Ecological Monographs*, vol. 88, no. 4, pp. 526–542, 2018.

[26] D. Donoho and A. Montanari, "High dimensional robust m-estimation: Asymptotic variance via approximate message passing," *Probability Theory and Related Fields*, vol. 166, no. 3, pp. 935–969, 2016.

[27] D. L. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," *IEEE transactions on information theory*, vol. 59, no. 11, pp. 7434–7464, 2013.

[28] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106 45, pp. 18 914–9, 2009.

[29] R. Dudeja, Y. M. Lu, and S. Sen, "Universality of approximate message passing with semi-random matrices," *arXiv preprint arXiv:2204.04281*, 2022.

[30] Z. Fan, "Approximate message passing algorithms for rotationally invariant matrices," *The Annals of Statistics*, vol. 50, no. 1, pp. 197–224, 2022.

[31] O. Y. Feng, R. Venkataramanan, C. Rush, R. J. Samworth, *et al.*, "A unifying tutorial on approximate message passing," *Foundations and Trends® in Machine Learning*, vol. 15, no. 4, pp. 335–536, 2022.

[32] A. Fletcher and P. Schniter, "Learning and free energies for vector approximate message passing," Mar. 2017, pp. 4247–4251.

[33] A. K. Fletcher, S. Rangan, and P. Schniter, "Inference in deep networks in high dimensions," in *2018 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2018, pp. 1884–1888.

[34] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman, "Visualization in bayesian workflow," *Journal of the Royal Statistical Society Series A: Statistics in Societ*, vol. 182, no. 2, pp. 389–402, 2019.

[35] A. Gelman, "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing," *International Statistical Review*, vol. 71, no. 2, pp. 369 –382, 2003.

[36] A. Gelman, "Two simple examples for understanding posterior p-values whose distributions are far from uniform," *Electronic Journal of Statistics*, vol. 7, no. none, pp. 2595 – 2602, 2013.

[37] A. Gelman, X.-L. Meng, and H. Stern, "Posterior predictive assessment of model fitness via realized discrepancies," *Statistica Sinica*, vol. 6, no. 4, pp. 733–760, 1996.

[38] A. Gelman *et al.*, *Bayesian workflow*, 2020.

[39] F. Gosselin, "A new calibrated bayesian internal goodness-of-fit method: Sampled posterior p-values as simple and general p-values that allow double use of the data," *PLOS ONE*, vol. 6, no. 3, pp. 1–10, Mar. 2011.

[40] S. Greenland, "Relaxation Penalties and Priors for Plausible Modeling of Nonidentified Bias Sources," *Statistical Science*, vol. 24, no. 2, pp. 195 –210, 2009.

[41] P. Gustafson, "What are the limits of posterior distributions arising from nonidentified models, and why should we care?" *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1682–1695, 2009.

[42] P. Gustafson *et al.*, "On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables [with comments and rejoinder]," *Statistical Science*, vol. 20, no. 2, pp. 111–140, 2005.

[43] M. Heidemanns, A. Gelman, and G. E. Morris, "An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election," *Harvard Data Science Review*, vol. 2, no. 4, 2020, https://hdsr.mitpress.mit.edu/pub/nw1dzd02.

[44] N. L. Hjort, F. A. Dahl, and G. H. Steinbakk, "Post-processing posterior predictive p values," *Journal of the American Statistical Association*, vol. 101, no. 475, pp. 1157–1174, 2006.

[45] T. Hou, Y. Liu, T. Fu, and J. Barbier, "Sparse superposition codes under vamp decoding with generic rotational invariant coding matrices," *arXiv preprint arXiv:2202.04541*, 2022.

[46] H. Hsu and F. Calmon, "Rashomon capacity: A metric for predictive multiplicity in classification," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 28 988–29 000.

[47] J. H. Huggins and J. W. Miller, *Robust inference and model criticism using bagged posteriors*, 2020. arXiv: `1912.07104 [stat.ME]`.

[48] J. H. Huggins and J. W. Miller, "Reproducible Model Selection Using Bagged Posteriors," *Bayesian Analysis*, vol. 18, no. 1, pp. 79 –104, 2023.

[49] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Information and Inference: A Journal of the IMA*, vol. 2, no. 2, pp. 115–144, 2013.

[50] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, pp. 620–630, 4 1957.

[51] H. Joe, *Dependence Modeling with Copulas* (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). Taylor & Francis, 2014, ISBN: 9781466583221.

[52] H. Joe, "Multivariate concordance," *Journal of Multivariate Analysis*, vol. 35, no. 1, pp. 12–30, 1990.

[53] V. E. Johnson, "A Bayesian χ^2 test for goodness-of-fit," *The Annals of Statistics*, vol. 32, no. 6, pp. 2361 –2384, 2004.

[54] V. E. Johnson, "Bayesian model assessment using pivotal quantities," *Bayesian Analysis*, vol. 2, no. 4, pp. 719 –733, 2007.

[55] V. E. Johnson, "Bayesian model assessment using pivotal quantities," *Bayesian Analysis*, vol. 2, no. 4, pp. 719 –733, 2007.

[56] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *Journal of Statistical Mechanics: Theory and Experiment*, 2012.

[57] S. Lele, K. Nadeem, and B. Schmuland, "Estimability and likelihood inference for generalized linear mixed models using data cloning," *Journal of the American Statistical Association*, vol. 105, pp. 1617–1625, Dec. 2010.

[58] G. Li, W. Fan, and Y. Wei, "Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization," *arXiv preprint arXiv:2302.03682*, 2023.

[59] G. Li and Y. Wei, "A non-asymptotic framework for approximate message passing in spiked models," *arXiv preprint arXiv:2208.03313*, 2022.

[60] L. Li, C. X. Feng, and S. Qiu, "Estimating cross-validatory predictiv p-values with integrated importance sampling for disease mapping models," *Statistics in Medicine*, 2017.

[61] L. Li, S. Qiu, B. Zhang, and C. X. Feng, "Approximating cross-validatory predictive evaluation in bayesian latent variable models with integrated IS and WAIC," *Statistics and Computing*, vol. 26, no. 4, pp. 881–897, 2015.

[62] H. Liu, C. Rush, and D. Baron, "Rigorous state evolution analysis for approximate message passing with side information," *IEEE Transactions on Information Theory*, 2022.

[63] J. Ma and L. Ping, "Orthogonal amp," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[64] Y. Ma, C. Rush, and D. Baron, "Analysis of approximate message passing with non-separable denoisers and markov random field priors," *IEEE Transactions on Information Theory*, vol. 65, no. 11, pp. 7367–7389, 2019.

[65] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová, "Multi-layer generalized linear estimation," *CoRR*, vol. abs/1701.06981, 2017. arXiv: 1701.06981.

[66] E. C. Marshall and D. J. Spiegelhalter, "Approximate cross-validatory predictive checks in disease mapping models," *Statistics in Medicine*, vol. 22, no. 10, pp. 1649–1660, 2003. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1403.

[67] X.-L. Meng, "Posterior predictive p-values," *The Annals of Statistics*, vol. 22, no. 3, pp. 1142–1160, 1994.

[68] T. P. Minka, *Expectation propagation for approximate bayesian inference*, 2013. arXiv: 1301.2294 [cs.AI].

[69] T. Minka *et al.*, "Divergence measures and message passing," Technical report, Microsoft Research, Tech. Rep., 2005.

[70] M. Mondelli, C. Thrampoulidis, and R. Venkataramanan, "Optimal combination of linear and spectral estimators for generalized linear models," *Foundations of Computational Mathematics*, vol. 22, no. 5, pp. 1513–1566, 2022.

[71] M. Mondelli and R. Venkataramanan, "Approximate message passing with spectral initialization for generalized linear models," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021, pp. 397–405.

[72] M. Mondelli and R. Venkataramanan, "Pca initialization for approximate message passing in rotationally invariant models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 616–29 629, 2021.

[73] A. Montanari, "Graphical models concepts in compressed sensing," in *Compressed Sensing*, Y. C. Eldar and G. Kutyniok, Eds., Cambridge University Press, 2012, pp. 394–438.

[74] M. Opper, B. Cakmak, and O. Winther, "A theory of solving tap equations for ising models with general invariant random matrices," *Journal of Physics A: Mathematical and Theoretical*, vol. 49, no. 11, p. 114 002, 2016.

[75] C. Prosser and J. Mellon, "The twilight of the polls? a review of trends in polling accuracy and the causes of polling misses," *Government and Opposition*, vol. 53, no. 4, 757–790, 2018.

[76] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 2168–2172, 2010.

[77] S. Rangan, A. K. Fletcher, V. K. Goyal, E. Byrne, and P. Schniter, "Vector approximate message passing," *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1588–1592, 2017.

[78] S. Rangan, A. K. Fletcher, P. Schniter, and U. S. Kamilov, "Inference for generalized linear models via alternating directions and bethe free energy minimization," vol. 63, no. 1, 676–697, 2017.

[79] G. O. Roberts and J. S. Rosenthal, "General state space markov chains and mcmc algorithms," *Probability Surveys*, vol. 1, no. none, Jan. 2004.

[80] J. M. Robins, A. van der Vaart, and V. Ventura, "Asymptotic distribution of p values in composite null models," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1143–1156, 2000.

[81] C. Rush and R. Venkataramanan, "Finite sample analysis of approximate message passing algorithms," *IEEE Transactions on Information Theory*, vol. 64, pp. 7264–7286, 2016.

[82] C. Rush and R. Venkataramanan, "The error exponent of sparse regression codes with amp decoding," in *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 2478–2482.

[83] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *IEEE Trans. Inf. Theor.*, vol. 66, no. 1, 302–323, 2020.

[84] P Schniter, S Rangan, and A. Fletcher, "Vector approximate message passing for the generalized linear model," *Proceedings of the 2016 50th Asilomar Conference on Signals, Systems and Computers*, 2017.

[85] G. H. Steinbakk and G. O. Storvik, "Posterior predictive p-values in bayesian hierarchical models," *Scandinavian Journal of Statistics*, vol. 36, no. 2, pp. 320–336, 2009.

[86] H. S. Stern and N. Cressie, "Posterior predictive model checks for disease mapping models," *Statistics in Medicine*, vol. 19, no. 17-18, pp. 2377–2397, 2000. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0258%2820000915/30%2919%3A17/18%3C2377%3A%3AAID-SIM576%3E3.0.CO%3B2-1`.

[87] P. Sur and E. J. Candès, "A modern maximum-likelihood theory for high-dimensional logistic regression," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14 516–14 525, 2019.

[88] K. Takeuchi, "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements," *IEEE Transactions on Information Theory*, vol. 66, Jan. 2017.

[89] R. van de Schoot *et al.*, "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, vol. 1, Jan. 2021.

[90] A. Vehtari, A. Gelman, and J. Gabry, "Practical bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2016.

[91] A. Vehtari and J. Lampinen, "Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities," *Neural Computation*, vol. 14, no. 10, pp. 2439–2468, Oct. 2002. eprint: `https://direct.mit.edu/neco/article-pdf/14/10/2439/815137/0899766260293292.pdf`.

[92] R. Venkataramanan, K. Kögler, and M. Mondelli, "Estimation in rotationally invariant generalized linear models via approximate message passing," in *International Conference on Machine Learning*, PMLR, 2022, pp. 22 120–22 144.

[93] T. Wang, X. Zhong, and Z. Fan, "Universality of approximate message passing algorithms and tensor networks," *arXiv preprint arXiv:2206.13037*, 2022.

[94] S. Watanabe and M. Opper, "Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory.," *Journal of machine learning research*, vol. 11, no. 12, 2010.

[95] Y. Xie and B. Carlin, "Measures of bayesian learning and identifiability in hierarchical models," *Journal of Statistical Planning and Inference*, vol. 136, pp. 3458–3477, Oct. 2006.

[96] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, 2521–2530, ISBN: 9781510860964.

[97] Y. Yano, S. Beal, and L. Sheiner, "Evaluating pharmacokinetic/pharmacodynamic models using the predictive check," *Journal of pharmacokinetics and pharmacodynamics*, vol. 28, pp. 171–92, Apr. 2001.

[98] Y. Yuan and V. Johnson, "Goodness-of-fit diagnostics for bayesian hierarchical models," *Biometrics*, vol. 68, pp. 156–64, Nov. 2011.

[99] J. L. Zhang, "Comparative investigation of three bayesian p values," *Computational Statistics & Data Analysis*, vol. 79, pp. 277–291, 2014.

[100] Y. Zhang, M. Mondelli, and R. Venkataramanan, "Precise asymptotics for spectral methods in mixed generalized linear models," *arXiv preprint arXiv:2211.11368*, 2022.

Appendix A: Appendices to chapter 2

A.1 Proof of lemma 5

To prove Lemma 5, we prove Lemma 6 below. Notice that part (a) and (b) of Lemma 5 are also part (a) and (b) of Lemma 6, only now Lemma 6 also has parts (c)-(g) as well.

Lemma 6. *Throughout the lemma, we use c and C to denote constants that do not depend on the iteration number, but which may vary in their exact value between usages. The values of C_k, C'_k, c_k, and c'_k for $k \geq 0$ are given in* (2.74)*. Let $X_n \ddot{=} c$ mean*

$$\mathbb{P}(|X_n - c| \geq \epsilon) \leq Ck^4 C'_{k-1} \exp\left(-cc'_{k-1} n\epsilon^2/k^7\right),$$

and let $X_n \dot{=} c$ mean

$$\mathbb{P}\left(|X_n - c| \geq \epsilon\right) \leq Ck^4 C_{k-1} \exp\left(-cc_{k-1} n\epsilon^2/k^7\right).$$

(a)

$$\mathbb{P}\left(\frac{1}{N}\|\boldsymbol{\Delta}_{qk}^{\text{in/out}}\|^2 \geq \epsilon\right) \leq CkC'_{k-1} \exp\left(-cc'_{k-1} n\epsilon/k^2\right), \tag{A.39}$$

$$\mathbb{P}\left(\frac{1}{N}\|\boldsymbol{\Delta}_{pk}^{\text{in/out}}\|^2 \geq \epsilon\right) \leq CkC_{k-1} \exp\left(-cc_{k-1} n\epsilon/k^2\right). \tag{A.40}$$

(b) Let $\phi : \mathbb{R}^{k+2} \to \mathbb{R}$ pseudo-Lipschitz, or satisfy the following bounded conditionally pseudo-Lipschitz condition:

   (a) ϕ is continuous in all inputs.

   (b) $\phi(\cdot, w)$ is pseudo-Lipschitz for each w with PL constants continuous in w.

(c) The domain of w is compact.

Then we have

$$\frac{1}{N}\sum_{i=1}^{N}\phi\left([\mathbf{q}_0^{\text{in}}]_i,...,[\mathbf{q}_{k+1}^{\text{in}}]_i,[\mathbf{q}_0^{\text{out}}]_i,...,[\mathbf{q}_{k+1}^{\text{out}}]_i,[\mathbf{w}]_i\right)\overset{..}{=}\mathbb{E}\left\{\phi\left(Q_0^{\text{in}},...,Q_{k+1}^{\text{in}},Q_0^{\text{out}},...,Q_{k+1}^{\text{out}},W\right)\right\}, \tag{A.41}$$

$$\frac{1}{N}\sum_{i=1}^{N}\phi\left([\mathbf{p}_0^{\text{in}}]_i,...,[\mathbf{p}_k^{\text{in}}]_i,[\mathbf{p}_0^{\text{out}}]_i,...,[\mathbf{p}_k^{\text{out}}]_i,[\mathbf{w}]_i\right)\overset{..}{=}\mathbb{E}\left\{\phi\left(P_0^{\text{in}},...,P_k^{\text{in}},P_0^{\text{out}},...,P_k^{\text{out}},W\right)\right\}, \tag{A.42}$$

where the $Q_i^{\text{in/out}}$ *are jointly Gaussian with* $Q_i^{\text{in/out}}\sim N(0,\tau_{qi}^{\text{in/out}})$ *and* $\mathbb{E}\{Q_i^{\text{in/out}}Q_j^{\text{in/out}}\}=[\mathbf{\Sigma}_{v,k+1}^{\text{in/out}}]_{ij}$*, the* $P_i^{\text{in/out}}$ *are jointly Gaussian with* $P_i^{\text{in/out}}\sim N(0,\tau_{pi}^{\text{in/out}})$ *and* $\mathbb{E}\{P_i^{\text{in/out}}P_j^{\text{in/out}}\}=[\mathbf{\Sigma}_{uk}^{\text{in/out}}]_{ij}$*, and* $\mathbf{w}$ *is any vector independent of the* $\mathbf{p}$ *and* $\mathbf{q}$ *iterates that concentrates exponentially fast (see Definition 3) around some limiting variable* W *with finite second moment.*

(c) For all $k\geq 0$*,*

$$\frac{1}{N}[\mathbf{q}_{k+1}^{\text{in/out}}]^T\mathbf{u}_0^{\text{in/out}}\overset{..}{=}0,\quad \frac{1}{N}[\mathbf{q}_k^{\text{in/out}}]^T\mathbf{w}_q^{\text{in/out}}\overset{..}{=}0, \tag{A.43}$$

$$\frac{1}{N}[\mathbf{p}_k^{\text{in/out}}]^T\mathbf{w}_p^{\text{in/out}}\overset{.}{=}0. \tag{A.44}$$

(d) For all $0\leq j\leq k$*,*

$$\frac{1}{N}[\mathbf{q}_j^{\text{in/out}}]^T\mathbf{q}_k^{\text{in/out}}\overset{..}{=}[\mathbf{\Sigma}_{v,k+1}^{\text{in/out}}]_{j+1,k+1}, \tag{A.45}$$

$$\frac{1}{N}[\mathbf{p}_j^{\text{in/out}}]^T\mathbf{p}_k^{\text{in/out}}\overset{.}{=}[\mathbf{\Sigma}_{uk}^{\text{in/out}}]_{j+1,k+1}. \tag{A.46}$$

(e) For all $1 \le j \le k+1$ *and for all* $0 \le i \le k$,

$$\frac{1}{N}[\mathbf{u}_j^{\text{in/out}}]^T \mathbf{u}_{k+1}^{\text{in/out}} \doteq [\mathbf{\Sigma}_{u,k+1}^{\text{in/out}}]_{j+1,k+2}, \quad \frac{1}{N}[\mathbf{u}_j^{\text{in/out}}]^T \mathbf{u}_0^{\text{in/out}} \doteq 0, \tag{A.47}$$

$$\frac{1}{N}[\mathbf{v}_i^{\text{in/out}}]^T \mathbf{v}_k^{\text{in/out}} \doteq [\mathbf{\Sigma}_{vk}^{\text{in/out}}]_{i+1,k+1}. \tag{A.48}$$

(f) For $0 \le j \le k$,

$$\begin{aligned}
\mathbb{P}\left(\left|\alpha_{qk}^{\text{in/out}} - \overline{\alpha}_{qk}^{\text{in/out}}\right| \ge \epsilon\right) &\le C' k^4 C'_{k-1} \exp\left(-nc' c'_{k-1}\epsilon^2/k^7\right), \\
\mathbb{P}\left(\left|\gamma_{p(k+1)}^{\text{in/out}} - \overline{\gamma}_{p(k+1)}^{\text{in/out}}\right| \ge \epsilon\right) &\le C' k^4 C'_{k-1} \exp\left(-nc' c'_{k-1}\epsilon^2/k^7\right), \\
\frac{1}{N}[\mathbf{q}_k^{\text{in/out}}]^T \mathbf{u}_{j+1}^{\text{in/out}} \doteq 0 \qquad & \frac{1}{N}[\mathbf{q}_j^{\text{in/out}}]^T \mathbf{u}_{k+1}^{\text{in/out}} \doteq 0.
\end{aligned} \tag{A.49}$$

For $0 \le j \le k$,

$$\begin{aligned}
\mathbb{P}\left(\left|\alpha_{pk}^{\text{in/out}} - \overline{\alpha}_{pk}^{\text{in/out}}\right| \ge \epsilon\right) &\le C k^4 C_{k-1} \exp\left(-ncc_{k-1}\epsilon^2/k^7\right), \\
\mathbb{P}\left(\left|\gamma_{qk}^{\text{in/out}} - \overline{\gamma}_{qk}^{\text{in/out}}\right| \ge \epsilon\right) &\le C k^4 C_{k-1} \exp\left(-ncc_{k-1}\epsilon^2/k^7\right), \\
\frac{1}{N}[\mathbf{p}_j^{\text{in/out}}]^T \mathbf{v}_k^{\text{in/out}} \doteq 0, \qquad & \frac{1}{N}[\mathbf{p}_k^{\text{in/out}}]^T \mathbf{v}_j^{\text{in/out}} \doteq 0.
\end{aligned} \tag{A.50}$$

(g) (i)

$$\mathbb{P}\left(\left|\frac{1}{N}\|[\mathbf{B}_{\mathbf{C}_{uk}^{\text{in/out}}}^{\perp}]^T \mathbf{q}_k^{\text{in/out}}\|^2 - \rho_{qk}^{\text{in/out}}\right| \ge \epsilon\right) \le C k^6 C'_{k-1} \exp\left(-ncc'_{k-1}/k^9\right), \tag{A.51}$$

$$\mathbb{P}\left(\left|\frac{1}{N}\|[\mathbf{B}_{\mathbf{C}_{vk}^{\text{in/out}}}^{\perp}]^T \mathbf{p}_k^{\text{in/out}}\|^2 - \rho_{pk}^{\text{in/out}}\right| \ge \epsilon\right) \le C k^6 C_{k-1} \exp\left(-ncc_{k-1}/k^9\right). \tag{A.52}$$

(ii) For $1 \le i, j \le 2(k+1)$,

$$\begin{aligned}
\mathbb{P}\left([\mathbf{C}_{q,k+1}^{\text{in/out}}]^T \mathbf{C}_{q,k+1}^{\text{in/out}} \textit{ singular}\right) &\le C k^6 C'_{k-1} \exp\left(-ncc'_{k-1}/k^9\right), \\
\mathbb{P}\left(\left|\left[\left([\mathbf{C}_{q,k+1}^{\text{in/out}}]^T \mathbf{C}_{q,k+1}^{\text{in/out}}\right)^{-1}\right]_{ij} - \left[\mathbf{\Pi}_{vk}^{\text{in/out}}\right]_{ij}\right| \ge \epsilon\right) &\le C k^6 C'_{k-1} \exp\left(-ncc'_{k-1}/k^9\right).
\end{aligned} \tag{A.53}$$

For $1 \leq i, j \leq 2k+1$,

$$\mathbb{P}\left([\mathbf{C}_{pk}^{\text{in/out}}]^T \mathbf{C}_{pk}^{\text{in/out}} \text{ singular}\right) \leq Ck^8 C_{k-1} \exp\left(-ncc_{k-1}/k^{11}\right),$$
$$\mathbb{P}\left(\left|\left[([\mathbf{C}_{pk}^{\text{in/out}}]^T \mathbf{C}_{pk}^{\text{in/out}})^{-1}\right]_{ij} - \left[\mathbf{\Pi}_{uk}^{\text{in/out}}\right]_{ij}\right| \geq \epsilon\right) \leq Ck^8 C_{k-1} \exp\left(-ncc_{k-1}/k^{11}\right). \tag{A.54}$$

(iii)

$$\mathbb{P}\left(\left|\frac{1}{N}\|[\mathbf{B}^{\perp}_{\mathbf{C}_{q,k+1}^{\text{in/out}}}]^T \mathbf{u}_{k+1}^{\text{in/out}}\|^2 - \rho_{p,k+1}^{\text{in/out}}\right| \geq \epsilon\right) \leq Ck^8 C'_{k-1} \exp\left(-ncc'_{k-1}\epsilon^2/k^{11}\right), \tag{A.55}$$
$$\mathbb{P}\left(\left|\frac{1}{N}\|[\mathbf{B}^{\perp}_{\mathbf{C}_{pk}^{\text{in/out}}}]^T \mathbf{v}_{k}^{\text{in/out}}\|^2 - \rho_{qk}^{\text{in/out}}\right| \geq \epsilon\right) \leq Ck^8 C_{k-1} \exp\left(-ncc_{k-1}\epsilon^2/k^{11}\right). \tag{A.56}$$

(iv) For $1 \leq i, j \leq 2k+3$,

$$\mathbb{P}\left([\mathbf{C}_{u,k+1}^{\text{in/out}}]^T \mathbf{C}_{u,k+1}^{\text{in/out}} \text{ singular}\right) \leq Ck^8 C'_{k-1} \exp\left(-ncc'_{k-1}/k^{11}\right),$$
$$\mathbb{P}\left(\left|\left[([\mathbf{C}_{u,k+1}^{\text{in/out}}]^T \mathbf{C}_{u,k+1}^{\text{in/out}})^{-1}\right]_{ij} - \left[\mathbf{\Pi}_{u(k+1)}^{\text{in/out}}\right]_{ij}\right| \geq \epsilon\right) \leq Ck^8 C'_{k-1} \exp\left(-ncc'_{k-1}/k^{11}\right). \tag{A.57}$$

For $1 \leq i, j \leq 2(k+1)$,

$$\mathbb{P}\left([\mathbf{C}_{v,k+1}^{\text{in/out}}]^T \mathbf{C}_{v,k+1}^{\text{in/out}} \text{ singular}\right) \leq Ck^8 C_{k-1} \exp\left(-ncc_{k-1}/k^{11}\right),$$
$$\mathbb{P}\left(\left|\left[([\mathbf{C}_{v,k+1}^{\text{in/out}}]^T \mathbf{C}_{v,k+1}^{\text{in/out}})^{-1}\right]_{ij} - \left[\mathbf{\Pi}_{vk}^{\text{in/out}}\right]_{ij}\right| \geq \epsilon\right) \leq Ck^8 C_{k-1} \exp\left(-ncc_{k-1}/k^{11}\right). \tag{A.58}$$

Remark 12. *In the course of the Lemma 6 proof, it often happens that the concentrating value of some quantity is a factor in the rate of concentration for another quantity (e.g. when we apply Lemma A.3). In theory, we must therefore take care about the varying magnitudes of these concentrating values as the iteration number changes. However, in our case, we can ignore these issues by bounding the relevant concentrating values independently of n and the concentration number.*

In particular, it suffices to find upper and lower bounds on $\tau^{\text{in/out}}_{(p/q)k}$ *and* $\rho^{\text{in/out}}_{(p/q)k}$ *and upper bounds on* $\overline{\alpha}^{\text{in/out}}_{(p/q)k}$ *for all* $k \geq 1$. *Our stopping criteria allows us to assume that the* $\tau^{\text{in/out}}_{(p/q)k}$ *and* $\rho^{\text{in/out}}_{(p/q)k}$ *are lower bounded independently of* k *in our concentration analysis. For an upper bound, observe that* $\rho^{\text{in/out}}_{(p/q)k} \leq \tau^{\text{in/out}}_{(p/q)k}$. *If either of the* $\tau^{\text{in}}_{(p/q)k}$ *is a not bounded, then by the translation* $\sigma_{1k} = \tau^{\text{in}}_{pk}$ *and* $\sigma_{2k} = \tau^{\text{in}}_{qk}$ *and Theorem 1 with* $\phi(x, y) = (x - y)^2$, *the state evolution prediction of the mean squared error of* $\hat{x}_{1k}$ *or* $\hat{x}_{2k}$ *will diverge as* $k \to \infty$. *Similarly, we can relate non-bounded behavior for* $\tau^{\text{out}}_{(p/q)k}$ *to predicted divergence of iterates in the state evolution.*

Finally, recall that by definition, $\overline{\alpha}^{\text{in/out}}_{pk} = \mathbb{E}\{(f^{\text{in/out}}_p)'(P^{\text{in}}_k, P^{\text{out}}_k, W^{\text{in/out}}_p, \overline{\gamma}^{\text{in}}_{pk}, \overline{\gamma}^{\text{out}}_{pk})\}$. *Now, under the strong concavity assumption (**Assumption 2**),* $(f^{\text{in/out}}_p)'$ *is bounded independently of* k, *so we trivially have an upper bound on the size of* $\overline{\alpha}^{\text{in/out}}_{pk}$ *tat is independent of* k, *and similarly for* $\overline{\alpha}^{\text{in/out}}_{qk}$.

The proof of Lemma 5 follows by induction on iteration k. We will refer to results (A.39), (A.41), (A.43), (A.45), (A.47), (A.49), (A.51), (A.53), (A.55), and (A.57) as $\mathbf{Q_{k+1}}.(\mathbf{a})$ - $\mathbf{Q_{k+1}}.(\mathbf{g})$ and results (A.40), (A.42), (A.44), (A.46), (A.48), (A.50), (A.52), (A.54), (A.56), and (A.58) as $\mathbf{P_k}.(\mathbf{a})$ - $\mathbf{P_k}.(\mathbf{g})$. Then the proof proceeds as follows:

1. Prove $\mathbf{P_0}$.

2. Assuming $\mathbf{P_0}$, prove $\mathbf{Q_1}$.

3. Assuming $\mathbf{P_{r-1}}$ and $\mathbf{Q_r}$ for $1 \leq \mathbf{r} \leq k$, prove $\mathbf{P_k}$.

4. Assuming $\mathbf{P_r}$ and $\mathbf{Q_r}$ for $1 \leq \mathbf{r} \leq k$, prove $\mathbf{Q_{k+1}}$.

A.1.1 Showing $\mathbf{P_0}$ holds

$\mathbf{P_0}.(\mathbf{a})$ We prove the result for $\mathbf{\Delta}^{\text{in}}_{p0}$ and the result for $\mathbf{\Delta}^{\text{out}}_{p0}$ follows in the same way. From Lemma 3, recall $\mathbf{\Delta}^{\text{in}}_{p0} = \left(\frac{\|\mathbf{u}^{\text{in}}_0\|}{\|\mathbf{Z}^{\text{in}}_{p0}\|} - \sqrt{\rho^{\text{in}}_{p0}}\right)\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{v0}}\mathbf{Z}^{\text{in}}_{p0}$. Therefore, we write

$$\mathbb{P}\left(\frac{\|\mathbf{\Delta}^{\text{in}}_{p0}\|}{\sqrt{N}} \geq \sqrt{\epsilon}\right) = \mathbb{P}\left(\frac{\|\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{v0}}\mathbf{Z}^{\text{in}}_{p0}\|}{\sqrt{N}}\left|\frac{\|\mathbf{u}^{\text{in}}_0\|}{\|\mathbf{Z}^{\text{in}}_{p0}\|} - \sqrt{\rho^{\text{in}}_{p0}}\right| \geq \sqrt{\epsilon}\right) \overset{(a)}{=} \mathbb{P}\left(\frac{\|\mathbf{Z}^{\text{in}}_{p0}\|}{\sqrt{N}}\left|\frac{\|\mathbf{u}^{\text{in}}_{p0}\|}{\|\mathbf{Z}^{\text{in}}_{p0}\|} - \sqrt{\rho^{\text{in}}_{p0}}\right| \geq \sqrt{\epsilon}\right)$$
$$\overset{(b)}{\leq} \mathbb{P}\left(\left|\frac{\|\mathbf{u}^{\text{in}}_{p0}\|}{\sqrt{N}} - \sqrt{\rho^{\text{in}}_{p0}}\right| \geq \frac{\sqrt{\epsilon}}{2}\right) + \mathbb{P}\left(\sqrt{\rho^{\text{in}}_{p0}}\left|\frac{\|\mathbf{Z}^{\text{in}}_{p0}\|}{\sqrt{N}} - 1\right| \geq \frac{\sqrt{\epsilon}}{2}\right), \tag{A.59}$$

where step (a) follows because $[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{v0}}]^T\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{v0}} = \mathbf{I}$ and step (b) from Lemma 9 using that

$$\frac{\|\mathbf{Z}^{\text{in}}_{p0}\|}{\sqrt{N}}\left|\frac{\|\mathbf{u}^{\text{in}}_{p0}\|}{\|\mathbf{Z}^{\text{in}}_{p0}\|} - \sqrt{\rho^{\text{in}}_{p0}}\right| = \left|\frac{\|\mathbf{u}^{\text{in}}_{p0}\|}{\sqrt{N}} - \sqrt{\frac{\rho^{\text{in}}_{p0}}{N}}\|\mathbf{Z}^{\text{in}}_{p0}\|\right| \leq \left|\frac{\|\mathbf{u}^{\text{in}}_{p0}\|}{\sqrt{N}} - \sqrt{\rho^{\text{in}}_{p0}}\right| + \sqrt{\rho^{\text{in}}_{p0}}\left|1 - \frac{\|\mathbf{Z}^{\text{in}}_{p0}\|}{\sqrt{N}}\right|.$$

Now, we bound the first term on the right side of (A.59) by $C\exp(-cN\epsilon)$ using **Condition 0** and Lemma 12 along with the fact that c depends on $\rho^{\text{in}}_{p0} > \epsilon'_1$ by the stopping criteria discussed in **Condition 5**. Similarly, using Lemmas 15 and 12 and that $\epsilon'_1 < \rho^{\text{in}}_{p0} < \infty$, we find

$$\mathbb{P}\left(\sqrt{\rho^{\text{in}}_{p0}}\left|\frac{\|\mathbf{Z}^{\text{in}}_{p0}\|}{\sqrt{N}} - 1\right| \geq \frac{\sqrt{\epsilon}}{2}\right) = \mathbb{P}\left(\left|\frac{\|\mathbf{Z}^{\text{in}}_{p0}\|}{\sqrt{N}} - 1\right| \geq \sqrt{\frac{\epsilon}{4\rho^{\text{in}}_{p0}}}\right) \leq C\exp\left(-cN\epsilon/\rho^{\text{in}}_{p0}\right).$$

$\mathbf{P_0}.(\mathbf{b})$ We prove the statement using a general vector $\mathbf{w}$, independent of $\mathbf{p}$ and $\mathbf{q}$. When this result is applied in later steps in the induction, this value will be $\mathbf{w}^{\text{in/out}}_p$. Let $\mathbf{w} = \left(\mathbf{w}^{(r)}, \mathbf{w}^{(d)}\right)$, where $\mathbf{w}^{(d)}$ and $\mathbf{w}^{(r)}$ are the deterministic and random components of $\mathbf{w}$ respectively.

Now for the proof, first we observe, using Lemma 3, that $\mathbf{p}^{\text{in}}_0|_{\mathcal{P}_0} \overset{d}{=} \sqrt{\rho^{\text{in}}_{p0}}\,\mathbf{O}^{\text{in}}_{p0}\,\overline{\mathbf{Z}}^{\text{in}}_{p0} + \mathbf{\Delta}^{\text{in}}_{p0}$, where we recall that $\overline{\mathbf{Z}}^{\text{in}}_{p0}$ is a length-N vector with independent, standard Gaussian entries that are independent of the conditioning sigma-algebra and $\mathbf{O}^{\text{in}}_{p0} \in \mathbb{R}^{N\times N}$ is a deterministic orthogonal matrix

(since we have $\mathbf{C}_{v0}^{\text{in}} = \emptyset$ as defined in (2.39)). Using the notation defined in Lemma 3,

$$\mathbf{O}_{p0}^{\text{in}} \overline{\mathbf{Z}}_{p0}^{\text{in}} = \mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp} \mathbf{Z}_{p0}^{\text{in}} + \mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}} \check{\mathbf{Z}}_{p0}^{\text{in}} = \mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp} \mathbf{Z}_{p0}^{\text{in}} \overset{d}{=} \mathbf{Z}_{p0}^{\text{in}}. \tag{A.60}$$

An analogous representation holds for $\mathbf{p}_0^{\text{out}}$. Therefore, using Lemma 9, we find

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} \phi\left([\mathbf{p}_0^{\text{in}}]_i, [\mathbf{p}_0^{\text{out}}]_i, [\mathbf{w}]_i\right) - \mathbb{E}\left\{\phi\left(P_0^{\text{in}}, P_0^{\text{out}}, W\right)\right\}\right| \geq \epsilon\right) \\
&\quad \leq \mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} \Big|\phi\left(\sqrt{\rho_{p0}^{\text{in}}}[\mathbf{Z}_{p0}^{\text{in}}]_i + [\mathbf{\Delta}_{p0}^{\text{in}}]_i, \sqrt{\rho_{p0}^{\text{out}}}[\mathbf{Z}_{p0}^{\text{out}}]_i + [\mathbf{\Delta}_{p0}^{\text{out}}]_i, [\mathbf{w}]_i\right) \\
&\qquad\qquad - \phi\left(\sqrt{\rho_{p0}^{\text{in}}}[\mathbf{Z}_{p0}^{\text{in}}]_i, \sqrt{\rho_{p0}^{\text{out}}}[\mathbf{Z}_{p0}^{\text{out}}]_i, [\mathbf{w}]_i\right)\Big| \geq \frac{\epsilon}{2}\Big) \\
&\quad + \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} \phi\left(\sqrt{\rho_{p0}^{\text{in}}}[\mathbf{Z}_{p0}^{\text{in}}]_i, \sqrt{\rho_{p0}^{\text{out}}}[\mathbf{Z}_{p0}^{\text{out}}]_i, [\mathbf{w}]_i\right) - \mathbb{E}\left\{\phi\left(P_0^{\text{in}}, P_0^{\text{out}}, W\right)\right\}\right| \geq \frac{\epsilon}{2}\right),
\end{aligned} \tag{A.61}$$

where to make the above well-defined, we continue our convention of letting $[\mathbf{Z}_{p0}^{\text{out}}]_i = [\mathbf{\Delta}_{v0}^{\text{out}}]_i = 0$ for $M < i \leq N$. Label the terms on the right side of (A.61) as T_1 and T_2. If ϕ is pseudo-Lipschitz over all inputs jointly, we can bound T_1 using the triangle inequality and Cauchy-Schwarz as follows:

$$\begin{aligned}
T_1 &\leq \mathbb{P}\Big(\frac{L}{N}\sum_{i=1}^{N}\Big[1 + 2\sqrt{\rho_{p0}^{\text{in}}}\left|[\mathbf{Z}_{p0}^{\text{in}}]_i\right| + 2\sqrt{\rho_{p0}^{\text{out}}}\left|[\mathbf{Z}_{p0}^{\text{out}}]_i\right| + 2\left|[\mathbf{w}]_i\right| + \left|[\mathbf{\Delta}_{p0}^{\text{in}}]_i\right| + \left|[\mathbf{\Delta}_{p0}^{\text{out}}]_i\right|\Big] \\
&\qquad\qquad \times \Big[\left|[\mathbf{\Delta}_{p0}^{\text{in}}]_i\right| + \left|[\mathbf{\Delta}_{p0}^{\text{out}}]_i\right|\Big] \geq \frac{\epsilon}{2}\Big) \\
&\leq \mathbb{P}\left(\frac{\|\mathbf{\Delta}_{p0}^{\text{in}}\| + \|\mathbf{\Delta}_{p0}^{\text{out}}\|}{\sqrt{N}}\left[1 + 2\sqrt{\rho_{p0}^{\text{in}}}\frac{\|\mathbf{Z}_{p0}^{\text{in}}\|}{\sqrt{N}} + 2\sqrt{\rho_{p0}^{\text{out}}}\frac{\|\mathbf{Z}_{p0}^{\text{out}}\|}{\sqrt{N}} + 2\frac{\|\mathbf{w}\|}{\sqrt{N}} + \frac{\|\mathbf{\Delta}_{p0}^{\text{in}}\|}{\sqrt{N}} + \frac{\|\mathbf{\Delta}_{p0}^{\text{out}}\|}{\sqrt{N}}\right] \geq \frac{\epsilon}{4L}\right) \\
&\leq \mathbb{P}\left(\left[\frac{\|\mathbf{\Delta}_{p0}^{\text{in}}\|}{\sqrt{N}} + \frac{\|\mathbf{\Delta}_{p0}^{\text{out}}\|}{\sqrt{M}}\right]\left[1 + \sqrt{\rho_{p0}^{\text{in}}}\frac{\|\mathbf{Z}_{p0}^{\text{in}}\|}{\sqrt{N}} + \sqrt{\rho_{p0}^{\text{out}}}\frac{\|\mathbf{Z}_{p0}^{\text{out}}\|}{\sqrt{M}} + \frac{\|\mathbf{w}\|}{\sqrt{N}} + \frac{\|\mathbf{\Delta}_{p0}^{\text{in}}\|}{\sqrt{N}} + \frac{\|\mathbf{\Delta}_{p0}^{\text{out}}\|}{\sqrt{M}}\right] \geq \frac{\epsilon}{8L}\right),
\end{aligned}$$

where the last inequality follows because $M \leq N$. Now we observe that $\frac{1}{\sqrt{N}}\|\mathbf{\Delta}_{p0}^{\text{in}}\|$ and $\frac{1}{\sqrt{M}}\|\mathbf{\Delta}_{p0}^{\text{out}}\|$ concentrate around 0 by $\mathbf{P_0}.(\mathbf{a})$, $\frac{1}{\sqrt{N}}\|\mathbf{Z}_{p0}^{\text{in}}\|$ and $\frac{1}{\sqrt{M}}\|\mathbf{Z}_{p0}^{\text{out}}\|$ concentrate around 1 by Lemmas 12 and

15, and $\frac{1}{\sqrt{N}}\|\mathbf{w}\|$ is further bounded by $\frac{1}{\sqrt{N}}\|\mathbf{w}^{(r)}\|+2S_{\max}$, which concentrates around $\sqrt{\mathbb{E}W^2}+2S_{\max}$ by the assumption of the part (b) statement. The overall concentration of term T_1 then follows from application of Lemmas 11 and 9. We note that the rate of concentration will depend on the values $\rho_{p0}^{\text{in}}, \rho_{p0}^{\text{out}}$, and $\mathbb{E}W^2$ all of which are assumed to be finite, with ρ_{p0}^{in} and ρ_{p0}^{out} not too small by the stopping criteria discussed in **Condition 5**; so all are absorbed into the constants.

If on the other hand ϕ satisfies the bounded conditionally pseudo-Lipschitz property, then the functions $f_i = \phi(\cdot, \cdot, [\mathbf{w}]_i)$ are each pseudo-Lipschitz with PL constants that satisfy a common upper bound (since the PL constants are continuous in the $[\mathbf{w}]_i$, which are themselves contained in a compact set by defintion). In this case, we can apply the exact same strategy as above with the f_i in place of the ϕ and obtain an analogous bound where the $\|\mathbf{w}\|/\sqrt{N}$ term is dropped and L is now an upper bound for the PL constants of the f_i.

For T_2, note that for all i for which the expression is defined, we have by part 1 of Lemma 4 that

$$[\overset{*}{\mathbf{p}}_0^{\text{in/out}}]_i = \sqrt{\rho_{p0}^{\text{in/out}}}[\widetilde{\mathbf{O}}_{p0}^{\text{in}}\bar{\mathbf{Z}}_{p0}^{\text{in/out}}]_i \overset{d}{=} \sqrt{\rho_{p0}^{\text{in/out}}}[\mathbf{Z}_{p0}^{\text{in/out}}]_i, \tag{A.62}$$

where the final equality uses that $\widetilde{\mathbf{O}}_{p0}^{\text{in}} = \mathbf{O}_{p0}^{\text{in}}$, due to the initialization of Algorithm 5 (see the discussion following the presentation of the algorithm) and (A.60); hence, $\widetilde{\mathbf{O}}_{p0}^{\text{in}}\bar{\mathbf{Z}}_{p0}^{\text{in/out}} \overset{d}{=} \mathbf{Z}_{p0}^{\text{in}}$. Moreover, by part 3 of Lemma 4,

$$[\overset{*}{\mathbf{p}}_0^{\text{in/out}}]_i \overset{d}{=} P_0^{\text{in/out}}. \tag{A.63}$$

Thus, if ϕ is jointly pseudo-Lipschitz in all inputs, then by Lemma 18, the $f_i = \phi\left(\cdot, \cdot, \cdot, \mathbf{w}^{(d)}\right)$ are again pseudo-Lipschitz with common constant (by the boundedness of $\mathbf{w}^{(d)}$). We can thus apply Lemma 17 to these f_i to get concentration for $T2$.

If instead ϕ satisfies the conditionally bounded pseudo-Lipschitz condition, then letting $\mathbb{E}_P$

denote the expectation with respect to $(P_0^{\text{in}}, P_0^{\text{out}})$ and using Lemma 9, we get the further inequality

$$\begin{aligned} T_2 &= \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\sqrt{\rho_{p0}^{\text{in}}}[\mathbf{Z}_{p0}^{\text{in}}]_i, \sqrt{\rho_{p0}^{\text{out}}}[\mathbf{Z}_{p0}^{\text{out}}]_i, [\mathbf{w}]_i\right) - \mathbb{E}\left\{\phi\left(P_0^{\text{in}}, P_0^{\text{out}}, W\right)\right\}\right| \geq \frac{\epsilon}{2}\right) \\ &\leq \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\sqrt{\rho_{p0}^{\text{in}}}[\mathbf{Z}_{p0}^{\text{in}}]_i, \sqrt{\rho_{p0}^{\text{out}}}[\mathbf{Z}_{p0}^{\text{out}}]_i, [\mathbf{w}]_i\right) - \mathbb{E}_P\left\{\phi\left(P_0^{\text{in}}, P_0^{\text{out}}, [\mathbf{w}]_i\right)\right\}\right| \geq \frac{\epsilon}{4}\right) \\ &\quad + \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_P\left\{\phi\left(P_0^{\text{in}}, P_0^{\text{out}}, [\mathbf{w}]_i\right)\right\} - \mathbb{E}\left\{\phi(P_0^{\text{in}}, P_0^{\text{out}}, W)\right\}\right| \geq \frac{\epsilon}{4}\right), \end{aligned} \tag{A.64}$$

Let the first term on the right side of (A.64) be denoted $T21$ and the second term be denoted $T22$. For $T21$, note that if $f_i(p_1, p_2) = \phi(p_1, p_2, [\mathbf{w}]_i)$, then by the conditionally bounded pseudo-Lipschitz condition, the f_i are pseudo-Lipschitz with constant continuous in $[\mathbf{w}]_i$. But since the $[\mathbf{w}]_i$ are bounded, so too are these constants. Thus, the f_i are all pseudo-Lipschitz with some common constant B, and Lemma 17 again gives concentration for T21.

For T22, the continuity of $x \mapsto \mathbb{E}_P\left\{\phi(P_0^{\text{in}}, P_0^{\text{out}}, x)\right\}$ and the fact that the $[\mathbf{w}]_i$ are bounded and either i.i.d. or deterministic implies that the terms of $T22$ are bounded and independent. Thus, the desired concentration rate is given by Hoeffding's inequality.

$\mathbf{P_0}.(\mathbf{c})$ The result follows using $\mathbf{P_0}.(\mathbf{b})$ with the pseudo-Lipschitz functions $\phi(p^{\text{in}}, p^{\text{out}}, w) = p^{\text{in}}w$ and $\phi(p^{\text{in}}, p^{\text{out}}, w) = p^{\text{out}}w$ along with the facts that $\mathbb{E}\{P_0^{\text{in}}W_p^{\text{in}}\} = 0$ and $\mathbb{E}\{P_0^{\text{out}}W_p^{\text{out}}\} = 0$, since $W_p^{\text{in/out}}$ and $P_0^{\text{in/out}}$ are independent.

$\mathbf{P_0}.(\mathbf{d})$ The result follows using $\mathbf{P_0}.(\mathbf{b})$ with the pseudo-Lipschitz functions $\phi(p^{\text{in}}, p^{\text{out}}, w) = (p^{\text{in}})^2$ and $\phi(p^{\text{in}}, p^{\text{out}}, w) = (p^{\text{out}})^2$.

$\mathbf{P_0}.(\mathbf{e}) - (\mathbf{f})$ These are proved in exactly the same way as $\mathbf{P_k}.(\mathbf{e})$ and $\mathbf{P_k}.(\mathbf{f})$ in Section A.1.3 with $j = k = 0$ throughout and where references to $\mathbf{Q_k}.(\mathbf{f})$ are replaced by our assumption that the initial sequences $\gamma_{p0}^{\text{in/out}}$ converges to $\overline{\gamma}_{p0}^{\text{in/out}}$. Thus, we omit the details here and refer to the proof of the more general case of $\mathbf{P_k}.(\mathbf{e})$ and $\mathbf{P_k}.(\mathbf{f})$.

$\mathbf{P_0}.(\mathbf{g}).(\mathbf{i})$ Because $\mathbf{C}_{v0}^{\text{in}} = \emptyset$ is empty, $\mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp}$ is a deterministic orthogonal matrix; therefore, $\|[\mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp}]^T\mathbf{p}_0^{\text{in}}\|^2 = \|\mathbf{p}_0^{\text{in}}\|^2 = \|\mathbf{u}_0^{\text{in}}\|^2$ where the second equality follows as $\mathbf{p}_0^{\text{in}} = \mathbf{V}\mathbf{u}_0^{\text{in}}$ in Algorithm 3. The result then follows from the assumed concentration of $\mathbf{u}_0^{\text{in}}$ and the definition of ρ_{p0}^{in} in (2.72).

$\mathbf{P_0}.(\mathbf{g}).(\mathbf{ii})$ By (2.37), $\mathbf{C}_{p0}^{\text{in/out}} = \mathbf{p}_0^{\text{in/out}}$. Thus, the invertibility of $[\mathbf{C}_{p0}^{\text{in/out}}]^T\mathbf{C}_{p0}^{\text{in/out}}$ just requires that $[\mathbf{p}_0^{\text{in/out}}]^T\mathbf{p}_0^{\text{in/out}} \neq 0$, which is true with probability one. Moreover, by $\mathbf{P_0}.(\mathbf{d})$, we know that $[\mathbf{p}_0^{\text{in/out}}]^T\mathbf{p}_0^{\text{in/out}}/N$ concentrates to $\mathbb{E}\left[P_0^{\text{in/out}}\right]^2$ and since, by $\mathbf{P_0}.(\mathbf{d})$ and (2.73),

$$\left[\mathbf{\Pi}_{u0}^{\text{in/out}}\right]^{-1} = \mathbb{E}\left[U_0^{\text{in/out}}\right]^2 = \lim_{N\to\infty}\frac{1}{N}\|\mathbf{u}_0^{\text{in/out}}\|^2 = \lim_{N\to\infty}\frac{1}{N}\|\mathbf{p}_0^{\text{in/out}}\|^2 = \mathbb{E}\left[P_0^{\text{in/out}}\right]^2,$$

the desired result follows by Lemma 13. Application of Lemma 13 will put a power of $\tau_{p0}^{\text{in/out}}$ in the numerator of the rate of concentration; this can be lower bounded by the stopping time assumption.

$\mathbf{P_0}.(\mathbf{g}).(\mathbf{iii})$ We prove the 'in' version of the result, while the 'out' version follows correspondingly. Recall, $\mathbf{C}_{p0}^{\text{in}} = \mathbf{p}_0^{\text{in}}$ by (2.37) and $\mathbf{B}_{\mathbf{C}_{p0}^{\text{in}}}^{\perp} \in \mathbb{R}^{N\times N-1}$ is an orthogonal matrix with columns that form an orthonormal basis for $\text{range}(\mathbf{C}_{p0}^{\text{in}})^{\perp} = [\mathbf{p}_0^{\text{in}}]^{\perp}$. Thus, the matrix $[\mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp}\ \mathbf{p}_0^{\text{in}}/\|\mathbf{p}_0^{\text{in}}\|] \in \mathbb{R}^{N\times N}$ is orthogonal and

$$\|\mathbf{v}_0^{\text{in}}\|^2 = \left\|\begin{bmatrix}[\mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp}]^T\\ \|\mathbf{p}_0^{\text{in}}\|^{-1}[\mathbf{p}_0^{\text{in}}]^T\end{bmatrix}\mathbf{v}_0^{\text{in}}\right\|^2 = \|[\mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp}]^T\mathbf{v}_0^{\text{in}}\|^2 + \frac{([\mathbf{p}_0^{\text{in}}]^T\mathbf{v}_0^{\text{in}})^2}{\|\mathbf{p}_0^{\text{in}}\|^2}. \tag{A.65}$$

Rearranging, this gives us

$$\|[\mathbf{B}_{\mathbf{C}_{p0}^{\text{in}}}^{\perp}]^T\mathbf{v}_0^{\text{in}}\|^2 = \|\mathbf{v}_0^{\text{in}}\|^2 - \frac{([\mathbf{v}_0^{\text{in}}]^T\mathbf{p}_0^{\text{in}})^2}{\|\mathbf{p}_0^{\text{in}}\|^2}. \tag{A.66}$$

Finally, recalling that $\rho_{q0}^{\text{in}} = \mathbb{E}\{[V_0^{\text{in}}]^2\}$ from (2.72) and using Lemma 9, we have

$$\mathbb{P}\left(\left|\frac{1}{N}\|[\mathbf{B}_{\mathbf{C}_{p0}^{\text{in}}}^{\perp}]^T\mathbf{v}_0^{\text{in}}\|^2 - \rho_{q0}^{\text{in}}\right| \geq \epsilon\right) \leq \mathbb{P}\left(\left|\frac{1}{N}\|\mathbf{v}_0^{\text{in}}\|^2 - \mathbb{E}\{[V_0^{\text{in}}]^2\}\right| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(\frac{([\mathbf{v}_0^{\text{in}}]^T\mathbf{p}_0^{\text{in}})^2}{\|\mathbf{p}_0^{\text{in}}\|^2} \geq \frac{\epsilon}{2}\right). \tag{A.67}$$

The first term concentrates by $\mathbf{P_0}.(\mathbf{e})$ since $\mathbf{\Sigma}_{v0}^{\text{in}} = \mathbb{E}\{[V_0^{\text{in}}]^2\}$ by definition. The second term concentrates by $\mathbf{P_0}.(\mathbf{d})$, $\mathbf{P_0}.(\mathbf{f})$, Lemma 13, and Lemma 11.

$\mathbf{P_0}.(\mathbf{g}).(\mathbf{iv})$ To show that $[\mathbf{C}_{v1}^{\text{in}}]^T\mathbf{C}_{v1}^{\text{in}}$ is invertible with high probability, recall that $\mathbf{C}_{v1}^{\text{in}} = [\mathbf{P}_0^{\text{in}}\ \mathbf{V}_0^{\text{in}}] = [\mathbf{p}_0^{\text{in}}\ \mathbf{v}_0^{\text{in}}]$ by (2.37); therefore, $\det\left([\mathbf{C}_{v1}^{\text{in}}]^T\mathbf{C}_{v1}^{\text{in}}\right) = \|\mathbf{p}_0^{\text{in}}\|^2\|\mathbf{v}_0^{\text{in}}\|^2 - ([\mathbf{p}_0^{\text{in}}]^T\mathbf{v}_0^{\text{in}})$. Not-

ing that $\|\mathbf{p}_0^{\text{in/out}}\| = \|\mathbf{u}_0^{\text{in/out}}\|$, we have by $\mathbf{P_0}.(\mathbf{e})(\mathbf{f})$, **Condition 0**, and Lemmas 9 and 10 that this concentrates around $(\mathbf{\Sigma}_{u0}^{\text{in}})(\mathbf{\Sigma}_{v0}^{\text{in}}) > 0$. Then we have

$$([\mathbf{C}_{v1}^{\text{in}}]^T\mathbf{C}_{v1}^{\text{in}})^{-1} = \frac{1}{\det\left([\mathbf{C}_{v1}^{\text{in}}]^T\mathbf{C}_{v1}^{\text{in}}\right)}\begin{bmatrix} [\mathbf{v}_0^{\text{in}}]^T\mathbf{v}_0^{\text{in}} & -[\mathbf{p}_0^{\text{in}}]^T\mathbf{v}_0^{\text{in}} \\ -[\mathbf{p}_0^{\text{in}}]^T\mathbf{v}_0^{\text{in}} & [\mathbf{p}_0^{\text{in}}]^T\mathbf{p}_0^{\text{in}} \end{bmatrix}. \tag{A.68}$$

By similar reasoning, namely via $\mathbf{P_0}.(\mathbf{d})(\mathbf{e})(\mathbf{f})$ and Lemma 10, the entries of (A.68) have the desired concentrate concentrating constant given below using the definition of $\mathbf{\Pi}_{v0}^{\text{in}}$ given in (2.73).

$$\begin{bmatrix} (\mathbf{\Sigma}_{u0}^{\text{in}})^{-1} & 0 \\ 0 & (\mathbf{\Sigma}_{v0}^{\text{in}})^{-1} \end{bmatrix} = \mathbf{\Pi}_{v0}^{\text{in}}.$$

A.1.2 Showing $\mathbf{Q_1}$ holds

$\mathbf{Q_1}.(\mathbf{a})$ First, recall that by Lemma 3,

$$\mathbf{\Delta}_{q0}^{\text{in}} = \frac{[\mathbf{p}_0^{\text{in}}]^T\mathbf{v}_0^{\text{in}}}{\|\mathbf{u}_0^{\text{in}}\|^2}\mathbf{u}_0^{\text{in}} + \left(\frac{\|[\mathbf{B}_{\mathbf{C}_{p0}^{\text{in}}}^{\perp}]^T\mathbf{v}_0^{\text{in}}\|}{\|\mathbf{Z}_{q0}^{\text{in}}\|} - \sqrt{\rho_{q0}^{\text{in}}}\right)\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp}\mathbf{Z}_{q0}^{\text{in}} - \sqrt{\rho_{q0}^{\text{in}}}\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}\check{\mathbf{Z}}_{q0}^{\text{in}}.$$

Labeling the above three terms T_1 - T_3, we have $\|\mathbf{\Delta}_{q0}^{\text{in}}\|^2 = \|T_1+T_2+T_3\|^2 \leq 3\|T_1\|^2+3\|T_2\|^2+3\|T_3\|^2$ by Lemma 16. Hence, it suffices to derive bounds for $\mathbb{P}(\frac{1}{N}\|T_i\|^2 \geq \frac{\epsilon}{9})$ for $i \in \{1,2,3\}$ by Lemma 9.

For T_1, we have $\|T_1\|^2 = \frac{([\mathbf{p}_0^{\text{in}}]^T\mathbf{v}_0^{\text{in}})^2}{\|\mathbf{u}_0^{\text{in}}\|^2}$ which concentrates by our **Condition 0** on $\mathbf{u}_0$, $\mathbf{P_0}.(\mathbf{f})$ and Lemma 13.

For T_2, observe that $\|\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp}\mathbf{Z}_{q0}^{\text{in}}\|^2 = [\mathbf{Z}_{q0}^{\text{in}}]^T[\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp}]^T\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp}\mathbf{Z}_{q0}^{\text{in}} = \|\mathbf{Z}_{q0}^{\text{in}}\|^2$ using $[\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp}]^T\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp} = \mathbf{I}$ and therefore

$$\|T_2\|^2 = \left|\frac{\|[\mathbf{B}_{\mathbf{C}_{p0}^{\text{in}}}^{\perp}]^T\mathbf{v}_0^{\text{in}}\|}{\|\mathbf{Z}_{q0}^{\text{in}}\|} - \sqrt{\rho_{q0}^{\text{in}}}\right|^2 \|\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp}\mathbf{Z}_{q0}^{\text{in}}\|^2 = \left|\frac{\|[\mathbf{B}_{\mathbf{C}_{p0}^{\text{in}}}^{\perp}]^T\mathbf{v}_0^{\text{in}}\|}{\|\mathbf{Z}_{q0}^{\text{in}}\|} - \sqrt{\rho_{q0}^{\text{in}}}\right|^2 \|\mathbf{Z}_{q0}^{\text{in}}\|^2$$
$$= \left|\|[\mathbf{B}_{\mathbf{C}_{p0}^{\text{in}}}^{\perp}]^T\mathbf{v}_0^{\text{in}}\| - \sqrt{\rho_{q0}^{\text{in}}}\|\mathbf{Z}_{q0}^{\text{in}}\|\right|^2.$$

By Lemma 9 along with $\mathbf{P_0}.(\mathbf{g}).(\mathbf{iii})$, Lemma 15, and Lemma 12, we have the desired bound using

$$\mathbb{P}\left(\frac{1}{\sqrt{N}}\|T_2\| \geq \sqrt{\frac{\epsilon}{9}}\right) \leq \mathbb{P}\left(\frac{1}{\sqrt{N}}\left|\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{p0}}]^T\mathbf{v}^{\text{in}}_0\| - \sqrt{\rho^{\text{in}}_{q0}}\right| + \sqrt{\frac{\rho^{\text{in}}_{q0}}{N}}\left|\|\mathbf{Z}^{\text{in}}_{q0}\| - 1\right| \geq \sqrt{\frac{\epsilon}{9}}\right).$$

Finally, notice that $\|T_3\| = \sqrt{\rho^{\text{in}}_{q0}}\|\mathbf{B}_{\mathbf{C}^{\text{in}}_{u0}}\check{\mathbf{Z}}^{\text{in}}_{q0}\|$. Observe that because rank $\left(\mathbf{C}^{\text{in}}_{u0}\right) = 1$, the matrix $\mathbf{B}_{\mathbf{C}^{\text{in}}_{u0}}$ is just a unit column vector. Furthermore, $\mathbf{Z}^{\text{in}}_{q0}$ has length 1 (by definition); thus, is a standard normal. Combined, these give us that $\|\mathbf{B}_{\mathbf{C}^{\text{in}}_{u0}}\check{\mathbf{Z}}^{\text{in}}_{q0}\| = |\check{\mathbf{Z}}^{\text{in}}_{q0}|\|\mathbf{B}_{\mathbf{C}^{\text{in}}_{u0}}\| = |\check{\mathbf{Z}}^{\text{in}}_{q0}| \overset{d}{=} |\check{Z}|$ with $\check{Z} \sim N(0,1)$. So the above concentrates by Lemma 14.

$\mathbf{Q_1}.(\mathbf{b})$ The first part of the proof of part (b) follows as in $\mathbf{P_0}$. Using Lemma (4) part 2, we have $[\mathbf{q}^{\text{in/out}}_0]_i \overset{d}{=} [\widetilde{\mathbf{q}}^{\text{in/out}}_0]_i$; hence, by Lemma 9 we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left([\mathbf{q}^{\text{in}}_0]_i, [\mathbf{q}^{\text{out}}_0]_i, [\mathbf{w}^{\text{in}}_q]_i\right) - \mathbb{E}\phi\left(Q^{\text{in}}_0, Q^{\text{out}}_0, W^{\text{in}}_q\right)\right| \geq \epsilon\right) \\
&= \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left([\widetilde{\mathbf{q}}^{\text{in}}_0]_i, [\widetilde{\mathbf{q}}^{\text{out}}_0]_i, [\mathbf{w}^{\text{in}}_q]_i\right) - \mathbb{E}\phi\left(Q^{\text{in}}_0, Q^{\text{out}}_0, W^{\text{in}}_q\right)\right| \geq \epsilon\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left([\widetilde{\mathbf{q}}^{\text{in}}_0]_i, [\widetilde{\mathbf{q}}^{\text{out}}_0]_i, [\mathbf{w}^{\text{in}}_q]_i\right) - \frac{1}{N}\sum_{i=1}^{N}\phi\left([\overset{*}{\mathbf{q}}{}^{\text{in}}_0]_i, [\overset{*}{\mathbf{q}}{}^{\text{out}}_0]_i, [\mathbf{w}^{\text{in}}_q]_i\right)\right| \geq \frac{\epsilon}{2}\right) \\
&\qquad + \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left([\overset{*}{\mathbf{q}}{}^{\text{in}}_0]_i, [\overset{*}{\mathbf{q}}{}^{\text{out}}_0]_i, [\mathbf{w}^{\text{in}}_q]_i\right) - \mathbb{E}\phi\left(Q^{\text{in}}_0, Q^{\text{out}}_0, W^{\text{in}}_q\right)\right| \geq \frac{\epsilon}{2}\right).
\end{aligned}$$

Now, using Lemma 17/Hoeffding's inequality and Lemma 4 part 3, we can bound the second term above using exactly the same strategy as in the proof of $\mathbf{P_0}.(\mathbf{b})$. For the first term, if ϕ is pseudo-Lipschitz in all inputs, we can apply the pseudo-Lipschitz property and Lemma 4 part 2 to

get

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left([\widetilde{\mathbf{q}}_0^{\text{in}}]_i,[\widetilde{\mathbf{q}}_0^{\text{out}}]_i,[\mathbf{w}_q^{\text{in}}]_i\right)-\frac{1}{N}\sum_{i=1}^{N}\phi\left([\overset{*}{\mathbf{q}}_0^{\text{in}}]_i,[\overset{*}{\mathbf{q}}_0^{\text{out}}]_i,[\mathbf{w}_q^{\text{in}}]_i\right)\right|\geq\frac{\epsilon}{2}\right)\\
&\quad\leq\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\left|\phi\left([\widetilde{\mathbf{q}}_0^{\text{in}}]_i,[\widetilde{\mathbf{q}}_0^{\text{out}}]_i,[\mathbf{w}_q^{\text{in}}]_i\right)-\phi\left([\overset{*}{\mathbf{q}}_0^{\text{in}}]_i,[\overset{*}{\mathbf{q}}_0^{\text{out}}]_i,[\mathbf{w}_q^{\text{in}}]_i\right)\right|\geq\frac{\epsilon}{2}\right)\\
&\quad\leq\mathbb{P}\left(\frac{L}{N}\sum_{i=1}^{N}[|[\boldsymbol{\Delta}_{q0}^{\text{in}}]_i|+|[\boldsymbol{\Delta}_{q0}^{\text{out}}]_i|]\left[1+2|[\overset{*}{\mathbf{q}}_0^{\text{in}}]_i|+2|[\overset{*}{\mathbf{q}}_0^{\text{out}}]_i|+2|[\mathbf{w}_q^{\text{in}}]_i|+|[\boldsymbol{\Delta}_{q0}^{\text{in}}]_i|+|[\boldsymbol{\Delta}_{q0}^{\text{out}}]_i|\right]\geq\frac{\epsilon}{2}\right)\\
&\quad\leq\mathbb{P}\left(L\sum_{i=1}^{N}\left[\frac{|[\boldsymbol{\Delta}_{q0}^{\text{in}}]_i|}{\sqrt{N}}+\frac{|[\boldsymbol{\Delta}_{q0}^{\text{out}}]_i|}{\sqrt{M}}\right]\left[1+2\frac{|[\overset{*}{\mathbf{q}}_0^{\text{in}}]_i|}{\sqrt{N}}+2\frac{|[\overset{*}{\mathbf{q}}_0^{\text{out}}]_i|}{\sqrt{M}}+2\frac{|[\mathbf{w}_q^{\text{in}}]_i|}{\sqrt{N}}+\frac{|[\boldsymbol{\Delta}_{q0}^{\text{in}}]_i|}{\sqrt{N}}+\frac{|[\boldsymbol{\Delta}_{q0}^{\text{out}}]_i|}{\sqrt{M}}\right]\geq\frac{\epsilon}{2}\right),
\end{aligned}$$

where the last inequality follows from the fact that $M \leq N$. Now this term can be shown to concentrate in exactly the same way as in $\mathbf{P_0}.(\mathbf{b})$, namely applying Cauchy-Schwarz to re-express the above in terms of norms, then applying $\mathbf{Q_0}.(\mathbf{a})$ to the $\|\boldsymbol{\Delta}_{q0}^{\text{in/out}}\|$ terms, our concentration assumption to $\|\mathbf{w}^q\|/\sqrt{N} \leq 2W_{\max} + \|[\mathbf{w}^{(r)}]^q\|/\sqrt{N}$, and Lemmas 4 part 3, 12, and 15 to the $\|\overset{*}{\mathbf{q}}_0^{\text{in/out}}\|$ as in (A.62)-(A.63).

If instead ϕ satisfies the conditionally bounded pseudo-Lipschitz condition, then we can apply the analogous strategy to the functions $f_i = \phi\left(\cdot,\cdot,\left[\mathbf{w}_q^{\text{in}}\right]_i\right)$, which are themselves pseudo-Lipschitz with respect to a common constant (since these constants are a continuous function of the bounded $[\mathbf{w}_q^{\text{in}}]_i$). The resulting bound drops the $2|[\mathbf{w}_q^{\text{in}}]_i|/\sqrt{N}$ term, but is otherwise unchanged.

Combining these inequalities, this completes the proof of $\mathbf{Q_0}.(\mathbf{b})$.

$\mathbf{Q_1}.(\mathbf{c}) - (\mathbf{f})$ For the inner product $[\mathbf{u}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}}$, note that $[\mathbf{u}_0^{\text{in}}]_i = [\mathbf{w}_q^{\text{in}}]_{2i}$ for $1 \leq i \leq N$, and $(\mathbf{q}_0^{\text{in}}, \mathbf{q}_0^{\text{out}}, \mathbf{w}_q^{\text{in}}) \mapsto [\mathbf{u}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}}$ is pseudo-Lipschitz of order 2, so by $\mathbf{Q}_1.(b)$, we have that this concentrates to $\mathbb{E}\left[Q_0U_0\right] = \mathbb{E}[Q_0]\mathbb{E}[U_0] = 0$ since Q_0 and U_0 are independent by hypothesis and $\mathbb{E}Q_0 = 0$.

For the inner product $[\mathbf{u}_0^{\text{in}}]^T\mathbf{u}_1^{\text{in}}$, we expand the definition of $\mathbf{u}_1^{\text{in}}$, obtaining

$$\frac{1}{1-\alpha_{qk}^{\text{in}}}\left[[\mathbf{u}_0^{\text{in}}]^T f_q^{\text{in}}\left(\mathbf{q}_0^{\text{in}},\mathbf{q}_0^{\text{out}},\mathbf{w}_q^{\text{in}},\gamma_{qk}^{\text{in}},\gamma_{qk}^{\text{out}}\right)-\alpha_{qk}^{\text{in}}[\mathbf{u}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}}\right].$$

Showing that this quantity concentrates around $\mathbb{E}U_0U_1$ uses the definition of U_0 and U_1 along with

the previously established concentration of α_{qk}^{in} in $\mathbf{Q_1}.(\mathbf{f})$, the previously established concentration of $[\mathbf{u}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}}$ (above), and the concentration of $[\mathbf{u}_0^{\text{in}}]^T f_q^{\text{in}}\left(\mathbf{q}_0^{\text{in}}, \mathbf{q}_0^{\text{out}}, \mathbf{w}_q^{\text{in}}, \gamma_{qk}^{\text{in}}, \gamma_{qk}^{\text{out}}\right)$, which follows from $\mathbf{Q_q}.(\mathbf{b})$ and the fact that the above is pseudo-Lipschitz in $(\mathbf{q}_0^{\text{in}}, \mathbf{q}_0^{\text{out}}, \mathbf{w}_q^{\text{in}})$. The details of the argument are entirely parallel to the details of proving $\mathbf{P_k}.(\mathbf{e})$, which we present explicitly as that is the more complicated case.

The rest of these follow by applications of $\mathbf{Q_1}.(\mathbf{b})$ and concentration of $\alpha_{q0}^{\text{in/out}}$ in the same way that $\mathbf{P_0}.(\mathbf{c}) - (\mathbf{f})$ were proved.

$\mathbf{Q_1}.(\mathbf{g}).(\mathbf{i})$ We will show that $\|[\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp}]^T\mathbf{q}_0^{\text{in}}\|^2$ concentrates around ρ_{q0}^{in}. First, observe that

$$\|[\mathbf{B}_{\mathbf{C}_{u0}^{\text{in}}}^{\perp}]^T\mathbf{q}_0^{\text{in}}\|^2 = [\mathbf{q}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}} - \frac{([\mathbf{u}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}})^2}{[\mathbf{u}_0^{\text{in}}]^T\mathbf{u}_0^{\text{in}}}. \tag{A.69}$$

Because $[\mathbf{q}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}} = [\mathbf{v}_0^{\text{in}}]^T\mathbf{v}_0^{\text{in}}$, using $\mathbf{P_0}.(\mathbf{e})$, $\mathbf{Q_1}.(\mathbf{e})$, and $\mathbf{Q_1}.(\mathbf{f})$, along with Lemmas 9 and 10, the above concentrates around $\mathbb{E}[V_0^{\text{in}}]^2 = \rho_{q0}^{\text{in}}$ as desired.

$\mathbf{Q_1}.(\mathbf{g}).(\mathbf{ii})$ First observe that $[\mathbf{C}_{q1}^{\text{in}}]^T\mathbf{C}_{q1}^{\text{in}}$ is invertible if and only if

$$\det([\mathbf{C}_{q1}^{\text{in}}]^T\mathbf{C}_{q1}^{\text{in}}) = ([\mathbf{u}_0^{\text{in}}]^T\mathbf{u}_0^{\text{in}})([\mathbf{q}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}}) - ([\mathbf{u}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}}) \neq 0.$$

Using the fact that $\|\mathbf{q}_0^{\text{in/out}}\| = \|\mathbf{v}_0^{\text{in/out}}\|$, we have by **Conditon 0**, $\mathbf{P_0}.(\mathbf{e})$, $\mathbf{Q_1}.(\mathbf{c})$ and Lemmas 9 and 10 that the above concentrates around $\mathbf{\Sigma}_{u0}^{\text{in}}\mathbf{\Sigma}_{v0}^{\text{in}} > 0$. Now, by Lemma 10, we have that

$$([\mathbf{C}_{q1}^{\text{in}}]^T\mathbf{C}_{q1}^{\text{in}})^{-1} = \frac{1}{\det([\mathbf{C}_{q1}^{\text{in}}]^T\mathbf{C}_{q1}^{\text{in}})}\begin{bmatrix} [\mathbf{q}_0^{\text{in}}]^T\mathbf{q}_0^{\text{in}} & -[\mathbf{q}_0^{\text{in}}]^T\mathbf{u}_0^{\text{in}} \\ -[\mathbf{q}_0^{\text{in}}]^T\mathbf{u}_0^{\text{in}} & [\mathbf{u}_0^{\text{in}}]^T\mathbf{u}_0^{\text{in}} \end{bmatrix},$$

will concentrate around the desired constants

$$\begin{bmatrix} (\mathbf{\Sigma}_{u0}^{\text{in}})^{-1} & 0 \\ 0 & (\mathbf{\Sigma}_{v0}^{\text{in}})^{-1} \end{bmatrix}.$$

$\mathbf{Q_1}.(\mathbf{g}).(\mathbf{iii})$ Recall that $\mathbf{C}_{q1}^{\text{in}} = [\mathbf{u}_0^{\text{in}}\ \mathbf{q}_0^{\text{in}}]$ by (2.37) and that $\mathbf{B}_{\mathbf{C}_{q1}^{\text{in}}}^{\perp} \in \mathbb{R}^{N\times N-2}$ is an orthogonal

matrix with columns that form an orthonormal basis for range$(\mathbf{C}_{q1}^{\text{in}})^{\perp} = [\mathbf{u}_0^{\text{in}}\ \mathbf{q}_0^{\text{in}}]^{\perp}$. Thus, the matrix $[\mathbf{B}_{\mathbf{C}_{q1}^{\text{in}}}^{\perp}\ \mathbf{B}_{\mathbf{C}_{q1}^{\text{in}}}] \in \mathbb{R}^{N\times N}$ is orthogonal and

$$\|\mathbf{u}_1^{\text{in}}\|^2 = \left\| \begin{bmatrix} [\mathbf{B}_{\mathbf{C}_{q1}^{\text{in}}}^{\perp}]^T \\ [\mathbf{B}_{\mathbf{C}_{q1}^{\text{in}}}]^T \end{bmatrix} \mathbf{u}_1^{\text{in}} \right\|^2 = \|[\mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}^{\perp}]^T \mathbf{u}_1^{\text{in}}\|^2 + \|[\mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}]^T \mathbf{u}_1^{\text{in}}\|^2. \tag{A.70}$$

Rearranging, and noticing that $\|[\mathbf{B}_{\mathbf{C}_{v0}^{\text{in}}}]^T \mathbf{u}_1^{\text{in}}\|^2 = [\mathbf{u}_1^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}} ([\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}})^{-1} [\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{u}_1^{\text{in}}$, this gives us

$$\|[\mathbf{B}_{\mathbf{C}_{q1}^{\text{in}}}^{\perp}]^T \mathbf{u}_1^{\text{in}}\|^2 = \|\mathbf{u}_1^{\text{in}}\|^2 - [\mathbf{u}_1^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}} ([\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}})^{-1} [\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{u}_1^{\text{in}}. \tag{A.71}$$

Now $\|\mathbf{u}_1^{\text{in}}\|^2$ concentrates around $(\mathbf{\Sigma}_{u1}^{\text{in}})_{22}$ by $\mathbf{Q_1}.(\mathbf{e})$, and we have

$$[\mathbf{u}_1^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}} ([\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}})^{-1} [\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{u}_1^{\text{in}} =$$
$$([\mathbf{u}_1^{\text{in}}]^T \mathbf{u}_0^{\text{in}})^2 \frac{[\mathbf{q}_0^{\text{in}}]^T \mathbf{q}_0^{\text{in}}}{\det([\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}})} + ([\mathbf{u}_1^{\text{in}}]^T \mathbf{q}_0^{\text{in}})^2 \frac{[\mathbf{u}_0^{\text{in}}]^T \mathbf{u}_0^{\text{in}}}{\det([\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}})} - 2([\mathbf{u}_1^{\text{in}}]^T \mathbf{u}_0^{\text{in}})([\mathbf{u}_1^{\text{in}}]^T \mathbf{q}_0^{\text{in}}) \frac{[\mathbf{u}_0^{\text{in}}]^T \mathbf{q}_0^{\text{in}}}{\det([\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}})}.$$

Label these terms $T_1 - T_3$. We first notice that

$$\det\left([\mathbf{C}_{q1}^{\text{in}}]^T \mathbf{C}_{q1}^{\text{in}}\right) = \|\mathbf{q}_0^{\text{in}}\|^2 \|\mathbf{u}_0^{\text{in}}\|^2 - ([\mathbf{q}_0^{\text{in}}]^T \mathbf{u}_0^{\text{in}})^2.$$

By $\mathbf{Q_1}.(\mathbf{d})(\mathbf{e})(\mathbf{f})$, with **Condition 0** and Lemmas 9 and 10, this concentrates to $(\mathbf{\Sigma}_{q1}^{\text{in}})_{22}(\mathbf{\Sigma}_{u1}^{\text{in}})_{11} > 0$.

Then T_1 concentrates around $\frac{(\mathbf{\Sigma}_{u1}^{\text{in}})_{12}^2}{(\mathbf{\Sigma}_{u1}^{\text{in}})_{11}}$ by the above, $\mathbf{Q_1}.(\mathbf{e})$, $\mathbf{Q_1}.(\mathbf{d})$, and Lemma 10. Next notice that by the above, $\mathbf{Q_1}.(\mathbf{f})$, and Lemma 10, we have that T_2 concentrates around zero. And similarly by the above, $\mathbf{Q_1}.(\mathbf{e})$, $\mathbf{Q_1}.(\mathbf{f})$, and Lemma 10, T_3 concentrates around zero. So we ultimately get that $\|[\mathbf{B}_{\mathbf{C}_{q1}^{\text{in}}}^{\perp}]^T \mathbf{u}_1^{\text{in}}\|^2$ concentrates around $(\mathbf{\Sigma}_{u1}^{\text{in}})_{11} - \frac{(\mathbf{\Sigma}_{u1}^{\text{in}})_{01}^2}{(\mathbf{\Sigma}_{u1}^{\text{in}})_{00}} = \rho_{p1}^{\text{in}}$, where the equality follows using (2.70), (2.71), and (2.72) and which say

$$\rho_{p1}^{\text{in}} = \mathbb{E}\{[U_1^{\text{in}}]^2\} - (\mathbf{b}_{u1}^{\text{in}})^T [\mathbf{\Sigma}_{u0}^{\text{in}}]^{-1} \mathbf{b}_{u1}^{\text{in}} = \mathbb{E}\{[U_1^{\text{in}}]^2\} - \frac{(\mathbb{E}[U_0^{\text{in}} U_1^{\text{in}}])^2}{\mathbf{\Sigma}_{u0}^{\text{in}}} = [\mathbf{\Sigma}_{u1}^{\text{in}}]_{11} - \frac{[\mathbf{\Sigma}_{u1}^{\text{in}}]_{10}^2}{[\mathbf{\Sigma}_{u1}^{\text{in}}]_{00}}.$$

Q1.(**g**).(**iv**) Notice that by (2.73) for $k \geq 1$, we have $[\mathbf{\Pi}^{\text{in}}_{uk}]^{-1} = \begin{bmatrix} \mathbf{\Sigma}^{\text{in}}_{uk} & 0 \\ 0 & \mathbf{\Sigma}^{\text{in}}_{v(k-1)} \end{bmatrix}$, so in particular, $[\mathbf{\Pi}^{\text{in}}_{u1}]^{-1} = \begin{bmatrix} \mathbf{\Sigma}^{\text{in}}_{u1} & 0 \\ 0 & \mathbf{\Sigma}^{\text{in}}_{v0} \end{bmatrix}$, where these matrices are defined in (2.71). Next, define $\mathbf{F} := [\mathbf{e}_2 \ \mathbf{e}_1 \ \mathbf{e}_3] \in \mathbb{R}^{3\times 3}$, where $\mathbf{e}_i$ is the i^{th} canonical vector, namely the vector of zeros with a one in the i^{th} position, so that for a matrix $\mathbf{A}$, the transformation $\mathbf{AF}$ swaps the second and first column of $\mathbf{A}$. Then, using that $\mathbf{b}^{\text{in}}_{u1} = \mathbb{E}[U^{\text{in}}_0 U^{\text{in}}_1]$ by (2.70),

$$\mathbf{F}^T[\mathbf{\Pi}^{\text{in}}_{u1}]^{-1}\mathbf{F} = \begin{bmatrix} \mathbb{E}[(U^{\text{in}}_1)^2] & [\mathbf{b}^{\text{in}}_{u1}\ 0] \\ \begin{bmatrix} \mathbf{b}^{\text{in}}_{u1} \\ 0 \end{bmatrix} & [\mathbf{\Pi}^{\text{in}}_{q0}]^{-1} \end{bmatrix}.$$

Then, using $\mathbf{F} = \mathbf{F}^T = \mathbf{F}^{-1}$, so that $(\mathbf{F}^T[\mathbf{\Pi}^{\text{in}}_{u0}]^{-1}\mathbf{F})^{-1} = \mathbf{F}\mathbf{\Pi}^{\text{in}}_{u0}\mathbf{F}^T$, the block matrix inversion formula gives

$$\mathbf{F}\mathbf{\Pi}^{\text{in}}_{u1}\mathbf{F}^T = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{\Pi}^{\text{in}}_{q0} \end{bmatrix} + \frac{1}{\rho^{\text{in}}_{p1}} \begin{bmatrix} 1 & -[\boldsymbol{\beta}^{\text{in}}_{p1}\ 0] \\ -\begin{bmatrix} \boldsymbol{\beta}^{\text{in}}_{p1} \\ 0 \end{bmatrix} & \begin{bmatrix} \boldsymbol{\beta}^{\text{in}}_{p1} \\ 0 \end{bmatrix} [\boldsymbol{\beta}^{\text{in}}_{p1}\ 0] \end{bmatrix}. \tag{A.72}$$

In the above, we have used that $\rho^{\text{in}}_{p1} = \mathbb{E}[(U^{\text{in}}_1)^2] - (\mathbf{b}^{\text{in}}_{u1})^2/\mathbb{E}[(U^{\text{in}}_0)^2] = \mathbb{E}[(U^{\text{in}}_1)^2] - [\mathbf{b}^{\text{in}}_{u1}\ 0]\mathbf{\Pi}^{\text{in}}_{q0}[\mathbf{b}^{\text{in}}_{u1}\ 0]^T$ and $\boldsymbol{\beta}^{\text{in}}_{p1} = \mathbf{b}^{\text{in}}_{u1}/\mathbb{E}[(U^{\text{in}}_0)^2]$ by (2.72).

Similarly, using that $\mathbf{C}^{\text{in}}_{u1} = [\mathbf{u}^{\text{in}}_0\ \mathbf{u}^{\text{in}}_1\ \mathbf{q}^{\text{in}}_0]$ and $\mathbf{C}^{\text{in}}_{q1} = [\mathbf{u}^{\text{in}}_0\ \mathbf{q}^{\text{in}}_0]$, we get

$$\mathbf{F}^T([\mathbf{C}^{\text{in}}_{u1}]^T\mathbf{C}^{\text{in}}_{u1})F = \begin{bmatrix} [\mathbf{u}^{\text{in}}_1]^T\mathbf{u}^{\text{in}}_1 & [\mathbf{u}^{\text{in}}_1]^T\mathbf{C}^{\text{in}}_{q1} \\ [\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{u}^{\text{in}}_1 & [\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{C}^{\text{in}}_{q1} \end{bmatrix}.$$

Letting $\boldsymbol{\nu} = ([\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{C}^{\text{in}}_{q1})^{-1}[\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{u}^{\text{in}}_1$, again by block matrix inversion, formula gives us

$$\mathbf{F}^T([\mathbf{C}^{\text{in}}_{u1}]^T\mathbf{C}^{\text{in}}_{u1})^{-1}\mathbf{F} = \begin{bmatrix} 0 & 0 \\ 0 & ([\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{C}^{\text{in}}_{q1})^{-1} \end{bmatrix} + \frac{1}{\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{q1}}]^T\mathbf{u}^{\text{in}}_1\|^2} \begin{bmatrix} 1 & -\boldsymbol{\nu}^T \\ -\boldsymbol{\nu} & -\boldsymbol{\nu}\boldsymbol{\nu}^T \end{bmatrix}, \tag{A.73}$$

where we use that $\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{q1}}]^T\mathbf{u}^{\text{in}}_1\|^2 = \|\mathbf{u}^{\text{in}}_1\|^2 - \|\mathbf{B}_{\mathbf{C}^{\text{in}}_{q1}}\mathbf{u}^{\text{in}}_1\|^2 = \|\mathbf{u}^{\text{in}}_1\|^2 - [\mathbf{u}^{\text{in}}_1]^T\mathbf{C}^{\text{in}}_{q1}([\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{C}^{\text{in}}_{q1})^{-1}[\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{u}^{\text{in}}_1$, using an argument similar to that in (A.70)-(A.71).

Now our goal is to show that, elementwise, (A.73) concentrates on (A.72). By Lemmas 9 and 10, it suffices to show three results. Namely, that $([\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{C}^{\text{in}}_{q1})^{-1}$ concentrates on $\mathbf{\Pi}^{\text{in}}_{v0}$, that $\|\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{q1}}\mathbf{u}^{\text{in}}_1\|^{-2}$ concentrates on $[\rho^{\text{in}}_{p1}]^{-1}$, and that $\boldsymbol{\nu}$ concentrates on $[\boldsymbol{\beta}^{\text{in}}_{p1}\ 0]^T$. Elementwise concentration for $([\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{C}^{\text{in}}_{q1})^{-1}$ follows from $\mathbf{Q_1}.(\mathbf{g}).(\mathbf{i})$, and concentration for $\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{q1}}]^T\mathbf{u}^{\text{in}}_1\|^{-2}$ follows from $\mathbf{Q_1}.(\mathbf{g}).(\mathbf{ii})$ along with Lemma 13. Note that the ρ terms are bounded above by our requirement that the state evolution not diverge in **Condition 5**.

Finally, let $\widetilde{\mathbf{C}}^{\text{in}}_{q1} = ([\mathbf{C}^{\text{in}}_{q1}]^T\mathbf{C}^{\text{in}}_{q1})^{-1}$ and observe that

$$\boldsymbol{\nu} = \widetilde{\mathbf{C}}^{\text{in}}_{q1}\begin{bmatrix}[\mathbf{u}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1\\ [\mathbf{q}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1\end{bmatrix} = \begin{bmatrix}[\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{11}([\mathbf{u}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1) + [\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{12}([\mathbf{q}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1)\\ [\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{21}([\mathbf{u}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1) + [\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{22}([\mathbf{q}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1)\end{bmatrix}. \tag{A.74}$$

Now we show concentration for the first element of (A.74). By $\mathbf{Q_1}.(\mathbf{e})$, $[\mathbf{u}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1$ concentrates around $\mathbb{E}[U^{\text{in}}_0U^{\text{in}}_1]$ and by $\mathbf{Q_1}.(\mathbf{g}).(\mathbf{ii})$, we have that $[\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{11}$ concentrates around $1/\mathbb{E}[(U^{\text{in}}_0)^2]$, so by Lemma 10, we have that $[\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{11}([\mathbf{u}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1)$ concentrates around $\frac{\mathbb{E}[U^{\text{in}}_0U^{\text{in}}_1]}{\mathbb{E}[(U^{\text{in}}_0)^2]} = \boldsymbol{\beta}^{\text{in}}_{q1}$ by (2.72). Similarly, we have that $[\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{12}$ concentrates around 0 and $[\mathbf{q}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1$ concentrates around 0 by $\mathbf{Q_1}.(\mathbf{f})$. Therefore, using so again by Lemmas 9, we get that $[\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{11}([\mathbf{u}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1) + [\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{12}([\mathbf{q}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1)$ concentrates around $\boldsymbol{\beta}^{\text{in}}_{q1}$.

Noting that $[\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{21}$ and $[\mathbf{q}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1$ concentrate around 0 and proceeding similarly gives that $[\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{21}([\mathbf{u}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1) + [\widetilde{\mathbf{C}}^{\text{in}}_{q1}]_{22}([\mathbf{q}^{\text{in}}_0]^T\mathbf{u}^{\text{in}}_1)$ concentrates around 0. This completes the proof that the entries of $([\mathbf{C}^{\text{in}}_{u1}]^T\mathbf{C}^{\text{in}}_{u1})^{-1}$ concentrate around the entries of $\mathbf{\Pi}^{\text{in}}_{u1}$.

A.1.3 Showing $\mathbf{P_k}$ holds

First we prove the following lemma.

Lemma 7. *For all* $1 \le i \le 2k$ *and for* $k \ge 1$,

$$\mathbb{P}\left(\left|\left[([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}[\mathbf{C}_{qk}^{\text{in}}]^T\mathbf{u}_k^{\text{in}}\right]_i - \begin{bmatrix}\boldsymbol{\beta}_{pk}^{\text{in}}\\ 0\end{bmatrix}_i\right| \ge \epsilon\right) \le CkC_{k-1}\exp\left(-ncc_{k-1}\epsilon^2/k^2\right).$$

Proof. Let $\widetilde{\mathbf{C}}_k^{\text{in}} = ([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}$ and recall by (2.37) that $\mathbf{C}_{qk}^{\text{in}} = [\mathbf{U}_{k-1}^{\text{in}}\,\mathbf{Q}_{k-1}^{\text{in}}]$. Then, for $1 \le i \le 2k$,

$$\left|\left[([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}[\mathbf{C}_{qk}^{\text{in}}]^T\mathbf{u}_k^{\text{in}}\right]_i - \begin{bmatrix}\boldsymbol{\beta}_{pk}^{\text{in}}\\ 0\end{bmatrix}_i\right| \le \left|\sum_{j=1}^{k}[\widetilde{\mathbf{C}}_k^{\text{in}}]_{ij}([\mathbf{u}_{j-1}^{\text{in}}]^T\mathbf{u}_k^{\text{in}}) - [\boldsymbol{\beta}_{pk}^{\text{in}}]_i\right| + \left|\sum_{j=1}^{k}[\widetilde{\mathbf{C}}_k^{\text{in}}]_{i,j+k}([\mathbf{q}_{j-1}^{\text{in}}]^T\mathbf{u}_k^{\text{in}})\right|. \tag{A.75}$$

Now by $\mathbf{P_{k-1}}.(\mathbf{g}).(\mathbf{iv})$, we know that the entries of $\widetilde{\mathbf{C}}_k^{\text{in}}$ concentrate on finite values. In particular, we know that if $1 \le i, j, \le k$, then $[\widetilde{\mathbf{C}}_k^{\text{in}}]_{ij}$ concentrates on $[\boldsymbol{\Pi}_{v,k-1}^{\text{in}}]_{ij} = \left[\left(\boldsymbol{\Sigma}_{u(k-1)}^{\text{in}}\right)^{-1}\right]_{ij}$ where the equality follows by (2.73). Moreover, by $\mathbf{Q_k}.(\mathbf{e})$ and $\mathbf{Q_k}.(\mathbf{f})$, we have that $[\mathbf{u}_{j-1}^{\text{in}}]^T\mathbf{u}_k^{\text{in}}$ and $[\mathbf{q}_{j-1}^{\text{in}}]^T\mathbf{u}_k^{\text{in}}$ concentrate on $[\boldsymbol{\Sigma}_{u,k}^{\text{in}}]_{j,k+1} = \mathbb{E}[U_{j-1}^{\text{in}}U_k^{\text{in}}] = [\mathbf{b}_{uk}^{\text{in}}]_j$ and 0, respectively.

Therefore, using Lemmas 9 and 10, $\sum_{j=1}^{k}[\widetilde{\mathbf{C}}_k^{\text{in}}]_{ij}([\mathbf{u}_{j-1}^{\text{in}}]^T\mathbf{u}_k^{\text{in}})$ concentrates on $\sum_{j=1}^{k}[(\boldsymbol{\Sigma}_{u(k-1)}^{\text{in}})^{-1}]_{ij}(\mathbf{b}_{uk}^{\text{in}})_j = \left[(\boldsymbol{\Sigma}_{u(k-1)}^{\text{in}})^{-1}(\mathbf{b}_{uk}^{\text{in}})\right]_i = [\boldsymbol{\beta}_{pk}^{\text{in}}]_i$ and $\sum_{j=1}^{k}[\widetilde{\mathbf{C}}_k^{\text{in}}]_{i,j+k}([\mathbf{q}_{j-1}^{\text{in}}]^T\mathbf{u}_k^{\text{in}})$ concentrates on 0. The rate of concentration depends inversely on the magnitudes of the components of $(\boldsymbol{\Sigma}_{u(k-1)}^{\text{in}})^{-1}$ and $\mathbf{b}_{uk}^{\text{in}}$, but these are bounded above independent of N as discussed in Section 2.3.5.

Now if $k+1 \le i \le 2k$, then we can apply the same analysis, but now for $1 \le j \le k$, we get $[\widetilde{\mathbf{C}}_k^{\text{in}}]_{i,j}$ concentrating on 0, so $\left[([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}[\mathbf{C}_{qk}^{\text{in}}]^T\mathbf{u}_k^{\text{in}}\right]_i$ concentrates on 0, again with the desired rate exponential in N, as needed. □

$\mathbf{P_k}$.(a) Letting $\boldsymbol{\mu}_{pk}^{\text{in}} = ([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}[\mathbf{C}_{qk}^{\text{in}}]^T\mathbf{u}_k^{\text{in}}$, recall by Lemma 3, that we have

$$\begin{aligned}\boldsymbol{\Delta}_{pk}^{\text{in}} &= \mathbf{C}_{vk}^{\text{in}}\left(\boldsymbol{\mu}_{pk}^{\text{in}} - \begin{bmatrix}\boldsymbol{\beta}_{pk}^{\text{in}}\\ 0\end{bmatrix}\right) + \left(\frac{\|[\mathbf{B}_{\mathbf{C}_{qk}^{\text{in}}}^{\perp}]^T\mathbf{u}_k^{\text{in}}\|}{\|\mathbf{Z}_{pk}^{\text{in}}\|} - \sqrt{\rho_{pk}^{\text{in}}}\right)\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp}\mathbf{Z}_{pk}^{\text{in}} - \sqrt{\rho_{pk}^{\text{in}}}\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}\check{\mathbf{Z}}_{pk}^{\text{in}}\\ &= \sum_{i=1}^{k}\left([\boldsymbol{\mu}_{pk}^{\text{in}}]_i - [\boldsymbol{\beta}_{pk}^{\text{in}}]_i\right)[\mathbf{C}_{vk}^{\text{in}}]_{(\cdot,i)} + \sum_{i=k+1}^{2k}[\boldsymbol{\mu}_{pk}^{\text{in}}]_i[\mathbf{C}_{vk}^{\text{in}}]_{(\cdot,i)}\\ &\qquad + \left(\frac{\|[\mathbf{B}_{\mathbf{C}_{qk}^{\text{in}}}^{\perp}]^T\mathbf{u}_k^{\text{in}}\|}{\|\mathbf{Z}_{pk}^{\text{in}}\|} - \sqrt{\rho_{pk}^{\text{in}}}\right)\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp}\mathbf{Z}_{pk}^{\text{in}} - \sqrt{\rho_{pk}^{\text{in}}}\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}\check{\mathbf{Z}}_{pk},\end{aligned}$$

where we have used the notation $[\mathbf{C}_{vk}^{\text{in}}]_{(\cdot,i)}$ to indicate the i^{th} column of the matrix $\mathbf{C}_{vk}^{\text{in}}$.

Therefore, by the triangle inequality, we find

$$\begin{aligned}\|\boldsymbol{\Delta}_{pk}^{\text{in}}\| &\le \sum_{i=1}^{k}\left|[\boldsymbol{\mu}_{pk}^{\text{in}}]_i - [\boldsymbol{\beta}_{pk}^{\text{in}}]_i\right|\|[\mathbf{C}_{vk}^{\text{in}}]_{(\cdot,i)}\| + \sum_{i=k+1}^{2k}\left|[\boldsymbol{\mu}_{pk}^{\text{in}}]_i\right|\|[\mathbf{C}_{vk}^{\text{in}}]_{(\cdot,i)}\|\\ &\qquad + \left|\frac{\|[\mathbf{B}_{\mathbf{C}_{qk}^{\text{in}}}^{\perp}]^T\mathbf{u}_k^{\text{in}}\|}{\|\mathbf{Z}_{pk}^{\text{in}}\|} - \sqrt{\rho_{pk}^{\text{in}}}\right|\|\mathbf{Z}_{pk}^{\text{in}}\| + \rho_{pk}^{\text{in}}\|\check{\mathbf{Z}}_{pk}^{\text{in}}\|,\end{aligned}$$

where we have used that $[\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp}]^T\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp} = [\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}]^T\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}} = \mathbf{I}$. Then, using Lemma 9, we find

$$\begin{aligned}\mathbb{P}\left(\frac{\|\boldsymbol{\Delta}_{pk}^{\text{in}}\|}{\sqrt{N}} \ge \sqrt{\epsilon}\right) &\le \sum_{i=1}^{k}\mathbb{P}\left(\left|[\boldsymbol{\mu}_{pk}^{\text{in}}]_i - [\boldsymbol{\beta}_{pk}^{\text{in}}]_i\right|\frac{\|[\mathbf{C}_{vk}^{\text{in}}]_{(\cdot,i)}\|}{\sqrt{N}} \ge \frac{\sqrt{\epsilon}}{4k}\right) + \sum_{i=k+1}^{2k}\mathbb{P}\left(\left|[\boldsymbol{\mu}_{pk}^{\text{in}}]_i\right|\frac{\|[\mathbf{C}_{vk}^{\text{in}}]_{(\cdot,i)}\|}{\sqrt{N}} \ge \frac{\sqrt{\epsilon}}{4k}\right)\\ &\qquad + \mathbb{P}\left(\left|\frac{\|[\mathbf{B}_{\mathbf{C}_{qk}^{\text{in}}}^{\perp}]^T\mathbf{u}_k^{\text{in}}\|}{\|\mathbf{Z}_{pk}^{\text{in}}\|} - \sqrt{\rho_{pk}^{\text{in}}}\right|\frac{\|\mathbf{Z}_{pk}^{\text{in}}\|}{\sqrt{N}} \ge \frac{\sqrt{\epsilon}}{4k}\right) + \mathbb{P}\left(\frac{\|\check{\mathbf{Z}}_{pk}^{\text{in}}\|}{\sqrt{N}} \ge \frac{\sqrt{\epsilon}}{4k\sqrt{\rho_{pk}^{\text{in}}}}\right).\end{aligned} \tag{A.76}$$

In what follows, we show upper bound for each of the terms on the right side of (A.76).

First term of (A.76)**.** Observe that by the definition of $\mathbf{C}_{vk}^{\text{in}}$ in (2.37), for $1 \le i \le k$, that

$\|[\mathbf{C}^{\text{in}}_{vk}]_{(\cdot,i)}\| = \|\mathbf{p}^{\text{in}}_{i-1}\|$. Then for $0 \le j \le k-1$,

$$\begin{aligned}
&\mathbb{P}\left(\left|[\boldsymbol{\mu}^{\text{in}}_{pk}]_{j+1} - [\boldsymbol{\beta}^{\text{in}}_{pk}]_{j+1}\right| \frac{\|[\mathbf{C}^{\text{in}}_{vk}]_{(\cdot,j+1)}\|}{\sqrt{N}} \ge \frac{\sqrt{\epsilon}}{4k}\right) \\
&\quad = \mathbb{P}\left(\left|[\boldsymbol{\mu}^{\text{in}}_{pk}]_{j+1} - [\boldsymbol{\beta}^{\text{in}}_{pk}]_{j+1}\right| \left[\left|\frac{\|\mathbf{p}^{\text{in}}_{j}\|}{\sqrt{N}} - (\boldsymbol{\Sigma}^{\text{in}}_{uk})_{jj}\right| + (\boldsymbol{\Sigma}^{\text{in}}_{uk})_{jj}\right] \ge \frac{\sqrt{\epsilon}}{4k}\right) \\
&\quad \overset{(a)}{\le} \mathbb{P}\left(\left|\frac{\|\mathbf{p}^{\text{in}}_{j}\|}{\sqrt{N}} - (\boldsymbol{\Sigma}^{\text{in}}_{uk})_{jj}\right| \ge \sqrt{\epsilon}\right) + \mathbb{P}\left(\left|[\boldsymbol{\mu}^{\text{in}}_{pk}]_{j+1} - [\boldsymbol{\beta}^{\text{in}}_{pk}]_{j+1}\right| \ge \frac{\sqrt{\epsilon}}{8k \max\{1, (\boldsymbol{\Sigma}^{\text{in}}_{uk})_{jj}\}}\right).
\end{aligned} \tag{A.77}$$

The first term concentrates by $\mathbf{P_j}.(\mathbf{d})$ $(0 \le j \le k-1)$ and the fact that

$$\mathbb{E}\left[P^{\text{in/out}}_i P^{\text{in/out}}_j\right] = \lim_{N\to\infty}\left[\mathbf{p}^{\text{in/out}}_i\right]^T \mathbf{p}^{\text{in/out}}_j = \lim_{N\to\infty}\left[\mathbf{u}^{\text{in/out}}_i\right]^T \mathbf{u}^{\text{in/out}}_j = \mathbb{E}\left[U^{\text{in/out}}_i U^{\text{in/out}}_j\right],$$

where the first and last equalities follow from $\mathbf{P_{k-1}}.(\mathbf{d})$ and $\mathbf{Q_{k-1}}.(\mathbf{e})$, and the middle equality follows from the fact that $\mathbf{p}^{\text{in/out}}_i = \mathbf{V}\mathbf{u}^{\text{in/out}}_i$.

Then the second term concentrates by Lemma 7. We note that the above also implies that $(\boldsymbol{\Sigma}^{\text{in}}_{uk})_{jj} = \tau^{\text{in}}_{pj}$, which is upper bounded independently of N and j by **Condition 5**.

Second term of (A.76)**.** For $k+1 \le i \le 2k$, we have $\|[\mathbf{C}^{\text{in}}_{vk}]_{(\cdot,i)}\| = \|\mathbf{v}^{\text{in}}_{i-(k+1)}\|$. Letting $i' = i-(k+1)$,

$$\begin{aligned}
\mathbb{P}\left(\left|[\boldsymbol{\mu}^{\text{in}}_{pk}]_i\right| \frac{\|[\mathbf{C}^{\text{in}}_{vk}]_{(\cdot,i)}\|}{\sqrt{N}} \ge \frac{\sqrt{\epsilon}}{4k}\right) &\le \mathbb{P}\left(\left|[\boldsymbol{\mu}^{\text{in}}_{pk}]_i\right| \left[\left|\frac{\|\mathbf{v}^{\text{in}}_{i'}\|}{\sqrt{N}} - (\boldsymbol{\Sigma}^{\text{in}}_{vk})_{i'i'}\right| + (\boldsymbol{\Sigma}^{\text{in}}_{vk})_{i'i'}\right] \ge \frac{\sqrt{\epsilon}}{4k}\right) \\
&\overset{(b)}{\le} \mathbb{P}\left(\left|\frac{\|\mathbf{v}^{\text{in}}_{i'}\|}{\sqrt{N}} - (\boldsymbol{\Sigma}^{\text{in}}_{vk})_{i'i'}\right| \ge \sqrt{\epsilon}\right) + \mathbb{P}\left(\left|[\boldsymbol{\mu}^{\text{in}}_{pk}]_i\right| \ge \frac{\sqrt{\epsilon}}{8k \max\{1, (\boldsymbol{\Sigma}^{\text{in}}_{uk})_{i'i'}\}}\right).
\end{aligned} \tag{A.78}$$

Again, the first term concentrates by $\mathbf{P_{i'}}.(\mathbf{e})$ where $0 \le i' \le k-1$, and the second term by Lemma 7. As before, $\tau^{\text{in}}_{pi'}$ is bounded above independently of N, as needed.

Third term of (A.76)**.** By Lemma 11, we find

$$
\begin{aligned}
&P\left(\left|\frac{\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{qk}}]^T\mathbf{u}^{\text{in}}_k\|}{\|\mathbf{Z}^{\text{in}}_{pk}\|}-\sqrt{\rho^{\text{in}}_{pk}}\right|\frac{\|\mathbf{Z}^{\text{in}}_{pk}\|}{\sqrt{N}}\geq\frac{\sqrt{\epsilon}}{4k}\right)\\
&\quad\leq\mathbb{P}\left(\left|\frac{\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{qk}}]^T\mathbf{u}^{\text{in}}_k\|}{\sqrt{N}}-\sqrt{\rho^{\text{in}}_{pk}}\right|\geq\frac{\sqrt{\epsilon}}{8k}\right)+\mathbb{P}\left(\left|\frac{\|\mathbf{Z}^{\text{in}}_{pk}\|}{\sqrt{N}}-1\right|\geq\frac{\sqrt{\epsilon}}{8k\max\{1,\sqrt{\rho^{\text{in}}_{pk}}\}}\right).
\end{aligned}
\tag{A.79}
$$

The second term concentrates by Lemmas 15 and 12, and the first term by $\mathbf{Q_k}.(\mathbf{g}).(\mathbf{iii})$. The quantity ρ^{in}_{pk} is bounded above independently of N per the discussion in Section 2.3.5.

Fourth term of (A.76)**.** Finally, recall from Lemma 3 that $\check{\mathbf{Z}}^{\text{in}}_{pk}$ is a length-$2k$ vector. Thus,

$$
\mathbb{P}\left(\frac{\|\check{\mathbf{Z}}^{\text{in}}_{pk}\|}{\sqrt{N}}\geq\frac{\sqrt{\epsilon}}{4k}\right)=\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{2k}[\check{\mathbf{Z}}^{\text{in}}_{pk}]_i^2\geq\frac{\epsilon}{16k^2}\right)\leq\sum_{i=1}^{2k}\mathbb{P}\left([\check{\mathbf{Z}}^{\text{in}}_{pk}]_i^2\geq\frac{N\epsilon}{32k^3}\right).
\tag{A.80}
$$

Now the last expression concentrates by Lemma 14.

$\mathbf{P_k}.(\mathbf{b})$ In what follows, the notation $\underline{x}_k$ means $(x_0, x_1, \ldots, x_k)$ (e.g., $\underline{P}^{\text{in/out}}_k = (P^{\text{in/out}}_0, \ldots, P^{\text{in/out}}_k)$ and $\left[\underline{\mathbf{p}}^{\text{in/out}}\right]_i = ([\mathbf{p}^{\text{in/out}}_0]_i, \ldots, [\mathbf{p}^{\text{in/out}}_k]_i)$). Using Lemma 4 result (2.59) and the triangle inequality,

$$
\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\left[\underline{\mathbf{p}}^{\text{in}}_k\right]_i,\left[\underline{\mathbf{p}}^{\text{out}}_k\right]_i,[\mathbf{w}]_i\right)-\mathbb{E}\left\{\phi(\underline{P}^{\text{in}}_k,\underline{P}^{\text{out}}_k,W)\right\}\right|\\
&\overset{d}{=}\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\left[\underline{\widetilde{\mathbf{p}}}^{\text{in}}_k\right]_i,\left[\underline{\widetilde{\mathbf{p}}}^{\text{out}}_k\right]_i,[\mathbf{w}]_i\right)-\mathbb{E}\left\{\phi(\underline{P}^{\text{in}}_k,\underline{P}^{\text{out}}_k,W)\right\}\right|\\
&\leq\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\left[\underline{\overset{*}{\mathbf{p}}}^{\text{in}}_k\right]_i,\left[\underline{\overset{*}{\mathbf{p}}}^{\text{out}}_k\right]_i,[\mathbf{w}]_i\right)-\mathbb{E}\left\{\phi(\underline{P}^{\text{in}}_k,\underline{P}^{\text{out}}_k,W)\right\}\right|\\
&\qquad+\frac{1}{N}\sum_{i=1}^{N}\left|\phi\left(\left[\underline{\widetilde{\mathbf{p}}}^{\text{in}}_k\right]_i,\left[\underline{\widetilde{\mathbf{p}}}^{\text{out}}_k\right]_i,[\mathbf{w}]_i\right)-\phi\left(\left[\underline{\overset{*}{\mathbf{p}}}^{\text{in}}_k\right]_i,\left[\underline{\overset{*}{\mathbf{p}}}^{\text{out}}_k\right]_i,[\mathbf{w}]_i\right)\right|.
\end{aligned}
\tag{A.81}
$$

We label the two terms on the right side of (A.81) as T1 and T2. First, we establish concentration for T1. We modify our usual convention and take $[\underline{\overset{*}{\mathbf{p}}}^{\text{out}}_k]_i$ for $i > M$ to be i.i.d. copies of $\underline{P}^{\text{out}}_k$. We can do this since $\phi(\cdot,\cdot,[\mathbf{w}]_i)$ does not depend on its second argument for $i > M$. This modifi-

cation implies that $\left\{\left(\left[\overset{*}{\mathbf{p}}{}_0^{\text{in}}\right]_i, \ldots, \left[\overset{*}{\mathbf{p}}{}_k^{\text{in}}\right]_i, \left[\overset{*}{\mathbf{p}}{}_0^{\text{out}}\right]_i, \ldots, \left[\overset{*}{\mathbf{p}}{}_k^{\text{out}}\right]_i\right)\right\}_{i\geq 1}$ are independent samples from a common multivariate normal distribution by part 3 of Lemma 4. If ϕ is jointly pseudo-Lipschitz in all inputs, then the $f_i : (p^{\text{in}}, p^{\text{out}}, w^{(r)}) \mapsto= \phi(p^{\text{in}}, p^{\text{out}}, w^{(r)}, w_i^{(d)})$ are pseudo-Lipschitz with common constant (independent of N) by Lemma 18 and the fact that the $\mathbf{w}^{(d)}$ components are bounded by $W_{\max}$. Thus, in this case, concentration for T1 already follows directly by Lemma 17. (Note, in particular, that the variances of the $\overset{*}{\mathbf{p}}$ components which appear in the denominator of the exponential rate of convergence are just the $\tau_{p^\cdot}^{\text{in/out}}$ iterates of the general state evolution defined in Section 2.3.4, and that they are thus bounded above independently of N per the discussion in that section).

If instead ϕ satisfies the conditionally bounded pseudo-Lipschitz property, we further decompose T1 as

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\left[\underline{\overset{*}{\mathbf{p}}}{}_k^{\text{in}}\right]_i, \left[\underline{\overset{*}{\mathbf{p}}}{}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right) - \mathbb{E}\left\{\phi(\underline{P}_k^{\text{in}}, \underline{P}_k^{\text{out}}, W)\right\}\right| \\
&\qquad \leq \left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\left[\underline{\overset{*}{\mathbf{p}}}{}_k^{\text{in}}\right]_i, \left[\underline{\overset{*}{\mathbf{p}}}{}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right) - \mathbb{E}_P\left\{\phi(\underline{P}_k^{\text{in}}, \underline{P}_k^{\text{out}}, [\mathbf{w}]_i)\right\}\right| \\
&\qquad\qquad + \left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_P\left\{\phi(\underline{P}_k^{\text{in}}, \underline{P}_k^{\text{out}}, [\mathbf{w}]_i)\right\} - \mathbb{E}\left\{\phi(\underline{P}_k^{\text{in}}, \underline{P}_k^{\text{out}}, W)\right\}\right|,
\end{aligned}$$

where $\mathbb{E}_P$ is the expectation with respect to the random variables $(\underline{P}_k^{\text{in}}, \underline{P}_k^{\text{out}})$. We label these terms T1a and T1b.

Now, taking $f_i = \phi(\cdot, \cdot, [\mathbf{w}]_i)$, Lemma 17 will give the desired concentration for the first term so long as the f_i are pseudo-Lipschitz with a common constant. But by the bounded conditionally pseudo-Lipschitz condition, the PL constant of f_i is continuous in $[\mathbf{w}]_i$. Since these are themselves bounded, the PL constants must also be bounded by some B, which we can take as our common constant. As in the last case, all relevant quantities are thus bounded independently of N, giving the desired rate of convergence.

To get concentration for T1b, we note that $x \mapsto \mathbb{E}_P\left\{\phi(\underline{P}_k^{\text{in}}, \underline{P}_k^{\text{out}}, x)\right\}$ is continuous in x (since ϕ is continuous in all inputs), and is thus bounded over the domain of $\mathbf{w}$ since this domain is compact

by definition. Thus, the terms in T1b can be written as $g_i([\mathbf{w}^{(r)}]_i)$ with $g_i = \mathbb{E}_P\phi(\underline{P}_k^{\text{in}}, \underline{P}_k^{\text{out}}, [\mathbf{w}^{(r)}]_i, [\mathbf{w}^{(d)}]_i)$. Since these are bounded and independent random variables, with common bound independent of N, concentration follows with desired rate by a direct application of Hoeffding's inequality.

Now we bound T2 of (A.81). First, the pseudo-Lipschitz property of ϕ gives the bound

$$\begin{aligned}
&\left|\phi\left(\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right) - \phi\left(\left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right)\right| \\
&\leq L\left[1 + \|\left(\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right)\| + \|\left(\left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right)\|\right] \|\left(\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i\right)\| \\
&\leq L\left[1 + 2\|\left(\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right)\| + \|\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i\| + \|\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i\|\right] \left[\|\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i\| + \|\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i\|\right],
\end{aligned}$$

where the final inequality uses that $\|(\underline{a}, \underline{b})\| \leq \|\underline{a}\| + \|\underline{b}\|$ and that

$$\|(\underline{a}, \underline{b}, w)\| = \|(\underline{a}, \underline{b}, w) - (\tilde{\underline{a}}, \tilde{\underline{b}}, w) + (\tilde{\underline{a}}, \tilde{\underline{b}}, w)\| \leq \|(\underline{a} - \tilde{\underline{a}}, \underline{b} - \tilde{\underline{b}})\| + \|(\tilde{\underline{a}}, \tilde{\underline{b}}, w)\|.$$

Next, applying the above bound, Cauchy-Schwarz, and Lemma 16, we get

$$\begin{aligned}
[T2]^2 &= \left[\frac{1}{N}\sum_{i=1}^{N}\left|\phi\left(\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right) - \phi\left(\left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right)\right|\right]^2 \\
&\leq \left[\frac{L}{N}\sum_{i=1}^{N}\left(1 + 2\|(\left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i)\| + \|[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}]_i\| + \|[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}]_i\|\right)\right. \\
&\qquad \left.\times\left(\|[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}]_i\| + \|[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}]_i\|\right)\right]^2 \\
&\leq \frac{8L^2}{N}\left[\sum_{i=1}^{N}\left(\|[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}]_i\|^2 + \|[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}]_i\|^2\right)\right] \\
&\qquad \times\left[1 + \frac{4}{N}\sum_{i=1}^{N}\|(\left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i)\|^2 + \frac{1}{N}\sum_{i=1}^{N}\|\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i\|^2 + \frac{1}{N}\sum_{i=1}^{N}\|\left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i\|^2\right].
\end{aligned}$$

Next, using the state evolution in (2.67) and the discussion around (2.69), if $i \leq M$, then $\mathbb{E}\|(\left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i)\|^2 = \sum_{j=0}^{k}(\tau_{pj}^{\text{in}} + \tau_{pj}^{\text{out}}) + \mathbb{E}[\|[\mathbf{w}]_i\|^2]$, and if $M < i \leq N$, then $\mathbb{E}\|(\left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i)\|^2 = \sum_{j=0}^{k}\tau_{pj}^{\text{in}} + \mathbb{E}[\|[\mathbf{w}]_i\|^2] \leq \sum_{j=0}^{k}(\tau_{pj}^{\text{in}} + \tau_{pj}^{\text{out}}) + \mathbb{E}[\|[\mathbf{w}]_i\|^2]$, using our convention that $\left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i = 0$ for $i > M$. Note that $\mathbb{E}[\|[\mathbf{w}]_i\|^2] \leq 2\mathbb{E}W^2 + 2W_{\max}$, and define $\mathbb{E}_k := \sum_{j=0}^{k}(\tau_{pj}^{\text{in}} + \tau_{pj}^{\text{out}}) + 2\mathbb{E}W^2 + 2W_{\max}$.

Using these facts,

$$\mathbb{P}\left([T2]^2 \geq \frac{\epsilon^2}{4}\right) = \mathbb{P}\left(\left[\frac{1}{N}\sum_{i=1}^{N}\left|\phi\left(\left[\underline{\widetilde{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\underline{\widetilde{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right) - \phi\left(\left[\underline{\overset{*}{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\underline{\overset{*}{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i\right)\right|\right]^2 \geq \frac{\epsilon^2}{4}\right)$$
$$\leq \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\|[\underline{\widetilde{\mathbf{p}}}_k^{\text{in}} - \underline{\overset{*}{\mathbf{p}}}_k^{\text{in}}]_i\|^2 + \|[\underline{\widetilde{\mathbf{p}}}_k^{\text{out}} - \underline{\overset{*}{\mathbf{p}}}_k^{\text{out}}]_i\|^2\right) \geq \frac{\min\{1, 1/(8L^2)\}\,(\epsilon^2/4)}{2 + 8\mathbb{E}_k}\right)$$
$$+ \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\|(\left[\underline{\overset{*}{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\underline{\overset{*}{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i)\|^2 \geq 2\mathbb{E}_k\right). \tag{A.82}$$

Observe that the second term of (A.82) can be bounded above by

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\|(\left[\underline{\overset{*}{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\underline{\overset{*}{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i)\|^2 - \mathbb{E}\|(\left[\underline{\overset{*}{\mathbf{p}}}_k^{\text{in}}\right]_i, \left[\underline{\overset{*}{\mathbf{p}}}_k^{\text{out}}\right]_i, [\mathbf{w}]_i)\|^2\right) \geq \mathbb{E}_k\right). \tag{A.83}$$

Concentration for this follows directly by Lemma 17 since the functions $f_i : (p_1, p_2, w^{(r)}) \mapsto \|(p_1, p_2, w^{(r)}, w_i^{(d)})\|^2$ are pseudo-Lipschitz with a common constant by Lemma 18 and the fact that the components of $\mathbf{w}^{(d)}$ are bounded. We note in particular that $\mathbb{E}_k$ appears in the numerator of the exponential rate of convergence, but this is bounded below independently of N by the stopping criterion for the $\tau_{pj}^{\text{in/out}}$.

To bound the first term of (A.82), observe that by Lemma 4 part 2,

$$\sum_{i=1}^{N}\|\left[\underline{\widetilde{\mathbf{p}}}_k^{\text{in}} - \underline{\overset{*}{\mathbf{p}}}_k^{\text{in}}\right]_i\|^2 = \sum_{i=1}^{N}\sum_{j=0}^{k}\left(\left[\widetilde{\mathbf{p}}_j^{\text{in}} - \overset{*}{\mathbf{p}}_j^{\text{in}}\right]_i\right)^2 = \sum_{i=1}^{N}\sum_{j=0}^{k}\left[\sum_{r=0}^{j}[\mathbf{c}_{pj}^{\text{in}}]_r\left[\widetilde{\mathbf{\Delta}}_{pr}^{\text{in}}\right]_i\right]^2.$$

Next, by Cauchy-Schwarz, we have

$$\sum_{i=1}^{N}\|\left[\underline{\widetilde{\mathbf{p}}}_k^{\text{in}} - \underline{\overset{*}{\mathbf{p}}}_k^{\text{in}}\right]_i\|^2 \leq \sum_{i=1}^{N}\sum_{j=0}^{k}\left[\sum_{r=0}^{j}[\mathbf{c}_{pj}^{\text{in}}]_r^2\right]\left[\sum_{r=0}^{j}\left(\left[\widetilde{\mathbf{\Delta}}_{pr}^{\text{in}}\right]_i\right)^2\right] = \sum_{j=0}^{k}\left[\sum_{r=0}^{j}[\mathbf{c}_{pj}^{\text{in}}]_r^2\right]\left[\sum_{r=0}^{j}\|\widetilde{\mathbf{\Delta}}_{pr}^{\text{in}}\|^2\right]. \tag{A.84}$$

Observe that part 3 of Lemma 4 implies that $\tau_{pk}^{\text{in}} = \mathbb{E}\left[([\overset{*}{\mathbf{p}}_k^{\text{in}}]_i)^2\right] = \sum_{r=0}^{k}\rho_{pr}^{\text{in}}[\mathbf{c}_{pk}^{\text{in}}]_r^2$,where the

second equality follows from part 3 of Lemma 4 as $([\overset{*}{\mathbf{p}}{}_k^{\text{in}}]_i)^2 = (\sum_{r=0}^k \sqrt{\rho_{pr}^{\text{in}}}\,[\mathbf{c}_{pk}^{\text{in}}]_r\,[\widetilde{\mathbf{O}}_{pr}^{\text{in}}\,\overline{\mathbf{Z}}_{pr}^{\text{in}}]_i)^2 \overset{d}{=} (\sum_{r=0}^k \sqrt{\rho_{pr}^{\text{in}}}\,[\mathbf{c}_{pk}^{\text{in}}]_r\,[\overline{\mathbf{Z}}_{pr}^{\text{in}}]_i)^2$ with $[\overline{\mathbf{Z}}_{pr}^{\text{in}}]_i$ being i.i.d. standard Gaussian; therefore,

$$\sum_{r=0}^{k}[\mathbf{c}_{pk}^{\text{in}}]_r^2 \leq \frac{\tau_{pk}^{\text{in}}}{\min_{0\leq i\leq k}(\rho_{pi}^{\text{in}})} \leq \frac{\tau_{pk}^{\text{in}}}{\epsilon_1^*},$$

using the stopping criterion. Now using this bound in (A.84), we find

$$\begin{aligned}\sum_{i=1}^{N} \| \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i \|^2 \leq \sum_{j=0}^{k}\left[\sum_{r=0}^{j}[\mathbf{c}_{pj}^{\text{in}}]_r^2\right]\left[\sum_{r=0}^{j}\|\widetilde{\boldsymbol{\Delta}}_{pr}^{\text{in}}\|^2\right] &\leq \sum_{j=0}^{k}\left[\sum_{r=0}^{k}[\mathbf{c}_{pj}^{\text{in}}]_r^2\right]\left[\sum_{r=0}^{j}\|\widetilde{\boldsymbol{\Delta}}_{pr}^{\text{in}}\|^2\right] \\ &\leq \left[\frac{\tau_{pk}^{\text{in}}}{\epsilon_1^*}\right]\sum_{j=0}^{k}\sum_{r=0}^{j}\|\widetilde{\boldsymbol{\Delta}}_{pr}^{\text{in}}\|^2.\end{aligned}$$

Now, combining the above, we get that

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} \| \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{in}} - \overset{*}{\underline{\mathbf{p}}}_k^{\text{in}}\right]_i \|^2 \geq \frac{\min\{1, 1/(8L^2)\}\,(\epsilon^2/4)}{2 + 8\sum_{j=0}^{k}(\tau_{pj}^{\text{in}} + \tau_{pj}^{\text{out}}) + 8\mathbb{E}[|W|^2]}\right) \tag{A.85}$$

$$\leq \mathbb{P}\left(\frac{1}{N}\sum_{j=0}^{k}\sum_{r=0}^{j}\|\widetilde{\boldsymbol{\Delta}}_{pr}^{\text{in}}\|^2 \geq \frac{\epsilon_1^*\min(1, \frac{1}{8L^2})\,(\epsilon^2/4)}{\tau_{pk}^{\text{in}}(2 + 8\sum_{j=1}^{k}(\tau_{pj}^{\text{in}} + \tau_{pj}^{\text{out}}) + 8\mathbb{E}[|W|^2])}\right) \tag{A.86}$$

$$\leq k\sum_{r=0}^{k}\mathbb{P}\left(\frac{\|\widetilde{\boldsymbol{\Delta}}_{pr}^{\text{in}}\|^2}{N} \geq \frac{c\epsilon^2}{k^3}\right) \overset{(a)}{\leq} Ck^4\exp\left(-c\epsilon^2/k^7\right), \tag{A.87}$$

where c, C are constants, and the inequality (a) follows from $\mathbf{P_i}.(\mathbf{a})$ for $0 \leq i \leq k$. Note in particular that c is independent of N since the terms appearing in the denominator of the exponential rate, specifically the $\tau_{pj}^{\text{in/out}}$ and $\mathbb{E}[W^2]$ are all bounded above independently of N (see Section 2.3.4 for details). Next, consider

$$\frac{1}{N}\sum_{i=1}^{N} \| \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i - \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i \|^2 = \delta\frac{1}{M}\sum_{i=1}^{M} \| \left[\widetilde{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i - \left[\overset{*}{\underline{\mathbf{p}}}_k^{\text{out}}\right]_i \|^2,$$

where $\delta = M/N$ and the equality follows from the convention for $i > N$. Concentration for this term follows exactly the same as above, with the only difference coming from the appearance of a

δ^{-1} factor in the exponential rate. Combining these with lemma 9 gives concentration for the first term above, which completes the proof of $\mathbf{P_k}.(\mathbf{b})$.

$\mathbf{P_k}.(\mathbf{c})$ Observe that, by Lemma 8,

$$\phi([\mathbf{p}_1^{\text{in}}]_i, \ldots, [\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_1^{\text{out}}]_i, \ldots, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i) = [\mathbf{p}_k^{\text{in/out}}]_i [\mathbf{w}_p^{\text{in}}]_i \in PL(2).$$

Thus, by $\mathbf{P_k}.(\mathbf{b})$, we have that $\frac{1}{N}[\mathbf{p}_k^{\text{in/out}}]^T \mathbf{w}_p^{\text{in}} \doteq \mathbb{E}[P_k^{\text{in/out}} W_p^{\text{in}}] = 0$ (as W^p is assumed independent of $(P_0^{\text{in/out}}, \ldots, P_k^{\text{in/out}})$), as needed.

$\mathbf{P_k}.(\mathbf{d})$ Again by Lemma 8, for all $0 \leq j \leq k$,

$$\phi([\mathbf{p}_1^{\text{in}}]_i, \ldots, [\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_1^{\text{out}}]_i, \ldots, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i) = [\mathbf{p}_j^{\text{in/out}}]_i [\mathbf{p}_k^{\text{in/out}}]_i \in PL(2).$$

By $\mathbf{P_k}.(\mathbf{b})$, we find $\frac{1}{N}[\mathbf{p}_j^{\text{in/out}}]^T \mathbf{p}_k^{\text{in/out}} \doteq \mathbb{E}[P_j^{\text{in/out}} P_k^{\text{in/out}}]$, as needed.

$\mathbf{P_k}.(\mathbf{f})$ We note here that we prove $\mathbf{P_k}.(\mathbf{f})$ and then prove $\mathbf{P_k}.(\mathbf{e})$ afterwards, as the proof of $\mathbf{P_k}.(\mathbf{f})$ relies on the result of $\mathbf{P_k}.(\mathbf{f})$.

Concentration for α terms. First, we show $\alpha_{pk}^{\text{in}} \doteq \overline{\alpha}_{pk}^{\text{in}}$. To this end, recall that

$$\alpha_{pk}^{\text{in}} = \mathcal{T} \frac{1}{N} \sum_{i=1}^{N} [f_p^{\text{in}}]'([\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}^{\text{in}}, \gamma_{pk}^{\text{out}}),$$

where $\mathcal{T}$ is the truncation operator defined as $\mathcal{T}x = \min(t_2, \max(t_1, x))$. Recall that the upper truncation is only necessary if $\log p(x)$ or $\log p(y \mid z)$ is not strongly concave. Now, adopting the notation $[\mathbf{p}_k]_i = \left([\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_k^{\text{out}}]_i\right)$ and $\gamma_{pk} = (\gamma_{pk}^{\text{in}}, \gamma_{pk}^{\text{out}})$, observe that by Lemma 9,

$$\begin{aligned}\mathbb{P}\left(\left|\alpha_{pk}^{\text{in}} - \overline{\alpha}_{pk}^{\text{in}}\right| \geq \epsilon\right) \leq \mathbb{P}\left(\left|\mathcal{T}\frac{1}{N}\sum_{i=1}^{N}[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}\right) - \mathcal{T}\frac{1}{N}\sum_{i=1}^{N}[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}\right)\right| \geq \frac{\epsilon}{2}\right) \\ + \mathbb{P}\left(\left|\mathcal{T}\frac{1}{N}\sum_{i=1}^{N}[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}\right) - \overline{\alpha}_{pk}^{\text{in}}\right| \geq \frac{\epsilon}{2}\right). \qquad \text{(A.88)}\end{aligned}$$

Label the two terms of (A.88) as $T1$ and $T2$.

For term $T2$ of (A.88), because $\overline{\alpha}^{\text{in}}_{pk} \in [t_{\min}, t_{\max}]$ (by **Condition 6**),

$$\mathbb{P}\left(\left|\mathcal{T}\frac{1}{N}\sum_{i=1}^{N}[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}\right) - \overline{\alpha}^{\text{in}}_{pk}\right| \geq \frac{\epsilon}{2}\right) \leq \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}\right) - \overline{\alpha}^{\text{in}}_{pk}\right| \geq \frac{\epsilon}{2}\right).$$

By the assumption that $[f_p^{\text{in}}]'$ is assumed either uniformly Lipschitz or bounded conditionally Lipschitz in **Condition 3**, we have concentration for the right side by $\mathbf{P_k}.(\mathbf{b})$.

For term $T1$ of (A.88), observe that for any $x, y \in \mathbb{R}$, $|\mathcal{T}x - \mathcal{T}y| \leq |x - y|$; therefore,

$$\begin{aligned}\mathbb{P}\Big(\Big|\mathcal{T}\frac{1}{N}\sum_{i=1}^{N}[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}\right) - \mathcal{T}\frac{1}{N}\sum_{i=1}^{N}[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}\right)\Big| \geq \frac{\epsilon}{2}\Big)& \\ \leq \mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N}\Big|[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}\right) - [f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}\right)\Big| \geq \frac{\epsilon}{2}\Big).& \qquad \text{(A.89)}\end{aligned}$$

Now, by **Condition 3**, the functions $[f_p^{\text{in}}]'$ are uniformly Lipschitz around $\overline{\gamma}_{pk}$. This guarantees that there exists some $\delta > 0$ such that if $\left|\gamma^{\text{in}}_{pk} - \overline{\gamma}^{\text{in}}_{pk}\right| < \delta$ and $\left|\gamma^{\text{out}}_{pk} - \overline{\gamma}^{\text{out}}_{pk}\right| < \delta$, then

$$\begin{aligned}&\left|[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}\right) - [f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}\right)\right| \\ &\qquad\qquad \leq L_p\left(1 + \|([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right)\left[\left|\gamma^{\text{in}}_{pk} - \overline{\gamma}^{\text{in}}_{pk}\right| + \left|\gamma^{\text{out}}_{pk} - \overline{\gamma}^{\text{out}}_{pk}\right|\right].\end{aligned}$$

In particular, this implies that if $\left|\gamma^{\text{in}}_{pk} - \overline{\gamma}^{\text{in}}_{pk}\right| < \delta$ and $\left|\gamma^{\text{out}}_{pk} - \overline{\gamma}^{\text{out}}_{pk}\right| < \delta$, and if

$$\left[\left|\gamma^{\text{in}}_{pk} - \overline{\gamma}^{\text{in}}_{pk}\right| + \left|\gamma^{\text{out}}_{pk} - \overline{\gamma}^{\text{out}}_{pk}\right|\right]\frac{L_p}{N}\sum_{i=1}^{N}\left(1 + \|([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right) \leq \frac{\epsilon}{2},$$

then

$$\frac{1}{N}\sum_{i=1}^{N}\left|[f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}\right) - [f_p^{\text{in}}]'\left([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}\right)\right| \leq \frac{\epsilon}{2}.$$

It then follows from (A.89) that $T1$ can be upper bounded by

$$\mathbb{P}\left(\left[\left|\gamma_{pk}^{\text{in}}-\overline{\gamma}_{pk}^{\text{in}}\right|+\left|\gamma_{pk}^{\text{out}}-\overline{\gamma}_{pk}^{\text{out}}\right|\right]\frac{L_p}{N}\sum_{i=1}^{N}\left(1+\|([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i)\|\right)\geq\frac{\epsilon}{2}\right) \tag{A.90}$$
$$+\mathbb{P}\left(\left|\gamma_{pk}^{\text{in}}-\overline{\gamma}_{pk}^{\text{in}}\right|\geq\min\{1,\delta\}\right)+\mathbb{P}\left(\left|\gamma_{pk}^{\text{out}}-\overline{\gamma}_{pk}^{\text{out}}\right|\geq\min\{1,\delta\}\right).$$

The second and third terms of (A.90) concentrate by $\mathbf{Q_k}.(\mathbf{f})$, and the first term concentrates using Lemma 11 along with $\mathbf{Q_k}.(\mathbf{f})$ and $\mathbf{P_k}.(\mathbf{b})$, because $1+\|([\mathbf{p}_k^{\text{in}}]_i,[\mathbf{p}_k^{\text{out}}]_i,[\mathbf{w}_p^{\text{in}}]_i)\|$ is pseudo-Lipschitz of order 2 and since the $\mathbb{E}\|(P_k^{\text{in}},P_k^{\text{out}},W_p^{\text{in}},\mathbf{w}_i^{(d)})\|$ are upper bounded by a universal constant independent of N (in particular, since the $\mathbf{w}^{(d)}$ components are bounded above).

Concentration for γ terms. We now show concentration of γ_{qk}^{in} around $\overline{\gamma}_{qk}^{\text{in}}$. We can assume that the clipping of the γ iterates is absorbed by the Γ functions, because clipping a Lipschitz function to bound its range within a compact interval gives another Lipschitz function with the same constant L_1. Observe that by (7) in Algorithm 3,

$$\mathbb{P}\left(\left|\gamma_{qk}^{\text{in}}-\overline{\gamma}_{qk}^{\text{in}}\right|\geq\epsilon\right)=\mathbb{P}\left(\left|\Gamma_q^{\text{in}}\left(\gamma_{pk}^{\text{in}},\alpha_{pk}^{\text{in}}\right)-\Gamma_q^{\text{in}}\left(\overline{\gamma}_{pk}^{\text{in}},\overline{\alpha}_{pk}^{\text{in}}\right)\right|\geq\epsilon\right)\overset{(a)}{\leq}\mathbb{P}\left(\left|\gamma_{pk}^{\text{in}}-\overline{\gamma}_{pk}^{\text{in}}\right|+\left|\alpha_{pk}^{\text{in}}-\overline{\alpha}_{pk}^{\text{in}}\right|\geq\frac{\epsilon}{L_1}\right)$$
$$\overset{(b)}{\leq}\mathbb{P}\left(\left|\gamma_{pk}^{\text{in}}-\overline{\gamma}_{pk}^{\text{in}}\right|\geq\frac{\epsilon}{2L_1}\right)+\mathbb{P}\left(\left|\alpha_{pk}^{\text{in}}-\overline{\alpha}_{pk}^{\text{in}}\right|\geq\frac{\epsilon}{2L_1}\right). \tag{A.91}$$

where (a) follows from the assumption that Γ_q^{in} is Lipschitz continuous on its domain along with the fact that $\|(\gamma_{pk}^{\text{in}}-\overline{\gamma}_{pk}^{\text{in}},\alpha_{pk}^{\text{in}}-\overline{\alpha}_{pk}^{\text{in}})\|\leq|\gamma_{pk}^{\text{in}}-\overline{\gamma}_{pk}^{\text{in}}|+|\alpha_{pk}^{\text{in}}-\overline{\alpha}_{pk}^{\text{in}}|$ and (b) by Lemma 9. The final terms in the above can be upper bounded by the work just above in (A.88)-(A.90) and $\mathbf{Q_k}.(\mathbf{f})$.

Concentration for inner product terms. We show concentration for $[\mathbf{v}_k^{\text{in}}]^T\mathbf{p}_j^{\text{in}}$ where $0\leq j\leq k$. Recall that by (5) in Algorithm 3, $[\mathbf{v}_k^{\text{in}}]_i=(1-\alpha_{pk}^{\text{in}})^{-1}g_p([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk},\alpha_{pk}^{\text{in}})$, where we have defined functions

$$g_p([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk},\alpha_{pk}^{\text{in}}):=\mathbf{f}_p^{\text{in}}([\mathbf{p}_k^{\text{in}}]_i,[\mathbf{p}_k^{\text{out}}]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk}^{\text{in}},\gamma_{pk}^{\text{out}})-\alpha_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i, \tag{A.92}$$

where to simplify notation, we write $[\mathbf{p}_k]_i = \left([\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_k^{\text{out}}]_i\right)$ and $\gamma_{pk} = \left(\gamma_{pk}^{\text{in}}, \gamma_{pk}^{\text{out}}\right)$. Observe that

$$[\mathbf{p}_j^{\text{in}}]_i g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \alpha_{pk}^{\text{in}}) = (1-\alpha_{pk}^{\text{in}})^{-1}\left[[\mathbf{p}_j^{\text{in}}]_i f_p^{\text{in}}([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}) - \alpha_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i[\mathbf{p}_j^{\text{in}}]_i\right].$$

Now by Lemma 11 it suffices to show concentration separately for $(1-\alpha_{pk}^{\text{in}})^{-1}$ and

$$[\mathbf{p}_j^{\text{in}}]_i f_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}) - \alpha_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i[\mathbf{p}_j^{\text{in}}]_i.$$

For $(1-\alpha_{pk}^{\text{in}})^{-1}$, we have,

$$\mathbb{P}\left(\left|(1-\alpha_{pk}^{\text{in}})^{-1} - (1-\overline{\alpha}_{pk}^{\text{in}})^{-1}\right| \geq \epsilon\right) \leq \mathbb{P}\left(\left|\alpha_{pk}^{\text{in}} - \overline{\alpha}_{pk}^{\text{in}}\right| \geq \frac{\epsilon}{L_2}\right), \tag{A.93}$$

where the inequality follows from the fact that $\frac{1}{1-a}$ is Lipschitz on $[t_1, t_2]$ with Lipschitz constant $L_2 := (1-t_2)^{-2}$ and the fact that the α_{pk}^{in} and $\overline{\alpha}_{pk}^{\text{in}}$ lie in $[t_1, t_2]$ by **Assumption 4**. The right hand side then concentrates by the work for α concentration above ($\mathbf{P_k}.(\mathbf{f})$; the first part).

Next, we observe that

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{N}\left|\sum_{i=1}^{N}\left[[\mathbf{p}_j^{\text{in}}]_i f_p^{\text{in}}([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}) - \alpha_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i[\mathbf{p}_j^{\text{in}}]_i\right]\right| \geq \epsilon\right) \\
&\overset{(a)}{\leq} \mathbb{P}\left(\left|\alpha_{pk}^{\text{in}} - \overline{\alpha}_{pk}^{\text{in}}\right|\left|\frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_k^{\text{in}}]_i[\mathbf{p}_j^{\text{in}}]_i\right| \geq \frac{\epsilon}{2}\right) \\
&\qquad + \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_j^{\text{in}}]_i\left(f_p^{\text{in}}([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}) - \overline{\alpha}_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i\right)\right| \geq \frac{\epsilon}{2}\right),
\end{aligned} \tag{A.94}$$

where (a) follows by adding and subtracting $\overline{\alpha}_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i[\mathbf{p}_j^{\text{in}}]_i$ from each term and applying the triangle inequality. Now the first term in (A.94) concentrates by Lemma 11, $\mathbf{P_k}.(\mathbf{f})$ (the first part), and $\mathbf{P_k}.(\mathbf{b})$ since $[\mathbf{p}_j^{\text{in}}]_i[\mathbf{p}_k^{\text{in}}]_i \in PL(2)$. We note that while the rate of concentration depends inversely on $\mathbb{E}P_j^{\text{in}}P_k^{\text{in}}$, this quantity is bounded above independent of N (see Section 2.3.5 for details).

Adding and subtracting $[\mathbf{p}_j^{\text{in}}]_i\left(f_p^{\text{in}}([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}) - \overline{\alpha}_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i\right)$, we bound the second term

in (A.94) by

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_j^{\text{in}}]_i\left(f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})-\overline{\alpha}_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i\right)\right|\geq\frac{\epsilon}{4}\right) \tag{A.95}$$
$$+\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_j^{\text{in}}]_i\left|f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})-f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk})\right|\geq\frac{\epsilon}{4}\right).$$

The second of these terms can be bounded in exactly the same way as $T1b$ in $\mathbf{P_k}.(\mathbf{e})$ below.

For the first term in (A.95), note that if $f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})$ is Lipschitz (in all inputs), then

$$\phi([\mathbf{p}_0^{\text{in}}]_i,\ldots,[\mathbf{p}_k^{\text{in}}]_i,[\mathbf{p}_0^{\text{out}}]_i,\ldots,[\mathbf{p}_k^{\text{out}}]_i,[\mathbf{w}_p^{\text{in}}]_i)=[\mathbf{p}_j^{\text{in}}]_i\left(f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})-\overline{\alpha}_{pk}^{\text{in}}[\mathbf{p}_k^{\text{in}}]_i\right),$$

is in $PL(2)$ by Lemma 8. If instead $f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})$ satisfies the bounded conditionally Lipschitz condition, then ϕ is bounded conditionally pseudo-Lipschitz. In either case, $\mathbf{P_k}.(\mathbf{b})$ tells us that this will concentrate around

$$\mathbb{E}\left\{P_j^{\text{in}}\left(f_p^{\text{in}}(P_k^{\text{in}},P_k^{\text{out}},W_p^{\text{in}},\overline{\gamma}_{pk}^{\text{in}},\overline{\gamma}_{pk}^{\text{out}})-\overline{\alpha}_{pk}^{\text{in}}P_k^{\text{in}}\right)\right\}. \tag{A.96}$$

Therefore, to complete the proof, we just need to show that the expression in (A.96) is equal to zero. To do this, recall that $\mathbb{E}[P_j^{\text{in}}]=0$, so (A.96) is equal to

$$\begin{aligned}&\text{Cov}\left(P_j^{\text{in}},f_p^{\text{in}}(P_k^{\text{in}},P_k^{\text{out}},W_p^{\text{in}},\overline{\gamma}_{pk}^{\text{in}},\overline{\gamma}_{pk}^{\text{out}})-\overline{\alpha}_{pk}^{\text{in}}P_k^{\text{in}}\right)\\&\qquad=\mathbb{E}\left\{P_j^{\text{in}}P_k^{\text{in}}\right\}\times\mathbb{E}\left\{[f_p^{\text{in}}]'(P_k^{\text{in}},P_k^{\text{out}},W_p^{\text{in}},\overline{\gamma}_{pk}^{\text{in}},\overline{\gamma}_{pk}^{\text{out}})-\overline{\alpha}_{pk}^{\text{in}}\right\}=0,\end{aligned}$$

where the first equality follows from Stein's lemma, and the second equality follows from the definition of $\overline{\alpha}_{pk}^{\text{in}}$ in the GVAMP state evolution (2.67).

Concentration for $[\mathbf{p}_k^{\text{in}}]^T\mathbf{v}_j^{\text{in}}$ is proved in the same way as the above.

$\mathbf{P_k}.(\mathbf{e})$ By (5) in Algorithm 3 and the definition of the function g_p in (A.92),

$$\mathbb{P}\left(\left|\frac{1}{N}[\mathbf{v}_j^{\text{in}}]^T\mathbf{v}_k^{\text{in}} - \left[\mathbf{\Sigma}_{vk}^{\text{in}}\right]_{(j+1)(k+1)}\right| \geq \epsilon\right) \tag{A.97}$$
$$= \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\frac{g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \alpha_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pj}, \alpha_{pj}^{\text{in}})}{(1-\alpha_{pk}^{\text{in}})(1-\alpha_{pj}^{\text{in}})} - \mathbb{E}[V_k^{\text{in}}V_j^{\text{in}}]\right| \geq \epsilon\right).$$

Using Lemma 10, we can separately show concentration for $(1-\alpha_{pk}^{\text{in}})^{-1}(1-\alpha_{pj}^{\text{in}})^{-1}$ and

$$\frac{1}{N}\sum_{i=1}^{N} g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \alpha_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pj}, \alpha_{pj}^{\text{in}}).$$

First observe that the function $f(a) = \frac{1}{1-a}$ is Lipschitz over the interval $[t_1, t_2]$ within which the α iterates must lie given our truncation assumption (**Assumption 4**) and we can take $L_2 = \frac{1}{(1-t_2)^2}$ to be the Lipschitz constant. Therefore,

$$\mathbb{P}\left(\left|\frac{1}{1-\alpha_{pk}^{\text{in}}} - \frac{1}{1-\overline{\alpha}_{pk}^{\text{in}}}\right| \geq \epsilon\right) \leq \mathbb{P}\left(\left|\alpha_{pk}^{\text{in}} - \overline{\alpha}_{pk}^{\text{in}}\right| \geq \frac{\epsilon}{L_2}\right). \tag{A.98}$$

The right side of (A.98) concentrates by $\mathbf{P_k}.(\mathbf{f})$. Now $\frac{1}{1-\alpha_{pj}^{\text{in}}}$ concentrates in the same way using $\mathbf{P_j}.(\mathbf{f})$, and concentration for $(1-\alpha_{pk}^{\text{in}})^{-1}(1-\alpha_{pj}^{\text{in}})^{-1}$ follows by Lemma 10. Since the $\overline{\alpha}_{pk}^{\text{in/out}} \leq 1$, they do not affect the exponential rate of concentration.

Next, using Lemma 9, we have that

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \alpha_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pj}, \alpha_{pk}^{\text{in}}) - \left(1-\overline{\alpha}_{pk}^{\text{in}}\right)\left(1-\overline{\alpha}_{pj}^{\text{in}}\right)\mathbb{E}[V_k^{\text{in}}V_j^{\text{in}}]\right| \geq \epsilon\right) \\
&\leq \mathbb{P}\Bigg(\Bigg|\frac{1}{N}\sum_{i=1}^{N} g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \alpha_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pj}, \alpha_{pk}^{\text{in}}) \\
&\qquad - \frac{1}{N}\sum_{i=1}^{N} g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \overline{\alpha}_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pj}, \overline{\alpha}_{pk}^{\text{in}})\Bigg| \geq \frac{\epsilon}{2}\Bigg) \\
&+ \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \overline{\alpha}_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pj}, \overline{\alpha}_{pk}^{\text{in}}) - \left[1-\overline{\alpha}_{pk}^{\text{in}}\right]\left[1-\overline{\alpha}_{pj}^{\text{in}}\right]\mathbb{E}[V_k^{\text{in}}V_j^{\text{in}}]\right| \geq \frac{\epsilon}{2}\right).
\end{aligned} \tag{A.99}$$

We label these latter two terms $T1$ and $T2$ and analyze them separately.

Concentration for T1 of (A.99)**.** Observe that repeated application of the triangle inequality the definition of the function g_p in (A.92) gives us

$$\begin{aligned}\Big|\frac{1}{N}\sum_{i=1}^{N}\Big[g_p([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk},\alpha_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj},\alpha_{pk}^{\text{in}})-g_p([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk},\overline{\alpha}_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj},\overline{\alpha}\\ \leq \frac{1}{N}\left|\alpha_{pk}^{\text{in}}\alpha_{pj}^{\text{in}}-\overline{\alpha}_{pk}^{\text{in}}\overline{\alpha}_{pj}^{\text{in}}\right|\left|\sum_{i=1}^{N}[\mathbf{p}_k^{\text{in}}]_i[\mathbf{p}_j^{\text{in}}]_i\right|+\frac{1}{N}\left|\alpha_{pk}^{\text{in}}-\overline{\alpha}_{pk}^{\text{in}}\right|\left|\sum_{i=1}^{N}[\mathbf{p}_k^{\text{in}}]_i f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj})\right|\\ +\frac{1}{N}\left|\alpha_{pj}^{\text{in}}-\overline{\alpha}_{pj}^{\text{in}}\right|\left|\sum_{i=1}^{N}[\mathbf{p}_j^{\text{in}}]_i f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk})\right|.\end{aligned}$$

By Lemmas 9 and 11, to establish concentration for these terms, it suffices to show concentration for each of the following:

$$\begin{aligned}&\left|\alpha_{pk}^{\text{in}}-\overline{\alpha}_{pk}^{\text{in}}\right|, && \left|\alpha_{pj}^{\text{in}}-\overline{\alpha}_{pj}^{\text{in}}\right|, && \left|\alpha_{pk}^{\text{in}}\alpha_{pj}^{\text{in}}-\overline{\alpha}_{pk}^{\text{in}}\overline{\alpha}_{pj}^{\text{in}}\right|,\\ &\frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_k^{\text{in}}]_i[\mathbf{p}_j^{\text{in}}]_i, && \frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_k^{\text{in}}]_i f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj}), && \frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_j^{\text{in}}]_i f_p^{k}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk}).\end{aligned}$$

We note that the rates of concentration will depend inversely on the limiting values of the last three quantities. However, each of these is a continuous function of the limiting quantities defined in Section 2.3.5 (which follows from the recursive definition (2.60)). Thus, since these limiting quantities are contained in an N-independent compact set, these limiting quantities are themselves bounded above independently of N and thus do not affect the exponential rate of concentration.

Concentration for the first two follows by $\mathbf{P_k}.(\mathbf{f})$ and $\mathbf{P_j}.(\mathbf{f})$ above, and concentration for the third follows from that and Lemma 10. Concentration for fourth term follows from $\mathbf{P_k}.(\mathbf{b})$ as

$$\phi([\mathbf{p}_0^{\text{in}}]_i,\ldots,[\mathbf{p}_k^{\text{in}}]_i,[\mathbf{p}_0^{\text{out}}]_i,\ldots,[\mathbf{p}_k^{\text{out}}]_i,[\mathbf{w}_p^{\text{in}}]_i)=[\mathbf{p}_k^{\text{in}}]_i[\mathbf{p}_j^{\text{in}}]_i,$$

is pseudo-Lipschitz of order 2 by Lemma 8.

Thus, it remains to show concentration for the last two terms above. For simplicity, we show

this only for the first of these, with the other following in the same way. First, defining $\mu := \mathbb{E}\left\{P_j^{\text{in}} f_p\left(P_k^{\text{in}}, P_k^{\text{out}}, W_p^{\text{in}}, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}\right)\right\}$ and writing $\overline{\gamma}_{pk} = \left(\overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}\right)$,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_j^{\text{in}}]_i f_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}) - \mu\right| \geq \epsilon\right) &\leq \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_j^{\text{in}}]_i f_p^{\text{in}}([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}) - \mu\right| \geq \frac{\epsilon}{2}\right) \\ &+ \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}[\mathbf{p}_j^{\text{in}}]_i \left(f_p^{\text{in}}([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}) - f_p^{\text{in}}([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk})\right)\right| \geq \frac{\epsilon}{2}\right).\end{aligned} \tag{A.100}$$

Let the terms on the right hand side of (A.100) be $T1a$ and $T1b$, respectively. First we analyze $T1a$. By assumption, $f_p^{\text{in}}(p^{\text{in}}, p^{\text{out}}, w, \gamma^{\text{in}}, \gamma^{\text{out}})$ is either uniformly Lipschitz or uniformly bounded conditionally Lipschitz (UBCL) at $(\gamma^{\text{in}}, \gamma^{\text{out}}) = (\overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}})$. By definition, this implies that $f_p^{\text{in}}(p^{\text{in}}, p^{\text{out}}, w, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}})$ is either Lipschitz in $(p^{\text{in}}, p^{\text{out}}, w)$ or bounded conditionally Lipschitz in $(p^{\text{in}}, p^{\text{out}})$ for each w^p. Thus, the function

$$\phi([\mathbf{p}_0^{\text{in}}]_i, \ldots, [\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_0^{\text{out}}]_i, \ldots, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i) = [\mathbf{p}_j^{\text{in}}]_i f_p([\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}})$$

is either pseudo-Lipschitz of order 2 or bounded conditionally pseudo-Lipschitz by Lemma 8. Therefore, concentration for $T1a$ follows from $\mathbf{P_k}.(\mathbf{b})$.

Now for $T1b$, we have by the definition of uniform Lipschitz continuity and UBCL that there is some $\delta > 0$ such that $\left|\gamma_{pk}^{\text{in}} - \overline{\gamma}_{pk}^{\text{in}}\right| < \delta$ and $\left|\gamma_{pk}^{\text{out}} - \overline{\gamma}_{pk}^{\text{out}}\right| < \delta$ implies that

$$\begin{aligned}&\left|f_p^{\text{in}}(p^{\text{in}}, p^{\text{out}}, w, \gamma_{pk}^{\text{in}}, \gamma_{pk}^{\text{out}}) - f_p^{\text{in}}(p^{\text{in}}, p^{\text{out}}, w, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}})\right| \\ &\qquad \leq L_p\left(1 + \|(p^{\text{in}}, p^{\text{out}}, w)\|\right)\left[\left|\gamma_{pk}^{\text{in}} - \overline{\gamma}_{pk}^{\text{in}}\right| + \left|\gamma_{pk}^{\text{out}} - \overline{\gamma}_{pk}^{\text{out}}\right|\right].\end{aligned}$$

Following the work in (A.90)- (A.89), we have the upper bound

$$T_1 b \leq \mathbb{P}\left(\left|\gamma_{pk}^{\text{in}} - \overline{\gamma}_{pk}^{\text{in}}\right| \geq \min\{1,\delta\}\right) + \mathbb{P}\left(\left|\gamma_{pk}^{\text{out}} - \overline{\gamma}_{pk}^{\text{out}}\right| \geq \min\{1,\delta\}\right) \tag{A.101}$$
$$+ \mathbb{P}\left(L_p\left[\left|\gamma_{pk}^{\text{in}} - \overline{\gamma}_{pk}^{\text{in}}\right| + \left|\gamma_{pk}^{\text{out}} - \overline{\gamma}_{pk}^{\text{out}}\right|\right]\frac{1}{N}\sum_{i=1}^{N}\left[|[\mathbf{p}_j^{\text{in}}]_i|\left(1 + \|([\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right)\right] \geq \frac{\epsilon}{2}\right).$$

The first two terms above concentrates by $\mathbf{Q_k}.(\mathbf{f})$, and the third concentrates (at a rate independent of N) using Lemma 11 combined with $\mathbf{Q_k}.(\mathbf{f})$ and $\mathbf{P_k}.(\mathbf{b})$ since

$$\phi([\mathbf{p}_0^{\text{in}}]_i, \ldots, [\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_0^{\text{out}}]_i, \ldots, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i) = |[\mathbf{p}_j^{\text{in}}]_i|\left(1 + \|([\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right)$$

is pseudo-Lipschitz of order 2 by Lemma 8, and since the $\mathbf{w}^{(d)}$ components are bounded independent of N. This establishes concentration for $T1b$.

Concentration for T2 of (A.99)**.** We have to show concentration for

$$T2 = \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \overline{\alpha}_{pk}^{\text{in}}) g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pj}, \overline{\alpha}_{pk}^{\text{in}}) - \left[1 - \overline{\alpha}_{pk}^{\text{in}}\right]\left[1 - \overline{\alpha}_{pj}^{\text{in}}\right]\mathbb{E}[V_k^{\text{in}} V_j^{\text{in}}]\right| \geq \frac{\epsilon}{2}\right). \tag{A.102}$$

Adding and subtracting $\frac{1}{N}\sum_{i=1}^{N} g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}, \overline{\alpha}_{pk}^{\text{in}}) g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pj}, \overline{\alpha}_{pk}^{\text{in}})$, we find that from Lemma 9,

$$T2 \leq \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}, \overline{\alpha}_{pk}^{\text{in}}) g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pj}, \overline{\alpha}_{pk}^{\text{in}}) - \left[1 - \overline{\alpha}_{pk}^{\text{in}}\right]\left[1 - \overline{\alpha}_{pj}^{\text{in}}\right]\mathbb{E}[V_k^{\text{in}} V_j^{\text{in}}]\right| \geq \frac{\epsilon}{2}\right)$$
$$+ \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\left[g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pk}, \overline{\alpha}_{pk}^{\text{in}}) g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \gamma_{pj}, \overline{\alpha}_{pk}^{\text{in}})\right.\right.$$
$$\left.\left. - \; g_p([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk}, \overline{\alpha}_{pk}^{\text{in}}) g_p([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pj}, \overline{\alpha}_{pk}^{\text{in}})\right]\right| \geq \frac{\epsilon}{2}\right). \tag{A.103}$$

By the same reasoning used above, $g_p(p^{\text{in}}, p^{\text{out}}, w, \overline{\gamma}_{pk}^{\text{in}}, \overline{\gamma}_{pk}^{\text{out}}, \overline{\alpha}_{pk}^{\text{in}})$ is Lipschitz in $(p^{\text{in}}, p^{\text{out}}, w)$ or UBCL in $(p^{\text{in}}, p^{\text{out}})$ for each w. Therefore, using Lemma 8 and $\mathbf{P_k}.(\mathbf{b})$, the function in the first term on the right side of (A.103) is either pseudo-Lipschitz or bounded conditionally pseudo-Lipschitz,

and thus the first term in (A.103) concentrates because

$$\left(1-\overline{\alpha}_{pk}^{\text{in}}\right)\left(1-\overline{\alpha}_{pj}^{\text{in}}\right)\mathbb{E}\{V_k^{\text{in}}V_j^{\text{in}}\}=\mathbb{E}\left\{g_p\left(P_k^{\text{in}},P_k^{\text{out}},W_p^{\text{in}},\overline{\gamma}_{pk}^{\text{in}},\overline{\gamma}_{pk}^{\text{out}},\overline{\alpha}_{pk}^{\text{in}}\right)g_p\left(P_j^{\text{in}},P_j^{\text{out}},W_p^{\text{in}},\overline{\gamma}_{pj}^{\text{in}},\overline{\gamma}_{pj}^{\text{out}},\overline{\alpha}_{pj}^{\text{in}}\right)\right\}.$$

Thus, we just need to show that the second term on the right side of (A.103) concentrates. Applying the definition of the function g_p in (A.92) and expanding terms, we see that

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\Big[g_p([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk},\overline{\alpha}_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj},\overline{\alpha}_{pk}^{\text{in}})-g_p([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk},\overline{\alpha}_{pk}^{\text{in}})g_p([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pj},\overline{\alpha}\right.\\
&\le\overline{\alpha}_{pk}^{\text{in}}\frac{1}{N}\sum_{i=1}^{N}|[\mathbf{p}_k^{\text{in}}]_i|\left|f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj})-f_p([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pj})\right|\\
&\quad+\overline{\alpha}_{pj}^{\text{in}}\frac{1}{N}\sum_{i=1}^{N}|[\mathbf{p}_j^{\text{in}}]_i|\left|f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk})-f_p([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})\right|\\
&\quad+\frac{1}{N}\sum_{i=1}^{N}\left|f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk})f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj})-f_p([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pj})\right|.
\end{aligned}\tag{A.104}$$

Notice that the first two terms on the right side of (A.104) concentrate in the same way as term T1b of (A.100). Namely, using the uniformly Lipschitz or UBCL property of f_p^{in}, we can bound the sums by sums of (bounded conditionally) pseudo-Lipschitz functions. Considering the third term on the right side of (A.104), we apply Lemma 9 and then to complete the proof, we obtain concentration for the following three quantities:

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N}\left|f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})\right|\left|f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj})-f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pj})\right|\ge\epsilon\Big),$$

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N}\left|f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pj})\right|\left|f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk})-f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})\right|\ge\epsilon\Big),$$

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N}\left|f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pk})-f_p^{\text{in}}([\mathbf{p}_k]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pk})\right|\left|f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\gamma_{pj})-f_p^{\text{in}}([\mathbf{p}_j]_i,[\mathbf{w}_p^{\text{in}}]_i,\overline{\gamma}_{pj})\right|\ge\epsilon\Big).$$

Let these probabilities be labeled as $T2a-T2c$.

Following the work in (A.90)- (A.89) and using the uniformly (bounded conditionally) Lipschitz property of f_p, we have the upper bound

$$T_2a \leq \mathbb{P}\left(\left|\gamma_{pk}^{\text{in}} - \overline{\gamma}_{pk}^{\text{in}}\right| \geq \min\{1,\delta\}\right) + \mathbb{P}\left(\left|\gamma_{pk}^{\text{out}} - \overline{\gamma}_{pk}^{\text{out}}\right| \geq \min\{1,\delta\}\right) \tag{A.105}$$
$$+ \mathbb{P}\left(\left[\left|\gamma_{pj}^{\text{in}} - \overline{\gamma}_{pj}^{\text{in}}\right| + \left|\gamma_{pj}^{\text{out}} - \overline{\gamma}_{pj}^{\text{out}}\right|\right] \frac{L_1}{N}\sum_{i=1}^{N}\left|f_p^{\text{in}}([\mathbf{p}_j]_i, [w_p^{\text{in}}]_i, \overline{\gamma}_{pk})\right|\left(1 + \|([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right) \geq \epsilon\right).$$

The first two terms can be bounded using $\mathbf{Q_k}.(\mathbf{f})$. The last term, in turn, can be upper bounded using Lemma 11 along with $\mathbf{Q_k}.(\mathbf{f})$ and $\mathbf{P_k}.(\mathbf{b})$ since by Lemma 8

$$\phi([\mathbf{p}_0^{\text{in}}]_i, \ldots, [\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_0^{\text{out}}]_i, \ldots, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}^p]_i) = \left|f_p^{\text{in}}([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i, \overline{\gamma}_{pk})\right|\left(1 + \|([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right)$$

is (bounded conditionally) pseudo-Lipschitz of order 2.

Term $T2b$ can be bounded in exactly the same way as $T2a$. We note that these rates of concentration are again independent of N since $\mathbb{E}f_p^{\text{in}}\left(P_j, W_p^{\text{in}}, \overline{\gamma}_{pk}\right)\left(1 + \|\left(P_k, W_p^{\text{in}}, \mathbf{w}^{(d)}\right)\|\right)$ can be bounded independently of N. In particular, this expectation depends (continuously) only on the deterministic $\mathbf{w}^{(d)}$ and the limiting quantities defined in Section 2.3.5, and these are all contained in N-independent compact sets.

For T_2c, we again use the strategy as in (A.90)- (A.89) to provide the following upper bound for term T_2c. We have

$$\begin{aligned} T_2c \leq\ & \mathbb{P}\left(\left|\gamma_{pk}^{\text{in}} - \overline{\gamma}_{pk}^{\text{in}}\right| \geq \min\{1,\delta\}\right) + \mathbb{P}\left(\left|\gamma_{pk}^{\text{out}} - \overline{\gamma}_{pk}^{\text{out}}\right| \geq \min\{1,\delta\}\right) \\ & + \mathbb{P}\left(\left|\gamma_{pj}^{\text{in}} - \overline{\gamma}_{pj}^{\text{in}}\right| \geq \min\{1,\delta\}\right) + \mathbb{P}\left(\left|\gamma_{pj}^{\text{out}} - \overline{\gamma}_{pj}^{\text{out}}\right| \geq \min\{1,\delta\}\right) \\ & + \mathbb{P}\Big(\left[\left|\gamma_{pj}^{\text{in}} - \overline{\gamma}_{pj}^{\text{in}}\right| + \left|\gamma_{pj}^{\text{out}} - \overline{\gamma}_{pj}^{\text{out}}\right|\right]\left[\left|\gamma_{pk}^{\text{in}} - \overline{\gamma}_{pk}^{\text{in}}\right| + \left|\gamma_{pk}^{\text{out}} - \overline{\gamma}_{pk}^{\text{out}}\right|\right] \\ & \qquad \times \frac{L_1^2}{N}\sum_{i=1}^{N}\left(1 + \|([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right)\left(1 + \|([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right) \geq \frac{\epsilon}{2}\Big). \end{aligned} \tag{A.106}$$

We know from $\mathbf{Q_k}.(\mathbf{f})$ and $\mathbf{Q_j}.(\mathbf{f})$ that the first four terms concentrate as needed. Then the last term in (A.106) can be upper bounded with repeated application of Lemmas 11 and 9 along with

$\mathbf{Q_k}.(\mathbf{f})$, $\mathbf{Q_j}.(\mathbf{f})$, $\mathbf{P_k}.(\mathbf{b})$, and the fact that

$$\phi([\mathbf{p}_0^{\text{in}}]_i, \ldots, [\mathbf{p}_k^{\text{in}}]_i, [\mathbf{p}_0^{\text{out}}]_i, \ldots, [\mathbf{p}_k^{\text{out}}]_i, [\mathbf{w}_p^{\text{in}}]_i) = \left(1 + \|([\mathbf{p}_k]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right)\left(1 + \|([\mathbf{p}_j]_i, [\mathbf{w}_p^{\text{in}}]_i)\|\right)$$

is pseudo-Lipschitz of order 2 by Lemma 8. The limiting terms which are inversely related to the rate of concentration can be bounded independently of N by the same argument given for the last two terms.

$\mathbf{P_k}.(\mathbf{g}).(\mathbf{i})$

We want to show that $\frac{1}{N}\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{vk}}]^T\mathbf{p}_k^{\text{in}}\|^2$ concentrates around ρ_{pk}^{in}. To this end, observe

$$\begin{aligned}\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{vk}}]^T\mathbf{p}_k^{\text{in}}\|^2 &= [\mathbf{p}_k^{\text{in}}]^T\mathbf{p}_k^{\text{in}} - [\mathbf{p}_k^{\text{in}}]^T\mathbf{B}_{\mathbf{C}^{\text{in}}_{vk}}[\mathbf{B}_{\mathbf{C}^{\text{in}}_{vk}}]^T\mathbf{p}_k^{\text{in}} \\ &= [\mathbf{u}_k^{\text{in}}]^T\mathbf{u}_k^{\text{in}} - [\mathbf{p}_k^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}}([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}[\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{p}_k^{\text{in}}, \end{aligned} \tag{A.107}$$

where the second equality follows because $\mathbf{p}_k^{\text{in}} = \mathbf{V}\mathbf{u}_k^{\text{in}}$ along with the fact that $\mathbf{B}_{\mathbf{C}^{\text{in}}_{vk}}[\mathbf{B}_{\mathbf{C}^{\text{in}}_{vk}}]^T$ and $\mathbf{C}_{vk}^{\text{in}}([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}[\mathbf{C}_{vk}^{\text{in}}]^T$ are both the orthogonal projection onto range $\left(\mathbf{C}_{vk}^{\text{in}}\right)$.

We know that $\frac{1}{N}[\mathbf{u}_k^{\text{in}}]^T\mathbf{u}_k^{\text{in}}$ concentrates around $\left(\mathbf{\Sigma}_{uk}^{\text{in}}\right)_{k+1,k+1}$ by $\mathbf{Q_k}.(\mathbf{e})$. Next observe that with $\widetilde{\mathbf{C}}_k^{\text{in}} := ([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}$, from the definition of $\mathbf{C}_{vk}^{\text{in}}$ in (2.37), we have

$$\begin{aligned}&[\mathbf{p}_k^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}}([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}[\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{p}_k^{\text{in}} \\ &\quad= \sum_{i,j=0}^{k-1}([\mathbf{p}_k^{\text{in}}]^T\mathbf{p}_i^{\text{in}})([\mathbf{p}_k^{\text{in}}]^T\mathbf{p}_j^{\text{in}})[\widetilde{\mathbf{C}}_k^{\text{in}}]_{i+1,j+1} + \sum_{i,j=0}^{k-1}([\mathbf{p}_k^{\text{in}}]^T\mathbf{v}_i^{\text{in}})([\mathbf{p}_k^{\text{in}}]^T\mathbf{v}_j^{\text{in}})[\widetilde{\mathbf{C}}_k^{\text{in}}]_{i+1+k,j+1+k} \\ &\qquad+ 2\sum_{i,j=0}^{k-1}([\mathbf{p}_k^{\text{in}}]^T\mathbf{p}_i^{\text{in}})([\mathbf{p}_k^{\text{in}}]^T\mathbf{v}_j^{\text{in}})[\widetilde{\mathbf{C}}_k^{\text{in}}]_{i+1+k,j+1}. \end{aligned} \tag{A.108}$$

Let these terms be labeled $T_1 - T_3$. For T_1, observe that because $\mathbf{p}_k^{\text{in}} = \mathbf{V}\mathbf{u}_k^{\text{in}}$,

$$([\mathbf{p}_k^{\text{in}}]^T\mathbf{p}_i^{\text{in}})([\mathbf{p}_k^{\text{in}}]^T\mathbf{p}_j^{\text{in}}) = ([\mathbf{u}_k^{\text{in}}]^T\mathbf{u}_i^{\text{in}})([\mathbf{u}_k^{\text{in}}]^T\mathbf{u}_j^{\text{in}}), \tag{A.109}$$

which concentrates around $[\mathbf{\Sigma}^{\text{in}}_{uk}]_{i+1,k+1}[\mathbf{\Sigma}^{\text{in}}_{uk}]_{j+1,k+1} = \mathbb{E}\left[U^{\text{in}}_i U^{\text{in}}_k\right]\mathbb{E}\left[U^{\text{in}}_j U^{\text{in}}_k\right] = [\mathbf{b}^{\text{in}}_{uk}]_{i+1}[\mathbf{b}^{\text{in}}_{uk}]_{j+1}$ by $\mathbf{Q_k}.(\mathbf{e})$ and Lemma 10, where the equality follows from (2.70) and (2.71). Furthermore, the entries of $\mathbf{\Sigma}^{\text{in}}_{uk}$ are bounded above independently of N, yielding the desired rate of concentration.

Furthermore, by $\mathbf{P_{k-1}}.(\mathbf{g}).(\mathbf{iv})$, $[\widetilde{\mathbf{C}}^{\text{in}}_k]_{i,j} = [([\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{C}^{\text{in}}_{vk})^{-1}]_{i,j}$ concentrates around $[\mathbf{\Pi}^{\text{in}}_{v(k-1)}]_{ij} = [(\mathbf{\Sigma}^{\text{in}}_{u(k-1)})^{-1}]_{i,j}$ (by (2.73)). Thus, again using Lemma 10 and Lemma 9, we get that T_1 concentrates around

$$\sum_{i,j=0}^{k-1}[\mathbf{b}^{\text{in}}_{uk}]_{i+1}[\mathbf{b}^{\text{in}}_{uk}]_{j+1}[(\mathbf{\Sigma}^{\text{in}}_{u(k-1)})^{-1}]_{i+1,j+1} = [\mathbf{b}^{\text{in}}_{uk}]^T[\mathbf{\Sigma}^{\text{in}}_{u(k-1)}]^{-1}\mathbf{b}^{\text{in}}_{uk},$$

and the components in each of the factors on the right-hand side are again bounded above independently of N.

For T_2, we proceed similarly, noting that $([\mathbf{p}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_i)([\mathbf{p}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_j)$ concentrates around 0 by $\mathbf{P_k}.(\mathbf{f})$ and Lemma 10. Because $[\widetilde{\mathbf{C}}^{\text{in}}_k]_{i+1+k,j+1+k}$ concentrates around $[(\mathbf{\Sigma}^{\text{in}}_{v(k-1)})^{-1}]_{i,j}$ by $\mathbf{P_{k-1}}.(\mathbf{g}).(\mathbf{iv})$, by Lemma 11 and 9 we find that T_2 concentrates around 0. By analogous reasoning, T_3 also concentrates around 0.

Thus, combining the above with Lemma 9 and using (2.72), we find $\|\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{vk}}\mathbf{p}^{\text{in}}_k\|^2$ concentrates on

$$\left[\mathbf{\Sigma}^{\text{in}}_{uk}\right]_{k+1,k+1} - [\mathbf{b}^{\text{in}}_{uk}]^T[\mathbf{\Sigma}^{\text{in}}_{u(k-1)}]^{-1}[\mathbf{b}^{\text{in}}_{uk}] = \rho^{\text{in}}_{pk}.$$

$\mathbf{P_k}.(\mathbf{g}).(\mathbf{ii})$ First we establish that $[\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk}$ is invertible with high probability. Before we can do this, we make a minor technical adjustment to simplify notation. First recall from (2.37) that $\mathbf{C}^{\text{in}}_{pk} = [\mathbf{P}^{\text{in}}_k\ \mathbf{V}^{\text{in}}_{k-1}]$. Next, define the $(2k-1)\times(2k-1)$ matrix $\mathbf{F} = [\mathbf{e}_k\ \mathbf{e}_1\ \mathbf{e}_2\ \cdots\ \mathbf{e}_{k-1}\ \mathbf{e}_{k+1}\ \cdots\ \mathbf{e}_{2k-1}]$. Observe that multiplying a matrix on the right by $\mathbf{F}$ has the effect of making the k^{th} column the first column. Also observe that $\mathbf{F}$ has an inverse, we call this $\mathbf{G}$ and note that $\mathbf{FG} = \mathbf{GF} = \mathbb{I}$. Now with this notation, let $\overline{\mathbf{C}}^{\text{in}}_{pk} := \mathbf{C}^{\text{in}}_{pk}\mathbf{F} = \left[\mathbf{p}^{\text{in}}_k\ \mathbf{C}^{\text{in}}_{vk}\right]$ where $\mathbf{C}^{\text{in}}_{vk} = [\mathbf{P}^{\text{in}}_{k-1}\ \mathbf{V}^{\text{in}}_{k-1}]$. Furthermore, we see

$$\mathbf{G}^T\left([\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk}\right)\mathbf{G} = \mathbf{G}^T\left([\mathbf{C}^{\text{in}}_{pk}\mathbf{F}]^T\mathbf{C}^{\text{in}}_{pk}\mathbf{F}\right)\mathbf{G} = [\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk}, \tag{A.110}$$

which implies that if $[\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk}$ is invertible, then so is $[\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk}$. Symmetrically, we see that $[\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk}$ is invertible if and only if $[\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk}$ is invertible. Moreover, we have from (A.110) that when $[\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk}$ is invertible,

$$([\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk})^{-1} = [\mathbf{G}^T\left([\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk}\right)\mathbf{G}]^{-1} = \mathbf{F}([\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk})^{-1}\mathbf{F}^T.$$

Now, the right hand side of the above differs from $([\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk})^{-1}$ by rearranging the columns and rows. In particular, the right hand side has the same entries as $([\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk})^{-1}$ up to rearrangement. Therefore it suffices to show that the entries of $([\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk})^{-1}$ concentrate.

To this end, we observe that

$$[\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk} = \begin{bmatrix} [\mathbf{p}^{\text{in}}_k]^T\mathbf{p}^{\text{in}}_k & [\mathbf{p}^{\text{in}}_k]^T\mathbf{C}^{\text{in}}_{vk} \\ [\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{p}^{\text{in}}_k & [\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{C}^{\text{in}}_{vk} \end{bmatrix}. \tag{A.111}$$

Now, by (A.107) in $\mathbf{P_k}.(\mathbf{g}).(\mathbf{i})$, we have that

$$\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{vk}}]^T\mathbf{p}^{\text{in}}_k\|^2 = [\mathbf{u}^{\text{in}}_k]^T\mathbf{u}^{\text{in}}_k - [\mathbf{p}^{\text{in}}_k]^T\mathbf{C}^{\text{in}}_{vk}([\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{C}^{\text{in}}_{vk})^{-1}[\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{p}^{\text{in}}_k.$$

Define $\boldsymbol{\nu} = ([\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{C}^{\text{in}}_{vk})^{-1}[\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{p}^{\text{in}}_k$. Using (A.111), the block matrix inversion formula gives us that, if $[\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{C}^{\text{in}}_{vk}$ is invertible and $\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{vk}}]^T\mathbf{p}^{\text{in}}_k\| \neq 0$, then

$$\left([\overline{\mathbf{C}}^{\text{in}}_{pk}]^T\overline{\mathbf{C}}^{\text{in}}_{pk}\right)^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & ([\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{C}^{\text{in}}_{vk})^{-1} \end{bmatrix} + \frac{1}{\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{vk}}]^T\mathbf{p}^{\text{in}}_k\|^2}\begin{bmatrix} 1 & -\boldsymbol{\nu}^T \\ -\boldsymbol{\nu} & \boldsymbol{\nu}\boldsymbol{\nu}^T \end{bmatrix}. \tag{A.112}$$

Now, by $\mathbf{P_k}.(\mathbf{g}).(\mathbf{i})$, we know that $\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{vk}}]^T\mathbf{p}^{\text{in}}_k\|^2$ concentrates around $\rho^{\text{in}}_{pk} > \epsilon^*_1 > 0$ (where the

first inequality is due to the stopping criterion). Then, we have

$$\begin{aligned}\mathbb{P}\left([\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}} \text{ singular}\right) &= \mathbb{P}\left([\overline{\mathbf{C}}_{pk}^{\text{in}}]^T\overline{\mathbf{C}}_{pk}^{\text{in}} \text{ singular}\right) \\ &\leq \mathbb{P}\left([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}} \text{ singular}\right) + \mathbb{P}\left(\left|\frac{1}{N}\|[\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp}]^T\mathbf{p}_k^{\text{in}}\|^2 - \rho_{pk}^{\text{in}}\right| \geq \frac{\epsilon_1^*}{2}\right).\end{aligned} \tag{A.113}$$

The first term concentrates by $\mathbf{P_{k-1}}.(\mathbf{g}).(\mathbf{iv})$, and the second concentrates by $\mathbf{P_k}.(\mathbf{g}).(\mathbf{i})$.

Next, we show entrywise concentration of $([\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}})^{-1}$ around $\mathbf{\Pi}_{uk}^{\text{in}} \doteq \begin{bmatrix}[\mathbf{\Sigma}_{uk}^{\text{in}}]^{-1} & 0 \\ 0 & [\mathbf{\Sigma}_{v(k-1)}^{\text{in}}]^{-1}\end{bmatrix}$, where the definition is originally given in (2.73). First, define two events $\mathcal{F}_k^1 := \left\{[\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}} \text{ singular}\right\}$ and $\mathcal{F}_{k-1}^2 := \left\{[\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}} \text{ singular}\right\}$. Then, notice that

$$\begin{aligned}&\mathbb{P}\left(\left|\left[([\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}})^{-1}\right]_{ij} - \left[\mathbf{\Pi}_{uk}^{\text{in}}\right]_{ij}\right| \geq \epsilon\right) \\ &\qquad \leq \mathbb{P}(\mathcal{F}_k^1) + \mathbb{P}(\mathcal{F}_{k-1}^2) + \mathbb{P}\left(\left|\left[([\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}})^{-1}\right]_{ij} - \left[\mathbf{\Pi}_{uk}^{\text{in}}\right]_{ij}\right| \geq \epsilon \;\middle|\; (\mathcal{F}_k^1)^c \cap (\mathcal{F}_{k-1}^2)^c\right),\end{aligned} \tag{A.114}$$

where $(\mathcal{F}_k^1)^c$ and $(\mathcal{F}_{k-1}^2)^c$ denote the complimentary events. We mention that the first two terms on the right side of (A.114) concentrate by (A.113) and $\mathbf{P_{k-1}}.(\mathbf{g}).(\mathbf{iv})$. Moving forward, we focus on the third term on the right side of (A.114).

As noted above, we can reorder the columns, so it suffices to replace $[\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}}$ by $[\overline{\mathbf{C}}_{pk}^{\text{in}}]^T\overline{\mathbf{C}}_{pk}^{\text{in}}$ and $\mathbf{\Pi}_{pk}^{\text{in}}$ by $\mathbf{F}\mathbf{\Pi}_{pk}^{\text{in}}\mathbf{F}^T$. Notice that $\mathbf{F}A^{-1}\mathbf{F}^T = (\mathbf{F}A\mathbf{F}^T)^{-1}$. Then, by the block matrix inversion

formula and the fact that $\mathbb{E}\left[P_k^{\text{in/out}}\right]^2 = \mathbb{E}\left[U_k^{\text{in/out}}\right]^2$ we find

$$\mathbf{F}\mathbf{\Pi}_{uk}^{\text{in}}\mathbf{F}^T = \left(\mathbf{F}\begin{bmatrix}\mathbf{\Sigma}_{uk}^{\text{in}} & 0\\ 0 & \mathbf{\Sigma}_{v(k-1)}^{\text{in}}\end{bmatrix}\mathbf{F}^T\right)^{-1} = \begin{bmatrix}\mathbb{E}\{[P_k^{\text{in}}]^2\} & \mathbf{b}_{pk}^{\text{in}} & 0\\ (\mathbf{b}_{pk}^{\text{in}})^T & \mathbf{\Sigma}_{u(k-1)}^{\text{in}} & 0\\ 0 & 0 & \mathbf{\Sigma}_{v(k-1)}^{\text{in}}\end{bmatrix}^{-1}$$
$$= \begin{bmatrix}0 & 0\\ 0 & \mathbf{\Pi}_{v(k-1)}^{\text{in}}\end{bmatrix} + \frac{1}{\rho_{pk}^{\text{in}}}\begin{bmatrix}1 & -[\boldsymbol{\beta}_{pk}^{\text{in}}\ 0]\\ -\begin{bmatrix}\boldsymbol{\beta}_{pk}^{\text{in}}\\ 0\end{bmatrix} & \begin{bmatrix}\boldsymbol{\beta}_{pk}^{\text{in}}\\ 0\end{bmatrix}[\boldsymbol{\beta}_{pk}^{\text{in}}\ 0]\end{bmatrix}. \tag{A.115}$$

In the above, we have used that $\rho_{pk}^{\text{in}} = \mathbb{E}\{[U_k^{\text{in}}]^2\} - (\mathbf{b}_{uk}^{\text{in}})^T\left[\mathbf{\Sigma}_{u(k-1)}^{\text{in}}\right]^{-1}\mathbf{b}_{uk}^{\text{in}}$ and $\boldsymbol{\beta}_{pk}^{\text{in}} = \left[\mathbf{\Sigma}_{u(k-1)}^{\text{in}}\right]^{-1}\mathbf{b}_{uk}^{\text{in}}$ by (2.72). Considering (A.114) and using Lemmas 9 and 10, it suffices to show that $([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}$ concentrates on $\mathbf{\Pi}_{v(k-1)}^{\text{in}}$, $\|[\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp}]^T\mathbf{p}_k^{\text{in}}\|^{-2}$ concentrates on $[\rho_{pk}^{\text{in}}]^{-1}$, and $\boldsymbol{\nu}$ concentrates on $[\boldsymbol{\beta}_{pk}^{\text{in}}\ 0]^T$. But $\mathbf{P_{k-1}}.(\mathbf{g}).(\mathbf{iv})$ gives us concentration for $([\mathbf{C}_{vk}^{\text{in}}]^T\mathbf{C}_{vk}^{\text{in}})^{-1}$, and $\mathbf{P_k}.(\mathbf{g}).(\mathbf{i})$ along with Lemma 13 gives concentration for $\|\mathbf{B}_{\mathbf{C}_{vk}^{\text{in}}}^{\perp}\mathbf{p}_k^{\text{in}}\|^{-2}$. Finally, observe that if $0 \le i \le k-1$, we have

$$\left|[\boldsymbol{\nu}]_i - [[\boldsymbol{\beta}_{pk}^{\text{in}}]^T\ 0]_i\right| \le \left|\sum_{j=0}^{k-1}[\widetilde{\mathbf{C}}_{vk}^{\text{in}}]_{ij}([\mathbf{p}_j^{\text{in}}]^T\mathbf{p}_k^{\text{in}}) - [\boldsymbol{\beta}_{pk}^{\text{in}}]_j\right| + \left|\sum_{j=0}^{k-1}[\widetilde{\mathbf{C}}_{vk}^{\text{in}}]_{i,j+(k-1)}([\mathbf{v}_i^{\text{in}}]^T\mathbf{p}_k^{\text{in}})\right|. \tag{A.116}$$

By $\mathbf{P_{k-1}}.(\mathbf{g}).(\mathbf{iv})$, we get concentration for the values of $\widetilde{\mathbf{C}}_{vk}^{\text{in}}$. In particular, if $0 \le i, j \le k-1$, then we have $[\widetilde{\mathbf{C}}_{vk}^{\text{in}}]_{i,j}$ concentrates around $[\mathbf{\Sigma}_{u(k-1)}^{\text{in}}]_{ij}^{-1}$. Moreover, by $\mathbf{P_k}.(\mathbf{d})$ and $\mathbf{P_k}.(\mathbf{f})$, we have that $[\mathbf{p}_j^{\text{in}}]^T\mathbf{p}_k^{\text{in}}$ concentrates around $[\mathbf{b}_{pk}^{\text{in}}]_j$ and $[\mathbf{v}_i^{\text{in}}]^T\mathbf{p}_k^{\text{in}}$ concentrates around 0. Thus by Lemmas 10 and 9, we get that the right hand term above concentrates around 0 and the left hand term concentrates around

$$\sum_{j=0}^{k-1}[\mathbf{\Sigma}_{u(k-1)}^{\text{in}}]_{ij}^{-1}[\mathbf{b}_{pk}^{\text{in}}]_j = \left[\left[\mathbf{\Sigma}_{u(k-1)}^{\text{in}}\right]^{-1}\mathbf{b}_{uk}^{\text{in}}\right]_i = [\boldsymbol{\beta}_{pk}^{\text{in}}]_i.$$

The same analysis applies if $k \le i \le 2(k-1)$ except that now for $0 \le i \le k-1$, we have

$[\widetilde{\mathbf{C}}^{\text{in}}_{vk}]_{ij}$ concentrates around 0 (again by $\mathbf{P_{k-1}}.(\mathbf{g}).(\mathbf{iv})$), which gives us that $\boldsymbol{\nu}_i$ concentrates around 0, as desired. Thus, we get that the entries of $([\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk})^{-1}$ concentrate around those of $\mathbf{\Pi}^{\text{in}}_{pk}$, as needed. Furthermore, all of the limiting quantities that affect the rate of concentration are bounded independently of N. We also need in this case that ρ^{in}_{pk} is bounded away from 0 (independently of N), but this is guaranteed by our stopping criterion.

$\mathbf{P_k}.(\mathbf{g}).(\mathbf{iii})$ We will show that $\frac{1}{N}\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{pk}}]^T\mathbf{v}^{\text{in}}_k\|^2$ concentrates to ρ^{in}_{qk}. To this end, observe that

$$\|[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{pk}}]^T\mathbf{v}^{\text{in}}_k\|^2 = [\mathbf{v}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_k - [\mathbf{v}^{\text{in}}_k]^T\mathbf{B}_{\mathbf{C}^{\text{in}}_{pk}}[\mathbf{B}_{\mathbf{C}^{\text{in}}_{pk}}]^T\mathbf{v}^{\text{in}}_k = [\mathbf{v}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_k - [\mathbf{v}^{\text{in}}_k]^T\mathbf{C}^{\text{in}}_{pk}([\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk})^{-1}[\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{v}^{\text{in}}_k,$$

where the second equality follows because $\mathbf{B}_{\mathbf{C}^{\text{in}}_{pk}}[\mathbf{B}_{\mathbf{C}^{\text{in}}_{pk}}]^T T$ and $\mathbf{C}^{\text{in}}_{pk}([\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk})^{-1}[\mathbf{C}^{\text{in}}_{pk}]^T$ are both the orthogonal projection onto range $\left(\mathbf{C}^{\text{in}}_{pk}\right)$.

First observe that $\frac{1}{N}[\mathbf{v}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_k$ concentrates around $[\mathbf{\Sigma}^{\text{in}}_{vk}]_{k+1,k+1}$ by $\mathbf{P_k}.(\mathbf{e})$. Next observe that defining $\widetilde{\mathbf{C}}^{\text{in}}_{pk} = ([\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk})^{-1}$ and using from the definition of $\mathbf{C}^{\text{in}}_{pk}$ in (2.37), we have

$$\begin{aligned}
&[\mathbf{v}^{\text{in}}_k]^T\mathbf{C}^{\text{in}}_{pk}([\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{C}^{\text{in}}_{pk})^{-1}[\mathbf{C}^{\text{in}}_{pk}]^T\mathbf{v}^{\text{in}}_k \\
&\quad= \sum_{i,j=0}^{k}([\mathbf{v}^{\text{in}}_k]^T\mathbf{p}^{\text{in}}_i)([\mathbf{v}^{\text{in}}_k]^T\mathbf{p}^{\text{in}}_j)[\widetilde{\mathbf{C}}^{\text{in}}_{pk}]_{i,j} + \sum_{i,j=0}^{k-1}([\mathbf{v}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_i)([\mathbf{v}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_j)[\widetilde{\mathbf{C}}^{\text{in}}_{pk}]_{i+1+k,j+1+k} \\
&\qquad\qquad + 2\sum_{i=0}^{k}\sum_{j=0}^{k-1}([\mathbf{v}^{\text{in}}_k]^T\mathbf{p}^{\text{in}}_i)([\mathbf{v}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_j)[\widetilde{\mathbf{C}}^{\text{in}}_{pk}]_{i+1+k,j}.
\end{aligned}$$

Let the three terms of (A.117) be labeled $T_1 - T_3$. For T_1, observe that $([\mathbf{v}^{\text{in}}_k]^T\mathbf{p}^{\text{in}}_i)([\mathbf{v}^{\text{in}}_k]^T\mathbf{p}^{\text{in}}_j)$ concentrates around 0 for $0 \le i, j \le k$ by $\mathbf{P_k}.(\mathbf{f})$. Because $[\widetilde{\mathbf{C}}^{\text{in}}_{pk}]_{i,j}$ concentrates around $[\mathbf{\Pi}^{\text{in}}_{uk}]_{i,j} = [(\mathbf{\Sigma}^{\text{in}}_{pk})^{-1}]_{i,j}$ when $0 \le i, j \le k$ by $\mathbf{P_k}.(\mathbf{g}).(\mathbf{ii})$ and (2.73), we get by Lemma 10 and 9 that T_1 concentrates around 0.

For T_2, observe that $([\mathbf{v}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_i)([\mathbf{v}^{\text{in}}_k]^T\mathbf{v}^{\text{in}}_j)$ concentrates around $[\mathbf{\Sigma}^{\text{in}}_{vk}]_{i+1,k+1}[\mathbf{\Sigma}^{\text{in}}_{vk}]_{j+1,k+1} = \mathbb{E}\left[V^{\text{in}}_i V^{\text{in}}_k\right]\mathbb{E}\left[V^{\text{in}}_j V^{\text{in}}_k\right]$ $[\mathbf{b}^{\text{in}}_{vk}]_i[\mathbf{b}^{\text{in}}_{vk}]_j$ by $\mathbf{P_k}.(\mathbf{e})$ and Lemma 10, where the equality follows from (2.70) and (2.71). Furthermore, $[\widetilde{\mathbf{C}}^{\text{in}}_{pk}]_{i+1+k,j+1+k}$ concentrates around $[\mathbf{\Pi}^{\text{in}}_{uk}]_{i+1+k,j+1+k} = [(\mathbf{\Sigma}^{\text{in}}_{v(k-1)})^{-1}]_{i,j}$ when $0 \le i, j \le k$ by $\mathbf{P_k}.(\mathbf{g}).(\mathbf{ii})$ and (2.73). Thus, again by Lemma 11 and Lemma 9, we get that T_2 concentrates

around $\sum_{i,j=0}^{k-1}[\mathbf{b}_{vk}^{\text{in}}]_i[\mathbf{b}_{vk}^{\text{in}}]_j[(\mathbf{\Sigma}_{v(k-1)}^{\text{in}})^{-1}]_{i,j} = [\mathbf{b}_{vk}^{\text{in}}]^T[\mathbf{\Sigma}_{v(k-1)}^{\text{in}}]^{-1}\mathbf{b}_{vk}^{\text{in}}$. Using analogous reasoning, T_3 also concentrates around 0.

Combining the above with Lemma 9 and using (2.72), we get that $\frac{1}{N}\|[\mathbf{B}_{\mathbf{C}_{pk}^{\text{in}}}^{\perp}]^T\mathbf{v}_k^{\text{in}}\|^2$ concentrates around the desired constant: $\left[\mathbf{\Sigma}_{vk}^{\text{in}}\right]_{k+1,k+1} - [\mathbf{b}_{vk}^{\text{in}}]^T[\mathbf{\Sigma}_{v(k-1)}^{\text{in}}]^{-1}\mathbf{b}_{vk}^{\text{in}} = \rho_{qk}^{\text{in}}$.

$\mathbf{P_k}.(\mathbf{g}).(\mathbf{iv})$ We want to show that $[\mathbf{C}_{v(k+1)}^{\text{in}}]^T\mathbf{C}_{v(k+1)}^{\text{in}}$ is invertible with high probability and that, when invertible, the entries of the inverse concentrate around

$$\mathbf{\Pi}_{vk}^{\text{in}} = \begin{bmatrix} [\mathbf{\Sigma}_{uk}^{\text{in}}]^{-1} & 0 \\ 0 & [\mathbf{\Sigma}_{vk}^{\text{in}}]^{-1} \end{bmatrix}.$$

This can be proved in much the same way as $\mathbf{P_k}.(\mathbf{g}).(\mathbf{ii})$. In particular, note that $\mathbf{C}_{v(k+1)}^{\text{in}} = [\mathbf{C}_{pk}^{\text{in}}\mathbf{v}_k^{\text{in}}]$. Then applying the block inversion formula to $[\mathbf{C}_{v(k+1)}^{\text{in}}]^T\mathbf{C}_{v(k+1)}^{\text{in}}$, we see that it is invertible if $[\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}}$ is invertible and $\|[\mathbf{B}_{\mathbf{C}_{pk}^{\text{in}}}^{\perp}]^T\mathbf{v}_k^{\text{in}}\|^2 \neq 0$. But the probabilities of both of these events concentrate by $\mathbf{P_k}.(\mathbf{g}).(\mathbf{ii})$ and $\mathbf{P_k}.(\mathbf{g}).(\mathbf{iii})$ respectively.

Now, analogously to $\mathbf{P_k}.(\mathbf{g}).(\mathbf{ii})$, to get concentration for the entries of $[\mathbf{C}_{v(k+1)}^{\text{in}}]^T\mathbf{C}_{v(k+1)}^{\text{in}}$, it suffices to get concentration for the entries of $([\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}})^{-1}$, $([\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{C}_{pk}^{\text{in}})^{-1}[\mathbf{C}_{pk}^{\text{in}}]^T\mathbf{v}_k^{\text{in}}$, and of $\|[\mathbf{B}_{\mathbf{C}_{pk}^{\text{in}}}^{\perp}]^T\mathbf{v}_k^{\text{in}}\|^{-2}$. For these, we can use $\mathbf{P_k}.(\mathbf{g}).(\mathbf{ii})$ and $\mathbf{P_k}.(\mathbf{g}).(\mathbf{iii})$ and apply the same arguments as above. This gives concentration for the entries of $[\mathbf{C}_{v(k+1)}^{\text{in}}]^T\mathbf{C}_{v(k+1)}^{\text{in}}$ around those of $\mathbf{\Pi}_{vk}^{\text{in}}$.

A.1.4 Showing $\mathbf{Q_{k+1}}$ holds

The arguments for this case are analogous to those in $\mathbf{P_k}$, so we omit them here.

A.2 Useful lemmas

Many of the lemmas used in this section were originally given in [81]. We do not include the proofs of these lemmas here, and instead refer the reader to [81] for the proofs.

Lemma 8 (Products of Lipschitz Functions)**.** *If $f, g : \mathbb{R}^p \to \mathbb{R}$ are Lipschitz continuous, then $h(x) = f(x)g(x) : \mathbb{R}^p \to \mathbb{R}$ is pseudo-Lipschitz of order 2.*

Lemma 9 (Concentration of Sums)**.** *If random variables $X_1, \ldots, X_M$ satisfy $P(|X_i| \geq \epsilon) \leq e^{-n\kappa_i\epsilon^2}$ for $1 \leq i \leq M$, then*

$$P\Big(|\sum_{i=1}^{M} X_i| \geq \epsilon\Big) \leq \sum_{i=1}^{M} P\left(|X_i| \geq \frac{\epsilon}{M}\right) \leq Me^{-n(\min_i \kappa_i)\epsilon^2/M^2}.$$

Lemma 10 (Concentration of Products)**.** *For random variables X, Y and non-zero constants c_X, c_Y, if $P(|X-c_X| \geq \epsilon) \leq Ke^{-\kappa n\epsilon^2}$, and $P(|Y-c_Y| \geq \epsilon) \leq Ke^{-\kappa n\epsilon^2}$, then the probability $P(|XY-c_Xc_Y| \geq \epsilon)$ is bounded by*

$$P\Big(|X - c_X| \geq \min\left(\sqrt{\frac{\epsilon}{3}}, \frac{\epsilon}{3c_Y}\right)\Big) + P\Big(|Y - c_Y| \geq \min\left(\sqrt{\frac{\epsilon}{3}}, \frac{\epsilon}{3c_X}\right)\Big) \leq 2Ke^{-\frac{\kappa n\epsilon^2}{9\max(1, c_X^2, c_Y^2)}}.$$

Lemma 11 (Concentration of Products)**.** *For random variables X, Y, and constant c_X, if*

$$P(|X - c_X| \geq \epsilon) \leq Ke^{-r_X\epsilon^2} \text{ and } P(|Y| \geq \epsilon) \leq Ke^{-r_Y\epsilon^2},$$

then if $c_X \neq 0$, we have

$$P(|XY| \geq \sqrt{\epsilon}) \leq P(|X - c_X| \geq \sqrt{\epsilon}) + P\Big(|Y| \geq \frac{\epsilon}{2\max\{1, |c_X|\}}\Big) \leq 2K\exp\Big\{-\frac{\epsilon^2\min\{r_Y, r_X\}}{4\max\{1, c_X^2\}}\Big\},$$

and if $c_X = 0$ and $\epsilon < 1$, we have $P(|XY| \geq \epsilon) \leq 2K\exp\Big\{-\epsilon^2\min\{r_Y, r_X\}\Big\}$.

Lemma 12 (Concentration of Square Roots)**.** *Let $c \neq 0$. If $P(|X_n^2 - c^2| \geq \epsilon) \leq e^{-\kappa n\epsilon^2}$, then $P(||X_n| - |c|| \geq \epsilon) \leq e^{-\kappa n|c|^2\epsilon^2}$.*

Lemma 13 (Concentration of Scalar Inverses)**.** *Assume $c \neq 0$ and $0 < \epsilon < 1$. If*

$$P(|X_n - c| \geq \epsilon) \leq e^{-\kappa n\epsilon^2}, \text{ then } P(|X_n^{-1} - c^{-1}| \geq \epsilon) \leq 2e^{-n\kappa\epsilon^2c^2\min\{c^2,1\}/4}.$$

Lemma 14. *For a random variable $Z \sim \mathcal{N}(0, 1)$ and $\epsilon > 0$, $P\Big(|Z| \geq \epsilon\Big) \leq 2e^{-\frac{1}{2}\epsilon^2}$.*

Lemma 15 (χ^2-concentration)**.** *For Z_i, $i \in [n]$ that are i.i.d. $\sim \mathcal{N}(0,1)$, and $0 \leq \epsilon \leq 1$,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - 1\right| \geq \epsilon\right) \leq 2e^{-n\epsilon^2/8}.$$

Lemma 16. *For any scalars $a_1, ..., a_t$ and positive integer m, we have $(|a_1| + \ldots + |a_t|)^m \leq t^{m-1}\sum_{i=1}^{t}|a_i|^m$. Consequently, for any vectors $\mathbf{u}_1, \ldots, \mathbf{u}_t \in \mathbb{R}^N$, $\|\sum_{k=1}^{t}\mathbf{u}_k\|^2 \leq t\sum_{k=1}^{t}\|\mathbf{u}_k\|^2$.*

Lemma 17. *Let $Z_1, \ldots, Z_t \in \mathbb{R}^N$ be random vectors such that $(Z_{1,i}, \ldots, Z_{t,i})$ are i.i.d. across $i \in [n]$, with $(Z_{1,i}, \ldots, Z_{t,i})$ being jointly Gaussian with zero mean, unit variance and covariance matrix $K \in \mathbb{R}^{t\times t}$. Let $G \in \mathbb{R}^N$ be a random vector with entries $G_1, \ldots, G_N$ i.i.d. $\sim p_G$, where p_G is sub-Gaussian with variance factor ν. Then for any sequence of pseudo-Lipschitz functions $\{f_i : \mathbb{R}^{t+1} \to \mathbb{R}\}_{i=1}^{N}$ with common pseudo-Lipschitz constant, non-negative constants $\sigma_1, \ldots, \sigma_t$, and $0 < \epsilon \leq 1$, we have*

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N} f(\sigma_1 Z_{1,i}, \ldots, \sigma_t Z_{t,i}, G_i) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[f_i(Z_{1,1}, \ldots, Z_{t,1}, G)]\right| \geq \epsilon\right)$$
$$\leq 2\exp\left\{\frac{-N\epsilon^2}{128L^2(t+1)^2(\nu + 4\nu^2 + \sum_{m=1}^{t}(\sigma_m^2 + 4\sigma_m^4))}\right\},$$

where $L > 0$ is an absolute constant. (L can be bounded above by three times the common pseudo-Lipschitz constant of the f_i.)

Proof. The statement of Lemma 17 is slightly more general than Lemma B.4 in [81] in that Lemma B.4 requires the functions f_i to all be equal. However, the proof of that lemma proves the above result exactly if we simply substitute the specific functions f_i for f throughout. □

Lemma 18. *Let $\phi : \mathbb{R}^d \times [a,b] \to \mathbb{R}$ be pseudo-Lipschitz of order 2 with constant L, and for $a \leq z_1, \ldots, z_k \leq b$ let functions $\psi_j : \mathbb{R}^d \to \mathbb{R}$ be given by*

$$\psi(\cdot) = \phi\left(\cdot, z_j\right),$$

for $1 \leq j \leq k$. *Then the* ψ_j *are pseudo-Lipschitz of order* 2 *with pseudo-Lipschitz constant* $L\,(1 + 2\max(|a|, |b|))$.

Proof. For $x, y \in \mathbb{R}^d$, we have that

$$\begin{aligned}
\left|\psi_j\,(x) - \psi_j\,(y)\right| &= \left|\phi(x, z_j) - \phi(y, z_j)\right| \\
&\leq L\left|(x, z_j) - (y, z_j)\right|\left[1 + \left|(x, z_j)\right| + \left|(y, z_j)\right|\right] \\
&\leq L\,|x - y|\left[1 + |x| + |y| + 2|z_j|\right] \\
&= L(1 + 2|z_j|)\,|x - y|\left[1 + \frac{|x| + |y|}{1 + 2|z_j|}\right] \\
&\leq L(1 + 2\max(|a|, |b|))\,|x - y|\,[1 + |x| + |y|]\,,
\end{aligned}$$

which completes the proof. □

A.3 Concentration is preserved by Haar matrices

Lemma 19. *Suppose that* $\mathbf{v}_N \in \mathbb{R}^N$ *is a (possibly random) vector-valued sequence such that*

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}\phi_i([\mathbf{v}_N]_i) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\phi_i(V)\Big| \geq \epsilon\Big) \leq C\exp\left(-cN\epsilon^2\right)$$

for any sequence $\{\phi_i\}_{i=1}^N \subset PL(2)$ *with a common pseudo-Lipschitz constant and some random variable* V *with finite second moment. If* $\mathbf{A}_N$ *is an* $N \times N$ *Haar-distributed orthogonal matrix independent of* $\mathbf{v}_N$ *for each* N, *using* $Z \sim N(0, \sigma^2)$ *for* $\sigma^2 = \lim_{N\to\infty}\|\mathbf{v}_N\|^2$, *then*

$$\mathbb{P}\Big(\frac{1}{N}\Big|\sum_{i=1}^{N}\phi_i([\mathbf{A}_N\mathbf{v}_N]_i) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\phi_i(Z)\Big| \geq \epsilon\Big) \leq C'\exp\left(-c'N\epsilon^2\right).$$

In particular, the above statement holds if $\mathbf{v}_N$ *is a deterministic sequence.*

Proof. As $\mathbf{A}_N$ is independent of $\mathbf{v}_N$ and $\mathbf{A}_N$ is Haar distributed, $\mathbf{B}\mathbf{A}_N\mathbf{v}_N \overset{d}{=} \mathbf{A}_N\mathbf{v}_N$ for any orthogonal matrix $\mathbf{B}$. Thus, the distribution of $\mathbf{A}_N\mathbf{v}_N$ is rotationally invariant and depends only on the

distribution of the norm $\|\mathbf{A}_N\mathbf{v}_N\| = \|\mathbf{v}_N\|$ Therefore, if $\mathbf{Z}_N \sim N(\mathbf{0}, \mathsf{I}_N)$ independent of $\mathbf{v}_N$,

$$\mathbf{A}_N\mathbf{v}_N \overset{d}{=} \frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|}\mathbf{Z}_N.$$

Then we find

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{n}\phi_i([\mathbf{A}_N\mathbf{v}_N]_i) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\phi_i(Z)\right| \overset{d}{=} \left|\frac{1}{N}\sum_{i=1}^{N}\phi_i\left(\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|}[\mathbf{Z}_N]_i\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\phi_i(Z)\right| \\
&= \left|\frac{1}{N}\sum_{i=1}^{N}\phi_i\left(\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|}[\mathbf{Z}_N]_i\right) - \frac{1}{N}\sum_{i=1}^{N}\phi_i\left(\sigma[\mathbf{Z}_N]_i\right)\right| + \left|\frac{1}{N}\sum_{i=1}^{N}\phi_i\left(\sigma[\mathbf{Z}_N]_i\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\phi_i(Z)\right|.
\end{aligned}$$

Now by Lemma 9, it suffices to show concentration for these two terms separately. For the first we use the triangle inequality, the pseudo-Lipschitz property of the ϕ_i, and Cauchy-Schwarz to get

$$\begin{aligned}
\left|\frac{1}{N}\sum_{i=1}^{N}\phi_i\left(\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|}[\mathbf{Z}_N]_i\right) - \frac{1}{N}\sum_{i=1}^{n}\phi_i\left(\sigma[\mathbf{Z}_N]_i\right)\right| &\leq \frac{1}{N}\sum_{i=1}^{N}\left|\phi_i\left(\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|}[\mathbf{Z}_N]_i\right) - \phi_i\left(\sigma[\mathbf{Z}_N]_i\right)\right| \\
&\leq \frac{L}{N}\sum_{i=1}^{N}\left|1 + \left(\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|} + \sigma\right)[\mathbf{Z}_N]_i\right|\left|\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|} - \sigma\right||[\mathbf{Z}_N]_i| \\
&\leq L\left|\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|} - \sigma\right|\left(\frac{\|\mathbf{Z}_N\|}{\sqrt{N}} + \left(\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|} + \sigma\right)\frac{\|\mathbf{Z}_N\|^2}{N}\right).
\end{aligned}$$

Now we know that $\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|}$ concentrates around σ using our assumptions on $\mathbf{v}_N$, Lemma 15, and Lemma 10. Then using this fact along with Lemmas 12, 10, and 15, we get that $\left(\frac{\|\mathbf{Z}_N\|}{\sqrt{N}} + \left(\frac{\|\mathbf{v}_N\|}{\|\mathbf{Z}_N\|} + \sigma\right)\frac{\|\mathbf{Z}_N\|^2}{N}\right)$ concentrates around a finite value (specifically $1 + 2\sigma$). Combining these with Lemma 10, we get that the the above concentrates with the desired rates (since each of our intermediate lemmas preserve this $N\epsilon^2$ rate). Finally concentration for

$$\left|\frac{1}{N}\sum_{i=1}^{N}\phi_i\left(\sigma[\mathbf{Z}_N]_i\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\phi_i(Z)\right|,$$

follows directly from Lemma 17 (again with the desired rate) since we assume that the ϕ_i share a common pseudo-Lipschitz constant. This completes the proof of the lemma. □

A.4 Concentration analysis for the matrix general recursion

In this appendix, we outline the adjustments that need to made to the main concentration arguments in order to get the analogous concentration result for the general recursion in Algorithm 3 with $d > 1$ (i.e. with matrix rather than vector iterates). The main differences lie in accounting for the distributions of the asymptotic variables $P_k^{\text{in/out}}$ and $Q_k^{\text{in/out}}$, which are now zero-mean Gaussian vectors rather than simply Gaussian random variables. As we will see, this adjustment follows directly from a straightforward extension of Lemma 3, which in turn relies on the following characterization of orthogonally invariant random matrices.

Lemma 20. *Suppose that* $\mathbf{x} = \mathbf{U}\mathbf{y} \in \mathbb{R}^{n\times d}$ *for* $\mathbf{U} \in \mathbb{R}^{n\times n}$ *a Haar distributed orthogonal matrix and* $\mathbf{y} \in \mathbb{R}^{n\times d}$ *an independent random matrix which has full rank with probability* 1. *Furthermore, let* $\mathbf{G}_\mathbf{x} = \mathbf{x}^T\mathbf{x} \in \mathbb{R}^{d\times d}$ *be the Gram matrix of* $\mathbf{x}$*, and let* $\mathbf{z} \in \mathbb{R}^{n\times d}$ *be a random matrix with independent standard Gaussian entries which is independent of* $\mathbf{U}$ *and* $\mathbf{x}$*. Then we have that*

$$\mathbf{x} \stackrel{d}{=} \mathbf{z}\mathbf{G}_\mathbf{z}^{-1/2}\mathbf{G}_\mathbf{y}^{1/2} \tag{A.117}$$

Proof. (Informal). The following proof is only informal because its claims rest on statements about probabilities of measure zero events. The difficulty of making these informal arguments more formal arises from the fact that the relevant random quantities take values in the space of $n \times d$ matrices with orthonormal columns. As this space is non-Euclidean, probability densities on this space do not satisfy the usual change of variables formula (since the normal definition of the Jacobian does not make sense). I believe this may be fixable using a change of variables formula from geometric measure theory that employs a suitable generalization of the Jacobian.

Let $\Phi(\mathbf{c}) = \mathbf{c}\mathbf{G}_\mathbf{c}^{-1/2}$ for any matrix $\mathbf{c}$. Then note that following facts:

1. $\Phi(\mathbf{x}) \in \mathbb{R}^{n\times d}$ has orthonormal columns.
2. $\mathbf{G}_{\mathbf{V}\mathbf{c}} = \mathbf{G}_\mathbf{c}$ for any orthogonal matrix $\mathbf{V}$.
3. $\mathbf{V}\mathbf{x} \stackrel{d}{=} \mathbf{x}$ for any fixed orthogonal matrix $\mathbf{V}$ since $\mathbf{V}\mathbf{U} \stackrel{d}{=} \mathbf{U}$ and $\mathbf{y}$ and $\mathbf{U}$ are independent.

Now let $\pi : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times d}$ be the map which simply drops the last $n - d$ columns of its input. Then let $\mathbf{V}' \in \mathbb{R}^{n\times d}$ be any matrix with orthonormal columns and let $\mathbf{V}$ be an orthogonal matrix such that $\pi(\mathbf{V}) = \mathbf{V}'$. Then by the above facts, we have

$$\mathbb{P}\left(\Phi(\mathbf{x}) = \mathbf{V}'\right) = \mathbb{P}\left(\Phi(\mathbf{V}\mathbf{x}) = \mathbf{V}'\right) = \mathbb{P}\left(\mathbf{V}\mathbf{x}\mathbf{G}_{\mathbf{x}}^{-1/2} = \mathbf{V}'\right) = \mathbb{P}\left(\mathbf{x}\mathbf{G}_{\mathbf{x}}^{-1/2}\pi(\mathbf{I})\right) = \mathbb{P}\left(\Phi(\mathbf{x}) = \rho(\mathbf{I})\right).$$

As this is true for any $\mathbf{V}'$, $\Phi(\mathbf{x})$ is uniform over $d \times n$ matrices with orthonormal columns.

Now since $\mathbf{z}$ has independent standard Gaussian entries, we have that $\mathbf{V}\mathbf{z} \overset{d}{=} \mathbf{z}$; thus,

$$\mathbb{P}\left(\mathbf{z}\mathbf{G}_{\mathbf{z}}^{-1/2} = \mathbf{V}'\right) = \mathbb{P}\left(\mathbf{V}\mathbf{z}\mathbf{G}_{\mathbf{V}\mathbf{z}}^{-1/2} = \mathbf{V}'\right) = \mathbb{P}\left(\mathbf{z}\mathbf{G}_{\mathbf{z}}^{-1/2} = \rho(\mathbf{I})\right).$$

This implies that $\mathbf{z}\mathbf{G}_{\mathbf{z}}^{-1/2} = \Phi(\mathbf{x})$. Finally, we note that the entirety of the above argument still goes through with $\mathbf{y}$ fixed, which implies that $\Phi(\mathbf{x})$ is uniformly distributed conditional of any particular of $\mathbf{y}$ and is thus independent of $\mathbf{y}$. Therefore, as claimed,

$$\mathbf{x} = \mathbf{x}\mathbf{G}_{\mathbf{x}}^{-1/2}\mathbf{G}_{\mathbf{x}}^{1/2} = \Phi(\mathbf{x})\mathbf{G}_{\mathbf{y}}^{1/2} \overset{d}{=} \mathbf{z}\mathbf{G}_{\mathbf{z}}^{-1/2}\mathbf{G}_{\mathbf{y}}^{1/2}.$$

□

Using this, we obtain a matrix version of Lemma 3. To state this lemma, first note the necessary adjustments to our notation. The matrices $\mathbf{C}_{qk}^{\text{in/out}}$, $\mathbf{C}_{pk}^{\text{in/out}}$, $\mathbf{C}_{uk}^{\text{in/out}}$, and $\mathbf{C}_{vk}^{\text{in/out}}$ are defined in exactly the same way as before, where now we understand the iterates (e.g. $\mathbf{p}_j^{\text{in/out}}$) as $(N/M)\times d$ matrices, so that, e.g., $\mathbf{C}_{pk}^{\text{in}}$ is now a $N\times d(2k+1)$ matrix. Likewise, the sigma algebras $\mathcal{P}_j$ and $\mathcal{Q}_j$ retain their definitions in terms of these matrix iterates. With $\mathbf{C}$ equal to any of these matrices, the definitions of the corresponding matrices $\mathbf{B}_{\mathbf{C}}$ and $\mathbf{B}_{\mathbf{C}}^{\perp}$ are unchanged. The Gaussian quantities $\mathbf{Z}_{pk}^{\text{in/out}}$, $\check{\mathbf{Z}}_{pk}^{\text{in/out}}$, and $\overline{\mathbf{Z}}_{pk}^{\text{in/out}}$ have i.i.d. standard Gaussian entries before, but are expanded to have d columns so that, e.g. $\overline{\mathbf{Z}}_{pk}^{\text{in/out}} \in \mathbb{R}^{(N/M)\times d}$. Finally, the scalar/vector parameters $\rho_{(p/q)k}^{\text{in/out}}$ and $\beta_{(p/q)k}^{\text{in/out}}$ are replaced by matrix parameters $\mathbf{P}_{(p/q)k}^{\text{in/out}} \in \mathbb{R}^{d\times d}$ and $\mathbf{B}_{(p/q)k}^{\text{in/out}} \in \mathbb{R}^{kd\times d}$ respectively. We introduce an additional indexing notation: for each $1 \le i \le k$ letting $[\mathbf{B}_{(p/q)k}^{\text{in/out}}]_{(i)} \in \mathbb{R}^{d\times d}$ be the $d \times d$ submatrix with

top-left entry $[\mathbf{B}^{\text{in/out}}_{(p/q)k}]_{d(i-1)+1,1}$. In other words, if we view $\mathbf{B}^{\text{in/out}}_{(p/q)k}$ as a k-dimensional vector with $d \times d$ matrix entries, then this is just accessing the index i component of this vector.

With this notation, we can now state our generalized lemma. To save space, we state this only for the **p** iterates, but the extension to the **q** iterates is completely symmetric.

Lemma 21. *If $[\mathbf{C}^{\text{in}}_{vk}]^T\mathbf{C}^{\text{in}}_{vk}$ has full rank for $0 \le k \le K$, then we have for all such k that*

$$\mathbf{p}^{\text{in}}_0|_{\mathcal{P}_0} \overset{d}{=} \mathbf{O}^{\text{in}}_{p0}\,\overline{\mathbf{Z}}^{\text{in}}_{p0}\,[\mathbf{P}^{\text{in}}_{p0}]^{1/2} + \mathbf{\Delta}^{\text{in}}_{p0}, \qquad \textit{and} \qquad \mathbf{p}^{\text{in}}_k|_{\mathcal{P}_k} \overset{d}{=} \sum_{\ell=0}^{k-1} \mathbf{p}^{\text{in}}_\ell\,[\mathbf{B}^{\text{in}}_{pk}]_{(\ell+1)} + \mathbf{O}^{\text{in}}_{pk}\,\overline{\mathbf{Z}}^{\text{in}}_{pk}\,[\mathbf{P}^{\text{in}}_{pk}]^{1/2} + \mathbf{\Delta}^{\text{in}}_{pk}, \tag{A.118}$$

where $\mathbf{O}^{\text{in}}_{pk}$ is defined in (2.39)*, and $\mathbf{\Delta}^{\text{in}}_{p0}$ is given in **Condition 0**, with*

$$\mathbf{\Delta}^{\text{in}}_{pk} = \mathbf{C}^{\text{in}}_{vk}\left(([\mathbf{C}^{\text{in}}_{qk}]^T\mathbf{C}^{\text{in}}_{qk})^{-1}[\mathbf{C}^{\text{in}}_{qk}]^T\mathbf{u}^{\text{in}}_k - \begin{bmatrix}\mathbf{B}^{\text{in}}_{pk}\\ \mathbf{0}_{dk\times d}\end{bmatrix}\right) + \mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{vk}}\mathbf{Z}^{\text{in}}_{pk}\left[\mathbf{G}^{-1/2}_{\mathbf{Z}^{\text{in}}_{pk}}\mathbf{G}^{1/2}_{[\mathbf{B}^{\perp}_{\mathbf{C}^{\text{in}}_{qk}}]^T\mathbf{u}^{\text{in}}_k} - [\mathbf{P}^{\text{in}}_{pk}]^{1/2}\right]$$
$$- \mathbf{B}_{\mathbf{C}^{\text{in}}_{vk}}\check{\mathbf{Z}}^{\text{in}}_{pk}[\mathbf{P}^{\text{in}}_{pk}]^{1/2}.$$

Furthermore, we have that $\mathbf{p}^{\text{in}}_k$ and $\mathbf{p}^{\text{out}}_k$ are conditionally independent given $\mathcal{P}_k$ for all $k \ge 0$.

Proof. The proof of this statement is essentially identical to that of Lemma 3, merely replacing applications of Lemma 1 with Lemma 20 and substituting our matrix parameters $\mathbf{P}^{\text{in/out}}_{(p/q)k} \in \mathbb{R}^{d\times d}$ and $\mathbf{B}^{\text{in/out}}_{(p/q)k} \in \mathbb{R}^{kd\times d}$ for $\rho^{\text{in/out}}_{(p/q)k}$ and $\beta^{\text{in/out}}_{(p/q)k}$. □

Next we state an analog of Lemma 4 that depends on a matrix version of the Gaussian equivalent recursion in Algorithm 5. We omit the explicit statement here, as it is defined exactly to make statement (2) in the following lemma true (hence, is entirely characterized by the lemma statement).

Lemma 22. *First define*

$$\begin{aligned}
\overset{*}{\mathbf{p}}{}_0^{\text{in}} &= \widetilde{\mathbf{O}}_{p0}^{\text{in}}\,\overline{\mathbf{Z}}_{p0}^{\text{in}}\,[\mathbf{P}_{p0}^{\text{in}}]^{1/2}, \qquad \textit{and} \qquad \overset{*}{\mathbf{p}}{}_k^{\text{in}} = \sum_{r=0}^{k-1} \overset{*}{\mathbf{p}}{}_r^{\text{in}}[\mathbf{B}_{pk}^{\text{in}}]_{(r+1)} + \widetilde{\mathbf{O}}_{pk}^{\text{in}}\,\overline{\mathbf{Z}}_{pk}^{\text{in}}[\mathbf{P}_{pk}^{\text{in}}]^{1/2}, \\
\overset{*}{\mathbf{p}}{}_0^{\text{out}} &= \widetilde{\mathbf{O}}_{p0}^{\text{out}}\,\overline{\mathbf{Z}}_{p0}^{\text{out}}\,[\mathbf{P}_{p0}^{\text{out}}]^{1/2}, \qquad \textit{and} \qquad \overset{*}{\mathbf{p}}{}_k^{\text{out}} = \sum_{r=0}^{k-1} \overset{*}{\mathbf{p}}{}_r^{\text{out}}[\mathbf{B}_{pk}^{\text{out}}]_{(r+1)} + \widetilde{\mathbf{O}}_{pk}^{\text{out}}\,\overline{\mathbf{Z}}_{pk}^{\text{out}}\,[\mathbf{P}_{pk}^{\text{out}}]^{1/2}.
\end{aligned} \tag{A.119}$$

Similarly define $\overset{*}{\mathbf{q}}{}_k^{\text{in}}$ *and* $\overset{*}{\mathbf{q}}{}_k^{\text{out}}$*. With these definitions, we have*

1. $$\overset{*}{\mathbf{p}}{}_k^{\text{in}} = \sum_{r=0}^{k} \widetilde{\mathbf{O}}_{pr}^{\text{in}}\,\overline{\mathbf{Z}}_{pr}^{\text{in}}\,[\mathbf{P}_{pk}^{\text{in}}]^{1/2}\,[\mathbf{c}_{pk}^{\text{in}}]_r, \qquad \overset{*}{\mathbf{p}}{}_k^{\text{out}} = \sum_{r=0}^{k} \widetilde{\mathbf{O}}_{pr}^{\text{out}}\,\overline{\mathbf{Z}}_{pr}^{\text{out}}\,[\mathbf{P}_{pk}^{\text{out}}]^{1/2}\,[\mathbf{c}_{pk}^{\text{out}}]_r \tag{A.120}$$

   where for $k \geq 1$ *and* $0 \leq r \leq k-1$, $[\mathbf{c}_{pk}^{\text{in}}]_r \in \mathbb{R}^{d\times d}$ *and* $[\mathbf{c}_{pk}^{\text{out}}]_r \in \mathbb{R}^{d\times d}$ *are defined recursively as* $[\mathbf{c}_{pk}^{\text{in}}]_r = \sum_{i=r}^{k-1}[\mathbf{B}_{pk}^{\text{in}}]_{(i+1)}[\mathbf{c}_{pi}^{\text{in}}]_r$*, and* $[\mathbf{c}_{pk}^{\text{out}}]_r = \sum_{i=r}^{k-1}[\mathbf{B}_{pk}^{\text{out}}]_{(i+1)}[\mathbf{c}_{pi}^{\text{out}}]_r$*, with* $[\mathbf{c}_{pk}^{\text{in}}]_k = [\mathbf{c}_{pk}^{\text{out}}]_k = \mathsf{I} \in \mathbb{R}^{d\times d}$ *for all* $k \geq 0$*.*

2. *Let* $\underline{\overset{*}{\mathbf{r}}}_k = \left(\overset{*}{\mathbf{r}}_0, \ldots, \overset{*}{\mathbf{r}}_k\right)$ *where* $\overset{*}{\mathbf{r}}_j = (\overset{*}{\mathbf{p}}{}_j^{\text{in}}, \overset{*}{\mathbf{p}}{}_j^{\text{out}}, \overset{*}{\mathbf{q}}{}_j^{\text{in}}, \overset{*}{\mathbf{q}}{}_j^{\text{out}})$ *are defined as above for the ideal variables* $\overset{*}{\mathbf{p}}$ *and* $\overset{*}{\mathbf{q}}$*. Then we have that* $\underline{\widetilde{\mathbf{r}}}_k - \underline{\overset{*}{\mathbf{r}}}_k = (\mathbf{d}_0, \ldots, \mathbf{d}_k)$*, where* $\underline{\widetilde{\mathbf{r}}}_k$ *is defined in* (2.53) *with*

   $$\mathbf{d}_k = \left(\sum_{r=0}^{k} \widetilde{\mathbf{\Delta}}_{pr}^{\text{in}}\,[\mathbf{c}_{pk}^{\text{in}}]_r,\ \sum_{r=0}^{k} \widetilde{\mathbf{\Delta}}_{pr}^{\text{out}}\,[\mathbf{c}_{pk}^{\text{out}}]_r,\ \sum_{r=0}^{k} \widetilde{\mathbf{\Delta}}_{qr}^{\text{in}}\,[\mathbf{c}_{qk}^{\text{in}}]_r,\ \sum_{r=0}^{k} \widetilde{\mathbf{\Delta}}_{qr}^{\text{out}}\,[\mathbf{c}_{qk}^{\text{out}}]_r\right). \tag{A.121}$$

   Furthermore, for all $k \geq 0$*, we have that* $\underline{\widetilde{\mathbf{r}}}_k \overset{d}{=} \underline{\mathbf{r}}_k$*.*

3. *For all* $i \geq 1$*, we have* $\left([\overset{*}{\mathbf{p}}{}_0^{\text{in}}]_i, \ldots, [\overset{*}{\mathbf{p}}{}_k^{\text{in}}]_i, [\overset{*}{\mathbf{p}}{}_0^{\text{out}}]_i, \ldots, [\overset{*}{\mathbf{p}}{}_k^{\text{out}}]_i\right) \overset{d}{=} \left(P_0^{\text{in}}, \ldots, P_k^{\text{in}}, P_0^{\text{out}}, \ldots, P_k^{\text{out}}\right)$ *where* $(P_0^{\text{in}}, \ldots, P_k^{\text{in}}) \in \mathbb{R}^{d(k+1)}$ *and* $(P_0^{\text{out}}, \ldots, P_k^{\text{out}}) \in \mathbb{R}^{d(k+1)}$ *are independent, zero-mean, jointly Gaussian vectors (and hence the right hand side is itself jointly Gaussian).*

Analogous statements for (1.)-(3.) hold for the $\overset{*}{\mathbf{q}}$ *variables.*

Proof. The proof of this statement is again essentially identical to that of Lemma 4 once all vector quantities are replaced by their matrix analogs. □

Finally, we define limiting quantities for the matrix form of the general recursion. Specifically,

$$\begin{aligned}\overline{U}_k^{\text{in/out}} &= \left([U_0^{\text{in/out}}]^T, \ldots, [U_k^{\text{in/out}}]^T\right)^T \\ &= \left(\left[U_0^{\text{in/out}}\right]_1, \ldots, \left[U_0^{\text{in/out}}\right]_d, \ldots, \left[U_k^{\text{in/out}}\right]_1, \ldots, \left[U_k^{\text{in/out}}\right]_d\right)^T \in \mathbb{R}^{d\times(k+1)},\end{aligned}$$

so we can define

$$\boldsymbol{\Sigma}_{pk}^{\text{in/out}} = \mathbb{E}\left\{\overline{U}_k^{\text{in/out}}[\overline{U}_k^{\text{in/out}}]^T\right\} \in \mathbb{R}^{d(k+1)\times d(k+1)}, \quad \mathbf{b}_{uk} = \mathbb{E}\left\{\overline{U}_{k-1}^{\text{in/out}}[U_k^{\text{in/out}}]^T\right\} \in \mathbb{R}^{kd\times d}. \tag{A.122}$$

In terms of these, we can then define

$$\mathbf{B}_{pk}^{\text{in/out}} = \left[\boldsymbol{\Sigma}_{p(k-1)}^{\text{in/out}}\right]^{-1}\mathbf{b}_{uk}, \quad \mathbf{P}_{pk}^{\text{in/out}} = \left[\boldsymbol{\Sigma}_{pk}^{\text{in/out}}\right]_{(k,k)} - [\mathbf{b}_{uk}^{\text{in/out}}]^T\left[\boldsymbol{\Sigma}_{p(k-1)}^{\text{in/out}}\right]^{-1}\mathbf{b}_{uk}^{\text{in/out}}, \tag{A.123}$$

where the (i, j) indexing notation is used to extract the $d \times d$ submatrix with top left index $((i - 1)d + 1, (j - 1)d + 1)$.

With these results, it is easy to see how the proof of Lemma 6 extend to the matrix case. For parts (g) and (a), which are concerned with the broad components of the $\boldsymbol{\Lambda}$ terms, the only differences arise from the presence of matrix operations replacing scalar operations (in particular due to the introduction of the Gram matrices). Concentration for matrix multiplications can be established as we have done for other matrix products in the proof of Lemma 6 (specifically using our concentration results for scalar products and sums). Concentration for matrix square roots is more complicated, but for $d \leq 4$, explicit formulas in terms of radicals are possible, to which are existing suite of concentration tools can be applied. In each case, the matrix operations which replace scalar operations are all of dimension d, which is fixed and independent of N and k. Hence, whatever penalty is incurred in the concentration rate we expect to be proportional to d^2. In particular, we can view this as being absorbed into the universal constants c and C.

Part (b) of the proof is almost entirely analogous as the same arguments now simply need to be applied to pseudo-Lipschitz functions with $2kd$ rather than $2k$ inputs. The remaining parts of the

proof are essentially just applications of (b) to establish concentration for various inner products, which can also be established in exactly the same way.

A.5 Notation and assumptions for GAMP case

Algorithm 7 presents the GAMP iteration. It is shown in [76] that, under appropriate assumptions, GAMP satisfies a state evolution given in (A.124).

Algorithm 7 GAMP

Require: Number of iterations K, design matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, observed $\mathbf{y} \in \mathbb{R}^{M}$, and denoisers g_k^{in} and g_k^{out}.

Initialize $\hat{\mathbf{x}}_1$, $\tau_k^x \geq 0$, and $\hat{\mathbf{s}}_0 = \mathbf{0}$.
for $k \leftarrow 1, \ldots, K$ **do**
  $\tau_k^p \leftarrow \frac{1}{M}\|\mathbf{A}\|_F^2 \times \tau_k^x$
  $\hat{\mathbf{p}}_k \leftarrow \mathbf{A}\hat{x}_k - \tau_k^p \hat{s}_{k-1}$
  $\hat{\mathbf{z}}_k = \mathbf{A}\hat{\mathbf{x}}_k$

  $\hat{\mathbf{s}}_k \leftarrow g_k^{\text{out}}\left(\hat{\mathbf{p}}_k, \mathbf{y}, \tau_k^p\right)$
  $\tau_k^s \leftarrow -\frac{1}{M}\text{div}\left[g_k^{\text{out}}\left(\hat{\mathbf{p}}_k, \mathbf{y}, \tau_k^p\right)\right]$

  $\tau_k^r \leftarrow \frac{N}{\|\mathbf{A}\|_F^2 \times \tau_k^s}$
  $\hat{\mathbf{r}}_k \leftarrow \hat{\mathbf{x}}_k + \tau_k^r \mathbf{A}^T \hat{\mathbf{s}}_k$

  $\hat{x}_{k+1} \leftarrow g_k^{\text{in}}\left(\hat{\mathbf{r}}_k, \tau_k^r\right)$
  $\tau_{k+1}^x \leftarrow \tau_k^r \times \frac{1}{N}\text{div}\left[g_k^{\text{in}}\left(\hat{\mathbf{r}}_k, \tau_k^r\right)\right]$
end for

$$
\begin{aligned}
\overline{\tau}_k^p &= \frac{M}{N}\overline{\tau}_k^x \\
\mathbf{K}_k^p &= \frac{M}{N}\mathbf{K}_k^x \\
\overline{\tau}_k^r &= \left[\mathbb{E}\left(-\frac{\partial}{\partial \hat{p}} g_k^{\text{out}}\left(\hat{P}_k, Y, \overline{\tau}_k^p\right)\right)\right]^{-1} \\
\xi_k^r &= \left[\overline{\tau}_k^r\right]^2 \mathbb{E}\left[\left[g_k^{\text{out}}\right]^2\left(\hat{P}_k, Y, \overline{\tau}_k^p\right)\right] \\
\alpha_k^r &= \overline{\tau}_k^r \times \mathbb{E}\left[\frac{\partial}{\partial z} g_k^{\text{out}}\left(\hat{P}_k, h(Z,W), \overline{\tau}_k^p\right)\right] \\
\overline{\tau}_{k+1}^x &= \overline{\tau}_k^r \mathbb{E}\left[\frac{\partial}{\partial \hat{r}} g_k^{\text{in}}\left(\hat{R}_k, \overline{\tau}_k^r\right)\right] \\
\mathbf{K}_{k+1}^x &= \text{Cov}\left(X, \hat{X}_{k+1}\right) \\
\hat{X}_{k+1} &= g_k^{\text{in}}\left(\hat{R}, \overline{\tau}_k\right)
\end{aligned}
\tag{A.124}
$$

Here, $\overline{\tau}_0^x$ is initialized as $\text{Cov}\left(X, \hat{X}_0\right)$, where the random variables X and $\hat{X}_0$ are empirical limits of the true signal $\mathbf{x}$ and initial estimates $\hat{\mathbf{x}}_0$. The random variable $\hat{R}_k$ is defined as

$$\hat{R}_k = \alpha_k^r X + Z_k^r,$$

where $Z^r \sim N(0, \xi_k^r)$ is independent of all other quantities. The random variables W and Z are likewise empirical limits of the true $\mathbf{z}$ and wv vectors (where $\mathbf{w}$ is as in **Assumption 0**, i.e. $\mathbf{y} = h(\mathbf{z}, \mathbf{w})$). The random variable Y is defined as $h(Z, W)$, and Z and $\hat{P}_k$ are jointly distributed as $N(0, \mathbf{K}_k^p)$.

In terms of these GAMP iterates and state evolution quantities, we make the following assumptions for our concentration result Theorem 2. We note that these assumptions also ensure the existence of all relevant empirical limits that arise in the definition of the state evolution.

Assumption 0. The initial estimate $\hat{\mathbf{x}}_0$ is taken i.i.d. with subgaussian entries and is independent of the matrix $\mathbf{A}$. This condition implies that there exist random variable $\hat{X}_0$ such that the initial estimates $\hat{\mathbf{x}}_0$ jointly concentrates on $\hat{X}_0$ by Lemma 19.

The output vector $\mathbf{y}$ can be expressed as $\mathbf{y} = h(\mathbf{z}_0, \mathbf{w})$ where $\mathbf{z}_0 = \mathbf{A}\mathbf{x}_0$ is the transformed input, $\mathbf{w}$ is an independent "disturbance" vector, and h is a measurable function. Next, the truth $\mathbf{x}_0$ is also independent of these other quantities. The components of $\mathbf{w}$ and $\mathbf{x}_0$ must also be either (1) sampled i.i.d. from a subgaussian distribution or (2) non-random sequences which are bounded independently of N and which converge empirically to a limiting random variable with finite second moment.

The initial precision estimates τ_0^x is positive for all N and converges to some $\overline{\tau}^x$. Finally, we assume that the sparsity level $\delta = \frac{M}{N}$ is bounded away from 0 independent of N.

Assumption 1. The design matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ has i.i.d. entries distributed as $N(0, M^{-1})$

Assumption 2. The log prior $\log q(x)$ and the log likelihood $\log f(y \mid z)$ (as a function of z) are log concave and β-smooth. The latter condition requires that, for some $\beta > 0$,

$$-\frac{\partial^2}{\partial x^2} \log q(x) < \beta, \qquad -\frac{\partial^2}{\partial z^2} \log f(y \mid z) < \beta. \tag{A.125}$$

Assumption 3. The estimating functions $g_{\text{in}}(k, \cdot, \tau^r)$ and $g_{\text{out}}(k, \cdot, h(\cdot, w), \tau^p)$ are separable, Lipschitz, and have Lipschitz derivatives for all $k \geq 0$, with Lipschitz constants independent of k.

Assumption 4. The τ_k^s and τ_k^x are truncated to lie in some interval $[t_{\min}, t_{\max}] \subset (0, 1)$. The τ_k^p and τ_k^r are also clipped so that they lie in some interval $[\tau_{\min}, \tau_{\max}]$ with $0 < \tau_{\min} < \tau_{\max} < \infty$ respectively.

Assumption 5. We terminate the algorithm according to the following stopping criteria. First, we stop the iteration if $1/\overline{\tau}_k^r < \epsilon_1$ or $\overline{\tau}_k^x < \epsilon_2$. To define our second stopping criterion, we consider the iterates $\hat{\mathbf{r}}_k$ and $\hat{\mathbf{p}}_k$. Let π_{rk} be the magnitude of the projection of $\hat{\mathbf{r}}_k$ onto the space spanned by the previous residual iterates $\{\hat{\mathbf{r}}_j\}_{j=0}^{k-1}$.

Let π_{pk} be defined similarly as the magnitude of the projection of $\hat{\mathbf{p}}_k$ on all previous $\mathbf{p}$ iterates. Our analysis will imply that these projection magnitudes will converge to certain limits as $N \to \infty$. We stop the iteration if any of the iterates are nearly spanned the previous iterates in this limit, i.e. when $\lim_{N\to\infty} \|\hat{\mathbf{r}}_k\|^2 - [\pi_{rk}]^2 < \epsilon_1'$ or $\lim_{N\to\infty} \|\mathbf{p}_k\|^2 - \left[\pi_{pk}\right]^2 < \epsilon_2'$.

Appendix B: Appendices to chapter 3

B.1 Proof of theorem 3

Proof. Replacing p^{post} by $\mathsf{joint}{-}p$, the inequality in (3.3) shows that $\mathsf{joint}{-}p$ is dominated in convex order by

$$\mathbb{E}_{f(\mathbf{y}_{\text{rep}}|\boldsymbol{\theta})} \mathbb{1}\left\{T_1(\mathbf{y}_{\text{rep}}) > T_1(\mathbf{y}) \text{ and } T_2(\mathbf{y}_{\text{rep}}) > T_2(\mathbf{y}) \text{ and } \cdots\ T_d(\mathbf{y}_{\text{rep}}) > T_d(\mathbf{y})\right\} \tag{B.1}$$

under $p(\mathbf{y}, \boldsymbol{\theta})$. However, unlike the case of p^{post}, (B.1) is no longer uniformly distributed. Nevertheless, the argument of Lemma 1 in [67] extends directly. Let G and F be the cumulative distribution functions of $\mathsf{joint}{-}p$ and (B.1) respectively. Then the established dominance in convex order is equivalent to the inequality

$$\int_s^1 [1 - G(t)]\, dt \leq \int_s^1 [1 - F(t)]\, dt \tag{B.2}$$

for all $s \in [0, 1]$. Next note that

$$\begin{aligned}\int_0^1 [1 - G(t)]\, dt &= \mathbb{E}_{f(\mathbf{y},\mathbf{y}_{\text{rep}})} \mathbb{1}\left\{T_1(\mathbf{y}_{\text{rep}}) > T_1(\mathbf{y}) \text{ and } T_2(\mathbf{y}_{\text{rep}}) > T_2(\mathbf{y}) \text{ and } \cdots\ T_d(\mathbf{y}_{\text{rep}}) > T_d(\mathbf{y})\right\} \\ &= \int_0^1 [1 - F(t)]\, dt. \end{aligned} \tag{B.3}$$

Combining (B.2) and (B.3), we get that

$$\int_0^s G(t)dt \leq \int_0^s F(t)dt \tag{B.4}$$

for all $s \in [0, \alpha]$. Next, since G is nondecreasing, we can get for any $\alpha \in [0, 1]$ the further bound

$$\int_0^s G(t)dt = \int_0^\alpha G(t)dt + \int_\alpha^s G(t)dt \geq \int_0^\alpha G(t)dt + G(\alpha)(s - \alpha). \tag{B.5}$$

For $\alpha \in [0, s]$, rearranging and combining with (B.4) gives

$$G(\alpha) \leq \frac{\int_0^s F(t)dt - \int_0^\alpha G(t)dt}{s - \alpha} \leq \frac{\int_0^s F(t)dt}{s - \alpha}. \tag{B.6}$$

Optimizing over $s \in [\alpha, 1]$ on the right then gives the desired bound. □

B.2 Details of sample quantile simulation example

For the calibrated p-values, the empirical CDF of p^{post} is estimated using 2000 prior predictive samples $\mathbf{y}_{\text{rep}}$, and each $p^{\text{post}}(\mathbf{y}_{\text{rep}})$ is estimated using 1000 samples from the posterior $p(\theta \mid \mathbf{y}_{\text{rep}})$. The estimated CDF is plotted in a neighborhood of the observed posterior predictive p-values in Figure B.1.

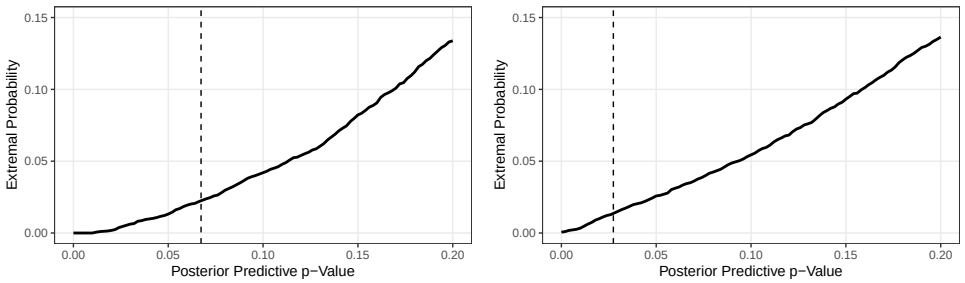

Figure B.1: Empirical CDF of $p_T^{\text{post}}(\mathbf{y}_{\text{rep}})$ for 2000 prior predictive draws of $\mathbf{y}_{\text{rep}}$ and for T equal to the 0.05 (left) and 0.95 (right) sample quantile statistics, respectively.

For the partial p-values, we calculate the partial posteriors on a grid of θ values between 0.5 an 4 (the support of the prior), estimating the likelihoods of the corresponding test statistics with kernel density estimators. These partial posteriors are displayed in Figure B.2. The kernel density estimates of the likelihoods for a given value of θ are computed from 2000000 samples of the 0.95 sample quantile and 800000 samples of the 0.05 sample quantile. These large sample sizes

reflect the need to have low estimation error in the tails of our kernel density estimates to prevent explosive errors when inverting them to form the partial posteriors. As B.2 shows, the estimation accuracy could still be improved further with greater sample sizes.

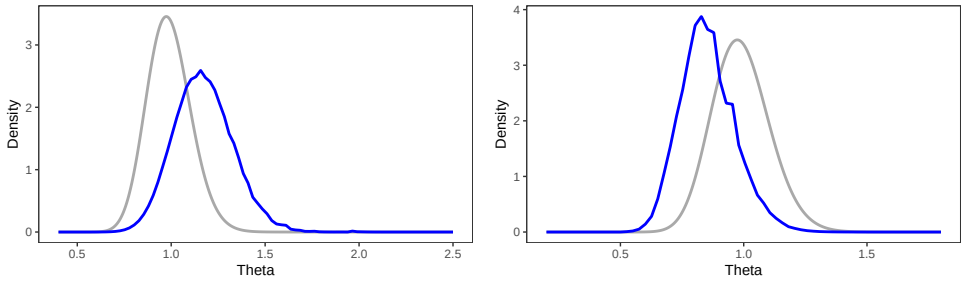

Figure B.2: Partial posteriors (in blue) computed for the 0.05 sample quantile (left) and 0.95 sample quantile (right), plotted against the full-data posterior (in gray).

The sampled p-values are estimated using 1000 draws from the sampling distribution $\mathsf{beta}(\theta, \theta)$ for each θ sampled from the posterior. The joint p-value bound is computed using $N_{\text{prior}} = 250$ samples θ_n from the prior distribution and $M_{\text{sampling}} = 50000$ samples from each corresponding sampling distribution $\mathsf{beta}(theta_n, \theta_n)$, out of which $L_{\text{estimate}} = 10000$ are randomly chosen for evaluation of the corresponding joint CDF (conditional on θ_n). By taking M_{sampling} larger than L_{estimate}, we are able to estimate all probabilities at higher resolution without needlessly multiplying the number of estimated probabilities for calculating the empirical CDF.

Appendix C: Appendices to chapter 4

C.1 Basic quantities and relations from information theory

In this section, we provide statements of the basic results from information theory that we make use of throughout this paper. Proofs of these results can be found in any introductory course on information theory. We state all results in terms of conditional entropies and mutual informations when appropriate since these contain the non-conditional statements as special cases. First we review relevant definitions.

Definition 7 (Basic Quantities of Information Theory)**.** *Let $q(\boldsymbol{\theta}, \mathbf{y})$ be some joint model. Then the entropy of $\boldsymbol{\theta}$ is defined as*

$$h_{q(\boldsymbol{\theta})}(\boldsymbol{\theta}) = -\mathbb{E}_{q(\boldsymbol{\theta})} \log q(\boldsymbol{\theta}). \tag{C.1}$$

The conditional entropy of $\boldsymbol{\theta}$ given $\mathbf{y}$ is just the average entropy of the conditional distributions:

$$h_{q(\boldsymbol{\theta},\mathbf{y})}\left(\boldsymbol{\theta} \mid \mathbf{y}\right) = \mathbb{E}_{f(\mathbf{y})} h_{q(\boldsymbol{\theta}|\mathbf{y})}\left(\boldsymbol{\theta}\right) = -\mathbb{E}_{q(\boldsymbol{\theta},\mathbf{y})} \log q(\boldsymbol{\theta} \mid \mathbf{y}). \tag{C.2}$$

The mutual information between $\boldsymbol{\theta}$ and $\mathbf{y}$ is the amount by which entropy is expected to decrease after conditioning $\mathbf{y}$:

$$\mathbf{I}_q\left(\boldsymbol{\theta}, \mathbf{y}\right) = h_{q(\boldsymbol{\theta})}\left(\boldsymbol{\theta}\right) - h_{q(\boldsymbol{\theta},\mathbf{y})}\left(\boldsymbol{\theta} \mid \mathbf{y}\right). \tag{C.3}$$

Finally, if we extend our joint model to $q(\boldsymbol{\theta}, \mathbf{y}, \mathbf{x})$ where $\mathbf{x}$ is any additional quantity, then the conditional mutual information given $\mathbf{x}$ is just the difference of the corresponding conditional entropies:

$$\mathbf{I}_q\left(\boldsymbol{\theta}, \mathbf{y} \mid \mathbf{x}\right) = h_{q(\boldsymbol{\theta},\mathbf{x})}\left(\boldsymbol{\theta} \mid \mathbf{x}\right) - h_{q(\boldsymbol{\theta},\mathbf{y},\mathbf{x})}\left(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}\right). \tag{C.4}$$

The first important result allows us to break up an entropy or mutual information expression

additive over the components of vector arguments.

Lemma 1 (Chain Rule for Entropy and Mutual Information)**.** *Let* $q(\mathbf{x}, \mathbf{y}, \mathbf{z})$ *be a joint model and suppose that* $\mathbf{x}$ *can be partitioned into sub-vectors* $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ *for some* $m \geq 1$*. Then we have that*

$$h_q(\mathbf{x} \mid \mathbf{y}) = \sum_{i=1}^{m} h(\mathbf{x}_i \mid \mathbf{x}_{<i}, \mathbf{y}),$$

where $\mathbf{x}_{<i} = (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1})$ *for* $i \geq 2$*, and* $\mathbf{x}_{<1} = \{\}$*. Furthermore, we have that*

$$\mathbf{I}_q(\mathbf{y}, \mathbf{x} \mid \mathbf{z}) = \sum_{i=1}^{m} \mathbf{I}(\mathbf{y}, \mathbf{x}_i \mid \mathbf{z}, \mathbf{x}_{<i}).$$

Next, it can be useful to express the (conditional) mutual information in terms of the KL divergence, which quantifies discrepancy between two probability distributions $p(\mathbf{y})$ and $q(\mathbf{y})$. In particular, the KL divergence is given as

$$D(p(\mathbf{y})||q(\mathbf{y})) = \mathbb{E}_{p(\mathbf{y})} \log \left[\frac{p(\mathbf{y})}{q(\mathbf{y})}\right].$$

The mutual information can be related to the KL divergence in two different ways.

Lemma 2 (Mutual Information as KL Divergence)**.** *Let* $q(\mathbf{x}, \mathbf{y}, \mathbf{z})$ *be a joint model. Then we have*

$$\mathbf{I}(\mathbf{y}, \mathbf{x} \mid \mathbf{z}) = \mathbb{E}_{q(\mathbf{z})} D(q(\mathbf{x}, \mathbf{y} \mid \mathbf{z})||q(\mathbf{x} \mid \mathbf{z})q(\mathbf{y} \mid \mathbf{z})) = \mathbb{E}_{q(\mathbf{y}, \mathbf{z})} D(q(\mathbf{x} \mid \mathbf{y}, \mathbf{z})||q(\mathbf{x} \mid \mathbf{z})).$$

It is of fundamental importance that the KL divergence is always nonnegative, which follows by an application of Jensen's inequality.

Lemma 3 (Nonnegativity of the KL Divergence)**.** *For any densities* $p(\mathbf{y})$ *and* $q(\mathbf{y})$*, we have*

$$D(p(\mathbf{y})||q(\mathbf{y})) \geq 0$$

with equality if and only if $p(\mathbf{y}) = q(\mathbf{y})$ *p-almost surely.*

This immediately implies nonnegativity of the mutual information, and in turn the fact that

$$h_{q(\mathbf{x},\mathbf{y})}(\mathbf{y} \mid \mathbf{x}) \leq h_{q(\mathbf{y})}(\mathbf{y})$$

for any joint distribution $q(\mathbf{x}, \mathbf{y})$.

It is often useful to know how these metrics operate under certain transformations of the random quantities in terms of which they are defined. This is characterized by the following result.

Lemma 4 (Entropy and Mutual Information Under Transformation)**.** *Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be any invertible matrix and let $\mathbf{y}' = \mathbf{A}\mathbf{y}$. Then we have*

$$h_{q(\mathbf{y}')}(\mathbf{y}') = h_{q(\mathbf{y})}(\mathbf{y}) + \log |\det \mathbf{A}| .$$

Furthermore, if $\mathbf{y}' = \mathbf{y} + c$ for any $c \in \mathbb{R}$, then $h(\mathbf{y}) = h(\mathbf{y}')$. Thus, the entropy is invariant under translations and orthogonal transformations. The mutual information satisfies the stronger property of invariance under arbitrary smooth reparametrizations of the individual arguments. Specifically, let ϕ, ψ be smooth, invertible maps, and define $\mathbf{y}' = \phi(\mathbf{y})$ and $\mathbf{x}' = \psi(\mathbf{x})$. Then we have that

$$\mathbf{I}(\mathbf{y}', \mathbf{x}' \mid \mathbf{z}) = \mathbf{I}(\mathbf{y}, \mathbf{x} \mid \mathbf{z}).$$

The general behavior of the mutual information under potentially noninvertible transformations is characterized by the data processing inequality.

Lemma 5 (Data Processing Inquality)**.** *Let $q(\mathbf{x}, \mathbf{y}, \mathbf{z})$ be any distribution, and suppose that $\mathbf{x}$ and $\mathbf{z}$ are conditionally independent given $\mathbf{y}$. Then we have that*

$$\mathbf{I}(\mathbf{x}, \mathbf{y}) \geq \mathbf{I}(\mathbf{x}, \mathbf{z}).$$

In particular, the above inequality holds if $\mathbf{z} = \psi(\mathbf{y})$ for any function ψ.

Finally, certain distributions maximize the entropy under certain conditions. For our purposes,

it suffices to note that normal distributions on $\mathbb{R}^d$ maximize the entropy among all distributions with fixed covariance matrix and support equal to $\mathbb{R}^d$.

Lemma 6 (Maximum Entropy of Normal)**.** *Let $q(\mathbf{y})$ be any probability distribution supported on $\mathbb{R}^d$, and let $p(\mathbf{y})$ be a normal distribution with any mean and covariance matrix equal to the covariance $\mathbf{\Sigma}_q$ of $q(\mathbf{y})$. Then we have*

$$h_{q(\mathbf{y})}(\mathbf{y}) \leq h_{p(\mathbf{y})}(\mathbf{y}) = \frac{1}{2}\log\left(\det\left(2\pi e \mathbf{\Sigma}_q\right)\right).$$

C.2 Weak identification simulation details

Algorithm 8 Posterior Bootstrap for Sampling Existing Subpopulations

Require: Observed data $\mathbf{y}$, # of replications R, # of new samples per subpopulation M^{new}, # of posterior samples S.

for $r \leftarrow 1, \ldots, R$ **do**
  Sample $(\theta_1, \ldots \theta_L, \sigma) \sim q(\boldsymbol{\theta}, \sigma \mid \mathbf{y})$.
  for $l \leftarrow 1, \ldots L$ **do**
    Sample $\mathbf{y}_{1l}^{\text{new}}, \ldots, \mathbf{y}_{M^{\text{new}}l}^{\text{new}} \overset{iid}{\sim} f(y \mid \theta_l, \sigma)$.
  end for
  $\mathbf{y}^{\text{rep}} \leftarrow \left[\begin{smallmatrix}\mathbf{y}\\ \mathbf{y}^{\text{new}}\end{smallmatrix}\right]$.
  Sample $\mu_{(1)}, \ldots, \mu_{(S)} \overset{iid}{\sim} q(\mu \mid \mathbf{y}^{\text{rep}})$
  $\hat{\sigma}_r \leftarrow \mathsf{sd}\left(\mu_{(1)}, \ldots, \mu_{(S)}\right)$
end for
return $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_R\}$.

Algorithm 9 Posterior Bootstrap for Sampling New Subpopulations

Require: Observed data $\mathbf{y}$, # of replications R, # of new subpopulations L^{new}, # of posterior samples S.

for $r \leftarrow 1, \ldots, R$ **do**
  Sample $(\mu, \tau, \sigma) \sim q(\mu, \tau, \sigma \mid \mathbf{y})$.
  Sample $(\theta_1, \ldots \theta_{L^{\text{new}}}) \sim q(\boldsymbol{\theta} \mid \mu, \tau)$.
  for $l \leftarrow 1, \ldots L^{\text{new}}$ **do**
    Sample $\mathbf{y}_{1l}^{\text{new}}, \ldots, \mathbf{y}_{Ml}^{\text{new}} \overset{iid}{\sim} f(y \mid \theta_l, \sigma)$.
  end for
  $\mathbf{y}^{\text{rep}} \leftarrow [\,\mathbf{y}\ \mathbf{y}^{\text{new}}\,]$.
  Sample $\mu_{(1)}, \ldots, \mu_{(S)} \overset{iid}{\sim} q(\mu \mid \mathbf{y}^{\text{rep}})$
  $\hat{\sigma}_r \leftarrow \mathsf{sd}\left(\mu_{(1)}, \ldots, \mu_{(S)}\right)$
end for
return $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_R\}$.

We simulated our three data sets in Section 5 from the following model.

$$\mathbf{y}_{ml} \mid \boldsymbol{\theta} \overset{iid}{\sim} \text{normal}(\theta_l, \sigma_*), \qquad \boldsymbol{\theta}_l \overset{iid}{\sim} \text{normal}(\mu_*, \tau_*), \tag{C.5}$$

The σ_*, μ_*, and τ_* are all hyperparameters. We fixed μ_* and τ_* for all data sets, but we varied σ_* from $\tau_*/2$ to $2\tau_*$. For our purposes, the key differentiator of the data sets is the ratio of the variance

of the subpopulation-level (sample) means to the overall (sample) variance. When this statistic is closer to 1, the between-subpopulation variation swamps the within-population variation, and the reverse is true when this statistic is close to 0. For the three resulting data sets, this statistic was approximately 0.45, 0.6, and 0.95 respectively. The complete data sets along with the R and Stan code used to generate and analyze them are available on the companion GitHub repository `https://github.com/collin-cademartori/BayesianModelExpansionPaper/`.

Algorithms 8 and 9 give detailed pseudo-code for the two resampling procedures we used to simulate the process of sampling data from (a) the same subpopulations from which the original data were sampled and (b) new subpopulations within the larger superpopulation. Each of these algorithms was run with $R = 500$ replications of the resampling scheme, $S = 2000$ samples drawn from each of the resulting posterior distributions, and with $M^{\text{new}} = 8$ and $L^{\text{new}} = 20$ respectively.

C.3 Details of the election forecasting model

The primary purpose of the forecasting model is to infer the level of support for the Democratic candidate over time and across states. This level of support is represented by a matrix parameter $\boldsymbol{\mu} \in \mathbb{R}^{S \times T}$ with rows representing the $S = 51$ states (including Washington DC) and columns representing the T days from the start of measurement until election day. This parameter is assigned a time series prior:

$$\boldsymbol{\mu}_t \mid \boldsymbol{\mu}_{t+1} \sim \text{normal}\left(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}^{\mu}\right) \text{ for } 1 \leq t \leq T-1, \text{ and } \boldsymbol{\mu}_T \sim \text{normal}\left(\mathbf{m}^{\text{f}}, \mathbf{S}^{\text{f}}\right). \tag{C.6}$$

Here $\boldsymbol{\Sigma}^{\mu} \in \mathbb{R}^{S \times S}$ is a hyperparameter encoding correlation between states and variation over time, constructed using demographic data, polling from previous elections, and domain knowledge. Likewise, $\mathbf{m}^{f} \in \mathbb{R}^{S}$ and $\mathbf{S}^{f} \in \mathbb{R}^{S \times S}$ are hyperparameters set using a 'fundamentals forecast' derived from variables known in political science to be good predictors of U.S. election outcomes.

Results of state and national polls are modeled with a binomial distribution that combines $\boldsymbol{\mu}$ with terms representing sources of polling bias. Letting $i = 1, \ldots, N_{\text{state}}$ index state polls and y_i

denote the number of respondents supporting the Democratic candidate out of n_i respondents in poll i, we have

$$y_i \sim \text{binomial}\left(\text{logit}^{-1}\left(\boldsymbol{\mu}_{s_i,t_i} + \beta_i\right), n_i\right), \tag{C.7}$$

where s_i, t_i denote the state and day for poll i, and β_i models various sources of bias. National polls are modeled similarly, except that state-level terms including $\boldsymbol{\mu}_{st}$ are averaged with weights accounting for each state's share of the national vote in the previous election.

The bias terms β_i in (C.7) are decomposed into several further terms which are designed to capture the following sources of polling bias:

- Pollster-level "house" effects, i.e. the observed phenomenon that most pollsters exhibit a nonzero, temporally stable bias toward one party or the other.
- Poll mode effects reflecting the observed phenomenon that the averages of phone- and internet-based polls tend to differ.
- Poll population effects reflecting the observed phenomenon that the averages of "registered voter" and "likely voter" polls tend to differ.
- Partisanship nonresponse effects that account for the fact that differential rates of response between members of the two parties can alter poll results. This term is only included for polls that do not attempt to adjust their results for the partisan composition of their sample. Unlike the previous terms, this term varies through time (and is assigned a time-series prior) in order to account for the fact that differential partisan nonresponse has been observed to vary over the course of an election cycle.
- State-level measurement error effects reflecting the observation that polls of some states have historically been more accurate than others.
- Poll-level random measurement error, i.e. a catch-all term representing any additional variance in the average support for the Democrat among the polls' sampling frames unaccounted for by the above terms.

A complete description of the polling model specification can be found in [43], and the code and data for the original model can be accessed at `https://github.com/TheEconomist/us-potus-model`. The code for our expanded models and our conditional posterior checks can be found at `https://github.com/collin-cademartori/BayesianModelExpansionPaper`.

C.4 Proof of theorem 4

In this section we prove our main result, Theorem 4. We build up the proof from many intermediate results, starting with a result that upper bounds the marginal Fisher information of an expanded model in terms of the (unmarginalized) Fisher information and the prior.

C.4.1 Bounding the marginal fisher information

For model $q(\mathbf{y}, \boldsymbol{\theta})$, the observed and Fisher information matrices are defined as

$$[\boldsymbol{\mathcal{J}}_q(\mathbf{y}, \boldsymbol{\theta})]_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\mathbf{y} \mid \boldsymbol{\theta}) \text{ for all } 1 \leq i, j \leq d,$$
$$\boldsymbol{\mathcal{I}}_q(\boldsymbol{\theta}) = \mathbb{E}_{f(\mathbf{y}|\boldsymbol{\theta})} \boldsymbol{\mathcal{J}}_q(\mathbf{y}, \boldsymbol{\theta}). \tag{C.8}$$

We drop the subscript when the model is clear from context. We now state a bound on the mutual information in terms of the Fisher information, which follows directly from Theorem 2 of [1].

Theorem 5 (Mutual Information Upper Bound)**.** *Let $q(\mathbf{y}, \boldsymbol{\theta})$ be a model such that the prior $q(\boldsymbol{\theta})$ is log-concave with covariance matrix $\boldsymbol{\Sigma}$, then*

$$\mathbf{I}(\boldsymbol{\theta}, \mathbf{y}) \leq d\psi\left(\frac{1}{d} \operatorname{tr}\left(\mathbb{E}_{q(\boldsymbol{\theta})} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \boldsymbol{\Sigma}^{1/2}\right)\right), \tag{C.9}$$

where $\psi(x)$ is the concave increasing function given by

$$\psi(x) = \begin{cases} \sqrt{x}, & 0 \leq x \leq 1 \\ 1 + \frac{1}{2}\log(x), & x > 1 \end{cases}.$$

Now let v_{pr} be the maximum eigenvalue of $\boldsymbol{\Sigma}$ (the covariance matrix over just the $\boldsymbol{\theta}$ parameters). Then we also clearly have

$$\mathbf{I}(\boldsymbol{\theta}, \mathbf{y}) \leq d\psi \left(\frac{v_{\text{pr}}}{d} \mathbb{E}_{q(\boldsymbol{\theta})} \operatorname{tr} \left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \right) \right). \tag{C.10}$$

If v_{pr} differs between a base model and expanded model, then we can rescale the prior over $\boldsymbol{\theta}$ in the expanded model so that they are equal. The only possible difficulty is that we may no longer have a $\boldsymbol{\lambda}_0$ for which $q_{\text{base}}(\boldsymbol{\theta}) = q(\boldsymbol{\theta} \mid \boldsymbol{\lambda}_0)$. Such situations can always be resolved however by passing to a larger model which includes a prior scale hyperparameter for $\boldsymbol{\theta}$ in $\boldsymbol{\lambda}$. With such a hyperparameter, we always have the ability to set both the marginal prior scale and $\boldsymbol{\lambda}_0$-conditional prior scale for $\boldsymbol{\theta}$ independently, allowing equality of v_{pr} and preservation of the model expansion property.

We also note that rescaling $\boldsymbol{\theta}$ leaves $\mathbf{I}(\boldsymbol{\theta}, \mathbf{y})$ unchanged since the mutual information is invariant to all invertible transformations of $\boldsymbol{\theta}$ and $\mathbf{y}$ separately. Thus, with loss of little generality, we henceforth assume that $v_{\text{pr}} = 1$ for all models. The weaker bound (C.10) will be useful for comparisons to other quantities in the next sections, and for deriving the following further upper bound, which applies more directly to model expansions, is easier to compute, and mirrors the decomposition of mutual information we obtained in Section 2.

Theorem 6. *Define the partial Hessian with respect to $\boldsymbol{\lambda}$ as $[\mathbf{H}(\boldsymbol{\lambda}; \boldsymbol{\theta}, \mathbf{y})]_{ij} = -\frac{\partial^2}{\partial\lambda_j \partial\lambda_k} \log q(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ for $1 \leq i, j \leq k$. Then under the regularity conditions in Appendix C.4,*

$$\mathbb{E} \operatorname{tr} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \leq \sum_{j=1}^{d} \left[\mathbb{E} \left\{ -\frac{\partial^2}{\partial\theta_j^2} \log f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \right\} - \Delta_j \right], \tag{C.11}$$

where we define

$$\Delta_j = \mathbb{E} \left\{ \frac{\partial^2}{\partial\theta_j^2} \log q(\boldsymbol{\lambda} \mid \boldsymbol{\theta}) \right\} + \frac{\left[\sum_{i=1}^{k} \mathbb{E} \frac{\partial}{\partial\lambda_i} \frac{\partial}{\partial\theta_j} \log q(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \right]^2}{\mathbb{E} \| \mathbf{H}(\boldsymbol{\lambda}; \boldsymbol{\theta}, \mathbf{y}) \|_{\text{op}}}.$$

The Δ_j terms compare prior- and posterior-based measures of dependence between $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. We have $\Delta_j \geq 0$ when $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are independent under the prior. In particular, when $q(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{y})$ is an

expansion of some base model and when $\sum_{i=1}^{d} \Delta_j > 0$, (C.11) again exhibits a downward bias on the Fisher information of the expanded model compared to the base model. We illustrate with two simple examples before turning to the proof.

1. Take a linear regression model q_{base} with response $\mathbf{y} \in \mathbb{R}^n$, predictors $\mathbf{X} \in \mathbb{R}^{n \times m}$, coefficients $\boldsymbol{\beta} \in \mathbb{R}^m$, intercept α, and log noise variance τ:

$$(2\pi \exp(\tau))^{-n/2} \exp\left[-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \alpha\mathbf{1})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \alpha\mathbf{1}) / 2\exp(\tau)\right].$$

   We consider an expansion p with additional predictor $\mathbf{z} \in \mathbb{R}^n$ and coefficient λ. Suppose the coefficients are assigned independent priors, and let $\boldsymbol{\theta} = (\sigma, \beta_1, \ldots, \beta_m, \alpha)$.

   We assume without loss of generality that all predictors are centered as this does not affect the posterior entropy. We then find for all $1 \leq j \leq m$,

$$\begin{aligned}
-\mathbb{E}_q \frac{\partial^2}{\partial \beta_j^2} \log f(\mathbf{y} \mid \lambda, \boldsymbol{\theta}) &= \mathbb{E}_{q(\tau)} \left\{ \frac{n \text{var}\left(\mathbf{x}_j\right)}{\exp(\tau)} \right\} = -\mathbb{E}_{q_{\text{base}}} \frac{\partial^2}{\partial \beta_j^2} \log q_{\text{base}}(\mathbf{y} \mid \boldsymbol{\theta}), \\
-\mathbb{E}_q \frac{\partial^2}{\partial \tau} \log f(\mathbf{y} \mid \lambda, \boldsymbol{\theta}) &= \frac{n}{2} = -\mathbb{E}_{q_{\text{base}}} \frac{\partial^2}{\partial \tau^2} \log q_{\text{base}}(\mathbf{y} \mid \boldsymbol{\theta}), \text{ and} \\
-\mathbb{E}_q \frac{\partial^2}{\partial \alpha^2} \log f(\mathbf{y} \mid \lambda, \boldsymbol{\theta}) &= \mathbb{E}_{q(\tau)} \left\{ \frac{n}{\exp(\tau)} \right\} = -\mathbb{E}_{q_{\text{base}}} \frac{\partial^2}{\partial \alpha^2} \log q_{\text{base}}(\mathbf{y} \mid \lambda, \boldsymbol{\theta})
\end{aligned}$$

   These computations show that the first term in (C.11) is just $\text{tr}\left(\mathbb{E}\boldsymbol{\mathcal{I}}_{q_{\text{base}}}(\boldsymbol{\theta})\right)$. Assuming $\mathbf{z}$ is also centered, computing the second term in (C.11) gives that

$$\mathbb{E}_{q_{\text{base}}} \text{Tr}\left(\boldsymbol{\mathcal{I}}_{q_{\text{base}}}\right) - \mathbb{E}_q \text{Tr}\left(\boldsymbol{\mathcal{I}}_q\right) \geq n^2 \mathbb{E}_{q(\tau)} \left\{ \frac{1}{\exp(\tau)} \right\} \sum_{j=1}^{m} \left[\frac{\text{cov}\left(\mathbf{x}_j, \mathbf{z}\right)}{\text{var}\left(\mathbf{z}\right)} \right]^2,$$

   which reflects the familiar fact that the identifiability of regression models is reduced by significant correlation between predictors.

2. Next consider an exchangeable Poisson base model with likelihood

$$\exp\left(\mu n \overline{\mathbf{y}} - n \exp(\mu)\right) \Big/ \left(y_1! \times y_2! \times \cdots \times y_n!\right).$$

This can be expanded to a negative binomial model with likelihood

$$\left[\prod_{i=1}^{n} \frac{\Gamma\left(y_i + \exp(\lambda)\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\exp(\lambda)\right)}\right]\left(\frac{\exp(\mu)}{\exp(\mu)+\exp(\lambda)}\right)^{\sum_{i=1}^{n} y_i}\left(\frac{\exp(\lambda)}{\exp(\mu)+\exp(\lambda)}\right)^{n\exp(\lambda)}.$$

This converges to the Poisson density as $\lambda \to \infty$, so this model is in fact an expansion of the Poisson model. Next observe that the second derivatives with respect to μ are given by

$$-\frac{\partial^2}{\partial \mu^2}\log f\left(\mathbf{y} \mid \mu, \lambda\right) = n\exp\left(\mu\right)\left[1 - \frac{\exp(\mu)}{\exp(\mu)+\exp(\lambda)}\right]\left[\frac{\overline{y} + \exp(\lambda)}{\exp(\mu)+\exp(\lambda)}\right],$$

which has expected value $n\exp\left(\mu\right)\left[1 - \frac{\exp(\mu)}{\exp(\mu)+\exp(\lambda)}\right]$ under $f(\mathbf{y} \mid \mu, \lambda)$. With this we can show using (C.11) that the Fisher information trace must fall in passing from the base to the expanded model:

$$\mathbb{E}\mathrm{Tr}\left(\boldsymbol{\mathcal{I}}_q\right) \leq \mathbb{E}\left\{n\exp\left(\mu\right)\left[\frac{\exp(\lambda)}{\exp(\mu)+\exp(\lambda)}\right]\right\} < \mathbb{E}\left\{n\exp(\mu)\right\} = \mathbb{E}\mathrm{Tr}\left(\boldsymbol{\mathcal{I}}_{q_{\text{base}}}\right),$$

We now turn to the proof of Theorem 6. The key idea in this proof is to rearrange the logarithm, derivative, and expectation operations to obtain a more tractable expression for the marginal score. First, we need a version of the Cramer-Rao bound that relates the resulting expression to the desired derivatives.

Lemma 7 (Cramer-Rao Lower Bound)**.** *Let $\phi = -\log p(\mathbf{x})$ be a differentiable potential function with $\mathbf{x} \in \mathbb{R}^d$, and let $f(\mathbf{x})$ be a differentiable function. Furthermore, assume that for any $\boldsymbol{\theta} \in \mathbb{R}^d$ fixed, $\lim_{\|\mathbf{x}\|\to\infty} f(\mathbf{x})p(\mathbf{x}-\boldsymbol{\theta}) = 0$.*

Then we have the following inequality:

$$\mathrm{Var}\left(f(\mathbf{x})\right) \geq \left[\mathbb{E}\nabla f(\mathbf{x})\right]^T\left[\mathbb{E}\mathbf{H}_\phi(\mathbf{x})\right]^{-1}\left[\mathbb{E}\nabla f(\mathbf{x})\right].$$

Proof. Assume without loss of generality that $p(\mathbf{x})$ has mean $\mathbf{0}$. Let $p_{\boldsymbol{\theta}}(\mathbf{x}) = p(\mathbf{x}-\boldsymbol{\theta})$, and consider the (biased) estimator of $(f(\boldsymbol{\theta}), 0, \ldots, 0) \in \mathbb{R}^d$ given by $T(\mathbf{x}) = (f(\mathbf{x}), 0, \ldots, 0)$. We note that,

under our assumptions, we have for $1 \leq i \leq d$ that

$$\frac{\partial}{\partial \theta_i} \mathbb{E}_{\boldsymbol{\theta}} f(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\theta}} \frac{\partial}{\partial x_i} f(\mathbf{x}),$$

and that

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathbb{E}\mathbf{H}_{\phi}(\mathbf{x} - \boldsymbol{\theta}),$$

where the Hessian is respect to the components of $\mathbf{x}$. Now the Cramer-Rao lower bound for $T(\mathbf{x})$ is just

$$\mathrm{Cov}_{\boldsymbol{\theta}}\left(T(\mathbf{x})\right) \succ \left[\mathbf{J}_{\mathbb{E}_{\boldsymbol{\theta}} T(\mathbf{x})}(\boldsymbol{\theta})\right] \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})^{-1} \left[\mathbf{J}_{\mathbb{E}_{\boldsymbol{\theta}} T(\mathbf{x})}(\boldsymbol{\theta})\right]^T.$$

Multiplying on the left and right by $\mathbf{e}_1^T$ and $\mathbf{e}_1$ respectively preserves the inequality (by definition of Loewner order), and so evaluating this multiplication and taking $\boldsymbol{\theta} = \mathbf{0}$ yields

$$\mathrm{Var}\left(f(\mathbf{x})\right) \geq \left[\mathbb{E}\nabla f(\mathbf{x})\right]^T \left[\mathbb{E}\mathbf{H}_{\phi}(\mathbf{x})\right]^{-1} \left[\mathbb{E}\nabla f(\mathbf{x})\right],$$

as claimed. □

For Theorem 6, we need the following regularity conditions:

1. $\frac{\partial}{\partial \theta_i} \int q\left(\mathbf{x}, \boldsymbol{\lambda} \mid \boldsymbol{\theta}\right) d\boldsymbol{\lambda} = \int \frac{\partial}{\partial \theta_i} q\left(\mathbf{x}, \boldsymbol{\lambda} \mid \boldsymbol{\theta}\right) d\boldsymbol{\lambda}$ for all $1 \leq i \leq d$.
2. $\mathrm{Var}_{q(\lambda \mid \mathbf{x}, \boldsymbol{\theta})}\left(\frac{\partial}{\partial \theta_i} \log f\left(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}\right)\right) < \infty$ for all $1 \leq i \leq d$.
3. For any fixed $\boldsymbol{\lambda}^* \in \mathbb{R}^k$, almost every $(\mathbf{x}, \boldsymbol{\theta})$, and all $1 \leq i \leq d$,

$$\lim_{\|\boldsymbol{\lambda}\| \to \infty} q(\boldsymbol{\lambda} - \boldsymbol{\lambda}^* \mid \boldsymbol{\theta}, \mathbf{x}) \left[\frac{\partial}{\partial \theta_i} \log f\left(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}\right)\right] = 0.$$

We can now proceed with the proof.

Proof. The i^{th} term of $\mathbb{E}\mathrm{Tr}\,(\boldsymbol{\mathcal{I}})$ is just

$$\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\left\{\left(\frac{\partial}{\partial\theta_i}\log f(\mathbf{x}\mid\boldsymbol{\theta})\right)^2\right\} \tag{C.12}$$

We can then rewrite

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\log f(\mathbf{x}\mid\boldsymbol{\theta}) &= \frac{1}{f(\mathbf{x}\mid\boldsymbol{\theta})}\frac{\partial}{\partial\theta_i}\int f(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})d\boldsymbol{\lambda}\\ &= \int\frac{q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})}{f(\mathbf{x}\mid\boldsymbol{\theta})}\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})d\boldsymbol{\lambda} \qquad \text{(C.13)}\\ &= \int q(\boldsymbol{\lambda}\mid\mathbf{x},\boldsymbol{\theta})\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})d\boldsymbol{\lambda}\\ &= \mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right\}\end{aligned}$$

where (C.13) follows by exchanging the integral and derivative and using the expression for the logarithmic derivative. Plugging this into the above and using the definition of the variance, we obtain the identity

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\left\{\left(\mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right\}\right)^2\right\}\\ &\quad= \mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\left\{\mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\left\{\left(\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right)^2\right\} - \mathrm{Var}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right\}\right\}\\ &\quad= \mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}\left\{\left(\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right)^2\right\} - \mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\left\{\mathrm{Var}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right\}\right\} \qquad \text{(C.14)}\end{aligned}$$

Now the usual Fisher information identity gives us that

$$\begin{aligned}\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}\left\{\left(\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right)^2\right\} &= \mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}\left\{-\frac{\partial^2}{\partial\theta_i^2}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right\}\\ &= \mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}\left\{-\frac{\partial^2}{\partial\theta_i^2}\log q(\mathbf{x}\mid\boldsymbol{\theta},\boldsymbol{\lambda})\right\}\\ &\quad+\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}\left\{-\frac{\partial^2}{\partial\theta_i^2}\log q(\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right\}.\end{aligned}$$

On the other hand, using the Cramer-Rao inequality above, we get that

$$\mathrm{Var}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right\} \geq \frac{\left[\mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\sum_{j=1}^{k}\frac{\partial}{\partial\lambda_j}\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})\right]^2}{\|\mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\mathbf{H}\left(\boldsymbol{\lambda};\boldsymbol{\theta},\mathbf{x}\right)\|_{\mathrm{op}}} \tag{C.15}$$

Plugging this in, the right-hand term in (C.14) can be bounded by

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\left\{\mathrm{Var}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\left\{\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\lambda}\mid\boldsymbol{\theta})\right\}\right\} \\
&\quad\geq \mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\left\{\left[\frac{\mathbb{E}_{q(\boldsymbol{\lambda}|\boldsymbol{\theta},\mathbf{x})}\sum_{j=1}^{k}\frac{\partial}{\partial\lambda_j}\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}{\|\mathbb{E}_{q(\boldsymbol{\lambda}|\boldsymbol{\theta},\mathbf{x})}\mathbf{H}(\boldsymbol{\lambda};\boldsymbol{\theta},\mathbf{x})\|_{\mathrm{op}}^{1/2}}\right]^2\right\} \\
&\quad\geq \left[\sum_{j=1}^{k}\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}\left\{\frac{\frac{\partial}{\partial\lambda_j}\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}{\|\mathbb{E}_{q(\boldsymbol{\lambda}|\boldsymbol{\theta},\mathbf{x})}\mathbf{H}(\boldsymbol{\lambda};\boldsymbol{\theta},\mathbf{x})\|_{\mathrm{op}}^{1/2}}\right\}\right]^2,
\end{aligned}$$

where the second inequality follows from Jensen's inequality applied to the outer expectation and the square. Combining this with the above and summing completes the proof of the stronger inequality.

To prove the weaker bound, we apply the reverse Holder's inequality to the expectation of (C.15) under $q(\mathbf{x},\boldsymbol{\theta})$ with $p=2$ to get the lower bound

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\frac{\left[\mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\sum_{j=1}^{k}\frac{\partial}{\partial\lambda_j}\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})\right]^2}{\|\mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\mathbf{H}\left(\boldsymbol{\lambda};\boldsymbol{\theta},\mathbf{x}\right)\|_{\mathrm{op}}} \\
&\quad\geq \frac{\left[\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\left|\mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\sum_{j=1}^{k}\frac{\partial}{\partial\lambda_j}\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})\right|\right]^2}{\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta})}\|\mathbb{E}_{q(\boldsymbol{\lambda}|\mathbf{x},\boldsymbol{\theta})}\mathbf{H}\left(\boldsymbol{\lambda};\boldsymbol{\theta},\mathbf{x}\right)\|_{\mathrm{op}}} \\
&\quad\geq \frac{\left[\sum_{j=1}^{k}\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}\frac{\partial}{\partial\lambda_j}\frac{\partial}{\partial\theta_i}\log q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})\right]^2}{\mathbb{E}_{q(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda})}\|\mathbf{H}\left(\boldsymbol{\lambda};\boldsymbol{\theta},\mathbf{x}\right)\|_{\mathrm{op}}},
\end{aligned}$$

completing the proof. □

C.4.2 Conditional mutual information bounds

In this section we state and prove a theorem which lower bounds the conditional mutual information in terms of the Fisher information. To establish the lower bound, we will need to use the following key fact on the monotonicity of the conditional mutual information under conditioning on successive replications.

Lemma 8 (Decreasing Conditional Mutual Information)**.** *Let $q(\boldsymbol{\theta}, \mathbf{y})$ be a joint model of parameters and data. For any $M, R \geq 1$, let*

$$(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M+R)}) \overset{iid}{\sim} f(\mathbf{y} \mid \boldsymbol{\theta}).$$

Then we have that $\mathbf{I}(\boldsymbol{\theta}, \mathbf{y}^{(M)} \mid \mathbf{y}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M-1)})$ is decreasing in M. Furthermore, we have that

$$\mathbf{I}(\boldsymbol{\theta}, \mathbf{y}^{(1)} \mid \mathbf{y}) \geq \frac{1}{M} \mathbf{I}(\boldsymbol{\theta}, (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}) \mid \mathbf{y}^{(M+1)}, \ldots \mathbf{y}^{(M+R)})$$

for all $M \geq 1$.

Proof. Letting $\overline{\mathbf{y}}_M = \left(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}\right)$ and $\widetilde{\mathbf{y}}_R = \left(\mathbf{y}^{(M+1)}, \ldots, \mathbf{y}^{(M+R)}\right)$, we have that

$$\begin{aligned}
\mathbf{I}\left(\boldsymbol{\theta}, \overline{\mathbf{y}}_M \mid \widetilde{\mathbf{y}}_R\right) &= h\left(\overline{\mathbf{y}}_M \mid \widetilde{\mathbf{y}}_R\right) - h\left(\overline{\mathbf{y}}_M \mid \widetilde{\mathbf{y}}_R, \boldsymbol{\theta}\right) \\
&= h\left(\overline{\mathbf{y}}_M \mid \widetilde{\mathbf{y}}_R\right) - h\left(\overline{\mathbf{y}}_M \mid \boldsymbol{\theta}\right) && \text{(C.16)} \\
&\leq h\left(\overline{\mathbf{y}}_M \mid \mathbf{y}^{(M+1)}\right) - h\left(\overline{\mathbf{y}}_M \mid \boldsymbol{\theta}\right) && \text{(C.17)} \\
&= h\left(\overline{\mathbf{y}}_M \mid \mathbf{y}^{(M+1)}\right) - h\left(\overline{\mathbf{y}}_M \mid \boldsymbol{\theta}, \mathbf{y}^{(M+1)}\right) && \text{(C.18)} \\
&= \mathbf{I}\left(\boldsymbol{\theta}, \overline{\mathbf{y}}_M \mid \mathbf{y}^{(M+1)}\right),
\end{aligned}$$

where (C.16) follows from the conditional independence of the $\mathbf{y}^{(i)}$ given $\boldsymbol{\theta}$, (C.17) follows from the fact that conditioning decreases entropy, and (C.18) again follows in the same way as (C.16).

Now, by the chain rule for mutual information and the same argument as above, we can bound

this latter conditional mutual information as

$$\begin{aligned}\mathbf{I}\left(\boldsymbol{\theta}, \overline{\mathbf{y}}_M \mid \mathbf{y}^{(M+1)}\right) &= \sum_{j=1}^{M} \mathbf{I}\left(\boldsymbol{\theta}, \mathbf{y}^{(j)} \mid \mathbf{y}^{(M+1)}, \overline{\mathbf{y}}_{j-1}\right) \\ &\leq \sum_{j=1}^{M} \mathbf{I}\left(\boldsymbol{\theta}, \mathbf{y}^{(j)} \mid \mathbf{y}^{(M+1)}\right) \\ &= M\mathbf{I}\left(\boldsymbol{\theta}, \mathbf{y}^{(1)} \mid \mathbf{y}\right),\end{aligned}$$

where $\overline{\mathbf{y}}_0 = \{\}$, and where the last inequality follows from the fact that the distributions of $\left(\boldsymbol{\theta}, \mathbf{y}^{(j)}, \mathbf{y}^{(M+1)}\right)$ are all equal to the distribution of $\left(\boldsymbol{\theta}, \mathbf{y}^{(1)}, \mathbf{y}\right)$. Now dividing both sides by M completes the proof. □

Before stating our main lemma, we first give a definition that extends the notion of a subexponential distribution to random symmetric matrices.

Definition 8 (Subexponential Random Matrix)**.** *A random symmetric matrix* $\mathbf{M}$ *is said to be* (α, β)*-subexponential if, for all* $k \geq 2$*, we have*

$$\mathbb{E}\mathbf{M}^k \prec \alpha \frac{k!}{2} \beta^{k-2} \mathbf{I},$$

where the inequality denotes the Loewner order.

We will also say that a random matrix is γ-subexponential if it is (α, β)-subexponential for some α, β such that $\gamma = \alpha + \beta$.

Theorem 7 (Fisher Information Lower Bound)**.** *For* $M, R \geq 1$*, define the* $(M + R)$*-replicated model:*

$$q\left(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M+R)}, \boldsymbol{\theta}\right) = q(\boldsymbol{\theta}) \prod_{i=1}^{M+R} q\left(\mathbf{y}^{(i)} \mid \boldsymbol{\theta}\right).$$

Suppose for R sufficiently large, we have that

- *the posterior distributions* $q\left(\boldsymbol{\theta} \mid \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(R)}\right)$ *are normal,*

- *the observed information matrix of $q(\mathbf{y}, \boldsymbol{\theta})$ is (α, β)-subexponential for some $\alpha, \beta > 0$ (i.e. the observed information does not have heavy tails),*

- $\mathbb{E}\lambda_d^2(\boldsymbol{\Sigma})$, $\mathbb{E}\lambda_1^{-2}(\boldsymbol{\Sigma})$, $\mathbb{E}\lambda_d^2\left(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right)$, *and* $\mathbb{E}\lambda_1^{-2}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})))$ *are bounded by some $B > 0$ where* $\boldsymbol{\Sigma} = \mathrm{Cov}\left(\boldsymbol{\theta} \mid \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}\right)$ *(i.e. the posterior covariance and Fisher information are neither too small nor too large on average).*

Then for C a constant depending on $\gamma = \alpha + \beta$ and B, we have for M sufficiently large that

$$\frac{1}{M}\mathbf{I}(\boldsymbol{\theta}, \overline{\mathbf{y}}_M \mid \overline{\mathbf{y}}_R) \geq \frac{C}{\log d}\,\mathrm{tr}\left(\mathbb{E}_{q(\boldsymbol{\theta},\mathbf{y})}\boldsymbol{\Sigma}_{\overline{\mathbf{y}}_R}^{1/2}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\boldsymbol{\Sigma}_{\overline{\mathbf{y}}_R}^{1/2}\right). \tag{C.19}$$

We give the proof in the case $R = 1$ and write $\mathbf{y}$ for $\mathbf{y}^{(M+1)}$ for simplicity, but the proof for $R > 1$ follows in the exact same way with $\left(\mathbf{y}^{(M+1)}, \ldots, \mathbf{y}^{(M+R)}\right)$ replacing $\mathbf{y}$ throughout.

Proof. By the invariance of the mutual information under invertible transformations, if we define $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\boldsymbol{\theta}$, we get

$$\mathbf{I}(\boldsymbol{\theta}, (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}) \mid \mathbf{y}) = \mathbf{I}(\widetilde{\boldsymbol{\theta}}, (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}) \mid \mathbf{y}) = \frac{1}{2}\mathbb{E}\log\left[\det\widetilde{\boldsymbol{\Sigma}}^{-1}_{\mathbf{y},\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(M)}}\right], \tag{C.20}$$

where the last equality follows by expressing the conditional mutual information as the entropy difference $h\left(\widetilde{\boldsymbol{\theta}} \mid \mathbf{y}\right) - h\left(\widetilde{\boldsymbol{\theta}} \mid \mathbf{y}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}\right)$, plugging in the expressions for the entropy of multivariate normal distributions, and simplifying.

Now for $1 \leq i \leq d$, let $\lambda_i^d(\cdot) : H_d \to \mathbb{R}$ be the map from $d \times d$ Hermitian matrices to their i^{th} eigenvalue under increasing order. We will usually write λ_i for λ_i^d when the matrix dimension is clear from context. We will henceforth write $\widetilde{\boldsymbol{\Sigma}}_M^{-1}$ for $\widetilde{\boldsymbol{\Sigma}}^{-1}_{\mathbf{y},\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(M)}}$ for convenience.

Furthermore, define $\mathbf{I}_{\text{obs}}(\overline{\mathbf{y}}, \widetilde{\boldsymbol{\theta}}) = \frac{1}{M}\sum_{i=1}^{M}\mathbf{H}_{\phi(\widetilde{\boldsymbol{\theta}}|\mathbf{y}^{(i)})}(\widetilde{\boldsymbol{\theta}})$ where $\phi(\widetilde{\boldsymbol{\theta}} \mid \mathbf{y}^{(i)}) = -\log f(\mathbf{y}^{(i)} \mid \widetilde{\boldsymbol{\theta}})$, and

$$\mathcal{E}_{\widetilde{\boldsymbol{\theta}}} = \left\{\left\|\mathbf{I}_{\text{obs}}(\overline{\mathbf{y}}, \widetilde{\boldsymbol{\theta}}) - \boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})\right\|_{\text{op}} < \frac{\delta}{2}\right\}. \tag{C.21}$$

Finally, define the event

$$\mathcal{G} = \{\lambda_1 (\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})) \geq \delta\}.$$

The proof will now proceed by analyzing the decomposition

$$\begin{aligned}\mathbb{E} \log \left[\det \widetilde{\boldsymbol{\Sigma}}_M^{-1}\right] = \sum_{i=1}^{d} \mathbb{E}\left[\log\left(\lambda_i\left(\widetilde{\boldsymbol{\Sigma}}_M^{-1}\right)\right) \mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}} \cap \mathcal{G}}\right] + \sum_{i=1}^{d} \mathbb{E}\left[\log\left(\lambda_i\left(\widetilde{\boldsymbol{\Sigma}}_M^{-1}\right)\right) \mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c \cap \mathcal{G}}\right] \\ + \sum_{i=1}^{d} \mathbb{E}\left[\log\left(\lambda_i\left(\widetilde{\boldsymbol{\Sigma}}_M^{-1}\right)\right) \mathbb{1}_{\mathcal{G}^c}\right], \qquad \text{(C.22)}\end{aligned}$$

where, on the right hand side, we now take the expectations over $\left(\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}, \widetilde{\boldsymbol{\theta}}\right)$. Label these terms T1 - T3.

Lower bound for T3.

First we note that

$$\begin{aligned}\mathbb{E}\left[\log\left(\lambda_i\left(\widetilde{\boldsymbol{\Sigma}}_M^{-1}\right)\right) \mathbb{1}_{\mathcal{G}^c}\right] &\geq \mathbb{E}\left[\log\left(\lambda_1\left(\widetilde{\boldsymbol{\Sigma}}_M^{-1}\right)\right) \mathbb{1}_{\mathcal{G}^c}\right] \\ &= -\mathbb{E}\left[\log\left(\lambda_d\left(\widetilde{\boldsymbol{\Sigma}}_M\right)\right) \mathbb{1}_{\{\lambda_1^{-1}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})) > \delta^{-1}\}}\right] \\ &\geq -\sqrt{\mathbb{E}\log\left(\lambda_d\left(\widetilde{\boldsymbol{\Sigma}}_M\right)\right)^2} \sqrt{\mathbb{P}\left(\lambda_1^{-1}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})) \geq \delta^{-1}\right)} \\ &\geq -\delta^{1/2} \sqrt{\mathbb{E}\log\left(\lambda_d\left(\widetilde{\boldsymbol{\Sigma}}_M\right)\right)^2 \mathbb{E}\lambda_1^{-1}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}))} \\ &\geq -\delta^{1/2} \sqrt{\left(\mathbb{E}\lambda_d\left(\widetilde{\boldsymbol{\Sigma}}_M\right) + \mathbb{E}\lambda_1^{-1}\left(\widetilde{\boldsymbol{\Sigma}}_M\right)\right) \mathbb{E}\lambda_1^{-1}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}))} \\ &\geq -\sqrt{2}\delta^{1/2} B.\end{aligned}$$

So T3 is lower bounded by $-\sqrt{2} d \delta^{1/2} B$.

Lower Bound for T2.

We now lower bound T2. First observe that

$$\begin{aligned}\mathbb{E}\left[\log\left(\lambda_i\left(\widetilde{\mathbf{\Sigma}}_M^{-1}\right)\right)\mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\cap\mathcal{G}}\right] &\geq \mathbb{E}\left[\log\left(\lambda_1\left(\widetilde{\mathbf{\Sigma}}_M^{-1}\right)\right)\mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\cap\mathcal{G}}\right] \\ &= -\mathbb{E}\left[\log\left(\lambda_d\left(\widetilde{\mathbf{\Sigma}}_M\right)\right)\mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\cap\mathcal{G}}\right] \\ &\geq -\sqrt{\mathbb{E}\left[\log\left(\lambda_d\left(\widetilde{\mathbf{\Sigma}}_M\right)\right)^2\right]}\sqrt{\mathbb{P}\left(\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\cap\mathcal{G}\right)}\end{aligned}$$

Now we observe that we can upper bound the probability factor as

$$\mathbb{P}\left(\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\cap\mathcal{G}\right) = \mathbb{E}_{q(\boldsymbol{\theta},\mathbf{y})}\mathbb{P}\left(\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\cap\mathcal{G}\mid\widetilde{\boldsymbol{\theta}},\mathbf{y}\right)\leq\mathbb{E}_{q(\boldsymbol{\theta},\mathbf{y})}\mathbb{P}\left(\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\mid\widetilde{\boldsymbol{\theta}},\mathbf{y}\right).$$

Since we assume that the observed information matrices $\mathbf{H}_{\phi(\widetilde{\boldsymbol{\theta}}|\mathbf{y}^{(i)})}(\widetilde{\boldsymbol{\theta}})$ are sub-exponential in the sense of (8), we can apply the matrix Bernstein inequality conditional on $(\widetilde{\boldsymbol{\theta}},\mathbf{y})$ with rank of the Fisher information $\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})$ to see that, if $\delta\leq 1$,

$$\mathbb{P}\left(\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\mid\widetilde{\boldsymbol{\theta}}\right)\leq 2d\exp\left(-M\frac{\delta^2}{4\alpha+2\beta\delta}\right)\leq 2d\exp\left(-M\frac{\delta^2}{4\gamma}\right), \tag{C.23}$$

where α,β are the sub-exponential parameters controlling the tails of the observed information, $\gamma=\alpha+\beta$. Since the right hand side is free of $(\widetilde{\boldsymbol{\theta}},\mathbf{y})$, we obtain the same upper bound for the marginal probability $\mathbb{P}(\mathcal{E}_{\widetilde{\boldsymbol{\theta}}}^c\cap\mathcal{G})$. This gives an overall bound for T2 of $-\sqrt{2}Bd\exp\left(-M\frac{\delta^2}{8\gamma}\right)$.

Lower Bound for T1.

Now, for any Gaussian distribution, the Hessian of the potential function is exactly the precision matrix. Combining this with the fact that derivatives of the log normalizing constant vanish and

the fact that $\mathrm{Cov}\left(\widetilde{\boldsymbol{\theta}} \mid \mathbf{y}\right) = \mathsf{I}$, we have that, for all $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^d$,

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}^{-1}_{\mathbf{y}^{(0)},\ldots,\mathbf{y}^{(M)}} &= \mathbf{H}_{-\log q(\widetilde{\boldsymbol{\theta}}|\mathbf{y}^{(0)},\ldots,\mathbf{y}^{(M)})}(\widetilde{\boldsymbol{\theta}}) \\
&= \mathbf{H}_{-\log q(\widetilde{\boldsymbol{\theta}}|\mathbf{y})}(\widetilde{\boldsymbol{\theta}}) + \sum_{i=1}^{M} \mathbf{H}_{\phi(\widetilde{\boldsymbol{\theta}}|\mathbf{y}^{(i)})}(\widetilde{\boldsymbol{\theta}}) \\
&= \mathsf{I} + \sum_{i=1}^{M} \mathbf{H}_{\phi(\widetilde{\boldsymbol{\theta}}|\mathbf{y}^{(i)})}(\widetilde{\boldsymbol{\theta}}). \qquad \text{(C.24)}
\end{aligned}$$

Now note that

$$\lambda_i\left(\mathsf{I} + \sum_{i=1}^{M} \mathbf{H}_{\phi(\widetilde{\boldsymbol{\theta}}|\mathbf{y}^{(i)})}(\widetilde{\boldsymbol{\theta}})\right) = 1 + \lambda_i\left(\sum_{i=1}^{M} \mathbf{H}_{\phi(\widetilde{\boldsymbol{\theta}}|\mathbf{y}^{(i)})}(\widetilde{\boldsymbol{\theta}})\right) \qquad \text{(C.25)}$$

Combining the observations of (C.25) and (C.24), we obtain

$$\begin{aligned}
\sum_{i=1}^{d} \mathbb{E}\left[\log\left(\lambda_i\left(\widetilde{\boldsymbol{\Sigma}}_M^{-1}\right)\right) \mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}} \cap \mathcal{G}}\right] &= \frac{1}{2}\mathbb{E}\sum_{i=1}^{d} \log\left(1 + \lambda_i\left(\sum_{i=1}^{M} \mathbf{H}_{\phi(\widetilde{\boldsymbol{\theta}}|\mathbf{y}^{(i)})}(\widetilde{\boldsymbol{\theta}})\right)\right) \mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}} \cap \mathcal{G}} \\
&= \frac{1}{2}\sum_{i=1}^{d} \mathbb{E}\log\left(1 + M\lambda_i\left(\mathbf{I}_{\mathrm{obs}}(\overline{\mathbf{y}}, \widetilde{\boldsymbol{\theta}})\right)\right) \mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}} \cap \mathcal{G}}. \qquad \text{(C.26)}
\end{aligned}$$

Now Weyl's inequalities imply that

$$\left|\lambda_i\left(\mathbf{I}_{\mathrm{obs}}(\overline{\mathbf{y}}, \widetilde{\boldsymbol{\theta}})\right) - \lambda_i\left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})\right)\right| \leq \left|\mathbf{I}_{\mathrm{obs}}(\overline{\mathbf{y}}, \widetilde{\boldsymbol{\theta}}) - \boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})\right|_{\mathrm{op}}. \qquad \text{(C.27)}$$

Using this, we have for all $1 \leq i \leq M$ that

$$\begin{aligned}
&\mathbb{E}_{p(\overline{\mathbf{y}}|\widetilde{\boldsymbol{\theta}},\mathbf{y})}\left[\log\left(1 + M\lambda_i\left(\mathbf{I}_{\mathrm{obs}}(\overline{\mathbf{y}}, \widetilde{\boldsymbol{\theta}})\right)\right) \mathbb{1}_{\mathcal{E}_{\widetilde{\boldsymbol{\theta}}} \cap \mathcal{G}}\right] \\
&\qquad \geq \log\left(1 + \frac{M}{2}\lambda_i\left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})\right)\right) \mathbb{1}_{\mathcal{G}}(\widetilde{\boldsymbol{\theta}}) \left(1 - \mathbb{P}_{f(\overline{\mathbf{y}}|\widetilde{\boldsymbol{\theta}},\mathbf{y})}(\mathcal{E}^c_{\widetilde{\boldsymbol{\theta}}})\right) \qquad \text{(C.28)}
\end{aligned}$$

where the expectations are over $\overline{\mathbf{y}} = \left(\mathbf{y}^{(1)}, \ldots \mathbf{y}^{(M)}\right)$ conditional on $(\widetilde{\boldsymbol{\theta}}, \mathbf{y})$.

Now combining (C.28), (C.26), (C.20), and our lower bounds for T2 and T3, and letting

$$M = 8k\frac{\gamma}{\delta^2}\log(2d),$$

we obtain an overall lower bound for the conditional mutual information

$$\begin{aligned}
&\mathbf{I}(\widetilde{\boldsymbol{\theta}}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)} \mid \mathbf{y}) \\
&\qquad \geq \frac{1}{4} \sum_{i=1}^{d} \mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \left[\log \left(1 + \frac{4k\gamma}{\delta^2} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \right) \mathbb{1}_{\mathcal{G}}(\widetilde{\boldsymbol{\theta}}) \right] - cd \left[\delta^{1/2} + 2^{-k} \right]. \qquad \text{(C.29)}
\end{aligned}$$

If we now define the function

$$\sigma^{\tau}(x) = \min(x, \tau) \qquad \text{(C.30)}$$

to be the τ-truncation of x, then we have for δ^{-1} and τ sufficiently large that the first term above is lower bounded by

$$\begin{aligned}
&\mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \sum_{i=1}^{d} \log \left[1 + \frac{4k\gamma}{\delta^2} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \right] \mathbb{1}_{\mathcal{G}}(\widetilde{\boldsymbol{\theta}}) \\
&\qquad \geq \mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \sum_{i=1}^{d} \log \left[1 + \frac{4k\gamma}{\delta^2} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \right] \mathbb{1}_{\{\delta < \lambda_1(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) \leq \lambda_d(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) < \tau\}} \\
&\qquad \geq \frac{\log\left(1 + 4k\gamma\tau/\delta^2\right)}{\tau} \sum_{i=1}^{d} \mathbb{E}_{p(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \mathbb{1}_{\{\delta < \lambda_1(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) \leq \lambda_d(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) < \tau\}} \\
&\qquad \geq \frac{1}{\tau} \sum_{i=1}^{d} \mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \mathbb{1}_{\{\delta < \lambda_1(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) \leq \lambda_d(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) < \tau\}}, \qquad \text{(C.31)}
\end{aligned}$$

where we have used the fact that $\log(1+x) \geq \frac{\log(1+c)}{c} x$ over the interval $[0, c]$ for any $c \geq 0$. Now we note that

$$\begin{aligned}
\mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \mathbb{1}_{\{\delta < \lambda_1(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) \leq \lambda_d(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) < \tau\}} &\geq \mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \\
&\quad - \mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \mathbb{1}_{\{\lambda_1(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) \leq \delta\}} - \mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \lambda_i \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right) \mathbb{1}_{\{\lambda_d(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})) \geq \tau\}}.
\end{aligned}$$

Now by applying Holder's inequality and Markov's inequality, the second to last expectation can be upper bounded by

$$\sqrt{\mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \lambda_d^2 \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right)} \sqrt{\mathbb{E}_{q(\widetilde{\boldsymbol{\theta}}, \mathbf{y})} \lambda_1^{-1} \left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) \right)} \delta^{1/2} \leq B\delta^{1/2}.$$

Similarly, the last expectation is upper bounded by $B\tau^{-1/2}$. Thus, the terms of our lower bound (C.31) can be further lower bounded as

$$\frac{1}{\tau}\mathbb{E}_{q(\widetilde{\boldsymbol{\theta}},\mathbf{y})}\lambda_i\left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})\right) - C\left(\tau^{-3/2} + \tau^{-1}\delta^{1/2} + 2^{-k}\right),$$

where we can take $C = \max(c, B)$. By Jensen's inequality, we also have a lower bound $\mathbb{E}_{q(\widetilde{\boldsymbol{\theta}},\mathbf{y})}\lambda_i \geq B^{-1}$. Taking τ, δ^{-1}, and k sufficiently large, we can ensure that the second term above is bounded by $B^{-1}/2\tau$, which yields an overall lower bound of

$$\frac{1}{2B\tau}\sum_{i=1}^{d}\mathbb{E}_{q(\widetilde{\boldsymbol{\theta}})}\lambda_i\left(\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})\right)$$

Putting this all together, we get that

$$\frac{1}{M}\mathbf{I}(\widetilde{\boldsymbol{\theta}}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)} \mid \mathbf{y}) \geq \frac{\delta^2}{16Bk\tau\gamma\log(2d)}\operatorname{tr}\left[\mathbb{E}\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}})\right], \tag{C.32}$$

which concludes the proof upon noting that

$$\boldsymbol{\mathcal{I}}(\widetilde{\boldsymbol{\theta}}) = \boldsymbol{\Sigma}_{\mathbf{y}}^{1/2}\boldsymbol{\mathcal{I}}(\theta)\boldsymbol{\Sigma}_{\mathbf{y}}^{1/2}.$$

□

C.4.3 Proof of theorem 4

Using the results of the last two sections, we can now proceed with the proof of Theorem 4. We first prove a pair of lemmas that give conditions under which we can further lower bound the cmi in terms of just the Fisher information. The key idea here is to exploit the Cramer-Rao bound to relate the posterior covariance matrix that appears in Theorem 7 to the Fisher information. The first step is to determine conditions under which we can get a lower bound on the cmi with the matrix product and expectation operations interchanged. To illustrate the idea, observe that for

$d = 1$, we have that

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_{\overline{\mathbf{y}}}\boldsymbol{\mathcal{I}}(\theta)\right) &= \mathbb{E}[\mathrm{Var}(\theta \mid \overline{\mathbf{y}})\,\boldsymbol{\mathcal{I}}(\theta)] \\ &= \mathrm{Cov}\left(\mathrm{Var}(\theta \mid \overline{\mathbf{y}}), \boldsymbol{\mathcal{I}}(\theta)\right) + \mathbb{E}\mathrm{Var}(\theta \mid \overline{\mathbf{y}})\mathbb{E}\boldsymbol{\mathcal{I}}(\theta) \\ &\geq \mathbb{E}\mathrm{Var}(\theta \mid \overline{\mathbf{y}})\mathbb{E}\boldsymbol{\mathcal{I}}(\theta) - \sqrt{\mathrm{Var}\left[\mathrm{Var}(\theta \mid \overline{\mathbf{y}})\right]\mathrm{Var}\left[\boldsymbol{\mathcal{I}}(\theta)\right]}.\end{aligned}$$

Thus, in the one-dimensional case, if $\sqrt{\mathrm{Var}\left[\mathrm{Var}(\theta \mid \overline{\mathbf{y}})\right]} < \delta\mathbb{E}\left[\mathrm{Var}(\theta \mid \overline{\mathbf{y}})\right]$ and $\sqrt{\mathrm{Var}\left[\boldsymbol{\mathcal{I}}(\theta)\right]} < \delta\mathbb{E}\left[\boldsymbol{\mathcal{I}}(\theta)\right]$ for some $\delta \in (0, 1)$, then we have that

$$\mathbb{E}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_{\overline{\mathbf{y}}}\boldsymbol{\mathcal{I}}(\theta)\right) \geq (1-\delta^2)\mathbb{E}\mathrm{Var}(\theta \mid \overline{\mathbf{y}})\mathbb{E}\boldsymbol{\mathcal{I}}(\theta) = (1-\delta^2)\,\mathrm{tr}\left(\mathbb{E}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbb{E}\boldsymbol{\mathcal{I}}(\theta)\right),$$

For a positive random variable X, the requirement that $\sqrt{\mathrm{Var}(X)} < \delta\mathbb{E}X$ is not a restriction on the variance X in an absolute sense, since the condition can be satisfied for distributions with arbitrarily large variances so long as the mean is correspondingly large. Rather, we argue that this is naturally seen as a condition on the skewness of X. Indeed, for positive random variables X with finite second moment, we have in general that

$$\mathbb{P}\left(X \geq (k+1)\sqrt{\mathbb{E}X^2}\right) \leq \frac{1}{k^2} \quad \text{and} \quad \mathbb{P}\left(X \geq k\mathbb{E}X\right) \leq \frac{1}{k}.$$

The above variance-mean inequality implies that

$$\mathbb{P}\left(X \geq (k+1)\mathbb{E}X\right) \leq \frac{\delta^2}{k^2}.$$

Since $\mathbb{P}(X \geq k\mathbb{E}X)$ is large when the distribution of X is skewed to the right, this variance-mean inequality primarily functions to limit the skew of the distribution of X. This skewness condition can be naturally generalized to the matrix case with $d \geq 1$, yielding the following lemma.

Lemma 9. *Under the conditions of Theorem 7 and Lemma 10, with relevant definitions taken from*

the same, we have that if

$$\sqrt{\mathrm{Var}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}))} < \delta\lambda_{\min}\left(\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right) \quad \textit{and} \quad \sqrt{\mathrm{Var}(\boldsymbol{\Sigma}_{\overline{\mathbf{y}}})} < \delta\lambda_{\min}\left(\mathbb{E}\boldsymbol{\Sigma}_{\overline{\mathbf{y}}}\right),$$

for some $\delta \in \left[0, 2^{-1/2}\right)$*, then we have that*

$$\mathrm{tr}\left(\mathbb{E}\boldsymbol{\Sigma}_{\overline{\mathbf{y}}}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right) \geq (1-\delta^2)\,\mathrm{tr}\left(\mathbb{E}\boldsymbol{\Sigma}_{\overline{\mathbf{y}}}\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right)$$

Proof. Our proof is given in terms of general positive-definite matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$. First observe that

$$\mathbb{E}\,\mathrm{tr}\left((\mathbf{A}-\mathbb{E}\mathbf{A})(\mathbf{B}-\mathbb{E}\mathbf{B})\right) = \mathrm{tr}\left(\mathbb{E}\mathbf{A}\mathbf{B}\right) - \mathrm{tr}\left(\mathbb{E}\mathbf{A}\mathbb{E}\mathbf{B}\right).$$

Therefore, it follows directly from von Neumann's trace inequalities that

$$\begin{aligned}
\left|\mathrm{tr}\left(\mathbb{E}\mathbf{A}\mathbf{B}\right) - \mathrm{tr}\left(\mathbb{E}\mathbf{A}\mathbb{E}\mathbf{B}\right)\right| &= \left|\mathbb{E}\,\mathrm{tr}\left((\mathbf{A}-\mathbb{E}\mathbf{A})(\mathbf{B}-\mathbb{E}\mathbf{B})\right)\right| \\
&\leq \mathbb{E}\left|\mathrm{tr}\left((\mathbf{A}-\mathbb{E}\mathbf{A})(\mathbf{B}-\mathbb{E}\mathbf{B})\right)\right| \\
&\leq d\left[\mathbb{E}||\mathbf{A}-\mathbb{E}\mathbf{A}||_{\mathrm{op}}||\mathbf{B}-\mathbb{E}\mathbf{B}||_{\mathrm{op}}\right] \\
&= d\Big[\mathrm{Cov}\left(||\mathbf{A}-\mathbb{E}\mathbf{A}||_{\mathrm{op}}, ||\mathbf{B}-\mathbb{E}\mathbf{B}||_{\mathrm{op}}\right) \\
&\qquad + \mathbb{E}||\mathbf{A}-\mathbb{E}\mathbf{A}||_{\mathrm{op}}\mathbb{E}||\mathbf{B}-\mathbb{E}\mathbf{B}||_{\mathrm{op}}\Big] \\
&\leq d\Big[\sqrt{\mathrm{Var}\left(||\mathbf{A}-\mathbb{E}\mathbf{A}||_{\mathrm{op}}\right)}\sqrt{\mathrm{Var}\left(||\mathbf{B}-\mathbb{E}\mathbf{B}||_{\mathrm{op}}\right)} \\
&\qquad + \mathbb{E}||\mathbf{A}-\mathbb{E}\mathbf{A}||_{\mathrm{op}}\mathbb{E}||\mathbf{B}-\mathbb{E}\mathbf{B}||_{\mathrm{op}}\Big] \\
&\leq 2d\sqrt{\mathbb{E}||\mathbf{A}-\mathbb{E}\mathbf{A}||_{\mathrm{op}}^2\mathbb{E}||\mathbf{B}-\mathbb{E}\mathbf{B}||_{\mathrm{op}}^2}.
\end{aligned}$$

It also follows from von Neumann's trace inequalities that

$$\mathrm{tr}\left(\mathbb{E}\mathbf{A}\mathbb{E}\mathbf{B}\right) \geq d\lambda_{\min}\left(\mathbb{E}\mathbf{A}\right)\lambda_{\min}\left(\mathbb{E}\mathbf{B}\right).$$

Letting $\text{Var}(\mathbf{A}) = \mathbb{E}||\mathbf{A} - \mathbb{E}\mathbf{A}||_{\text{op}}^2$, it follows that if there is a $\delta^2 \in (0,1)$ such that

$$\sqrt{\text{Var}(\mathbf{A})}\sqrt{\text{Var}(\mathbf{B})} < \frac{\delta}{2}\lambda_{\min}(\mathbb{E}\mathbf{A})\,\lambda_{\min}(\mathbb{E}\mathbf{B}),$$

then we have that $\text{tr}(\mathbb{E}\mathbf{A}\mathbf{B}) \geq (1-\delta^2)\,\text{tr}(\mathbb{E}\mathbf{A}\mathbb{E}\mathbf{B})$. Of course, it further suffices that there is some $\delta \in \left(0, \frac{1}{\sqrt{2}}\right)$ such that

$$\sqrt{\text{Var}(\mathbf{A})} < \delta\lambda_{\min}(\mathbb{E}\mathbf{A}) \quad \text{and} \quad \sqrt{\text{Var}(\mathbf{B})} < \delta\lambda_{\min}(\mathbb{E}\mathbf{B}),$$

which is guaranteed by our skewness conditions. □

Relative to the one-dimensional case, we note two clear defects of this result. Specifically, the requirement that $\delta < 2^{-1/2}$ rather than $\delta < 1$ and the fact that the variance is given in terms of the maximum singular value whereas the corresponding mean matrix is measured in terms of its minimal singular value make this bound more stringent than in the scalar case. Nevertheless, the qualitative requirement is essentially the same - that the distributions of the spectra of the posterior covariance and Fisher information are not too skewed.

Now we can use this lemma along with the Cramer-Rao bound to get a lower bound in terms of (only) the expected Fisher information.

Lemma 10. *If $\mathbb{E}_{q(\boldsymbol{\theta})}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ has spectrum $\{\iota_i\}_{i=1}^d$ and $p(\boldsymbol{\theta}) = \text{normal}(\boldsymbol{\theta}|\boldsymbol{\mu}, \mathsf{I})$, then*

$$\text{tr}\left([\mathbb{E}\boldsymbol{\Sigma}_{\bar{\mathbf{y}}_R}]^{1/2}[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})][\mathbb{E}\boldsymbol{\Sigma}_{\bar{\mathbf{y}}_R}]^{1/2}\right) \geq \sum_{i=1}^{d}\frac{\iota_i}{1+R\iota_i}, \tag{C.33}$$

Proof. We give the proof for general $R \geq 1$. Let $\mathbf{J}_{\boldsymbol{\theta},\mathbf{y}}$ be the observed information matrix. Then we note that, by the Cramer-Rao bound, Jensen's inequality, and our assumptions on the prior, we

have that

$$\begin{aligned}
\mathbb{E}\boldsymbol{\Sigma}_{\overline{\mathbf{y}}_R} &\succ \mathbb{E}_{f(\overline{\mathbf{y}})}\left[\mathbb{E}_{q(\boldsymbol{\theta}|\overline{\mathbf{y}})}\left\{\sum_{j=1}^{R}\mathbf{J}_{\boldsymbol{\theta},\mathbf{y}_j}+\mathbf{I}\right\}\right]^{-1} \\
&\succ \left[\mathbb{E}_{q(\boldsymbol{\theta},\overline{\mathbf{y}})}\left\{\sum_{j=1}^{R}\mathbf{J}_{\boldsymbol{\theta},\mathbf{y}_j}+\mathbf{I}\right\}\right]^{-1} \\
&= \left[R\mathbb{E}_{q(\boldsymbol{\theta})}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})+\mathbf{I}\right]^{-1},
\end{aligned}$$

where the inequalities represent the Loewner partial ordering of PSD matrices whereby $\mathbf{A} \geq \mathbf{B}$ iff $\mathbf{A} - \mathbf{B}$ is PSD. Using the fact that the trace is symmetric and Loewner order is preserved under conjugation by any other positive definite matrix, we have that

$$\begin{aligned}
\operatorname{tr}\left(\left[\mathbb{E}\boldsymbol{\Sigma}_{\overline{\mathbf{y}}}\right]^{1/2}\left[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]\left[\mathbb{E}\boldsymbol{\Sigma}_{\overline{\mathbf{y}}}\right]^{1/2}\right) &= \operatorname{tr}\left(\left[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]^{1/2}\left[\mathbb{E}\boldsymbol{\Sigma}_{\overline{\mathbf{y}}}\right]\left[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]^{1/2}\right) \\
&\geq \operatorname{tr}\left(\left[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]^{1/2}\left[R\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})+\mathbf{I}\right]^{-1}\left[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]^{1/2}\right) \\
&= \operatorname{tr}\left(\left[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]\left[R\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})+\mathbf{I}\right]^{-1}\right).
\end{aligned}$$

Now, writing the spectral decomposition of $\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ as $\mathbf{U}^T\boldsymbol{\Lambda}\mathbf{U}$, where $\boldsymbol{\Lambda} = \operatorname{diag}(\iota_1, \ldots, \iota_d)$, and substituting this into the above, we get that

$$\begin{aligned}
\operatorname{tr}\left(\left[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]\left[R\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})+\mathbf{I}\right]^{-1}\right) &= \operatorname{tr}\left(\mathbf{U}^T\boldsymbol{\Lambda}\left[R\boldsymbol{\Lambda}+\mathbf{I}\right]^{-1}\mathbf{U}\right) \\
&= \operatorname{tr}\left(\boldsymbol{\Lambda}\left[R\boldsymbol{\Lambda}+\mathbf{I}\right]^{-1}\right) \\
&= \sum_{i=1}^{d}\frac{\iota_i}{1+R\iota_i},
\end{aligned}$$

which completes the proof. □

We make a few observations about this result:

- Since the trace in (C.33) is invariant under orthogonal transformations of $\boldsymbol{\theta}$, the unit covariance assumption only imposes that the parameters have unit prior scale, which can always

be achieved by rescaling.

- The normality assumption may be relaxed by replacing the identity matrix by the Hessian $\mathbf{H}$ of $-\log p(\boldsymbol{\theta})$ in the proof. If $\mathbb{E}_{q(\boldsymbol{\theta})}\mathbf{H} \prec c\mathbf{I}$, then we get

$$\operatorname{tr}\left([\mathbb{E}\boldsymbol{\Sigma}_{\mathbf{y}}]^{1/2}[\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})][\mathbb{E}\boldsymbol{\Sigma}_{\mathbf{y}}]^{1/2}\right) \geq \sum_{i=1}^{d} \frac{\iota_i}{c+\iota_i}.$$

- If $R > 1$, we get a similar lower bound with terms $\iota_i/(1+R\iota_i)$.

We also note that this lower bound shares many features with the cmi: (i) a dimension dependence through the sum of d terms, (ii) dependence on the likelihood curvature through the eigenvalues ι_i, and (iii) a self-limiting behavior since increasing ι_i are offset by the decreasing curvature of $\frac{x}{1+x}$.

Finally we can prove Theorem 4.

Proof. The bound on $\mathbf{I}_{\text{base}}(\boldsymbol{\theta}, \mathbf{y})$ follows directly from Theorem 5 taking $\psi_1 = d\psi(d^{-1}\cdot)$ (noting that d is a constant here as it is the dimension of the shared parameters θ). Likewise, the bound on $\mathbf{I}(\boldsymbol{\theta}, \mathbf{y})$ follows from Theorem 5 and Theorem 6 with the same ψ_1 and letting $\boldsymbol{\Delta}^i = (\Delta_1, \ldots, \Delta_d)$. Taking $\psi_2(x) = (1-\delta^2)C\frac{x}{1+Rx}$, the bound on $\mathbf{I}_{\text{base}}(\mathbf{y}_{\text{rep}}, \boldsymbol{\theta} \mid \mathbf{y})$ follows from Theorem 7 along with Lemmas 9 and 10. Letting $\{\iota_i^{\text{exp}}\}_{i=1}^{d^{\text{exp}}}$ be the eigenvalues of $\mathbb{E}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$, the same argument gives the lower bound

$$\mathbf{I}\left(\mathbf{y}_{\text{rep}}, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \mid \mathbf{y}\right) \geq \sum_{i=1}^{d_{\text{exp}}} \psi_2\left(\iota_i^{\text{exp}}\right).$$

To get the bound in terms of the ι_i^{cond}, we observe that

$$\sum_{i=1}^{d} \psi_2\left(\iota_i^{\text{cond}}\right) \leq \sum_{i=1+d^{\text{exp}}-d}^{d^{\text{exp}}} \psi_2\left(\iota_i^{\text{exp}}\right) \leq \sum_{i=1}^{d^{\text{exp}}} \psi_2\left(\iota_i^{\text{exp}}\right),$$

where the first inequality follows from the eigenvalue interlacing theorem and the fact that ψ_2 is

increasing. Thus, the bound on $\mathbf{I}(\mathbf{y}_{\text{rep}}, \boldsymbol{\theta} \mid \mathbf{y})$ follows with

$$\Delta^f = \sum_{i=1}^{d^{\text{exp}}} \psi_2\left(\iota_i^{\text{exp}}\right) - \sum_{i=1}^{d} \psi_2\left(\iota_i^{\text{cond}}\right).$$

The final inequalities of Theorem 4 follow immediately from the above and the fact that the assumed inequality in Loewner order implies the same inequality with respect to the eigenvalues. □