Topics on Machine Learning under Imperfect Supervision

Gan Yuan

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
Abstract
Topics on Machine Learning under Imperfect Supervision
Gan Yuan

This dissertation comprises several studies addressing supervised learning problems where the supervision is imperfect.

Firstly, we investigate the margin conditions in active learning. Active learning is characterized by its special mechanism where the learner can sample freely over the feature space and exploit mostly of the limited labelling budget by querying the most informative labels. Our primary focus is to discern critical conditions under which certain active learning algorithms can outperform the optimal passive learning minimax rate. Within a non-parametric multi-class classification framework, our results reveal that the uniqueness of Bayes labels across the feature space serves as the pivotal determinant for the superiority of active learning over passive learning.

Secondly, we study the estimation of central mean subspace (CMS), and its application in transfer learning. We show that a fast parametric convergence rate of form $C_d \cdot n^{-1/2}$ is achievable via estimating the expected smoothed gradient outer product, for a general class of covariate distribution that admits Gaussian or heavier distributions. When the link function is a polynomial with a degree of at most $r$ and the covariates follows the standard Gaussian, we show that the prefactor depends on the ambient dimension $d$ as $C_d \propto d^r$. Furthermore, we show that under a transfer learning setting, an oracle rate of prediction error as if the CMS is known is achievable, when the source training data is abundant.

Finally, we present an innovative application involving the utilization of weak (noisy) labels
for addressing an Individual Tree Crown (ITC) segmentation challenge. Here, the objective is to
delineate individual tree crowns within a 3D LiDAR scan of tropical forests, with only 2D noisy
manual delineations of crowns on RGB images available as a source of weak supervision. We
propose a refinement algorithm designed to enhance the performance of existing unsupervised
learning methodologies for the ITC segmentation problem.
# Table of Contents

Acknowledgments ................................................. vi

Chapter 1: Introduction .......................................... 1
  1.1 Machine Learning with Imperfect Supervision ............... 1
  1.2 The Organization of This Dissertation ....................... 3
  1.3 Related Publications ....................................... 5

Chapter 2: Critical Regimes in Multi-class Active Learning .... 6
  2.1 Introduction .................................................. 6
  2.2 Problem Setting ............................................. 8
    2.2.1 Active Learning ........................................... 9
    2.2.2 Margin Condition .......................................... 10
  2.3 No Gain under Tsybakov’s Margin Condition ................. 10
  2.4 Gain under a Refined Margin Condition ..................... 13
    2.4.1 An Adaptive Procedure ................................... 15
    2.4.2 Rates Under Strong Density Condition ................. 17
    2.4.3 Rates for General Densities ............................. 18
  2.5 Proofs of Main Theorems .................................... 18
    2.5.1 Proof of Theorem 2.1 ..................................... 18
Chapter 4: Individual Tree Crown Delineation in Tropical Rain Forest

4.1 Introduction

4.2 Experiment Site and Data sets

4.2.1 Luquillo Forest Dynamic Plot

4.2.2 Remote Sensing Data

4.2.3 Manual Delineation of Crowns

4.3 Methods

4.3.1 Incorrect segmentation detection

4.3.2 Crown merging and splitting

4.4 Implementation Details and Results

4.4.1 Baseline 3D LiDAR segmentation

4.4.2 Refinement algorithm

4.4.3 Performance Evaluation

References

References
List of Figures

1.1 An illustration of three types of weakly supervised learning .................. 2
1.2 An illustration of transfer/multi-task learning ................................. 2

2.1 Different types of margin over space ........................................ 13
2.2 Construction of lower-bound .................................................. 20
2.3 Regression functions with non-unique Bayes classes in simulation design .... 40
2.4 Excess risk of the active learner and passive learner versus the sampling budget 41
2.5 The performance ratio of the active learner and passive learner versus $\varepsilon_0$ .... 42

3.1 The subspace estimation error $d(\hat{U}, U)$ v.s. sampling budget $n$ ........... 61
3.2 The subspace estimation error $d(\hat{U}, U)$ v.s. the choice of $m$ ............... 62
3.3 An example plot of the link function $f$ for which SAVE is not exhaustive. ..... 64

4.1 The Map of the El Yunque rainforest ....................................... 97
4.2 Illustration of shadow removal ............................................... 99
4.3 Illustration of the use of field measurement for aiding manual segmentation ... 100
4.4 Examples of manual delineation on the RGB imagery ........................... 101
4.5 Overall Workflow of crown segmentation .................................. 102
4.6 ROC curves for under-segmentation and over-segmentation detection ......... 106
4.7 Similarity metrics: boundary similarity and Jarccard index .................... 107
4.8 Histogram of absolute crown-stem distances and distance rankings .......... 110
4.9 Allometric plots by species .................................................. 110
List of Tables

1.1 Comparison of minimax rates under different regimes .................... 4
4.1 Under- and over-segmentation detection results ............................. 106
4.2 Percentage of predicted crowns with a matched manual crown .......... 108
4.3 Percentage of manual crowns with a matched predicted crown ............ 108
4.4 Number of stems used to derive species allometries for the target species . 109
Acknowledgements

First of all, I would like to express my deepest gratitude to my PhD supervisors, Professor Samory Kpotufe and Professor Tian Zheng. This dissertation would not be completed without their invaluable guidance and suggestions. Their unwavering support and encouragement have been, and will continue to be, indispensable throughout my academic journey. It is a profound honor to have been mentored by two exemplary scholars.

I am also immensely grateful to Professor Bodhisattva Sen, Professor Cynthia Rush and Professor Daniel Hsu for serving on my dissertation committee and providing enlightening feedback and sharp insights. I extend heartfelt thanks to Professor Daniel Hsu, Professor Maria Uriarte and Doctor Antonio Ferraz, for their exceptional mentorship during our collaboration, which has broadened my intellectual horizons and introduced me to exciting interdisciplinary fields beyond Statistics. To my friends and collaborators, Roi Ankori-Karlinsky, Dingyi Fang, Chengliang Tang, Mingyue Xu, Yunfan Zhao among many others, I express sincere appreciation for their intellectual support and the countless memorable moments we have shared during my time at Columbia University.

Last but not least, special thanks to my families and friends, without whom I would not be the person who I am today. To my parents, your unconditional love and unwavering belief in me makes me stronger. To my beloved girlfriend, thank you for your boundless love and support.
Chapter 1: Introduction

1.1 Machine Learning with Imperfect Supervision

The past two decades have witnessed the rapid expansion and success of machine learning. From self-driving cars [1] to virtual assistants [2], complicated machine learning models such as deep neural networks (DNNs) have demonstrated their potential for leveraging today’s big data and affordable high-performance computational resources in producing predictions that are comparable to human performance. However, training such algorithms to attain impressive performance relies heavily on a large volume of training data with high-quality labels, which are often expensive or even unavailable in many real-world applications. In particular, such perfect supervision becomes substantially scarcer in application domains that are more specialized, such as healthcare [3] and ecological studies [4, 5], where domain expertise is vital in data labeling. As a result, practical challenges due to the lack of perfect supervision in many real-world applications significantly limit the applicability and generalization of machine learning models.

Much efforts has been devoted trying to address the issue of the scarcity of perfect training data. Here are some examples.

Example 1.1 (Weakly Supervised Learning). Weakly supervised learning (WSL)[6, 7] refers to the construction of predictive models under the scenario when supervision is available but is somewhat “weak”. Roughly speaking, the weakness here can be categorized into three types: incomplete supervision, inexact supervision, and inaccurate supervision. See Figure 1.1 adapted from [6] for an illustration. The study of WSL expands the reach of conventional supervised learning and has garnered a lot of interests in applications.

Example 1.2 (Transfer Learning and Domain Adaption). Transfer learning [8, 9, 10, 11] are techniques of transferring the knowledge across domains. They aim to improve the predictive
Figure 1.1: An illustration of three types of weakly supervised learning [6]

Figure 1.2: An illustration of transfer/multi-task learning [8]
performance of in the target domain by exploiting the well-labelled training data from one or more source domains, especially when training data in the target task is scarce. The rough idea of transfer learning can be illustrated by Figure 1.2.

1.2 The Organization of This Dissertation

In this dissertation, I will present some of my works during my PhD career that are related to machine learning techniques with imperfect supervision.

In Chapter 2, we consider a non-parametric multi-class active learning problem. Active learning (see Section 2.2.1 for a formal definition) refers to the setting of machine learning where given a sampling budget $n$, the learner can sample sequentially at any part of in the feature space. It is generally believed to fall in the category of incomplete supervision under the larger umbrella of weak supervision [6], since only a fixed number of queries can be made within a large pool of unlabelled data. Compared to the passive learning setting where the learner only has access to i.i.d. training samples, active learner can often provide better predictive performance under the same sampling budget. In our works [12, 13], we strive to answer the following question:

What is the key condition that determines whether the minimax rate of the active learners is strictly faster than that of the passive learners?

We show that the whether such gain in active learning exists has to do with the uniqueness of Bayes classifiers over the space. On one hand, we show that the active minimax rate matches the passive one for a class of smooth distributions without forcing the uniqueness of Bayes classifiers over the space. On the other hand, we integrate the uniqueness of Bayes classification into a novel notion of margin condition, namely the refined margin condition (see Definition 2.10). We show that under this more restricted condition, there exists an active learning algorithm whose worst-case rate is faster than the passive minimax rate. In summary, the comparison of worst-case performance for passive and active learners can be presented as in Table 1.1.
<table>
<thead>
<tr>
<th>Margin Condition (i)</th>
<th>Margin Condition (ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearly Uniform $P_X$</td>
<td>$\mathcal{E}_A \ll \mathcal{E}_P$ (Theorem 2.3)</td>
</tr>
<tr>
<td>General $P_X$</td>
<td>$\mathcal{E}_A \ll \mathcal{E}_P$ (Theorem 2.1)</td>
</tr>
</tbody>
</table>

Table 1.1: Comparison of minimax rates for active learning and passive learning under different regimes. Here, $\mathcal{E}_A$ and $\mathcal{E}_P$ are the active and passive minimax excess risk rates, respectively.

In Chapter 3, we propose an efficient estimator for the central mean subspace. Note that the estimation of such lower-dimensional subspace allows for efficient transfer learning (see Section 3.6.4 for related discussion). In particular, We consider the following multi-index model for $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$:

$$
\mathbb{E}[Y \mid X] = f(U^T X);
$$

where $U \in \mathbb{R}^{d \times k}$ with $k \ll d$ is some unknown orthogonal matrix, and $f: \mathbb{R}^k \rightarrow \mathbb{R}$ are some unknown link functions. Our target is to learn the subspace $\text{col}(U)$ both statistically efficient\(^1\) and computationally tractable\(^2\). For a general class of distributions $P_X$ admitting Gaussians and heavier tail distributions, our estimator converges at a parametric rate of the form $C_d \cdot n^{-1/2}$, where $C_d$ is a constant related to the data dimension $d$. To elucidate the leading constant $C_d$, we consider the special case of polynomial links $f$—which has received much recent interest—and show that it behaves as $d^{O(r)}$ for polynomials of degree at most $r$ (Corollary 3.7), in line with recent lower-bounds of [14] for certain classes of computationally tractable procedures.

In Chapter 4, we will present the use of weakly supervised learning and transfer learning through the task of individual tree crown (ITC) delineation in a tropical rainforest in Puerto Rico. In particular, our task is to delineate the 3D point cloud representing a laser scanning of the rainforest into individual crowns. The first challenge for the ITC delineation is that there are no ground truth labels. Therefore, most works (e.g., [15, 16]) in the literature adopt unsupervised approaches—such as clustering algorithms—to get the individual crowns. As the first contribution, we create a manual delineation of individual crowns on the RGB imagery for a 16-ha plot, namely the Luquillo Forest Dynamics Plot (LFDP). Note that such human annotation is still a weak form of supervi-

\(^1\) $d(\hat{U}, U) = O(n^{-1/2})$ for sample size $n$ and some notion of distance $d$ between subspaces.

\(^2\) The algorithm runs in polynomial time.
sion, since our target is to delineate crowns from the 3D point cloud, but the supervision is on the 2D images. Using the manual delineation as a source of weak supervision, we then propose a weakly supervised refinement over the state-of-the-art AMD3D algorithm in the ITC delineation literature[16]. Finally, we quantify the improvement of the segmentation performance through:

- a congruence analysis that compares the segmentation to the manual delineation;
- a paired-tree analysis that compares the segmentation with the ground-level observations.

1.3 Related Publications

Parts of this dissertation appears in the published or submitted works listed below:

- Results in Chapter 2 appear in:
  
  Yuan G., Zhao Y., Kpotufe S., (2024+), Regimes of No Gain in Multi-class Active Learning. *Journal of Machine Learning Research*, Accepted.;


- Portion of Results in Chapter 3 appear in:


- Portion of Results in Chapter 4 appear in:

2.1 Introduction

We consider the setting of nonparametric classification with smooth regression functions, i.e., one where \( \eta_y(x) = \mathbb{P}(Y = y | X = x) \) is \( \alpha \)- Hölder continuous for every label \( y \in [L] \), where \( [L] = \{1, \ldots, L\} \). Two main notions of margin have appeared interchangeably in passive learning in this setting; assume for now, for simplicity that \( y = 1 \) or \( 2 \):

(i) \( \mathbb{P}(|\eta_1 - \eta_2| \leq \tau) \lesssim \tau^\beta \),  (ii) \( \mathbb{P}(0 < |\eta_1 - \eta_2| \leq \tau) \lesssim \tau^\beta \),

for some margin parameter \( \beta > 0 \). Both definitions are termed Tsybakov’s low noise or margin condition without distinction in the literature (e.g., [17, 18] for (i), and [19] for (ii)). However, excluding 0 as in (ii) is more natural since any classifier \( \hat{h} \) has the same error as Bayes in those regions where \( \eta_1 = \eta_2 \), i.e., where the Bayes classifier is not unique. On the other hand, (i) implies uniqueness (up to measure 0) of the Bayes classes, as seen by letting \( \tau \to 0 \). As such, (ii) admits more general settings with non-unique Bayes classes, and is thus preferred in the seminal result of [19] on margins in nonparametrics. More generally, in multi-class, the distinction is whether we view the margin as 0 if the Bayes classifier is not unique (corresponding to (i)), or consider the margin between unique values of \( \{\eta_y\}_{y \in [L]} \) (corresponding to (ii), and which seems more natural).

Interestingly, using (i) or (ii), the minimax risk is the same in passive learning, e.g. when \( P_X \) is uniform, the minimax risk is \( O(n^{-\alpha(\beta+1)/(2\alpha+d)}) \) see [19]. However, as we show, a sharp distinction emerges in active learning, where condition (ii) leads to two regimes in terms of savings:

• Under the common strong density assumption, relaxing uniform \( P_X \), no active learner can achieve a better rate—beyond constants—than the minimax passive rate (Theorem 2.1). In
contrast, as first shown in [18], condition (i) always leads to strictly faster rates than passive. While this was shown for binary classification, this is also true in multi-class.

• For general $P_X$, active learners always gain over the worst case passive rate under either conditions (Theorem 2.4). As it turns out, the active learning rates are the same under both conditions, matching known rates under (i) in [20] (when $L = 2$).

Previous work in nonparametric active learning invariably adopted condition (i) which makes sense in light of our results since savings cannot be shown otherwise. Our results in fact further highlight two sources of savings in active learning, owing to the distinction between the above two bulleted regimes: a), an active learner can evenly sample the decision boundary while i.i.d. samples might miss it under general $P_X$, and b), an active learner can quickly stop sampling in those regions where there is little to gain in excess error over the Bayes classifier, having discovered some label with sufficiently low excess error. Under near uniform $P_X$, the first source of saving a) does not apply since even i.i.d. data has good coverage of the decision boundary, while b) remains, although in a limited form: an active learner can only significantly benefit from regions where a single label is clearly better, i.e., has margin as in (i), while it cannot effectively identify regions where multiple labels are nearly equivalent (e.g., non-unique Bayes), where it may end up wasting queries while it should preferably give up; in fact we show in our main Theorem 2.1 that no active learning procedure can automatically decide when to give up on such a priori unknown regions, which forces a label complexity of the same order as in passive learning.

Here we emphasize that our results do not preclude limited gains in practice under uniform $P_X$, since minimax rates fail to fully capture constants. In fact, we can refine the margin conditions to account for regions with non-unique Bayes classes—where an active learner may still save over passive, and derive a refined upper-bound, under uniform $P_X$, that highlight such limited gains over passive learning (Theorem 2.3). Our upper-bounds require minor modification over past algorithms (e.g., those in [20]), namely additional book-keeping (Section 2.4), and refined correctness arguments that are required in the multi-class setting.

Our main hardness result of Theorem 2.1 requires considerable new technicality over usual
lower-bound arguments for active learning, involving careful randomization of hard regions of space (see discussion in Section 2.3). We strive to accurately capture rates dependence on the number of classes, both in lower and upper-bounds, a dependence which appears to be missing in the literature; we are thus able to match upper and lower-bounds, by introducing some new parametrization that captures a sense of effective classes (c.f., Definition 2.4 and Remark 2.8).

2.2 Problem Setting

We consider a joint distribution \( P_{X,Y} \) on \([0, 1]^d \times [L]\), where the short notation \([L]\) \(=\) \{1, \ldots, L\} for \(L \in \mathbb{N}\). Let \( P_X \) be the marginal for \( X \) with \( \text{Support}(P_X) \subset [0, 1]^d \). Define the regression function \( \eta(x) = (\eta_y(x))_{y \in [L]} \), where \( \eta_y(x) = P(Y = y | X = x) \) for \( y \in [L] \).

**Definition 2.1.** The function \( \eta \) is said to be \((\lambda, \alpha)\)-Hölder continuous, \( \alpha \in (0, 1], \lambda > 0 \), if:

\[
\forall x, x' \in \text{Support}(P_X), \quad \|\eta(x) - \eta(x')\|_\infty \leq \lambda \|x - x'\|_\infty^\alpha,
\]

where \(\|\cdot\|_\infty\) is the maximum element of a vector.

**Remark 2.1.** For simplicity of presentation, we assume \(\alpha \leq 1\). The case of \(\alpha > 1\), can be handled simply by replacing the averaging in each cell with higher order polynomial regression (as done e.g. in [20]), but does not add much to the main message despite the added technicality. As in prior work [20, 18], we assume access to \(\lambda\) or any upper-bound thereof.

**Definition 2.2.** For \( r = 2^{-k}, k \in \mathbb{N} \), define the partition \( \mathcal{C}_r \) of \([0, 1]^d\) as the collection of hypercubes of the form \( \prod_{i \in [d]} [(l_i - 1)r, l_i r) \), \( l_i \in [1/r] \). We call \( \mathcal{C}_r \) a dyadic partition at level \( r \).

Now, we are ready to define the strong density condition. The following definition is adapted from other works on active learning [20, 18].

**Definition 2.3.** \( P_X \) is said to satisfy a strong density condition if there exists some \( c_d > 0 \) such
that \( \forall r \in \{2^{-k} : k \in \mathbb{N}\} \) and \( C \in \mathcal{C}_r \) with \( P_X(C) > 0 \), we have

\[
P_X(C) \geq c_d \cdot r^d.
\]

The strong density condition clearly holds for \( P_X = \mathcal{U}[0,1]^d \), or simply has lower-bounded density. Note that it allows a disconnected support \( \mathcal{X} \), such as in our lower-bound construction in Section 2.5.1.

Finally, we note that the labels with low probabilities are less relevant to the difficulty of the classification problem. Thus, we introduce the notion of effective classes that filter out these low-probability labels.

**Definition 2.4.** A class \( y \in [L] \) is an effective class at \( x \) if \( \eta_y(x) \geq \max_{l \in [L]} \eta_l(x)/2 \). We let \( L^*(x) \) denote the number of effective classes at \( x \).

The number of effective classes differentiates real multi-class situations and the degenerate one where all but two classes have positive probabilities.

### 2.2.1 Active Learning

We consider active learning under a fixed budget \( n \) of queries. At each sampling step, the learner may query the label of any point \( x \in [0,1]^d \), and a label \( Y \) is returned according to the conditional \( P_{Y|X=x} \). We let \( S \equiv \{(X_i, Y_i)\}_{i=1}^n \) denote the resulting sample. A classifier \( \hat{h}_n = \hat{h}_n(S) : [0,1]^d \rightarrow [L] \) is then returned.

We evaluate the performance of an active learner by the excess risk of the final classifier \( \hat{h}_n \) it outputs. Throughout the chapter, we use the notation \( \hat{h} \) for the active learning algorithm, and \( \hat{h}_n \) for the final classifier the algorithm \( \hat{h} \) returns.

**Definition 2.5.** We consider the 0-1 risk of a classifier \( h : [0,1]^d \rightarrow [L] \), namely \( R(h) = \mathbb{P}(h(X) \neq Y) \), which is minimized by the so-called Bayes classifier \( h^*(x) \in \arg\max_y \mathbb{P}(Y = y|X = x) \). The
excess risk $\mathcal{E}(h) \triangleq R(h) - R(h^*)$ is then given by:

$$
\mathcal{E}(h) = \mathbb{E} \left[ \max_{y \in [L]} \eta_y(X) - \eta_h(X) \right].
$$

2.2.2 Margin Condition

We start with a notion of soft margin.

**Definition 2.6.** Let $\eta^{(1)} \geq \cdots \geq \eta^{(L)}$ denote order statistics on $\eta_y$, $y \in [L]$. The margin at $x$ is defined as $\mathcal{M}(x) \triangleq \eta^{(1)}(x) - \max_{\forall y: \eta_y(x) \neq \eta^{(1)}(x)} \eta_y(x)$. In the case where $\forall y \in [L], \eta_y(x) = 1/L$, we use the convention that max of empty set is $-\infty$ so that $\mathcal{M}(x) = \infty$.

**Definition 2.7.** $P_{X,Y}$ satisfies the Tsybakov’s margin condition (TMC) with $C_\beta > 0$, $\beta \geq 0$, if:

$$
\forall \tau > 0, \quad P_X \left( \{ x : \mathcal{M}(x) \leq \tau \} \right) \leq C_\beta \tau^\beta. \quad (2.1)
$$

The above extends TMC for $L = 2$ to general $L$: when $L = 2$, the margin $\mathcal{M}(x) = |\eta_1(x) - \eta_2(x)|$ when $\eta_1(x) \neq \eta_2(x)$ and $\mathcal{M}(x) = \infty$ when $\eta_1(x) = \eta_2(x) = 1/2$. The above thus coincides with condition (ii) of the introduction, i.e., admits non-unique Bayes as in [19], but here we allows general $L \geq 2$.

2.3 No Gain under Tsybakov’s Margin Condition

Surprisingly, under TMC, no active learner can gain in excess risk rate over their passive counterparts when we assume the strong density condition for $P_X$. We start our discussion by defining the family of distributions for which active learning has no gains.

**Definition 2.8.** Let $c_d, \lambda, C_\beta > 0$, $\alpha \in [0, 1), \beta \geq 0, \gamma \geq 1, 2 \leq L^* \leq L$, and $\mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma)$ denote the family of joint distributions $P_{X,Y}$ on $[0, 1]^d \times [L]$ where:

(i) $P_X$ satisfies a strong density condition with $c_d$;
(ii) the regression function \( \eta(x) \) is \((\lambda, \alpha)\)-Hölder;

(iii) \( P_{X,Y} \) satisfies TMC with parameter \((\beta, C_\beta)\):

(iv) \( \forall x \in \text{Support}(P_X), L^* \leq L^*(x) \leq \min\{L, (L^*)^\gamma\} \);

(v) \( \forall x \in \text{Support}(P_X), \) every class \( y \in [L] \) has a positive probability.

Remark 2.2. Condition (iv) of Definition 2.8 implies that: \( \forall x \in \text{Support}(P_X) \), the log number of effective classes at \( x \) satisfies: \( \log L^* \leq \log L^*(x) \leq \gamma \log L^* \). As we will see later in Theorem 2.1, 2.3, and 2.4, such condition for changing \( L^*(x) \) over the space ensures the same rate, despite a difference in the leading constant.

Theorem 2.1. Let \( 2 \leq L^* \leq L, \; c_d \in (0, 1], \alpha \in (0, 1], \lambda, \beta, C_\beta > 0 \) with \( \alpha \beta \leq d \), and \( \mathcal{P} = \mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, \gamma) \). Suppose that \( n \geq \lambda^{(1/\alpha-2)(2\alpha+d)}(L^*)^{(\alpha+d)/\alpha} \log L^* \), then \( \exists C_1 > 0 \), independent of \( n, L, \text{and } L^* \), such that

\[
\inf_{\hat{h}} \sup_{P_{X,Y} \in \mathcal{P}} \mathbb{E} \mathcal{E}(\hat{h}_n) \geq C_1 \left( \frac{\log L^*}{L^*} \right)^{\alpha(\beta+1)/2\alpha+d} \left( \frac{1}{n} \right)^{\alpha(\beta+1)/2\alpha+d}; \tag{2.2}
\]

where the infimum is taken over all (potentially active) learners, and the expectation is taken over the sample distribution, determined by \( P_{X,Y} \) and \( \hat{h} \) jointly.

Remark 2.3. The proof of Theorem 2.1 requires a more involved construction of difficult distributions than the binary construction in [12]. Here, we incorporate our novel concept of effective class in the lower-bound, and try to capture the full dependence on \( L^* \) that may arise. This means that we cannot use constructions with degenerate multi-class situations where only two classes have non-zero probabilities for some \( x \in [0, 1]^d \) as previous work by [21] does. Furthermore, when \( (L^*)^\gamma < L \), our result shows that active learning has no gain if some but not all classes are Bayes classes in parts of the space with a positive mass.

Remark 2.4. As a corollary to the Theorem 2.1, since \( \hat{h} \) is any learner, including the passive ones, the rate in (2.2) is also a lower-bound on passive learning for multi-class classification. As such we
know of no other work that so clearly integrates the number of (effective) classes into the learning rate.

**Remark 2.5.** We will show later in Theorem 2.3 an upper-bound that matches the rate in (2.2) up to logarithmic terms in $n$ and $L^*$. With other factors (e.g., margin conditions and smoothness) fixed, the rate in (2.2) gets faster when $L^*$ increases (the dependence is of form $\log L^*/L^*$). This might be counter-intuitive at first, but can be simply explained by the fact that a larger $L^*$ will lead to a smaller variance in the estimation of the regression function $\eta_y(x)$ (see more in Remark 2.7 and Lemma 2.2).

The details of the proof of Theorem 2.1 are presented in Section 2.5. Our main arguments depart from usual lower-bounds arguments in active learning [17, 18, 20]. If we followed such information-theoretic constructions where the regions with non-unique Bayes classes are fixed in parts of the space, the learner would know the location of these regions and can achieve a fast rate by simply giving up on sampling in such parts of the space. Instead of working directly on constructing a suitable subset of $\mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma)$, we move to a larger class $\Sigma$ with non-empty intersection $\Sigma_\beta$ with $\mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma)$. We then randomize the construction by putting a suitable measure on $\Sigma$ that concentrates on $\Sigma_\beta$. Importantly, this measure also encodes regions of $[0, 1]^d$ where the Bayes classifier is unique. We show that for any fixed sampling mechanism of $\hat{h}$, the excess error of the classifier $\hat{h}_n$ is lower-bounded as in Theorem 2.1, in expectation under our measure on $\Sigma$, implying the statement of Theorem 2.1 by concentration on $\Sigma_\beta$. A main difficulty remains in removing dependencies inherent in the observed sample $S$: this is done by decoupling the sampling $\hat{h}$ from the eventual classifier $\hat{h}_n$ by a reduction to simpler Neyman-Pearson type classifier $\hat{h}_n^*$—with the same sampling mechanism as $\hat{h}$—whose error can be localized to regions of $[0, 1]^d$ and depends just on local $Y$ values, thanks to our choice of distributions in $\Sigma$ where little information is leaked across regions of space.
Figure 2.1: Different types of margin over space. Here, we consider an example with \( d = 1 \), and \( L = 3 \). The marginal \( P_X \) is supported on \( \mathcal{X} = \bigcup_{i=1}^{3} \mathcal{X}_i \) (the thick lines on the x-axis). We have \( \{ x : \mathcal{M}(x) \leq \tau \} = \mathcal{X}_3 \), while \( \{ x : \mathcal{M}'(x) \leq \tau \} = \mathcal{X} \). In particular, RMC is satisfied with \( \varepsilon_0 = P_X(\mathcal{X}_1 \cup \mathcal{X}_2) \), as \( \{ \mathcal{M}' = 0 \} = \mathcal{X}_1 \cup \mathcal{X}_2 \).

### 2.4 Gain under a Refined Margin Condition

Theorem 2.1 indicates that the classical TMC is not enough to guarantee gains over passive learning, under strong density. Nonetheless, some gain can be shown under a refined margin condition that better isolates regions of space with unique Bayes class (Theorem 2.3). Furthermore, under more general \( P_X \), we show in Theorem 2.4 that a better rate than passive can always be attained even under classical TMC. Both results are established using the same procedure proposed in Section 2.4.1. We start with the following definition.

**Definition 2.9.** The **sharp margin** on \( \eta \) is defined as \( \mathcal{M}'(x) = \eta_{(1)}(x) - \eta_{(2)}(x) \), where we have \( \eta_{(1)} = \eta_{(2)} \) when the Bayes class is not unique at \( x \).

**Definition 2.10.** \( P_{X,Y} \) is said to satisfy a **refined margin condition** (RMC) with \( \varepsilon_0 \geq 0, C\beta > 0, \beta' \geq \beta \geq 0 \) if:

\[
\forall \tau > 0, \quad P_X(\{ x : \mathcal{M}(x) \leq \tau \}) \leq C\beta \tau^\beta; \quad \text{and} \\
\forall \tau > 0, \quad P_X(\{ x : \mathcal{M}'(x) \leq \tau \}) \leq \varepsilon_0 + C\beta \tau^{\beta'}. 
\]

**Remark 2.6.** The two conditions in Definition 2.10 differ only when there are non-unique Bayes classes in parts of the space with positive mass, i.e., when \( \mathbb{P}(\mathcal{M}' = 0) = \varepsilon_0 > 0 \), otherwise
\( M = M', \ P_X\text{-a.s.}, \) and we may choose \( \beta = \beta' \) and both conditions are equivalent. See the example of Figure 2.1 for a detailed illustration.

Next, we define the classes of distributions for upper-bounds.

**Definition 2.11.** Let \( 1 \leq L^*_{\min} \leq L^* \leq L, \ c_d, \lambda, C_\beta > 0, \ \alpha \in [0, 1), \beta' \geq \beta \geq 0, \ \varepsilon_0 \geq 0. \) We use \( \mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L^*_{\min}, L^*_{\max}) \) to denote the family of distributions on \([0, 1]^d \times [L]\) where:

(i) the regression function \( \eta(x) \) is \((\lambda, \alpha)\)-Hölder;

(ii) \( P_{X,Y} \) satisfies TMC with parameter \((\beta, C_\beta)\);

(iii) \( L^*_{\min} \leq L^*(X) \leq L^*_{\max} \) holds \( P_X\text{-a.s.} \);

We use \( \mathcal{P}_2(c_d, \lambda, \alpha, \varepsilon_0, C_\beta, \beta', L^*_{\min}, L^*_{\max}) \) to denote the subclass of \( \mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L^*_{\min}, L^*_{\max}) \) where the following additional conditions are satisfied:

(iv) \( P_X \) satisfies a strong density condition with \( c_d \);

(v) \( P_{X,Y} \) satisfies RMC with parameter \((\varepsilon_0, C_\beta, \beta, \beta')\).

**Remark 2.7.** The parameters \( L^*_{\min} \) and \( L^*_{\max} \) control the hardness of the problem. First, \( L^*_{\max} \) is an upper bound for the number of effective classes \( L^*(x) \) at any \( x \in X \), which is also the number of class probabilities \( \eta_y(x) \) one needs to estimate simultaneously at each \( x \). Second, as it turns out the magnitude of each \( \eta_y(x) \) also affects how well we may estimate it. In particular, note that estimating the regression function \( \eta_y(x) \) is essentially estimating a Bernoulli distribution with variance \( \eta_y(x)(1 - \eta_y(x)) \leq O(1/L^*_{\min}) \). Therefore, larger \( L^*_{\min} \) yields a smaller variance bound, and hence a faster concentration rate for estimating each \( \eta_y(x) \). More specifically, Lemma 2.2 below shows that we need a sample size of at least \( 3 \log(2/\delta)/(\varepsilon^2 L^*_{\min}) \) to guarantee that one can estimate each \( \eta_y(x) \) within an error of \( \varepsilon \) with \( 1 - \delta \) for any \( \varepsilon, \delta > 0 \).

**Lemma 2.2.** (Chernoff bound with small deviation [22]) Let \( Z_1, Z_2, \ldots, Z_m \) be i.i.d random variables taking values 0 or 1, and \( P(Z_i = 1) = p. \) Then, for any \( 0 \leq \varepsilon \leq mp, \)

\[
P\left( \left| \frac{1}{m} \sum_{i=1}^{m} Z_i - p \right| > \varepsilon \right) \leq 2 \exp\left( -\frac{m\varepsilon^2}{3p} \right).
\]
Algorithm 1 Meta Algorithm

1: Input: $n, \delta, \lambda$

2: Initialization:

3: • Set $\alpha_0 = 0, n_0 = \frac{n}{\lceil \log(n) \rceil}, \delta_0 = \frac{\delta}{\lceil \log(n) \rceil}$

4: • Set minimum level $r_0 = 2 \lceil \log_2 (n_0^{1/d}) \rceil$

5: • Set final candidate labels $\mathcal{L}_C = [L], \forall C \in \mathcal{C}_{r_0}$

6: for $i = 1, \ldots, \lceil \log(n) \rceil$ do

7: // Run the non-adaptive subroutine

8: Set $\alpha_i = \frac{i}{\lceil \log(n) \rceil}$

9: Run Algorithm 2 with $(n_0, \delta_0, \alpha_i, \lambda, r_0)$ to obtain candidate labels $\{\mathcal{L}_C^{\alpha_i} \} C \in \mathcal{C}_r$

10: // Aggregate candidate labels

11: if $\forall C \in \mathcal{C}_{r_0}, \mathcal{L}_C \cap \mathcal{L}_C^{\alpha_i} \neq \emptyset$ then

12: $\forall C \in \mathcal{C}_{r_0}$, set $\mathcal{L}_C = \mathcal{L}_C \cap \mathcal{L}_C^{\alpha_i}$

13: end if

14: end for

15: Output: $\hat{h}_n(x) = \min \mathcal{L}_c$ for $x \in C \in \mathcal{C}_{r_0}$

2.4.1 An Adaptive Procedure

The detailed approach is presented in Algorithm 1, and follows an adaptation strategy of [20, 23] for unknown smoothness $\alpha$. This procedure repeatedly calls a non-adaptive subroutine, Algorithm 2, for a sequence of increasing values of $\alpha$, i.e. $\{\alpha_i\}_{i=1}^{\lceil \log(n) \rceil}$ with $\alpha_i = i/\lceil \log(n) \rceil$.

In a departure from the binary case ($L = 2$) studied in prior work, both procedures operate by maintaining a set of candidate labels via local elimination (requiring new book-keeping), and remaining labels are then aggregated at the end to return a final classifier.

The budget is tracked throughout, by sampling $n_{r,\alpha}$ points in each $C \in \mathcal{C}_r$, with

$$n_{r,\alpha} = \min \left\{1, \frac{1}{|\mathcal{L}_C^{\alpha}|} + \tau_{r,\alpha} \right\} \log \left( \frac{8|\mathcal{L}_C^{\alpha}|}{\delta_0 r^{d+1}} \right) / 2(\lambda r^\alpha)^2$$ (2.3)
Algorithm 2 Non-adaptive Subroutine

1: Input: $n_0, \delta_0, \alpha, \lambda, r_0$

2: Initialization:

3: ● Initial level: $r = 1/2$

4: ● Active cells: $A_r = \mathcal{C}_r$

5: ● Budget up to level $r$: $m_r = |A_r| n_{r,\alpha}$ (see (2.3))

6: ● Candidate labels: $\mathcal{L}^\alpha_C = [L], \forall C \in \mathcal{C}_r$

7: while $(m_r \leq n_0)$ and $(|A_r| > 0)$ do

8:    // Eliminate bad labels

9:    for each $C \in A_r$ do

10:       Samples $(X^C_i, Y^C_i)_{i \leq n_{r,\alpha}}$ in cell $C$ and compute $\{\hat{\eta}_y(C)\}_{y \in [L]}$ by (2.4)

11:       Set $\mathcal{L}^\alpha_C = \mathcal{L}^\alpha_C \setminus \{y : \hat{\eta}(1)(C) - \hat{\eta}_y(C) \geq \tau_{r,\alpha}\}$ (2.5)

12:    end for

13:    // Pass information to the next level

14:    \forall C' \in \mathcal{C}_{r/2} \text{ with } C' \subset C, \text{ set } \mathcal{L}^\alpha_{C'} = \mathcal{L}^\alpha_C$

15:    Set $A_{r/2} = \bigcup\{C' \in \mathcal{C}_{r/2} : C' \subset C \text{ for some } C \in A_r \text{ with } |\mathcal{L}^\alpha_C| \geq 2\}$

16:    Set $r = r/2, m_{r/2} = m_r + |A_r| n_{r,\alpha}$ // Go to next level and update the budget used

17: end while

18: Set $r_{min} = 2r$ // The minimum level reached

19: Set $\mathcal{L}^\alpha_C = \mathcal{L}^\alpha_C, \forall C \in \mathcal{C}_{r_{min}}$ with $C \subset C' \in \mathcal{C}_{r_{min}}$

20: Output: $\{\mathcal{L}^\alpha_C\}_{C \in \mathcal{C}_{r_{min}}}$

where $\tau_{r,\alpha} \equiv 6 \lambda r^\alpha$. This sample is used to estimate $\eta$ in each cell $C$ as

$$\hat{\eta}_y(C) = n_{r,\alpha}^{-1} \sum_{i=1}^{n_{r,\alpha}} \mathbb{1}(Y^C_i = y),$$

(2.4)

and eliminate labels $y$ whenever $\hat{\eta}(1)(C) - \hat{\eta}_y(C) \geq \tau_{r,\alpha}$, where we define

$$\hat{\eta}(1)(C) \triangleq \max_{y \in [L]} \hat{\eta}_y(C)$$

(2.5)
2.4.2 Rates Under Strong Density Condition

We first consider the excess risk rate for the adaptive algorithm under the strong density condition (Definition 2.3).

**Theorem 2.3.** Let \( 1 \leq L^*_\text{min} \leq L^*_\text{max} \leq L, c_d, \lambda, C_\beta > 0, \alpha \in [0, 1), \beta' \geq \beta \geq 0, \varepsilon_0 \geq 0, \) and \( \alpha \beta' \leq d. \) Let \( \hat{h}_n \) denote the classifier returned by Algorithm 1 with input \( n > 0, \lambda > 0 \) and \( 0 < \delta < 1. \) Suppose \( P_{X,Y} \in \mathcal{P}_2(c_d, \lambda, \alpha, \varepsilon_0, C_\beta, \beta, \beta', L^*_\text{min}, L^*_\text{max}) \), then with probability at least \( 1 - \delta, \)

\[
\mathcal{E}(\hat{h}_n) \leq C_2 \left( \frac{\alpha(\beta + 1)}{2\alpha + d} \left( \log \frac{L^*_\text{max}}{L^*_\text{min}} \right) \frac{\lambda \varepsilon_0^3 (n) \log \left( \frac{8\lambda^2n}{\delta} \right)}{n} \right)^{\frac{\alpha(\beta + 1)}{2\alpha + d}}
\]

\[
+ \left( \frac{\log L^*_\text{max}}{L^*_\text{min}} \right)^{\frac{\alpha(\beta' + 1)}{2\alpha + d - \alpha \beta'}} \left( \frac{\lambda \varepsilon_0^3 (n) \log \left( \frac{8\lambda^2n}{\delta} \right)}{n} \right)^{\frac{\alpha(\beta' + 1)}{2\alpha + d - \alpha \beta'}}
\]

for some constant \( C_2 > 0 \) independent of \( n, \delta, \lambda, \varepsilon_0, L, L^*_\text{min}, L^*_\text{max}. \)

**Remark 2.8.** The upper-bound shown in Theorem 2.3 depends on \( \varepsilon_0, \) and recovers existing bounds (for the binary case) when \( \varepsilon_0 = 0, \) namely \( \tilde{O}(n^{-\alpha(\beta + 1)/(2\alpha + d - \alpha \beta')}) \) as shown e.g. in [20, 18].

This is an improvement over the passive learners and matches the active lower-bound in [18] under the strong density condition with \( \alpha \beta \leq d. \) For \( \varepsilon_0 \geq \tilde{O}((nL^*_\text{min})^{-\alpha \beta/(2\alpha + d - \alpha \beta'))}, \) the first term dominates and the upper-bound matches the passive minimax rate. In particular, note that the distribution class \( \mathcal{P} \equiv \mathcal{P}(c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma) \) (c.f., Definition 2.8) is a subset of \( \mathcal{P}_2 \equiv \mathcal{P}_2(c_d, \lambda, \alpha, 1, C_\beta, \beta, \infty, L^*, (L^*)^\gamma) \). The upper-bound here for the class \( \mathcal{P}_2 \) almost matches the one in Theorem 2.1 for \( \mathcal{P}, \) ignoring the polylogarithmic terms.

A main novelty in the analysis is to separately consider parts of space with unique Bayes classes, determined by \( \varepsilon_0 \) and \( \beta', \) and those parts of space where the Bayes classes might not be unique, but which still have margin, determined by \( \beta. \) Furthermore, our consideration of general multi-class, together with non-unique Bayes, brings in a bit of added technicality due largely to
additional book-keeping. In particular, while in [20, 18], the main correctness argument involved showing that all labeled parts of space (i.e. cells with a single label left) have 0 excess error w.h.p., we additionally have to show that in fact, remaining labels in most active cells are close in error to Bayes.

2.4.3 Rates for General Densities

For general $P_X$, on the other hand, Algorithm 2 has an excess risk rate $\tilde{O}(n^{-(\alpha(\beta+1))/(2\alpha+d)})$, which is always faster than the lower minimax rate $O(n^{-(\alpha(\beta+1))/(2\alpha+d+\alpha\beta)})$ for passive learning of $[19]$ under the same conditions.

In other words, under TMC, which allows non-unique Bayes classifiers, active learning guarantees savings over the worst-case rate of passive learning, given the ability to evenly sample the decision boundary.

**Theorem 2.4.** Let $1 \leq L_{\text{min}}^* \leq L_{\text{max}}^* \leq L$, $\lambda, C, \beta > 0$, $\alpha, \beta' \in [0, 1)$, $\beta' \geq 0$, and $\alpha \beta' \leq d$. Let $\hat{h}_n$ denote the classifier returned by Algorithm 1 with input $n > 0$, $\lambda > 0$ and $0 < \delta < 1$. Suppose that $P_{X,Y} \in \mathcal{P}_1(\lambda, \alpha, C, \beta, L_{\text{min}}^*, L_{\text{max}}^*)$, then with probability at least $1 - \delta$,

$$E(\hat{h}_n) \leq C_3 \left( \frac{\log L_{\text{max}}^*}{L_{\text{min}}^*} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} \left( \frac{\log^2(n) \lambda^d \log \left( \frac{8\lambda^2 n}{\delta} \right)}{n} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}$$

for some constant $C_3 > 0$ that does not depend on $n, \delta, \lambda, L, L_{\text{min}}^*, L_{\text{max}}^*$.

The proof ideas follow similar outlines as for Theorem 2.3, though more direct.

2.5 Proofs of Main Theorems

2.5.1 Proof of Theorem 2.1

**Construction of the Difficult Distributions** We operate over a dyadic partition $\mathcal{C}_r$ of the unit cubes $[0, 1]^d$. Let $r = (c_1 \log L^*/(nL^*))^{\frac{1}{2\alpha+d}}$, where $c_1 = \frac{8}{9\lambda^2}$. Without loss of generality, we assume that $-\log_2 r \in \mathbb{N}$ and $2 \leq L^* \leq L - 1$. The case where $L^* = L$ can be done using a
similar proving strategy with minor adjustments. Furthermore, we denote the barycenter of any $C \in \mathcal{C}_r$ as $x_C$.

The marginal distribution $P_X$ has the density with respect to the Lebesgue measure:

$$f(x) = \begin{cases} 4^d & \text{if } \|x - x_C\| < r/8 \text{ for some } C \in \mathcal{C}_r; \\ 0 & \text{otherwise}. \end{cases}$$

where $\|\cdot\|$ is the supnorm. Let $z = (z_C)_{C \in \mathcal{C}_r} \in \{0,1\}^{|\mathcal{C}_r|}$ and $\sigma = (\sigma_C)_{C \in \mathcal{C}_r} \in [L^* - 1]^{|\mathcal{C}_r|}$. Define the regression function $\eta_{z,\sigma}(x) = (\eta_{z,\sigma}^1(x), \ldots, \eta_{z,\sigma}^{L^*}(x))$, with

$$\eta_{z,\sigma}^y(x) = \begin{cases} \kappa/L^* + c_\eta \sum_{C \in \mathcal{C}_r} z_C (1(y = \sigma_C) - 1(y = L^*)) \phi_C(x) & \text{for } y \in [L^*], \\ (1 - \kappa)/(L - L^*) & \text{for } y \in [L] \setminus [L^*]. \end{cases}$$

where $c_\eta = \lambda/8$, $\max\{3L^*/(L + 2L^*), 1/2\} \leq \kappa < 1$, and

$$\phi_C(x) = \min \left\{ (2r^\alpha - 8r^{\alpha - 1}\|x - x_C\|)_+, r^\alpha \right\}.$$ 

Here, we adopt the notation $v_+ = \max(0, v), \forall v \in \mathbb{R}$.

Note that $\{\eta_{z,\sigma}^y\}_{y \in [L]}$ indeed defines a proper regression function with only the first $L^*$ classes being effective classes. For all $x \in [0,1]^d$, and $n \geq \lambda^{(1/\alpha - 2)/(2\alpha + d)} \cdot (L^*)^{(\alpha + d)/\alpha} \cdot \log L^*$, we have $\phi_C(x) \leq r^\alpha = (8/(9\lambda^2))^{\alpha/2} (\log L^*/(nL^*))^{\alpha/(\alpha + d)} \leq 1/(\lambda L^*)$, and hence

$$\eta_{z,\sigma}^{L^*}(x) \geq \kappa/L^* - c_\eta/(\lambda L^*) > 3\kappa/(4L^*); \quad \eta_{z,\sigma}^{\sigma_C}(x) \geq \kappa/L^* + c_\eta/(\lambda L^*) < 5\kappa/(4L^*).$$

Also, we have by the choice of $\kappa$ that $0 < (1 - \kappa)/(L - L^*) \leq \kappa/(3L^*)$. Therefore, for any $y^* \in [L^* - 1] \setminus \{\sigma_C\}$ and $y \in [L] \setminus [L^*]$,

$$0 < \eta_{z,\sigma}^y(x) < \eta_{z,\sigma}^{\sigma_C}(x)/2 \leq \eta_{z,\sigma}^{L^*}(x) \leq \eta_{z,\sigma}^{y^*}(x) \leq \eta_{z,\sigma}^{\sigma_C}(x).$$
Figure 2.2: Construction of $\eta_{z,\sigma}^k$ for Theorem 2.1 on a partition $\mathcal{G}_r$ of $[0,1]^d$ for $d = 2$, $L^* = 3$, and a critical $r = r(n, \alpha, \lambda)$. The marginal distribution $P_X$ is uniformly distributed over its support (highlighted regions in grey). Two coins are thrown in each cell $C$, one $z_C \in \{0,1\}$ with some bias determining whether the Bayes is unique, the other $\sigma_C \in [L^* - 1] \equiv \{1,2\}$ determining the Bayes class. In a cell $C$ with $z_C = 1$, $\eta_{z,\sigma}^C$ has a upside bump of height $c \eta r^{\alpha}$; and $\eta_{z,\sigma}^3$ has a downside bump of same size. Note that we omit the $L - L^*$ non-effective classes with lower class probabilities.

On the other hand, one can easily verify that $\sum_{y \in [L]} \eta_{z,\sigma}^y(x) = 1$ for any $z, \sigma$ and $x$, by noticing $\sum_{y \in [L]} 1(y = \sigma_C) - 1(y = L^*) = 0$.

For each pair $(z, \sigma)$, one can define a joint probability distribution $P_{z,\sigma}$ characterized by $P_X$ and $P[Y = y|X = x] = \eta_{z,\sigma}^y(x)$. See Figure 2.2 for an example of $P_{z,\sigma}$ for $L^* = 3$, and $d = 2$. In particular, $P_X$ is uniformly distributed within its support, which is the area shaded in gray. In a cell $C \in \mathcal{G}_r$ where $z_C = 1$, there is a small bump in the regression function $\eta_{z,\sigma}^C$ of size $c \eta r^{\alpha}$. By construction, $\eta_{z,\sigma}$ is always a constant in the intersection of any single cell $C$ and the support of $P_X$.

Let $\Sigma = \{P_{z,\sigma} : (z, \sigma) \in \{0,1\}^{\mathcal{G}_r} \times [L]^{\mathcal{G}_r}\}$, and $\Sigma_\beta = \{P_{z,\sigma} : (z, \sigma) \in \Theta_\beta\}$ with

$$
\Theta_\beta = \{(z, \sigma) : \forall \tau > 0, P_X(\{x : M_{z,\sigma}(x) \leq \tau\}) \leq C_\beta \tau^\theta\},
$$

where $M_{z,\sigma}(x)$ is the margin at $x$ with the regression function being $\eta_{z,\sigma}(x)$. For simplicity, we use the short notation,

$$
\Xi = (c_d, \lambda, \alpha, C_\beta, \beta, L^*, \gamma),
$$

to represent all of the parameters for the distribution class from Definition 2.8.
Establishing the Lower-bound In this part, we will show that no active learners \( \hat{h} \) has excess risk rate faster than
\[
C \cdot \left( \frac{\log L^*}{L^*} \right)^{\frac{\alpha (\beta + 1)}{2(\alpha + d)}} \cdot \left( \frac{1}{n} \right)^{\frac{\alpha (\beta + 1)}{2(\alpha + d))},
\]
with respect to \( n \) and \( L^* \).

First, we show that \( \Sigma_{\beta} \), i.e., the subset of distributions in \( \Sigma \) that satisfies the TMC with parameters \((\beta, C_{\beta})\), is contained in the distribution class \( \mathcal{P}(\Xi) \) as defined in Definition 2.8. In other words, \( \Sigma_{\beta} \) is the non-empty intersection of \( \Sigma \) and \( \mathcal{P}(\Xi) \). Therefore, a minimax lower-bound for the class \( \Sigma_{\beta} \) is also a minimax lower-bound for the class \( \mathcal{P}(\Xi) \) as required in Theorem 2.1.

**Proposition 2.5.** \( \Sigma_{\beta} \subset \mathcal{P}(\Xi) \). Consequently,
\[
\inf_{\hat{h}} \sup_{P_{X,Y} \in \mathcal{P}(\Xi)} \mathbb{E} \mathcal{E}(\hat{h}_n) \geq \inf_{\hat{h}} \sup_{P_{X,Y} \in \Sigma_{\beta}} \mathbb{E} \mathcal{E}(\hat{h}_n),
\]
where the infimum is taken over all active learners.

**Proof.** Let \( P_{z,\sigma} \in \Sigma_{\beta} \). The TMC is satisfied by definition. It can be easily verified that strong density condition holds for \( c_d = 1 \). Clearly, only the first \( L^* \) classes are effective classes, and all classes have positive probabilities. It is left to show that \( \eta_{z,\sigma} \) is \((\lambda, \alpha)-\text{Hölder}\). In fact, this hold for all \( P_{z,\sigma} \in \Sigma_{\beta} \).

Let \( x, x' \in [0,1]^d \). If they are in a common cell \( C \), then for all \( y \in [L] \):
\[
|\eta_{z,\sigma}^y(x) - \eta_{z,\sigma}^y(x')| \leq c_\eta(8r^{\alpha-1}\|x - x'\|) \\
\leq \lambda\|x - x'\|^{\alpha},
\]
where the last inequality is due to the fact \( r/\|x - x'\| \geq 1 \) and \( \alpha - 1 < 0 \). If they are in different cells, \( |\eta_{z,\sigma}(x) - \eta_{z,\sigma}(x')| = 0 \) if \( \|x - x'\| < r/4 \). Therefore, for all \( y \in [L] \):
\[
|\eta_{z,\sigma}^y(x) - \eta_{z,\sigma}^y(x')| \leq 2c_\eta r^{\alpha} \leq \lambda\|x - x'\|^{\alpha}.
\]
Therefore, $\eta z, \sigma$ is $(\lambda, \alpha) -$ Hölder and we can conclude the proof.

Next, we “randomize” the distribution class $\Sigma$ by letting $z \in \{0, 1\}^{\lvert \mathcal{C} \rvert} \overset{\text{i.i.d}}{\sim} \text{Ber}(r^{\alpha \beta})$, and $\sigma \in [L^{*}]^{\lvert \mathcal{C} \rvert} \overset{\text{i.i.d}}{\sim} \text{Unif}([L^{*}])$, $z \perp \sigma$. The following proposition shows that $\Sigma$ concentrates on $\Sigma_\beta$ under such randomness. Consequently, the mean risk (w.r.t. the randomness of $z, \sigma$) over the larger distribution class $\Sigma$ can only be larger than the worst-case risk over the subclass $\Sigma_\beta$ by a negligibly small quantity.

**Proposition 2.6.** Let $\hat{h}$ be any active learner. Then,

$$
\sup_{P_{X,Y} \in \Sigma_\beta} \mathbb{E}_{S \mid z, \sigma, \hat{h}} \mathcal{E}(\hat{h}_n) \geq \mathbb{E}_{z, \sigma} \mathbb{E}_{S \mid z, \sigma, \hat{h}} \mathcal{E}(\hat{h}_n) \geq \exp(-c_2 r^{-(d-\alpha \beta)}),
$$

for some $c_2 > 0$, where $\mathbb{E}_{S \mid z, \sigma, \hat{h}} (\cdot)$ is expectation taken over sample $S$, under the sampling distribution $P_{S \mid z, \sigma, \hat{h}}$ determined by the data distribution $P_{z, \sigma}$ and active sampling strategy $\hat{h}$ jointly, and $\mathbb{E}_{z, \sigma} (\cdot)$ is the expectation taken over $P_{z, \sigma}$.

**Proof.** By construction, $\mathcal{M}_{z, \sigma}(x)$ is bounded from below by $2c_2 r^{\alpha}$ almost surely. Thus, we only need to consider $\tau = tc_2 r^{\alpha}$ for some $t \geq 2$. For given $z$, $P_X(\{x : \mathcal{M}_{z, \sigma}(X) \leq t c_2 r^{\alpha}\}) \leq r^d 1^\top z$.

By Chernoff bound (Lemma B.1),

$$
\mathbb{P}_z (r^d 1^\top z \leq C_\beta r^\beta) = \mathbb{P}_z (r^d 1^\top z \leq C_\beta (tc_2)^\beta r^{\alpha \beta}) \geq 1 - \exp\left(-c_2 r^{-(d-\alpha \beta)}\right),
$$

where $c_2 = (C_\beta (tc_2)^\beta - 1)^2 / 3$. Therefore, $\mathbb{P}_{z, \sigma} ((z, \sigma) \in \Theta_\beta) \geq 1 - \exp(-c_2 r^{-(d-\alpha \beta)})$ and

$$
\sup_{P_{z, \sigma} \in \Sigma_\beta} \mathbb{E} \mathcal{E}(\hat{h}_n) \geq \mathbb{E}_{z, \sigma} \left[ \mathbb{E}_{S \mid z, \sigma, \hat{h}} \mathcal{E}(\hat{h}_n) \mid (z, \sigma) \in \Theta_\beta \right] \\
\quad \geq \mathbb{E}_{z, \sigma} \mathbb{E}_{S \mid z, \sigma, \hat{h}} \mathcal{E}(\hat{h}_n) - \mathbb{P}_{z, \sigma} ((z, \sigma) \notin \Theta_\beta) \\
\quad \geq \mathbb{E}_{z, \sigma} \mathbb{E}_{S \mid z, \sigma, \hat{h}} \mathcal{E}(\hat{h}_n) - \exp(-c_2 r^{-(d-\alpha \beta)}).
$$
Note that the uncertainties in an active classifier are in both its sampling decision and label prediction. These two types of uncertainties can be de-coupled by considering one single optimal label prediction rule given any sampling decision, if such an optimal rule exists. Formally, we introduce a class of learners with a certain labelling rule in Definition 2.12 and show that they are indeed optimal in Proposition 2.7.

**Definition 2.12.** The conditional Neyman-Pearson learner \( \hat{h}^* \) is the active learner that makes the same sampling decision \( \pi_{\hat{h}} \) as \( \hat{h} \), and labels according to the following rules for each \( C \in C_r \).

Conditional on the sample \( S_C = (X_C^i, Y_C^i)_{i=1}^{n_C} \) in \( C \),

\[
\hat{h}_n^*(x) = \arg\max_{\sigma \in [L^*]} \prod_{i=1}^{n_C} P_{z_C=1, \sigma_C=\sigma}(Y_i^C | X_i^C),
\]

for all \( x \in C \), where \( P_{z_C,\sigma_C}(Y_i^C | X_i^C) \) is the probability of \( Y_i^C \) given \( X_i^C \), \( z_C \) and \( \sigma_C \).

**Proposition 2.7.** Let \( \hat{h} \) be any active learner, and \( \hat{h}^* \) be the corresponding conditional Neyman-Pearson learner, then

\[
\mathbb{E}_{z,\sigma | \hat{h}} \mathbb{E}_{S | \hat{h}} \mathcal{E}(\hat{h}_n) \geq \mathbb{E}_{z,\sigma | \hat{h}} \mathbb{E}_{S | \hat{h}} \mathcal{E}(\hat{h}^*_n).
\]

**Proof.** Let \( \hat{h}_n \) be the classifier returned by active learner \( \hat{h} \), we can decompose its excess risk as:

\[
\mathcal{E}(\hat{h}_n) = \sum_{C \in C_r} \mathcal{E}_C(\hat{h}_n); \tag{2.6}
\]

with \( \mathcal{E}_C(\hat{h}_n) \doteq \int_{C \cap (\hat{h}_n \neq \sigma_C)} \left[ \eta_{z,\sigma}^C(x) - \eta_{z,\sigma}^C(\hat{h}_n(x)) \right] dP_X(x) \). Thus, we only need to show that for any \( C \in C_r \),

\[
\mathbb{E}_{S | \hat{h}} \mathbb{E}_{S | \hat{h}} \mathcal{E}_C(\hat{h}^*_n) \leq \mathbb{E}_{S | \hat{h}} \mathbb{E}_{S | \hat{h}} \mathcal{E}_C(\hat{h}_n),
\]

where \( \mathbb{E}_{S | \hat{h}} \) is the expectation taken over the distribution of sample \( S \) given active sampling strategy.
\( \hat{h} \) and \( \mathbb{E}_{z, \sigma | S, \hat{h}} \) is the taken over the conditional distribution of \((z, \sigma)\) given \(S\) and \(\hat{h}\). Note that:

\[
\mathbb{E}_{z, \sigma | S, \hat{h}} \left[ \mathcal{E}(\hat{h}_n) | z_C = 0, \sigma_C \right] = \left( \frac{\kappa r^d}{L^*} - \frac{1 - \kappa}{L - L^*} \right) \mathbb{I}(\hat{h}_n > L^*) ; \text{ and} \\
\mathbb{E}_{z, \sigma | S, \hat{h}} \left[ \mathcal{E}(\hat{h}_n) | z_C = 1, \sigma_C \right] = \left( \frac{\kappa r^d}{L^*} - \frac{1 - \kappa}{L - L^*} + c \eta^r \alpha + d \right) \mathbb{I}(\hat{h}_n > L^*) + 2c \eta^r \alpha + d \mathbb{I}(\hat{h}_n = L^*) + c \eta^r \alpha + d \mathbb{I}(\hat{h}_n < L^*, \hat{h}_n \neq \sigma_C).
\]

Clearly, an optimal learner should never predict labels that are larger than or equal to \(L^*\). For those learners with \(\hat{h}_n \in [L^* - 1]\),

\[
\mathbb{E}_{S|h z, \sigma | S, \hat{h}} \mathcal{E}(\hat{h}_n) = c \eta^r \alpha + d \mathbb{E}_{S|h} \left[ \sum_{\sigma \in [L^* - 1]} \mathbb{I}(\hat{h}_n = \sigma) \mathbb{P}(z_C = 1, \sigma_C \neq \sigma | S) \right]
\]

is minimized if \(\hat{h}_n(x) = \sigma\) when

\[
\frac{\mathbb{P}(z_C = 1, \sigma_C = \sigma | S, \hat{h})}{\mathbb{P}(z_C = 1, \sigma_C = \sigma' | S, \hat{h})} \geq 1,
\]

for any \(\sigma' \neq \sigma\). Finally, notice that

\[
\frac{\mathbb{P}(z_C = 1, \sigma_C = \sigma | S, \hat{h})}{\mathbb{P}(z_C = 1, \sigma_C = \sigma' | S, \hat{h})} = \frac{dP_{S|z_C=1,\sigma_C=\sigma,\hat{h}}(S)}{dP_{S|z_C=1,\sigma_C=\sigma',\hat{h}}(S)} = \frac{\prod_{i=1}^{n_{C}} P_{z_C=1,\sigma_C=\sigma}(Y_{C_i}^C|X_{C_i}^C)}{\prod_{i=1}^{n_{C}} P_{z_C=1,\sigma_C=\sigma'}(Y_{C_i}^C|X_{C_i}^C)}.
\]

The last step is clear from the definition

\[
dP_{S|z_C,\sigma_C,\hat{h}}(S) = \prod_{i=1}^{n} \pi_{\hat{h}}(X_i|\{X_j, Y_j\}_{j<i}) \\
\cdot \mathbb{E}_{z(\zeta), \sigma(\zeta)} \prod_{C' \subset \zeta, i=1}^{n_{C'}} P_{z_{C'}, \sigma_{C'}}(Y_{C_i}^{C'}|X_{C_i}^{C'}) dS
\]

where \(z(\zeta), \sigma(\zeta)\) are the \(z, \sigma\) vectors with \(z_C, \sigma_C\) removed, and \(\pi_{\hat{h}}\) is the sampling distribution
according to active learner \( \hat{h} \). Therefore, the labeling decision of \( \hat{h}^* \) minimizes \( \mathbb{E}_{z, \sigma} \mathbb{E}_{S|z, \sigma, \hat{h}} E_C(\hat{h}) \) for each \( C \), hence also minimizes \( \mathbb{E}_{z, \sigma} E(\hat{h}) \).

By Proposition 2.7, a conditional Neyman-Pearson learner always has as small an error rate as any other active learner with the same sampling rule. Therefore, we can conclude the proof of Theorem 2.1 by establishing a lower-bound for the class of all conditional Neyman-Pearson learners.

**Proposition 2.8.** Let \( \hat{h}^* \) be any conditional Neyman-Pearson learner. Then,

\[
\mathbb{E}_{z, \sigma} \mathbb{E}_{S|z, \sigma, \hat{h}^*} E(\hat{h}_{n}^*) \geq C_1 \left( \frac{\log L^*}{nL^*} \right)^{\alpha(d+1) / (2d + d)}.
\]

for some \( C_1 > 0 \) independent of \( n, L^* \).

**Proof.** Since \( E_C(\hat{h}) \) is a function of \( z_C, \sigma_C \) and \( S_C \), we have

\[
\mathbb{E}_{z, \sigma} \mathbb{E}_{S|z, \sigma, \hat{h}^*} E_C(\hat{h}_{n}^*) = \mathbb{E}_{z_C, \sigma_C, S_C|z_C, \sigma_C, \hat{h}} E_C(\hat{h}_{n}^*),
\]

where \( \mathbb{E}_{S_C|z_C, \sigma_C, \hat{h}} \) is the expectation over the distribution \( P_{S_C|z_C, \sigma_C, \hat{h}} \) of \( S_C \) given \( z_C, \sigma_C \) (where we have marginalized out the randomness in other cells). Furthermore, one can decompose \( P_{S_C|z_C, \sigma_C, \hat{h}} \) into the sampling location decision \( P_{X|h} \) and the labeling distribution \( P_{Y|X, z_C, \sigma_C} \):

\[
dP_{S_C|z_C, \sigma_C, \hat{h}}(S_C) = \prod_{j=1}^{n_C} dP_{X|h}^C(X_j | \{X_i^C, Y_i^C\}_{i < j}^C) P_{Y|X, z_C, \sigma_C}(Y_j^C | X_j^C).
\]

By (2.6), we have

\[
\mathbb{E}_{z, \sigma} \mathbb{E}_{S|z, \sigma, \hat{h}^*} E(\hat{h}_{n}^*) = \sum_{C \in \mathcal{C}} \mathbb{E}_{z_C, \sigma_C, S_C|z_C, \sigma_C, \hat{h}} \mathbb{E}_{S_C} E_C(\hat{h}_{n}^*).
\]

Let \( m = 2r^d n \equiv (c_n r^\alpha)^{-2} \log L^*/(36L^*) \). By the choice of \( m \), there are at \( r^{-d}/2 \) cells in \( \mathcal{C} \) with less than \( m \) labeled samples in them. Next, we will establish a lower-bound of the total excess risk
in these cells. Note that,

\[
\mathbb{E}_{z, \sigma} \mathbb{E}_{S|z, \sigma, \hat{h}} \mathcal{E}_c(\hat{h}_n^*) \geq \mathbb{E}_{z, \sigma} \sum_{n_c=1}^{m} \mathbb{E}_{S|z, \sigma, \hat{h}} [\mathcal{E}_c(\hat{h}_n^*) | |S_c| = n_c] \mathbb{P}_{S|z, \sigma, \hat{h}}(|S_c| = n_c) \\
\geq c_3 r^{d+\alpha} \mathbb{E}_{z, \sigma} \mathbb{P}_{S|z, \sigma, \hat{h}} (z_c = 1; |S_c| \leq m),
\]

where the last inequality follows from Lemma 2.11. Furthermore,

\[
\sum_{C \in \mathcal{C}} \mathbb{E}_{z, \sigma} \mathbb{P}_{S|z, \sigma, \hat{h}} (z_c = 1; |S_c| \leq m) = \sum_{C \in \mathcal{C}} \mathbb{P}(|S_c| \leq m) \mathbb{P}(z_c = 1| |S_c| \leq m)) \geq \frac{r^{\alpha \beta \alpha - d}}{1 + c_4} \sum_{C \in \mathcal{C}} \mathbb{P}(|S_c| \leq m) \geq \frac{r^{\alpha \beta - d}}{2(1 + c_4)},
\]

where the second last inequality is due to Lemma 2.12, and the last inequality is from the choice of \(m\) and a union bound. Finally,

\[
\mathbb{E}_{z, \sigma} \mathbb{E}_{S|z, \sigma, \hat{h}} \mathcal{E}(\hat{h}_n^*) = \sum_{C \in \mathcal{C}} \mathbb{E}_{z, \sigma} \mathbb{E}_{S|z, \sigma, \hat{h}} [\mathcal{E}_c(\hat{h}_n^*)] \\
= (c_3 r^{d+\alpha}) \left( \frac{r^{\alpha \beta - d}}{2(1 + c_4)} \right) \geq C_1 \left( \frac{\log L^*}{n \log L^*} \right)^{\alpha (\beta + 1) \frac{2\alpha + d}{2\alpha + d}},
\]

where \(C_1 = \frac{c_3(8 \lambda^{-2} / 9)^{2\alpha + d}}{2(1 + c_4)} > 0.\)

Now, we present some technical lemmas that have been used in the proof of Proposition 2.8. The following Lemma 2.9 shows that the labels in a cell \(C\) are independently and identically distributed given \(z_C, \sigma_C\), i.e., no information leak among the cells.

**Lemma 2.9.** Conditional on \(z_C, \sigma_C\) and \(|S_C| = n_c\), \(Y_C = \{Y_j^C\}_{j=1}^{n_c}\) are independently and identically distributed as \(P_{Y|X, z_C, \sigma_C}.\)
Proof. The conditional probability mass of $Y_C$ is

$$P_{Y_C|z_C,\sigma_C,\hat{h}}(Y_C)$$

$$= \prod_{j=1}^{n_c} dP_{X_j}^C(X_j^C|\{X_i^C, Y_i^C\}_{i<j}) P_{Y|X,z_C,\sigma_C}(Y_j^C|X_j^C)$$

$$= \prod_{j=1}^{n_c} P_{Y|X,z_C,\sigma_C}(Y_j^C|X_j^C),$$

which concludes the proof.

The next lemma Lemma 2.10 is an anti-concentration inequality for multi-classes. It shows that the labeling rule of conditional Neyman-Pearson learners fails to identify the Bayes class in a cell $C$ with a positive probability bounded away from zero, under an insufficient labeling budget.

**Lemma 2.10 (Anti-concentration).** Let $l \in [L^* - 1]$, $Y_1, \ldots, Y_m$ i.i.d. with

$$P_l(Y_1 = y) = \begin{cases} 
\kappa/L^* + c_{\eta}r^\alpha & \text{if } y = l; \\
\kappa/L^* & \text{if } y < L^* \text{ and } y \neq l; \\
\kappa/L^* - c_{\eta}r^\alpha & \text{if } y = L^*; \\
(1 - \kappa)/(L - L^*) & \text{if } y > L^*.
\end{cases}$$

where $1/2 \leq \kappa \leq 1$. If $m \leq (c_{\eta}^{-2}r^{-2\alpha} \log L^*)/(36L^*)$, then

$$P_l(\exists l' \neq l, m_{l'} > m_l) \geq c_5,$$

for some absolute constant $c_5 > 0$, where $m_l = \sum_{i=1}^m \mathbb{I}(Y_i = l)$.

Proof. For the case $L^* = 2$, we omit the proof as it follows from the same spirit of binary anti-concentration inequality, e.g., Theorem 2 (ii) of [22]. For $L^* > 2$, consider a finite collection of
distributions \( \{P_0, P_1, \ldots, P_{L^* - 1}\} \) where

\[
P_0(Y = y) = \begin{cases} 
\frac{\kappa}{L^*} & \text{if } y \leq L^*; \\
\frac{1 - \kappa}{(L - L^*)} & \text{if } y > L^*.
\end{cases}
\]

and \( \{P_l\}_{l \in \{L^* - 1\}} \) is as defined in the statement. We first find an upper bound for the KL divergence between \( P_0 \) and \( P_l \). Let \( \varepsilon := c_\eta r^{2\alpha} < \kappa/(8L^*) \) for the choice of large \( n \). By definition of KL divergence,

\[
KL(P_0|P_l) = \frac{\kappa}{L^*} \cdot \log \left( \frac{\kappa/L^*}{\kappa/L^* + \varepsilon} \right) + \frac{\kappa}{L^*} \cdot \log \left( \frac{\kappa/L^*}{\kappa/L^* - \varepsilon} \right)
\]

\[
= \frac{\kappa}{L^*} \cdot \left( \log \left( 1 + \frac{\varepsilon}{\kappa/L^* - \varepsilon} \right) - \log \left( 1 + \frac{\varepsilon}{\kappa/L^*} \right) \right)
\]

\[
= \frac{\kappa}{L^*} \cdot \left( \frac{\varepsilon}{\kappa/L^*} - \varepsilon \right) + o(\varepsilon^2)
\]

\[
= \frac{\varepsilon^2}{\kappa/L^* - \varepsilon} + o(\varepsilon^2)
\]

\[
\leq 2L^* \varepsilon = 2L^* c_\eta r^{2\alpha}.
\]

Therefore, the product measures \( P_l^m \) and \( P_0^m \) satisfies \( KL(P_0^m|P_l^m) \leq 2mL^* c_\eta^2 r^{2\alpha} \), for \( 1 \leq l \leq L^* - 1 \). Let \( \hat{P}_m = P_{\tilde{l}_m} \), with \( \tilde{l}_m = \arg\max_l m_l \) being the maximum likelihood estimator of \( l \). Define the discrete metric \( d \), with \( d(P, P') = 0 \) if \( P \equiv P' \) and \( d(P, P') = 1 \) otherwise. By Theorem 2.5 of [24], if \( m \leq (c_\eta^{-2}r^{-2\alpha} \log L^*)/(36L^*) \), then \( 2mL^* c_\eta^2 r^{2\alpha} \leq \log L^*/18 \leq \log(L^* - 1)/10 \) when \( L^* > 2 \), and we have

\[
\inf_P \sup_{P \in \{P_1, \ldots, P_{L^* - 1}\}} \mathbb{P}(d(P, \hat{P}) = 1) \geq c_5 > 0,
\]

for some \( c_5 > 0 \). By symmetry, \( \mathbb{P}(d(P, \hat{P}_m) = 1) \) is the same for all the choice of \( P \in \{P_1, \ldots, P_{L^* - 1}\} \). Therefore, for any \( l \in [L^* - 1] \),

\[
\mathbb{P}(d(P_l, \hat{P}_m) = 1) \geq c_5.
\]
Hence, the proof is complete by noticing \( \{ \exists l', m' > m_l \} = \{ d(P_l, \hat{P}_m) = 1 \} \).

The following Lemma 2.11 is an immediate corollary of Lemma 2.10 which provides the excess risk in a cell, under an insufficient labelling budget.

**Lemma 2.11.** Let \( n_c \leq m = (c_\eta^{-2} r^{-2\alpha} \log L^*)/(36 L^*) \) and \( \hat{h}^* \) be a conditional Neyman-Pearson learner. Then, in cell \( C \), for any combination of \((z_C, \sigma_C)\),

\[
\mathbb{E}_{S_c|z_C, \sigma_C, \hat{h}} \left[ \mathcal{E}_c(\hat{h}_n^*) \mid |S_c| = n_c \right] \geq c_3 r^{d+\alpha} \mathbbm{1}(z_C = 1).
\]

for some \( c_3 > 0 \).

**Proof.** When \( z_C = 0 \), the inequality holds trivially. When \( z_C = 1 \), let \( n_c^\sigma = \sum_{j=1}^{n_c} \mathbbm{1}(Y_j^C = \sigma) \), then

\[
\mathcal{E}_c(\hat{h}_n^*) = c_\eta r^{d+\alpha} \mathbbm{1} \left( \sigma_C \neq \arg\max_{\sigma \in [L^* - 1]} n_c^\sigma \right).
\]

Therefore,

\[
\mathbb{E}_{S_c|z_C, \sigma_C, \hat{h}} \left[ \mathcal{E}_c(\hat{h}_n^*) \mid |S_c| = n_c \right] = c_\eta r^{d+\alpha} \mathbb{P}_{S_c|z_C, \sigma_C, \hat{h}} \left( \exists \sigma, n_c^\sigma < n_c^\sigma \mid |S_c| = n_c \right)
\]

\[
\geq c_3 r^{d+\alpha}.
\]

where the last equality holds by Lemma 2.9 and 2.10, with \( c_3 = c_5 c_\eta \).

Finally, Lemma 2.12 is a technical lemma that shows the distribution of \( z_C \), conditioned on small sample size, is not far from the unconditional distribution.

**Lemma 2.12.** Let \( S_c = (X_j^C, Y_j^C) \) be the sample falls in \( C \). Then,

\[
\frac{\int_{|S_c| = n_c} dP_{S_c|z_C = 0}(S_c)}{\int_{|S_c| = n_c} dP_{S_c|z_C = 1}(S_c)} \leq c_4,
\]
for \( n_c \leq m \) and some absolute constant \( c_4 > 0 \). Consequently,

\[
\mathbb{P}(z_C = 1 | |S_C| \leq m) \geq \frac{r^\alpha \beta}{1 + c_4}.
\]

**Proof.** By definition,

\[
dP_{S_C|z_C=0}(S_C) \bigg/ \prod_{j=1}^{n_c} dP_{X_j}(X_j \mid (X_i^C, Y_i^C)_{i < j}) = \left( \frac{\kappa}{L^*} \right)^{\sum_{i=1}^{n_c} 1(Y_i^C \leq L^*)} \left( \frac{1 - \kappa}{L - L^*} \right)^{\sum_{i=1}^{n_c} 1(Y_i^C > L^*)}
\]

\[
dP_{S_C|z_C=1}(S_C) \bigg/ \prod_{j=1}^{n_c} dP_{X_j}(X_j \mid (X_i^C, Y_i^C)_{i < j}) = \frac{1}{L^* - 1} \left( \sum_{\sigma=1}^{L^* - 1} \left( \frac{\kappa}{L^*} c_4 \right)^n_{\bar{\xi}} \left( \frac{\kappa}{L^*} - c_4 \right)^n_{\bar{\xi}} \left( \frac{\kappa}{L^*} \right)^{\sum_{i=1}^{n_c} 1(Y_i^C < L^*, Y_i^C \neq \sigma)} \left( \frac{1 - \kappa}{L - L^*} \right)^{\sum_{i=1}^{n_c} 1(Y_i^C > L^*)} \right)
\]

For a fixed \( S_C \),

\[
\frac{dP_{S_C|z_C=0}(S_C)}{dP_{S_C|z_C=1}(S_C)} = \frac{L^* - 1}{\sum_{\sigma=1}^{L^* - 1} \left( \frac{L^* c_4 r^\alpha}{\kappa} \right)^n_{\bar{\xi}} \left( 1 + \frac{L^* c_4 r^\alpha}{\kappa} \right)^n_{\bar{\xi}} \left( 1 - \frac{L^* c_4 r^\alpha}{\kappa} \right) \sum_{\sigma=1}^{L^* - 1} \frac{n_{\bar{\xi}}}{(L^* - 1)} \right]^{-1}
\]

where the last step is by Jensen’s inequality. Therefore,

\[
\int_{|S_C| = n_c} dP_{S_C|z_C=0}(S_C) \leq 2 \int_{|S_C| = n_c} \sum_{\sigma=1}^{L^* - 1} \frac{n_{\bar{\xi}}}{(L^* - 1) \geq n_{\bar{\xi}}} \cdot \frac{dP_{S_C|z_C=0}(S_C)}{dP_{S_C|z_C=1}(S_C)} \leq 2 \left( 1 - \frac{(L^*)^2 c_4^2 r^{2\alpha}}{\kappa^2} \right)^{-n_{\bar{\xi}}} \leq 2 \left( 1 - \frac{L^*}{9m} \right)^{n_{\bar{\xi}}} \leq c_4,
\]

30
for $c_4 = 2 \exp(1/9)$. Consequently,
\[
\mathbb{P}(z_C = 1| |S_C| \leq m) = \frac{\mathbb{P}(z_C = 1, |S_C| \leq m)}{\mathbb{P}(|S_C| \leq m)} \\
= \frac{\mathbb{P}(z_C = 1, |S_C| \leq m)}{\mathbb{P}(z_C = 1, |S_C| \leq m) + \mathbb{P}(z_C = 0, |S_C| \leq m)} \\
= \frac{\mathbb{P}(|S_C| \leq m| z_C = 1) \mathbb{P}(z_C = 1)}{\mathbb{P}(|S_C| \leq m| z_C = 1) \mathbb{P}(z_C = 1) + \mathbb{P}(|S_C| \leq m| z_C = 0) \mathbb{P}(z_C = 0)} \\
\geq \frac{r\alpha\beta}{1 + c_4}.
\]

### 2.5.2 Proof of Theorem 2.3 and 2.4

First, we define some quantities and notions that will be used in the lemmas.

**Definition 2.13.** Let $A$ be any measurable subset of $[0, 1]^d$ and $y \in [L]$. We define the regression function in $A$ for label $y$ as $\eta_y(A) = \int_A \eta_y(x) \, dx / \int_A dx$.

Given $n_A$ independent samples $\{(X_j^A, Y_j^A)\}_{j=1}^{n_A}$ in $A$, an unbiased estimator of $\eta_y(A)$ is
\[
\hat{\eta}_y(A) = \frac{1}{n_A} \sum_{j=1}^{n_A} \mathbb{1}(Y_j^A = y).
\]

To get a high probability bound, we focus the discussion on a subset under which the estimation error of $\hat{\eta}$ at each cell is small enough throughout the proof. We consider a favorable event $\xi_\alpha = \bigcap_{r \in \mathcal{I}} \bigcap_{C \in \mathcal{C}_r} \xi_{C, r, \alpha}$, where
\[
\mathcal{I} = \{2^{-k} : k \in \mathbb{N}\},
\]
and
\[
\xi_{C, r, \alpha} = \bigcap_{y \in \mathcal{L}_C^\alpha} \{|\hat{\eta}_y(C) - \eta_y(C)| \leq \lambda r^\alpha\}
\]
where $\mathcal{L}_C^\alpha$ is the remaining candidate labels for cell $C$ before elimination as defined in Algorithm 2.

The following lemma shows that $\xi_\alpha$ is indeed a high probability event.
Lemma 2.13. $\mathbb{P}(\xi_\alpha) \geq 1 - \delta_0$.

Proof. First, we show that $\mathbb{P}(\xi_{C_{r,\alpha}}^C) \leq \delta_0 r^{d+1}$. By union bound and small deviation version of Chernoff bound (Lemma 2.2),

$$\mathbb{P}(\xi_{C_{r,\alpha}}^C) \leq |\mathcal{L}_C^\alpha| \mathbb{P}(|\hat{\eta}_y(C) - \eta_y(C)| > \lambda r^\alpha) \leq |\mathcal{L}_C^\alpha| \exp\left[-2n_{r,\alpha}(\lambda r^\alpha)^2/p_{\max}\right]$$

where $p_{\max} \doteq \min\{1, 1/|\mathcal{L}_C^\alpha| + \tau_{2r,\alpha}\}$ is an upper bound for $\max_{y \in \mathcal{L}_C^\alpha} \eta_y(C)$. Hence, by the choice of $n_{r,\alpha}$:

$$\mathbb{P}(\xi_{C_{r,\alpha}}^C) \leq 2|\mathcal{L}_C^\alpha| \exp\left(-2n_{r,\alpha}(\lambda r^\alpha)^2/p_{\max}\right) \leq \delta_0 r^{d+1}.$$

Another application of union bound yields $\mathbb{P}(\xi_\alpha) \geq 1 - \sum_{r \in \mathcal{I}} r^{-d}\delta_0 r^{d+1} \geq 1 - \delta_0$.

Next, we show some desired properties of Algorithm 2 on the favorable event $\xi_\alpha$. In particular, Lemma 2.14 shows that Algorithm 2 never eliminates Bayes classes; Lemma 2.15 shows that Algorithm 2 predicts only Bayes classes in the area where the soft margin is large enough; Lemma 2.16 shows that the algorithm will at least reach some certain level $r_{\min}$ of partition.

Lemma 2.14. On the event $\xi_\alpha$, suppose that Algorithm 2 is in the depth that the partition is of sidelength $r$. For any $x \in [0, 1]^d$, we have $\eta_y(x) < \eta_y(x)$ for any $y \notin S_C$, where $x \in C \in \mathcal{C}_r$.

That is, the algorithm never eliminates Bayes classes.

Proof. Let $y^*$ be a Bayes class and $y^* \in \mathcal{L}_C^\alpha$ before elimination. By definition of $\xi_\alpha$ and smoothness assumption, we have

$$\hat{\eta}_{y^*}(C) \geq \eta_{y^*}(x) - |\eta_{y^*}(x) - \eta_{y^*}(C)| - |\hat{\eta}_{y^*}(C) - \eta_{y^*}(C)| \geq \eta_{y^*}(x) - 2\lambda r^\alpha,$$

Similarly

$$\max_{y \in \mathcal{L}} \hat{\eta}_y(C) \leq \max_{y \in \mathcal{L}} \{\eta_y(x) + |\eta_y(C) - \eta_y(x)| + |\hat{\eta}_y(C) - \eta_y(C)|\} \leq \eta_{y^*}(x) + 2\lambda r^\alpha.$$
Therefore, \( \hat{\eta}(1)(x) - \hat{\eta}_y(x) \geq 6\lambda r^\alpha \). Therefore, \( \max_{y \in [L]} \hat{\eta}_y(C) - \hat{\eta}_y^*(C) < \tau_{r,\alpha} \) and \( y^* \) will not be eliminated.

**Lemma 2.15.** On the event \( \xi_\alpha \), suppose that Algorithm 2 is in the depth that the partition is of side length \( r \). If \( \eta(1)(x) - \eta_y(x) \geq \Delta_r = 10\lambda r^\alpha \) for some \( x \in [0,1]^d \) and \( y \in [L] \), then for the cell \( C \in \mathcal{C}_r \) that contains \( x \), the label \( y \) will be eliminated. Consequently, for any \( x \in [0,1]^d \) with \( M(x) > \Delta_r \), \( \mathcal{L}^\alpha_\beta \) contains only Bayes classes.

**Proof.** By smoothness assumption, \( \eta(1)(C) - \eta_y(C) \geq \eta(1)(x) - \eta_y(x) - 2\lambda r^\alpha = 8\lambda r^\alpha \). Let \( y \) be any label in \( \mathcal{L}^\alpha_\beta \) before elimination, we have \( |\eta_y(C) - \hat{\eta}_y(C)| \leq \lambda r^\alpha \), and hence

\[
\hat{\eta}(1)(C) - \hat{\eta}_y(C) \geq |\eta(1)(C) - \eta_y(C)| - |\eta_y(C) - \hat{\eta}_y(C)| - |\eta(1)(C) - \hat{\eta}(1)(C)| \\
\geq |\eta(1)(C) - \eta_y(C)| - 2\lambda r^\alpha \geq 6\lambda r^\alpha.
\]

**Lemma 2.16.** On the event \( \xi_\alpha \),

i) If \( P_{X,Y} \in \mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L_{\min}^*, L_{\max}^*) \), then the finest partition Algorithm 2 can reach satisfies

\[
r_{\min} \leq \left( \frac{c_6}{n_0 \lambda^2 L_{\min}^*} \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right) \right)^{1/(2\alpha + d)};
\]

for some \( c_6 > 0 \);

ii) If \( P_{X,Y} \in \mathcal{P}_2(c_d, \lambda, \alpha, \varepsilon_0, C_\beta, \beta', L_{\min}^*, L_{\max}^*) \), then

\[
r_{\min} \leq \max \left\{ \left( \frac{c_7 \varepsilon_0 \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)}{n_0 \lambda^2 L_{\min}^*} \right)^{1/2\alpha + d}, \left( \frac{c_7 \lambda \beta' \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)}{n_0 \lambda^2 L_{\min}^*} \right)^{1/2\alpha + d + \alpha \beta'} \right\},
\]

for some \( c_7 > 0 \).
Proof. i) Without loss of generality, we assume that \( n \) is large enough so that we can at least reach the level \( r \) s.t. \( 1/(2L) \geq \tau_{r,\alpha} = 6\lambda r^\alpha \). Note that the probability gap between the Bayes class and any non-effective label is at least \( 1/(2L) \), therefore \( L^*_C \) only contains effective class labels and \( L^*_{\min} \leq |L^*_C| \leq L^*_{\max} \). Hence,

\[
n_{r,\alpha} \leq \frac{1}{L^*_\min} \log \left( \frac{8L^*_\max}{\delta_0 r^{d+1}} \right) / (\lambda r^\alpha)^2.
\]

Suppose \( r_{\min} \) is the finest partition the algorithm can reach, i.e., the total budget is not sufficient for length \( r_{\min}/2 \). Then,

\[
n_0 \leq \sum_{r \in I: r \geq r_{\min}/2} |\mathcal{A}_r| n_r
\]

\[
\leq \sum_{r \in I: r \geq r_{\min}/2} \frac{r^{-d}}{L^*_\min} \log \left( \frac{8L^*_\max}{\delta_0 r^{d+1}} \right) / (\lambda r^\alpha)^2
\]

\[
\leq \frac{\log \left( \frac{8L^*_\max}{(\delta_0 r_{\min}^{d+1})} \right)}{\lambda^2 L^*_\min} \sum_{r \in I: r \geq r_{\min}/2} r^{-(2\alpha + d)}
\]

\[
\leq c_8 \log \left( \frac{8L^*_\max}{(\delta_0 r_{\min}^{d+1})} \right) \frac{r_{\min}}{r_{\min}^2}
\]

where \( c_8 > 0 \) is independent of \( r_{\min}, \delta_0, L^*_\min, L^*_\max, \lambda \) and \( n_0 \), the last inequality is due to the geometric growth of the summands as \( r < 1 \). We now prove an upper bound for \( \log \left( \frac{8L^*_\max}{(\delta_0 r_{\min}^{d+1})} \right) \).

Use the trivial bound

\[
\frac{r_{\min}^{-d}}{L^*_\min} \cdot \frac{\log \left( \frac{8L^*_\max}{(\delta_0 r_{\min}^{d+1})} \right)}{\lambda^2 r_{\min}^{2\alpha}} = n_{r_{\min},\alpha} \leq n_0,
\]

and \( \frac{r_{\min}^{-d}}{L^*_\min} > 1 \), \( 8L^*_\max/(\delta_0 r_{\min}^{d+1}) > 8/\delta_0 > 1 \), we have \( (\lambda^2 r_{\min}^{2\alpha})^{-1} \leq n_0 \), which implies \( r_{\min} \geq (\lambda^2 n_0)^{-1/2\alpha} \), and therefore

\[
\log \left( \frac{8L^*_\max}{(\delta_0 r_{\min}^{d+1})} \right) \leq \log \left( \frac{8L^*_\max}{\delta_0} (\lambda^2 n_0)^{(d+1)/2\alpha} \right) \leq \frac{d + 1}{2\alpha} \log \left( \frac{8L^*_\max}{\delta_0} \lambda^2 n_0 \right).
\] (2.7)

where the latter step is due to \( (d + 1)/2\alpha > 1 \). With this upper bound (2.7), we now proceed to
upper bound $r_{\min}$. Clearly,

$$n_0 \leq \frac{c_6}{\lambda^2 L^*_\min} \log \left( \frac{8L^*_\max \lambda^2 n_0}{\delta_0} \right) r_{\min}^{-(2\alpha+d)}$$

where $c_6 = \frac{c_8(d+1)}{2\alpha}$. Therefore,

$$r_{\min} \leq \left( \frac{c_6}{n_0 \lambda^2 L^*_\min} \log \left( \frac{8L^*_\max \lambda^2 n_0}{\delta_0} \right) \right)^{1/(2\alpha+d)}.$$  

ii) From the strong density condition and Lemma 2.15, we have a tighter bound on the number of active cells:

$$|A_r| \leq \varepsilon_0 + C_\beta(6\lambda r^\alpha)^{3\beta'} \frac{c_7}{cd^{r^d}}.$$  

Using a similar argument as in i), we have

$$n_0 \leq \sum_{r \in \mathcal{I}: r \geq r_{\min}/2} |A_r| n_r$$

$$\leq \sum_{r \in \mathcal{I}: r \geq r_{\min}/2} \varepsilon_0 + C_\beta(6\lambda r^\alpha)^{3\beta'} \frac{r^{-d} \log(8L^*_\max / (\delta_0 r^{d+1}))}{\lambda^2 L^*_\min r^{2\alpha}}$$

$$\leq \frac{c_7 \lambda^2 L^*_\min}{n_0 \lambda^2 L^*_\min} \log \left( \frac{8L^*_\max \lambda^2 n_0}{\delta_0} \right) \max \left\{ \varepsilon_0 r_{\min}^{-(2\alpha+d)}, \lambda^{\beta'} r_{\min}^{-(2\alpha+d-\alpha\beta')} \right\}.$$  

where $c_7 > 0$ is independent of $r_{\min}, \delta_0, L^*_\min, L^*_\max, \lambda$ and $n_0$. Therefore,

$$r_{\min} \leq \max \left\{ \left( \frac{c_7 \varepsilon_0 \log \left( \frac{8L^*_\max \lambda^2 n_0}{\delta_0} \right)}{n_0 \lambda^2 L^*_\min} \right)^{\frac{1}{2\alpha+d}}, \left( \frac{c_7 \lambda^{\beta'}}{n_0 \lambda^2 L^*_\min} \log \left( \frac{8L^*_\max \lambda^2 n_0}{\delta_0} \right) \right)^{\frac{1}{2\alpha+d-\alpha\beta'}} \right\}.$$  

Now we establish an upper-bound for the excess risk rate of the non-adaptive subroutine.

**Proposition 2.17 (Guarantees for Algorithm 2).** Let $n_0 \in \mathbb{N}$ and $\alpha \beta' \leq d$. Let $\{S_C\}_{C \in \mathcal{C}_{\delta_0}}$ be the outputs of Algorithm 2 with input $n_0$, $\lambda$, $\alpha$ and $\delta_0 \in (0, 1)$, and $\hat{h}_{n_0,\alpha}$ be any classifier that satisfies $\hat{h}_{n_0,\alpha}(x) \in S_C, \forall x \in C \in \mathcal{C}_{\delta_0}$.
i) Suppose that \( P_{X,Y} \in \mathcal{P}_1(\lambda, \alpha, C_\beta, \beta, L_{\min}^*, L_{\max}^*) \). With probability at least \( 1 - \delta_0 \),

\[
\mathcal{E} \left( \hat{h}_{n_0, \alpha} \right) \leq C_4 \left( \frac{\lambda^\frac{d}{\alpha} \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)}{n_0 L_{\min}^*} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}}
\]

ii) Suppose that \( P_{X,Y} \in \mathcal{P}_2(c_d, \lambda, \alpha, \epsilon_0, C_\beta, \beta, \beta', L_{\min}^*, L_{\max}^*) \). With probability at least \( 1 - \delta_0 \),

\[
\mathcal{E} \left( \hat{h}_{n_0, \alpha} \right) \leq C_5 \left( \frac{\alpha(\beta+1)}{2\alpha+d} \right) \left( \frac{\lambda^\frac{d}{\alpha} \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)}{n_0 L_{\min}^*} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}} + \left( \frac{\lambda^\frac{d}{\alpha} \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)}{n_0 L_{\min}^*} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d-\alpha \beta'}}
\]

for some constant \( C_4, C_5 > 0 \), which are independent of \( n_0, \lambda, L, \epsilon_0 \) and \( \delta_0 \).

**Proof.** (Proof of Proposition 2.17)

i) On the event \( \xi_\alpha \) with probability at least \( 1 - \delta_0 \), we have by Part i) of Lemma 2.16,

\[
\Delta_{r_{\min}} = 10 \lambda r_{\min}^\alpha \leq 10 \lambda \left( \frac{c_6 \lambda^{-2} \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)}{n_0 L_{\min}^*} \right)^{\frac{\alpha}{2\alpha+d}} \leq 10 \left( \frac{c_6 \lambda^{-2} \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)}{n_0 L_{\min}^*} \right)^{\frac{\alpha}{2\alpha+d}}.
\]

By Lemma 2.15, the classifier \( \hat{h}_{n_0, \alpha} \) makes no error at \( \{ x : \mathcal{M}(x) > \Delta_{r_{\min}} \} \), and thus

\[
\mathcal{E}(\hat{h}_{n_0, \alpha}) \leq \mathbb{P}(\mathcal{M}(X) \leq \Delta_{r_{\min}}) \cdot \Delta_{r_{\min}} \leq C_\beta \Delta_{r_{\min}}^{\beta+1} \leq C_4 \left( \frac{\lambda^\frac{d}{\alpha} \log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)}{n_0 L_{\min}^*} \right)^{\frac{\alpha(\beta+1)}{2\alpha+d}},
\]

where \( C_4 = C_\beta 10^{\beta+1} c_6^{\frac{\alpha(\beta+1)}{2\alpha+d}} \). ii) On \( \xi_\alpha \) with probability at least \( 1 - \delta_0 \), we have by Part ii) of Lemma 2.16,
\[ \Delta_{r_{\min}} \leq 10 \max \left\{ \frac{\varepsilon_0 \lambda^d}{\log \left( \frac{8L_{\max}^* \lambda^2 n_0}{\delta_0} \right)} \right\}^{\frac{\alpha}{2\alpha+d}}, \left( \frac{c_7 \lambda^d}{n_0 L_{\min}^*} \right)^{\frac{\alpha}{2\alpha+d}} \leq 10 \max \{Q_1, Q_2\} \]

Case 1: \( Q_1 \leq Q_2 \)

Under this case, it is clear that \( \varepsilon_0 \leq c_9 \Delta_{r_{\min}}^{\beta'} \) for some \( c_9 > 0 \). Therefore,

\[ \mathcal{E}(\hat{h}_{n,\alpha}) \leq \mathbb{P}(\mathcal{M}(X) \leq \Delta_{r_{\min}}) \Delta_{r_{\min}} \leq \mathbb{P}(\mathcal{M}'(X) \leq \Delta_{r_{\min}}) \Delta_{r_{\min}} \leq C_{\beta} (\varepsilon_0 + \Delta_{r_{\min}}^{\beta'}) \Delta_{r_{\min}} \leq C_{\beta} (c_9 + 1) \Delta_{r_{\min}}^{\beta'+1} \]

\[ \leq C_5' \left( \frac{\lambda^d}{n_0 L_{\min}^*} \right)^{\frac{\alpha}{2\alpha+d}} \leq C_5'' \left( \frac{16 \lambda^2 n_0}{P_{\min} \delta_0} \right)^{\frac{\alpha}{2\alpha+d}}, \]

where \( C_5'' = C_{\beta} (c_9 + 1) 10^{\beta'+1} c_7^{\frac{\alpha}{2\alpha+d}} \).

Case 2: \( Q_1 > Q_2 \)

Under this case,

\[ \mathcal{E}(\hat{h}_{n_0,\alpha}) \leq \mathbb{P}(\mathcal{M}(X) \leq \Delta_{r_{\min}}) \Delta_{r_{\min}} \leq C_{\beta} \Delta_{r_{\min}}^{\beta+1} \leq C_5'' \left( \frac{\lambda^d}{n_0 L_{\min}^*} \right)^{\frac{\alpha}{2\alpha+d}} \leq C_5'' \left( \frac{16 \lambda^2 n_0}{P_{\min} \delta_0} \right)^{\frac{\alpha}{2\alpha+d}}, \]

where \( C_5'' = C_{\beta} 10^{\beta+1} c_7^{\frac{\alpha}{2\alpha+d}} \). Finally, set \( C_5 = \max\{C_5', C_5''\} \) we get the desired result.

Proof. (Proof of Theorem 2.3 and 2.4). Due to their similarity, we only prove Theorem 2.3, and omit the proof of Theorem 2.4. The bound is trivial for \( \alpha < \frac{1}{\log(n)} \), since \( n^{-\alpha} \geq n^{-1/\log(n)} \geq \frac{1}{e} \). Thus, we will consider \( \alpha \geq \frac{1}{\log(n)} \). Let \( \delta_0 = \delta / ([\log(n)]^3) \) and \( \alpha_i = i / [\log(n)]^3 \) for \( i \in \mathbb{Z} \).
\[ \lceil \log(n) \rceil^3 \], as defined in Algorithm 2. Let \( i^* \) be the largest integer \( i \in \lceil \log(n) \rceil^3 \) such that \( \alpha_i \leq \alpha \). By Lemma 2.13 and 2.14, on \( \xi_{\alpha_i} \) with probability at least \( 1 - \delta_0 \), we have

\[ \forall C \in \mathcal{C}_r, \forall x \in C, \arg\max_y \eta_y(x) \in L_{\alpha_i}^C. \]

By a union bound, with probability at least \( 1 - \lceil \log(n) \rceil^3 \delta_0 = 1 - \delta \), above holds jointly for all \( i \leq i^* \). Thus, with probability at least \( 1 - \delta \),

\[ \forall C \in \mathcal{C}_r, \forall x \in C, \arg\max_y \eta_y(x) \subseteq \cap_{i \leq i^*} L_{\alpha_i}^C, \]

and hence \( \cap_{i \leq i^*} L_{\alpha_i}^C \neq \emptyset \). Therefore, \( L_C \subseteq L_{\alpha_i}^C \) for any \( C \in \mathcal{C}_r \). By proposition 2.17 and the fact that budget for each \( \alpha_i \) is \( n_0 = \frac{n}{\lceil \log(n) \rceil^3} \), we have

\[
\mathcal{E}\left( \hat{h}_n \right) \leq C_5 \left( \varepsilon_{0}^{\alpha_{i^*}(\beta + 1) \over 2 \alpha_{i^*} + d} \left( \frac{d}{\alpha_{i^*}} \log \left( \frac{8\lambda_{\max}^{2}n_0^{\alpha_{i^*}}}{\delta_{0}^{\alpha_{i^*}}} \right) \right) + \left( \frac{\alpha_{i^*}(\beta + 1) \over 2 \alpha_{i^*} + d - \alpha_{i^*} \beta} {2 \alpha_{i^*} + d - \alpha_{i^*} \beta} \right) \log \frac{3(n)}{\lceil \log(n) \rceil^3} \right)
\]

It remains to argue that going from \( \alpha_i^* \) to \( \alpha \), we add at most a constant multiplicative factor to the excess risk bound. Notice that

\[
\frac{\alpha(1 + \beta)}{2 \alpha + d} - \frac{\alpha_i^*(1 + \beta)}{2 \alpha_{i^*} + d} \leq \frac{1 + \beta}{2 \alpha \lceil \log(n) \rceil^3} \leq \frac{1 + \beta}{2 \log^2(n)} \cdot \frac{\log^3(n)}{\lceil \log(n) \rceil^3}
\]

where the last step is due to \( \alpha \geq \frac{1}{\lceil \log(n) \rceil^3} \). Similarly,

\[
\frac{\alpha(1 + \beta')}{2 \alpha + d - \alpha \beta'} - \frac{\alpha_i^*(1 + \beta')}{2 \alpha_{i^*} + d - \alpha_{i^*} \beta'} \leq \frac{(1 + \beta')(2 \alpha + d)}{2 \alpha + d - \alpha \beta'} \leq \frac{(1 + \beta')(2 \alpha + d)}{\log^3(n)(2 \alpha + d - \alpha \beta')} \frac{\log^3(n)}{\lceil \log(n) \rceil^3}
\]

\[
\leq \frac{(1 + \beta')(2 \alpha + d) \cdot \log^3(n)}{\log^3(n)(2 \alpha + d - \alpha \beta')} \frac{\log^3(n)}{\lceil \log(n) \rceil^3} \leq \frac{(1 + \beta')(2 + d)}{4 \log^2(n) \alpha^2} \cdot \frac{\log^3(n)}{\lceil \log(n) \rceil^3}
\]

\[
\leq \frac{(1 + \beta')(2 + d)}{4 \log(n)} \cdot \frac{\log^3(n)}{\lceil \log(n) \rceil^3}
\]
where the last step is due to $\alpha \geq \frac{1}{\log(n)}$. Therefore, for $n$ sufficiently large,

$$\left( \log^3(n) \lambda^{\frac{d}{n}} \log \left( \frac{8L_{\max}^2 \lambda^2 n}{\delta} \right) \right)^{-\frac{\alpha_{\star}(1+\beta)}{2\alpha + d} + \frac{\alpha_{\star}(1+\beta')}{2\alpha_{\star} + d}} \leq 2e^{\frac{1+\beta}{\log(n)}},$$

$$\left( \log^3(n) \lambda^{\frac{d}{n}} \log \left( \frac{8L_{\max}^2 \lambda^2 n}{\delta} \right) \right)^{-\frac{\alpha_{\star}(1+\beta')}{2\alpha + d-\alpha_{\star}d} + \frac{\alpha_{\star}(1+\beta')}{2\alpha_{\star} + d-\alpha_{\star}d}} \leq 2e^{(1+\beta')(2+d)/4}$$

and hence Theorem 2.3 holds with for $C_2 = 2e^{(1+\beta')(2+d)/4}C_5$.

### 2.6 Experiments

In this section, we demonstrate through a simulation study how non-unique Bayes classes affect the gain in active learning over passive learning.

**Data Distribution** The joint distribution $P_{X,Y}$ is supported on $[0, 1] \times \{1, 2, 3\}$, characterized by the marginal distribution $P_X \sim \text{Unif}(0, 1)$, and the regression function:

$$\eta_1(x) = \begin{cases} 
3x/(3 - 4\varepsilon_0), & 0 \leq x \leq (3 - 4\varepsilon_0)/8 \\
3/8, & (3 - 4\varepsilon_0)/8 < x \leq 3/8 \\
3/8 + \sin(40\pi x), & 3/8 < x \leq 5/8 \\
3/8, & 5/8 < x \leq (5 + 4\varepsilon_0)/8 \\
1 + 3(x - 1)/(3 - 4\varepsilon_0), & (5 + 4\varepsilon_0)/8 < x \leq 1 
\end{cases}$$

$\eta_2(x) = 3/4 - \eta_1(x)$ and $\eta_3(x) = 1/4$. Here, one can easily verify that the parameter $\varepsilon_0 = P_X(\eta_1(X) = \eta_2(X) > \eta_3(X))$ is the mass of region where the Bayes classes are non-unique. See Figure 2.3 for the plot of the regression function.
Figure 2.3: The plot of the regression function. The classification task is relatively easy when \( x \) is near the two endpoints of the \([0, 1]\) interval, where the Bayes class is unique with a large margin. The task is more challenging when \( x \) is around the center of the interval, where the regression function oscillates rapidly. The Bayes classes are not unique on the two orange dash-dot intervals, where both Class 1 and 2 are Bayes. The total mass of the region where Bayes classes are non-unique is controlled by the parameter \( \varepsilon_0 \).

Classifiers  We compare the performance of the non-adaptive active learner defined as in Algorithm 2 and its passive counterpart. The passive learner samples uniformly on the \([0, 1]\) interval, then partitions the interval into \( n^{1/(2\alpha+1)} \) sub-intervals and takes majority votes as its prediction within each partition. The choice of the number of partitions is known to be optimal, see [25, 19]. Throughout, we assume that both the active and passive learners know the smoothness parameters \( \alpha = 1, \lambda = 15\pi \).

Evaluation and Results  The classifiers are trained with different sampling budgets under multiple levels of \( \varepsilon_0 \). A test dataset of size 100,000 is generated and reserved for the evaluation of the classifiers.

Figure 2.4 shows how the gain of the active learner evolves with changing sampling budgets under two extreme choices of \( \varepsilon_0 \). When \( \varepsilon_0 = 0 \), we observe a much faster downward trend for the empirical excess risk of the active learner than that of the passive learner. In this case, the active learner can quickly decide the Bayes class in those regions with large margin and use the
Figure 2.4: The empirical excess risk of the active learner and passive learner versus the sampling budget, for \( \varepsilon_0 = 0 \) (left) and \( \varepsilon_0 = 0.6 \) (right). Each dot represents the mean empirical excess risk over 10 replications, and the error bars stand for the standard deviation.

The majority of the budget where the regression function is highly-oscillated. When \( \varepsilon_0 = 0.6 \), the gain of the active learner is significantly reduced. The active learner cannot differentiate between the real difficult region and the region where the Bayes classes are non-unique, and hence fails to save budget as efficiently. Nonetheless, there is still limited improvement in the small \( n \) regime, which agrees with Theorem 2.3.

Figure 2.5 presents the effect of \( \varepsilon_0 \) on the gain of the active learner on a finer scale. For each different value of \( \varepsilon_0 \), ranging from 0 to 0.6, we calculate the ratios of the active empirical excess risk to the passive one. We observe an upward trend in the ratio with respect to \( \varepsilon_0 \). From the level around \( \varepsilon_0 = 0.5 \) and above, the active learner has almost no advantage over the passive learner.
Figure 2.5: The performance ratio of the active learner and passive learner versus $\varepsilon_0$. Each dot represents the mean performance ratio over 10 replications, and the error bars stand for the standard deviation. The horizontal reference line of level 1 represents the case where the active learner has absolutely no gain over the passive learner.
Chapter 3: Efficient Estimation of Central Mean Subspace

3.1 Introduction

We consider a regression model for data \((X, Y) \in \mathbb{R}^d \times \mathbb{R}\) in which the regression function 
\[ g(x) = E[Y \mid X = x] \] 
has the form

\[ g(x) = f(U^T x) \]

for some unknown matrix \(U \in \mathbb{R}^{d \times k}\) of rank \(k\) and some unknown function \(f: \mathbb{R}^k \rightarrow \mathbb{R}\). The goal of this work is to provide computationally tractable methods for estimating the column space of \(U\), denoted \(\text{col}(U)\), from a finite sample of \((X, Y)\).

The model described above is called a multi-index model in the literature, and \(\text{col}(U)\) is referred to as the index space. The index space is also known as the central mean subspace in the context of sufficient dimension reduction [26, 27] under mild identifiability conditions on \(\text{supp}(X)\). An accurate estimate of the index space has many potential applications, including the following:

- The index space has, in the context of regression, an explanatory role analogous to that of the leading principal components of the data. In particular, it reveals which (linear combinations of) covariates are able to best predict the response \(Y\).

- If \(k \ll d\), then the achievable rates for estimating the regression function \(g\) may be greatly improved by first projecting the covariates \(x\) to the index space before performing estimation. For instance, if \(g\) is only known to be Lipschitz, then we may hope to improve the rate of estimation from \(n^{-1/(d+2)}\) to \(n^{-1/(k+2)}\).

1See Section 3.3.2 for a discussion of these conditions.
The problem of estimating the index space has received much attention, but important gaps remain in terms of what approximation error rates are achievable by computationally tractable estimators. Results may be cataloged into two main directions, one emphasizing fast convergence rates\(^2\) but requiring strict distributional conditions, the other preferring general distributional settings but admitting significantly worse convergence rates. For example, on one hand, so-called inverse regression approaches—which estimate the index space via functionals of \(\mathbb{E}[X \mid Y]\)—often admit fast convergence rates of the form \(O(n^{-1/2})\) however at the cost of restrictive conditions tying the marginal distribution \(P_X\) to \(\text{col}(U)\); furthermore, outside of these conditions on \(P_X\), only a strict subspace of \(\text{col}(U)\) may be recovered. On the other hand, forward regression methods—which leverage functionals of the regression function \(g := \mathbb{E}[Y \mid X]\), e.g., its derivatives—can ensure full recovery of the index space, importantly without un-naturally forcing a relation between \(P_X\) and \(\text{col}(U)\); however such milder distributional conditions often come at the cost of significantly slower estimation rates, often of a nonparametric form \(n^{-1/\Omega(k)} \gg n^{-1/2}\).

In this work, we try to elucidate the following middle-ground: we show that fast rates of the form \(C_d \cdot n^{-1/2}\) remain possible for a reweighted variant of forward methods, for a prefactor \(C_d\) depending on data distribution. While our distributional conditions—knowledge of the marginal distribution \(P_X\) and certain moment assumptions—are stricter than for some forward regression methods, our rate guarantees are of a parametric form even for nonparametric link \(g\), similar to usual convergence rates for inverse regression methods. Yet, unlike inverse regression methods, we guarantee full recovery of the index space \(\text{col}(U)\) under qualitatively milder distributional conditions, where in particular, we impose no condition on the relation between \(P_X\) and the unknown index space \(\text{col}(U)\) as is common in that literature. Our result relies on estimating outer products of a smoothed version of \(\nabla f\), where the smoothing itself—induced by careful reweighting of the training sample—leverages a first-order Stein’s Lemma to guarantee full recovery of \(\text{col}(U)\). Similar to other existing analyses, we use knowledge of \(P_X\) to properly reweighting of the training sample.

\(^2\)Under suitable subspace approximation measures, e.g., angles between spaces (see Section 3.3.2).
**Detailed Contributions**  We introduce an object called the *Expected Smoothed Gradient Outer Product (ESGOP)* $\overline{M}$, which can be viewed as a surrogate for the EGOP. The ESGOP is based on a *smoothed* version of the gradient $\nabla g(X)$, as well as a subtle change of distribution (from $P_X$ to a Gaussian) that enables application of Stein’s inequality.

We first show that $\overline{M}$ is *exhaustive* in that $\text{col}(\overline{M}) = \text{col}(U)$ (Proposition 3.1). The surrogate $\overline{M}$ can then be estimated from a finite sample from $P_{X,Y}$ via a reweighting scheme; in particular, the estimation of the *smoothed* gradients may be interpreted as local linear estimator with a data-dependent kernel (Remark 3.3).

Our main result states that, for a general class of distributions $P_X$ admitting Gaussians and heavier tail distributions, our estimator converges at a parametric rate of the form $C_d \cdot n^{-1/2}$. Finally, to elucidate the leading constant $C_d$, we consider the special case of polynomial links $g$—which has received much recent interest—and show that it behaves as $d^{O(r)}$ for polynomials of degree at most $r$ (Corollary 3.7), in line with recent lower-bounds of [14] for certain classes of computationally tractable procedures.

The remainder of the chapter is organized as follows. We start with a detailed review of related theoretical works in Section 3.2. In Section 3.3, we introduce the basic setups and some notations. In Section 3.4, we propose a novel estimator for central mean subspace based on the expected smoothed gradient outer product, and provide upper-bounds for the estimation errors for general link functions and polynomial link functions. In Section 3.5, we demonstrate the choice of hyperparameters in the main algorithm through a simulation study. In Section 3.6, we discuss a number of related issues that are not detailed in the main results. In Section 3.7, we provide the full analysis of the main results presented.

### 3.2 Prior Theoretical Works

To better situate our results, we expand next on the state-of-the-art on the subject, followed by a more detailed overview of the results.

The problem of index space estimation has received much attention over the last few decades
from both statistics and machine learning communities [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39]. For ease of comparison, we first note that a prevalent approach in the literature, also adopted here, is to proceed by defining and estimating a population object $M \in \mathbb{R}^{d \times d}$, whose column space estimates (part of) $\text{col}(U)$; the object $M$ is typically a functional of either the inverse or forward conditional distribution $X \mid Y$ or $Y \mid X$.

**Inverse regression methods** As the name suggests, the class of inverse regression methods leverage functionals of the inverse relation $X \mid Y$ that may reveal the index space or a subspace thereof. Although these methods are initially designed to estimate the central subspace, they can also serve the purpose of estimating the central mean subspace when the covariates $X$ and the error $\varepsilon = Y - g(X)$ are independent, under which case these two subspaces are identical. For instance, Sliced Inverse Regression (SIR) by [28] estimates $M = \text{Cov}[E[X \mid Y]]$ and Sliced Average Variance Estimator (SAVE) by [30] estimates $M = E[(I_d - \text{Cov}[X \mid Y])^2]$. By carefully estimating means or covariances over slices, i.e., partitions of the $Y$ space, parametric convergence rates of the form $O(n^{-1/2})$ may be guaranteed, albeit under restrictive distributional conditions. For instance, SIR imposes a so-called linear conditional mean condition (where $E[X \mid U^T X]$ is linear in $U^T X$) in order to ensure $\text{col}(M) \subset \text{col}(U)$, i.e., in order to guarantee partial recovery of $\text{col}(U)$; if in addition we impose a so-called random conditional projection condition (requiring for all $v \in \mathbb{R}^d$ that $E[v^T X \mid Y]$ is not constant as a r.v.) then one can ensure $\text{col}(M) = \text{col}(U)$. These distributional conditions, by restricting the relation between $P_X$ and the index space $\text{col}(U)$, rule out some rather benign settings: for instance, for $k = 1$, it is well known that under Gaussian $P_X$, these conditions cannot hold for even link functions such as $f(z) = z^2$ (also referred as phase retrieval problem). Similar examples of failure modes are discussed for SAVE in Appendix 3.6.1. Additional discussions and examples for non-exhaustiveness issues can be found in [37] and [36].

**Forward regression methods** In contrast to inverse regression methods, the class of forward regression methods leverage functionals of the regression function $E[Y \mid X]$. An important line of work relies on Stein’s Lemma in constructing $M$. The earliest work along this line by [40]
considered a single-index model ($k = 1$) estimates the index space direction, say $u \in \mathbb{R}^d$, via $M = \mathbb{E}[Y \cdot X]$: suppose that $P_X$ were standard Gaussian, and assuming an additive noise model, then Stein’s Lemma implies that $\mathbb{E}[Y \cdot X] = \mathbb{E}[f'(u^T X)] \cdot u$. For the general case $k > 1$, [41] proposed an estimator based on a second-order Stein’s Lemma that is $\sqrt{n}$-consistent. Similar to our work, they assume the setting on known $P_X$, and admit a broad class of distributions including Gaussian and more heavy-tailed distributions. However, their theoretical results relies on the non-singularity of the expected Hessian $\mathbb{E}[\nabla^2 g(X)]$, again ruling out some apparently benign settings: for instance for $k = 1$, and $\mathbb{E}X = 0$, simple link functions such as $f(z) = z^3$ are not admitted. In contrast, the approach analyzed here relies on a first-order Stein’s Lemma (Lemma 3.3) and thus imposes no condition on the Hessian or any other higher order derivatives of $g$.

Another class of forward regression methods are those that rely on functionals of the gradient of $g$. In the simplest case of single index models ($k = 1$), we may take $M = \mathbb{E}[\nabla g(X)] = \mathbb{E}[f'(u^T X)] \cdot u$, for an index direction $u$; thus, provided $\mathbb{E}[f'(u^T X)]$ does not vanish (similar to [40] discussed above), [33] designs an estimator of $u$ with convergence rate $n^{-1/2}$. In the more general case $k \geq 1$, one may consider a so-called expected gradient outer product (EGOP), defined as $M = \mathbb{E}[\nabla g(X)\nabla g(X)^T] = U \mathbb{E}[\nabla f(U^T X)\nabla f(U^T X)^T]U^T$. This approach is quite appealing as it holds that, under quite mild conditions (see Section 3.3.2), one can show that $\text{col}(M) = \text{col}(U)$. Thus, a number of estimators of the EGOP $M$ have been proposed over the years, e.g., [42, 31, 43, 44]. However, the convergence guarantees available for these methods are, at best, of the nonparametric form $n^{-1/\Omega(k)}$, without additional assumptions on $P_X$. For the restricted case of $k \leq 3$, [32, 34] shows that the parametric rate $n^{-1/2}$ may be achieved under known Gaussian noise and in a fixed design setting (under some conditions on the observed $X$ data itself). We have no such restriction on the index space dimension $k$, and operate in the more general random design setting, however while assuming known $P_X$.

Finally, the popular MAVE approach of [31], while still a forward regression method, differs considerably from all the above in that it does not explicitly define a surrogate object $M$ for $U$; instead, it works by optimizing a suitable objective over all possible projections $U$. MAVE is
appealing in that, similar to EGOP, it can guarantee full recovery of the index space under mild conditions. However, this also comes at the cost of a nonparametric rate of the form $n^{-1/\Omega(k)}$.

### 3.3 Problem Setting

#### 3.3.1 Notations

Throughout, we use bold font letters (e.g. $x$, $X$) to refer to vectors and matrices. Let $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote Euclidean inner product and norm on $\mathbb{R}^d$. For a probability measure $\nu$ on $\mathbb{R}^d$, we let $\langle f, g \rangle_\nu := \mathbb{E}_{X \sim \nu}[f(X) \cdot g(X)]$ and $\|f\|_\nu := \sqrt{\langle f, f \rangle_\nu}$ denote the inner-product and norm on the function space $L_2(\nu)$. The operator norm of a matrix $M$ (with respect to Euclidean norm) is denoted by $\|M\|_{\text{op}}$. For a positive integer $d$, we use $[d]$ to denote the set $\{1, \ldots, d\}$.

#### 3.3.2 Central Mean Subspace

Let $(X, Y) \sim P_{X,Y}$, with $P_X$ the marginal distribution assumed to be known, and let $g(X) := \mathbb{E}[Y \mid X]$ denote the regression function.

**Definition 3.1** (Mean dimension-reduction subspaces and central mean subspace [45]). A mean dimension-reduction subspace for $\mathbb{E}[Y \mid X]$ is a subspace $W \subseteq \mathbb{R}^d$ such that $\mathbb{E}[Y \mid X]$ is a function only of $V^T X$ for some matrix $V$ whose columns form a basis for $W$. Define

$$U := \bigcap \{W \subseteq \mathbb{R}^d : W \text{ is a mean dimension-reduction subspace for } \mathbb{E}[Y \mid X] \}.$$ 

We say $U$ is the central mean subspace (CMS) for $\mathbb{E}[Y \mid X]$ if $U$ is a mean dimension-reduction subspace. (If $U$ is not a mean dimension-reduction subspace, then the CMS does not exist.)

The CMS is known to exist under mild conditions on $\text{supp}(X)$, as discussed by [45]; an example sufficient condition is that $\text{supp}(X)$ is open and convex. Henceforth, we shall assume that $\text{supp}(X)$ satisfies a sufficient condition that guarantees the existence of the CMS. (Such an assumption will be explicit given later in Assumption 3.1.)
Let $k = \dim(U)$, and let the columns of $U := [u_1, \ldots, u_k] \in \mathbb{R}^{d \times k}$ be an orthonormal basis for $U$. Because $U$ is the CMS for $E[Y \mid X]$, the regression function has the form of a $k$-index model, specifically

$$g(X) = f(U^T X)$$

for some function $f : \mathbb{R}^k \to \mathbb{R}$, which we call the link function. Our goal is to estimate $U$ (or equivalently, $\text{col}(U)$) from an i.i.d. sample from $P_{X,Y}$. We evaluate our estimate using the following measure of error.

**Definition 3.2 (Distance with Optimal Rotation, Adapted from [46]).** The distance with optimal rotation between $U$ and its estimates $\hat{U}$ is defined as:

$$d(\hat{U}, U) := \min_{R \in O_{k \times k}} \|\hat{U}R - U\|_{\text{op}},$$

where $O_{k \times k}$ is the set of all $k \times k$ orthogonal matrices.

The distance with optimal rotation in Definition 3.2 turns out to be equivalent to the so-called $\sin\Theta$ distance, another popular notion of distance between subspaces given as follows. Let $U_\perp \in \mathbb{R}^{d \times (d-k)}$ denote a matrix whose columns form an orthonormal basis for the orthogonal complement $U^\perp$ of $U$. Then, the $\sin\Theta$ distance between $U$ and $\hat{U}$ is defined as $\sin_{\Theta}(\hat{U}, U) := \|\hat{U}^T U_\perp\|_{\text{op}}$.

It can be shown (Lemma 2.6 from [46]) that there exists $c, C > 0$ such that

$$c \cdot d(\hat{U}, U) \leq \sin_{\Theta}(\hat{U}, U) \leq C \cdot d(\hat{U}, U).$$

The distance with optimal rotation is appealing in that it readily yields control on regression error when data is first projected to $\text{col}(\hat{U})$ rather than $U$. For instance, suppose $f$ is $L$-Lipschitz (e.g., as in linear regression, or as in polynomial regression with a bounded domain), and let $\tilde{f} :=$
\( f \circ R^T \), where \( R \in \mathbb{O}^{k \times k} \) is such that \( d(\hat{U}, U) = \| \hat{U} R - U \|_{\text{op}} \), then:

\[
\| \tilde{f}(\hat{U}^T X) - f(U^T X) \|_{P_X} := \| f((\hat{U} R)^T X) - f(U^T X) \|_{P_X} \leq L \cdot \| X \|_{P_X} \cdot d(\hat{U}, U).
\]

To estimate \( \tilde{f} \), one can apply the classical Nadaraya-Watson estimator on the projected sample of \((\hat{U}^T X, Y)\). If \( d(\hat{U}, U) \lesssim n^{-1/(k+2)} \), the regression error rate is the same as if \( U \) is known.

### 3.4 Main Results

We first introduce a surrogate object \( \mathcal{M} \) in Section 3.4.1 below, and show that it indeed is 

\textit{exhaustive}, i.e., it recovers all of \( \text{col}(U) \). Following this, we derive a separate surrogate object \( \tilde{\mathcal{M}} \) in Section 3.4.3 which approximates \( \mathcal{M} \) and turns out to be simpler to estimate and retains the

\textit{same exhaustiveness properties}. Our main algorithm then estimates \( \tilde{\mathcal{M}} \), and returns the column

\textit{space} \( \hat{U} \) of this estimate. Our main result of Theorem 3.6 which bounds the estimation error is

presented in Section 3.4.4, followed by an instantiation for polynomial link functions in Section 3.4.5.

We will need the following few assumptions.

**Assumption 3.1.** The joint distribution \( P_{X,Y} \) satisfies the following basic conditions.

1. The link function \( f \) is differentiable, with gradient \( \nabla f(z) \), \( \forall z \in \mathbb{R}^k \);

2. The marginal \( P_X \) admits a Lebesgue density \( p \) supported on \( \mathbb{R}^d \), \( \text{i.e., } p > 0 \) everywhere.

The assumption of differentiability above implies that of the regression function \( g \), and is needed as our approach relies on gradient estimates. The second assumption above will be re-

quired in designing the estimation procedure, and also ensures the existence of \( U \) as previously

remarked.
3.4.1 Expected Smoothed Gradient Outer Product

We now present a smoothed version of the expected gradient outer product

\[ M := \mathbb{E}_{X \sim P_X} [\nabla g(X) \nabla g(X)^\top], \]

which was discussed in the introduction.

**Definition 3.3.** The Expected Smoothed Gradient Outer Product (ESGOP) \( \overline{M} \) is defined as follows. First, let \( h > 0 \) be given, and define the smoothed gradient of \( g \) at \( \theta \in \mathbb{R}^d \) as:

\[ \beta_h(\theta) := U \cdot \nabla_h f(\theta), \quad \text{where} \quad \nabla_h f(\theta) := \mathbb{E}_{Z \sim N(0, h^2 I_d)} [\nabla f(U^\top (Z + \theta))]. \]

Let \( \sigma_\theta > 0 \). We then have:

\[ \overline{M} := \mathbb{E}_{\theta \sim N(0, \sigma_\theta^2 I_d)} [\beta_h(\theta) \beta_h(\theta)^\top] = U \left( \mathbb{E}_{\theta \sim N(0, \sigma_\theta^2 I_d)} [\nabla_h f(\theta) \nabla_h f(\theta)^\top] \right) U^\top. \quad (3.2) \]

The ESGOP \( \overline{M} \) differs from the EGOP \( M \) in two important ways. First, the gradient \( \nabla f(U^\top X) \) is replaced by a smoothed gradient \( \nabla_h f(\theta) \). Second, the expectation in the definition of \( \overline{M} \) is taken over a Gaussian distribution (i.e., \( \theta \sim N(0, \sigma_\theta^2 I_d) \)), instead of the data distribution \( P_X \) as in the EGOP. As we will see later in Section 3.4.3, such a design will allow us to invoke Stein’s Lemma (Lemma 3.3) for Gaussians.

**Exhaustiveness of the ESGOP \( \overline{M} \)** From Equation (3.2) it is clear that \( \text{col}(\overline{M}) \) is contained in \( \mathcal{U} := \text{col}(U) \). We argue here that the two column spaces are in fact equal. From the same equation, this boils down to arguing that

\[ \mathbb{E}_{\theta \sim N(0, \sigma_\theta^2 I_d)} [\nabla_h f(\theta) \nabla_h f(\theta)^\top]. \]
is full-rank. To gain some intuition, let's first revisit a similar argument for the EGO

\[ M = U (E_{X \sim P_X} \nabla f(U^T X) \nabla f(U^T X)^T) U^T. \]

Suppose \( M \) were not exhaustive, then one can find a direction \( v \in \mathbb{R}^k \) with \( \|v\| = 1 \), such that \( E_{X \sim P_X} [(v^T \nabla f(U^T X))^2] = 0 \). Equivalently, \( v^T \nabla f(U^T x) = 0 \) for \( \mathcal{P}_X \)-almost every \( x \). Then, \( Uv \in \mathbb{R}^d \) is a direction along which the directional derivative of \( g \) is zero almost surely. In other words, \( \mathcal{U} \) admits an irrelevant direction for the regression function \( g \), which contradicts the assumption that \( \mathcal{U} \) is the CMS and hence cannot be further reduced.

The exhaustiveness of \( \overline{M} \) is established in a similar way as above, with the added difficulty that we now have to instead argue about directional *smoothed* derivatives \( v^T \nabla_h f \) and how they might similarly lead to a contradiction on the minimality of \( \mathcal{U} \). This is established by relating \( v^T \nabla_h f \) to \( v^T \nabla f \) by a Hermite expansion of \( f \). Note that, now the contradiction on minimality of \( \mathcal{U} \) can only be established \( \mathcal{N} \)-a.s., which fortunately implies \( \mathcal{P}_X \)-a.s. since \( \mathcal{P}_X \) has a density and is therefore \( \mathcal{N} \) dominated. Such arguments lead to the following proposition whose proof is given in Section 3.7.1. The result requires an additional minor condition on square integrability of \( f \) so that Hermite expansions are well defined.

**Proposition 3.1 (Exhaustiveness of \( \overline{M} \)).** Suppose that Assumption 3.1 holds, and that the link function \( f \) satisfies \( E_{Z \sim \mathcal{N}(0, h^2 I_h)} [f(Z)^2] < \infty \). Then, for \( h, \sigma_h > 0 \), the ESGOP \( \overline{M} \) is exhaustive, i.e., we have \( \text{col} (\overline{M}) = \mathcal{U} \).

### 3.4.2 Average Smoothed Gradient Outer Product

We now derive a first approximation to \( \overline{M} \) by approximating the outer-expectation in Equation (3.2) from a random sample of \( \theta \) locations.

**Definition 3.4.** Let \( m, h > 0 \), and let \( \theta_1, \ldots, \theta_m \overset{i.i.d.}{\sim} \mathcal{N}(0_d, \sigma_\theta^2 I_d) \). The Average Smoothed Gradient Outer Product (ASGOP) is obtained as:

\[
\tilde{M} := \frac{1}{m} \sum_{j=1}^{m} \beta_h(\theta_j) \beta_h(\theta_j)^T = U \left( \frac{1}{m} \sum_{j=1}^{m} [\nabla_h f(\theta_j) \nabla_h f(\theta_j)^T] \right) U^T. \tag{3.3}
\]
We will require that \( m \) is sufficiently large so that \( \tilde{M} \) inherits the exhaustiveness of \( M \). This is established in the following Corollary 3.2 via a concentration argument on eigenvalues (see Section 3.7.2 for the detailed proof). We require the following moment condition for matrix concentration.

**Assumption 3.2.** The smooth gradient \( \nabla_h f \) satisfies:

\[
\mu_\nabla := \left( \mathbb{E}_{\theta \sim \mathcal{N}(0_d, \sigma^2 I_d)} \left[ \left\| \nabla_h f (\theta) \right\|^4 \right] \right)^{1/4} < \infty.
\]

Examples of link function \( f \) for which the condition \( \mu_\nabla < \infty \) is satisfied include Lipschitz functions (in which case \( \nabla f \) is bounded and hence so is \( \nabla_h f \)) and polynomials (see Lemma 3.19 (ii)).

**Corollary 3.2 (Exhaustiveness of \( \tilde{M} \)).** Suppose that Assumption 3.2 and the condition of Proposition 3.1 holds, and Let \( 0 < \delta < 1/2 \), if \( m \geq 8 \mu_\nabla^4 \lambda_k(M)^{-2} \delta^{-1} (\log (4d/\delta))^2 \), then with probability at least \( 1 - \delta \), the ASGOP \( \tilde{M} \) satisfies \( \text{col}(\tilde{M}) = \mathcal{U} \).

**Remark 3.1.** The above corollary is important in that we can now focus on estimating the simpler object \( \tilde{M} \), which turns out to be exhaustive given just \( m = O(1) \) locations \( \theta_j \)'s. Such intuition leads to the estimator of the next section, which our main theorem relies upon (Theorem 3.6). Note that, estimating \( M \) itself appears to require more resources, namely, \( m = O(n) \) (see Remark 3.7).

### 3.4.3 Estimation of CMS

We now present a simple estimator \( \hat{M} \) of the \( \tilde{M} \) from i.i.d. samples \( \mathcal{D} := \{ X_i, Y_i \}_{i=1}^n \sim P_{X,Y} \); let \( \hat{U} \) denote the top \( k \) eigenvectors of \( \hat{M} \), then \( \text{col}(\hat{U}) \) estimates \( \mathcal{U} \). We assume access to the density \( p \) of \( P_X \).

To this end, we first propose an unbiased estimator of the smoothed gradient \( \beta_h(\theta) \) of the regression function \( g \), by relying on the following Stein’s Lemma:

**Lemma 3.3 (Stein’s Lemma, [47]).** Let \( Z \sim \mathcal{N}(\theta, h^2 I_d) \) and suppose \( g : \mathbb{R}^d \to \mathbb{R} \) is differentiable
Algorithm 3 Estimation of the CMS $U$

1: **Input:** $h > 0$, $\sigma > 0$, $m \in \mathbb{Z}$, dataset $\mathcal{D} \subset \mathbb{R}^d \times \mathbb{R}$

2: Randomly sample $\{\theta_j\}_{j=1}^m$ from $\mathcal{N}(0_d, \sigma^2 I_d)$

3: Split $\mathcal{D}$ into $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ with equal sizes

4: Split each $\mathcal{D}_j$ into $\mathcal{D}_{j,1}$ and $\mathcal{D}_{j,2}$ with equal sizes

5: $\forall j \in [m]$, let $\hat{\beta}_{j,1} \leftarrow \hat{\beta}_h(\theta_j, \mathcal{D}_{j,1})$, and $\hat{\beta}_{j,2} \leftarrow \hat{\beta}_h(\theta_j, \mathcal{D}_{j,2})$ $\triangleright$ As defined in Equation (3.5).

6: $\hat{M} \leftarrow \frac{1}{2m} \sum_{j=1}^m (\hat{\beta}_{j,1} \hat{\beta}_{j,2}^\top + \hat{\beta}_{j,2} \hat{\beta}_{j,1}^\top)$

7: **Return:** $\hat{U} \in \mathbb{R}^{d \times k}$, the top-$k$ eigenvectors of $\hat{M}$

with $\mathbb{E}\|\nabla g(Z)\| < \infty$. We then have

$$\mathbb{E}[g(Z)(Z - \theta)] = h^2 \mathbb{E}[\nabla g(Z)].$$

For intuition, notice that we have $\beta_h(\theta) = \mathbb{E}_{Z \sim \mathcal{N}(0, h^2 I_d)} \nabla g(Z + \theta) = \mathbb{E}_{Z \sim \mathcal{N}(\theta, h^2 I_d)} \nabla g(Z)$.

Thus, suppose for a moment we were to sample $X \sim \mathcal{N}(\theta, h^2 I_d)$, then we would have

$$\mathbb{E} Y \cdot (X - \theta) = \mathbb{E} g(X) \cdot (X - \theta) = h^2 \beta_h(\theta),$$

that is $h^{-2} Y \cdot (X - \theta)$ would be unbiased for $\beta_h(\theta)$. Our main insight therefore is to proceed by importance-weighting, that is, reweighting the estimator as $\rho_h(X; \theta) \cdot Y \cdot (X - \theta)$ so as to recover expectation under $\mathcal{N}(\theta, h^2 I_d)$. In particular, recalling $p$ the density of $P_X$, we let

$$\rho_h(x; \theta) := \frac{\varphi_h(x - \theta)}{p(x)}, \quad \text{where } \varphi_h(\cdot) \text{ denotes the density of } \mathcal{N}(0_d, h^2 I_d). \quad (3.4)$$

**Estimating $\beta_h(\theta_j)$’s** Consider the $m$ values $\{\theta_j\}_{j \in [m]}$ drawn from $\mathcal{N}(0_d, \sigma^2 I_d)$, upon which $\hat{M}$ is defined (Definition 3.4). In order to evaluate $\beta_h(\theta_j)$’s, assume (without loss of generality) that the sample size $n$ is divisible by $2m$, and partition the dataset $\mathcal{D}$ into independent $\{\mathcal{D}_{j,\ell}\}_{j \in [m], \ell = 1, 2}$.
of equal size. We then define the following estimators:

\[
\hat{\beta}_h(\theta_j; D_{j,\ell}) := \frac{h^{-2}}{|D_{j,\ell}|} \sum_{(X,Y) \in D_{j,\ell}} \rho_h(X; \theta_j) \cdot Y \cdot (X - \theta_j).
\] (3.5)

Thus, for each \( \theta_j \), we have two independent estimates \( \hat{\beta}_{j,\ell} := \hat{\beta}_h(\theta_j; D_{j,\ell}), \ell = 1, 2 \). Each such estimate is unbiased, following the above intuition, as formalized below.

**Proposition 3.4.** We have \( \mathbb{E} \hat{\beta}_{j,\ell} = \beta_h(\theta_j) \) for each \( j \in [m] \) and \( \ell = 1, 2 \).

**Remark 3.2.** Note that \( P_X \) being fully supported is a necessary condition for importance reweighting to recover exactly the expectation under Gaussian measure by sampling from \( P_X \). If \( P_X \) has a bounded support, one can still approximate the expectation by choosing the variance parameter \( h \) small, see Section 3.6.3 for a detailed discussion.

Notice that, by the independence of the two estimates of \( \beta_h(\theta_j) \), we also immediately have that \( \hat{\beta}_{j,1}\hat{\beta}_{j,2}^T = \beta_h(\theta_j)\beta_h(\theta_j)^T \). This leads to the estimator of \( \tilde{M} \) given below:

\[
\tilde{M} := \frac{1}{2m} \sum_{j=1}^m \left[ \hat{\beta}_{j,1}\hat{\beta}_{j,2}^T + \hat{\beta}_{j,2}\hat{\beta}_{j,1}^T \right].
\]

The form of \( \tilde{M} \) ensures that it is symmetric. The full procedure is detailed in Algorithm 3.

**Remark 3.3 (Connection to Local Linear Regression).** Finally, we remark that the estimator \( \hat{\beta}_h(\theta) \) is tightly related to a local linear regression estimator as they can be shown to match asymptotically (see Section 3.6.2).

### 3.4.4 Upper-bound on Estimation Error

We start with some further assumptions, some of which imply the various conditions laid out so far in Proposition 3.1 and Corollary 3.2, as explained below. These various conditions are required for concentration arguments on \( \hat{\beta}_h \) and \( \tilde{M} \).
Assumption 3.3 (Tail Conditions for Noise). The noise has a subgaussian tail, i.e., \( \exists C, \sigma_Y^2 > 0 \) such that \( \forall x \in \mathbb{R}^d, \mathbb{P}\{Y - \mathbb{E}[Y \mid X = x] > t \mid X = x\} \leq C \cdot \exp(-t^2/(2\sigma_Y^2)) \).

Assumption 3.4 (Moment Condition on Density Ratio). The density ratio \( \rho_h(X; 0_d) \), as defined in Equation (3.4), satisfies:

\[
\mu_\rho := \left( \mathbb{E}_{X \sim P_X} [\rho_h(X; 0_d)^5] \right)^{1/5} < \infty.
\]

Remark 3.4. Assumption 3.4 holds when \( P_X \) has a heavier tail than \( N(0_d, h^2 I_d) \), since \( \rho_h(X; 0_d) \) is then bounded. Thus, intuitively, lighter tail \( P_X \) are admitted by smaller choices of \( h \). This is illustrated in the following proposition.

Proposition 3.5 (Example of \( \mu_\rho \) and choice of \( h \)). Suppose \( P_X = N(0_d, \sigma^2 I_d) \). Then,

\[
\mu_\rho = \begin{cases} 
\left( \frac{5h^8}{\sigma^8} - \frac{4h^{10}}{\sigma^{10}} \right)^{-d/10}, & \text{when } 0 < h < \sqrt{5}\sigma/2 \\
\infty, & \text{when } h \geq \sqrt{5}\sigma/2
\end{cases}
\]

Next, we require some additional moment conditions on the link function \( f \).

Assumption 3.5 (Moment Condition on Link Function). The link function \( f \) satisfies:

\[
\mu_f := \left( \mathbb{E}_{X \sim P_X} [f(U^T X)^6] \right)^{1/6} < \infty.
\]

Remark 3.5. We note that the condition of Proposition 3.1 (exhaustiveness of \( \bar{M} \)), i.e., \( f \) is square integrable with respect to \( N(0_k, h^2 I_k) \), holds under Assumption 3.4 and 3.5.

The following result establishes an upper-bound of order \( O(n^{-1/2}) \) on the estimation error for the CMS \( \mathcal{U} \), and holds for general link functions \( f \) satisfying the minor moment conditions discussed above. In particular, this main result allows nonparametric, i.e., highly complex link and regression functions \( f \) and \( g \), and yet admits fast convergence to \( \mathcal{U} \).
Theorem 3.6 (Main Result). Let $\mathcal{D}$ be a random sample of size $n$ from $P_{X,Y}$, $h, \delta > 0$, $0 < \sigma_\theta < h/\sqrt{20}$ and $16\mu_f^4\lambda_k(\mathcal{M})^{-2}\delta^{-1}(\log (8d/\delta))^2 \leq m \leq n/(2d)$. Let $\hat{U}$ be the output from Algorithm 3 with inputs $(h, \sigma_\theta, m, \mathcal{D})$. Under Assumption 3.1-3.5, there exists some absolute constant $C_1 > 0$, such that with probability at least $1 - \delta$,

$$d(\hat{U}, U) \leq C_1 \left[ C_{h,\sigma_\theta,1} \cdot d(\mu_f + \sigma_Y)^2 \cdot \frac{\sqrt{m}}{n} + C_{h,\sigma_\theta,2} \cdot d^{1/2}(\mu_f + \sigma_Y) \cdot \frac{1}{\sqrt{n}} \right] \cdot \frac{\log(8d/\delta)}{\sqrt{\delta}},$$

(3.6)

where $C_{h,\sigma_\theta,1}$ and $C_{h,\sigma_\theta,2}$ are given by:

$$C_{h,\sigma_\theta,1} := (1 - 20\sigma_\theta^2/h^2)^{-d/2} \cdot \mu_f^{5/3} \cdot (\lambda_k(\mathcal{M}))^{-1}, \text{ and}$$

$$C_{h,\sigma_\theta,2} := (1 - 10\sigma_\theta^2/h^2)^{-d/2} \cdot \mu_\nabla \cdot \mu_\rho^{5/6} \cdot (\lambda_k(\mathcal{M}))^{-1}. \quad (3.7)$$

Remark 3.6 (Constants in Rate). Since the term $\sqrt{m}/n \leq 1/\sqrt{n}$, Theorem 3.6 establishes that $U$ can be estimated at a parametric rate $C_d \cdot n^{-1/2}$, for a constant

$$C_d \leq C \cdot \max\{C_{h,\sigma_\theta,1} \cdot d, C_{h,\sigma_\theta,2} \cdot d^{1/2}\}.$$ 

This prefactor $C_d$ depends indirectly on the condition number $\lambda_1/\lambda_k$ of $\mathcal{M}$ since the largest eigenvalue $\lambda_1(\mathcal{M})$ could be bounded by some of the other moment parameters appearing in $C_d$; for instance it is easy to see that $\lambda_1(\mathcal{M}) \leq \mathbb{E}_{\theta \sim N(0, \sigma_\theta^2 I_d)} \left\| \nabla h f(\theta) \right\|^2 \leq \mu_\nabla^2$. For a more explicit dependence, see Remark 3.14 relating to the proof of Lemma 3.16.

Clearly, the various moment parameters, i.e., $\mu_f, \mu_\nabla, \mu_\rho$ appearing in $C_d$ depend tightly on $P_{X,Y}$, and most importantly on the behavior of the link function $f$. In Section 3.4.5 below we consider explicit polynomial conditions on $f$ along with the example of Gaussian $P_X$, and illustrate $C_d$ via explicit bounds on moment conditions.

In the above result, while the lower-bound on $m$ was required for the exhaustiveness of $\tilde{M}$
(Corollary 3.2), the upper-bound \( m \leq n/(2d) \) ensures that the datasets used to estimate smoothed gradients \( \nabla_h f(\theta_j) \) are not empty.

**Remark 3.7** (Estimating \( \hat{M} \)). As it appears, estimation of \( \hat{M} \) may require larger values of \( m \) than required in the above main theorem for estimation of \( U \). For instance consider upper-bounding

\[
\| \hat{M} - M \|_{op} \leq \| \hat{M} - \tilde{M} \|_{op} + \| \tilde{M} - M \|_{op}.
\]

The second term \( \| \tilde{M} - M \|_{op} \) would necessarily introduce an \( O(1/\sqrt{m}) \) term which is of order \( O(n^{-1/2}) \) only for large \( m = \Omega(n) \).

**Remark 3.8** (Dependence on \( \delta \)). The dependence on the failure probability \( \delta \) is of order \( \tilde{O}(\delta^{-1/2}) \) which may appear large in light of usual high-probability results. This is due to the fact that we make minimal distributional assumptions, which in particular result in lower order moment conditions on the random matrices \( \hat{\beta}_{j,1}, \hat{\beta}_{j,2} \), preventing exponential concentration guarantees. Nonetheless, heavy tail approaches such as median-of-means (see e.g. [48]) may be employed to improve such dependence on \( \delta \).

### 3.4.5 Instantiation for Polynomial Links

In this section, we instantiate our main Theorem 3.6 under the assumptions of a polynomial link \( f \) and standard Gaussian \( P_X \). Such assumptions have been considered extensively in the recent literature on CMS estimation (see e.g. [37, 38, 14]) and are formalized below.

**Assumption 3.6.** We consider the following conditions on \( P_{X,Y} \):

1. The marginal density \( P_X \) is \( \mathcal{N}(0_d, I_d) \);

2. The link function \( f \) is a polynomial of degree at most \( r \). Furthermore, we assume w.l.o.g that \( f \) is normalized, i.e., \( \mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)}[f(Z)^2] = 1 \);

3. Assumption 3.3 holds, namely the noise is subgaussian with parameter \( \sigma_Y^2 > 0 \).
Assumption 3.6 above supersedes all previous Assumption 3.1-3.5 as will be evident in the proof of Corollary 3.7 (see Section 3.7.5). In particular, as per Remark 3.6, we are interested in understanding the constant \(C_d\) in the rate of Theorem 3.6, by explicitly bounding all the various moment parameters \(\mu_f, \mu_p, \mu_{\nabla}\) involved.

We adopt the following notion of minimum signal strength from [37], which will serve to characterize \(\lambda_k(M)\).

**Definition 3.5 (Minimum Signal Strength).** Let \(\tau > 0\) be defined as:

\[
\tau := \min_{\eta \in \mathbb{R}^k} \mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} \left[ (\eta^T \nabla f(Z))^2 \right].
\]

**Remark 3.9.** The fact that \(\tau \neq 0\) is guaranteed by the minimality of \(U\), since otherwise, for some \(\eta \in \mathbb{R}^k, U\eta \in U\) would be an irrelevant direction for the regression function \(g\).

The following corollary establishes an upper-bound on subspace estimation error, with an explicit dependence on \(d, h\) and \(\sigma_\theta\).

**Corollary 3.7.** Suppose that Assumption 3.6 holds. Let \(\hat{U}\) be the output from Algorithm 3 with inputs \(0 < h < \sqrt{5}/2, 0 < \sigma_\theta < h/\sqrt{20}\) and \(m = n/(2d)\). Then, there exists \(C_2 > 0\), independent of \(n, d, \delta, h\) and \(\sigma_\theta\), such that with probability at least \(1 - \delta\),

\[
d(\hat{U}, U) \leq C_2 \cdot A_{\sigma_\theta^2/h^2, d, r} \cdot B_{h, d, r} \cdot \frac{d}{\tau \sqrt{n}} \cdot \frac{\log(8d/\delta)}{\sqrt{\delta}}.
\]

where \(A_{\sigma_\theta^2/h^2, d, r}, B_{h, d, r}\) are given as:

\[
A_{\sigma_\theta^2/h^2, d, r} := (\sigma_\theta^2/h^2)^{(r-1)} \cdot (1 - 20\sigma_\theta^2/h^2)^{-d/2};
\]

\[
B_{h, d, r} := \max\{h^{-2r+1}, h^{r-1}\} \cdot (5h^8 - 4h^{10})^{-d/6}.
\]

**Remark 3.10 (Choice of \(h, \sigma_\theta\)).** In order to minimize the prefactor \(C_2 \cdot A_{\sigma_\theta^2/h^2, d, r} \cdot B_{h, d, r}\), we can minimize \(A_{\sigma_\theta^2/h^2, d, r}\) (with respect to \(\sigma_\theta^2/h^2\)) and \(B_{h, d, r}\) (with respect to \(h\)) separately. These terms...
are minimized at $\sigma_{\theta}^2/h^2 = (r - 1)/(20(r - 1) + 10d)$ and $h = 1$. With such optimal choices, we have $A_{\sigma_{\theta}^2/h^2,d,r} \leq (20 + 30d/(r - 1))^{-1}$, and $B_{h,d,r} = 1$. Hence $d(\hat{U}, U) \leq C_d \cdot n^{-1/2}$ where $C_d \propto d^r$, ignoring the logarithmic term and $\delta$.

**Remark 3.11 (Bound Optimality).** The dependence on $d$ in the bound of Corollary 3.7 is hard to improve since our estimator falls in the class of so-called Correlational Statistical Query (CSQ) learners, i.e., learners that rely solely on statistics of form $\sum_i Y_i \cdot \Phi(X_i)$ for some function $\Phi$ (see [49, 50, 51] for an exact definition). Namely, such CSQ learners have been shown in the recent work of [14] to require a sample complexity $n = \Omega(d^{r/2})$ to achieve nontrivial error in subspace estimation. In comparison, our method requires $n = O(d^r)$, which corresponds to a constant factor mismatch in the exponent of $d$. It remains unclear how such a gap may be closed for CSQ learners as the result of [14] also leaves this open.

We note however that, outside of CSQ procedures, much more benign sample size requirements are possible, for instance $n = O(d)$ was shown for a recent method of [38].

### 3.5 Simulations

In this section, we demonstrate how to choose the hyperparameters $h$ (the “radius” of smoothing) and $m$ (the number of $\theta_j$ sampled in ASGOP) properly through a simple example with polynomial link function and standard Gaussian design. The related theoretical discussions can be found in Remark 3.1, 3.4, 3.7 and 3.10.

**Experiment Settings** Throughout the experiments, we consider a multi-index model Equation 3.1 with $d = 10, k = 3$, and $U = [e_1, \ldots, e_k] \in \mathbb{R}^{d \times k}$, where $e_i$ denotes the $i$-th column of the $d \times d$ identity matrix. That is, only the first three coordinates of $X$ are relevant directions for $Y$. The link function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ is a polynomial: $f(z) = z_1^2 + z_2 z_3$.

**Choices of $h$ under Different $P_X$** Recall from Remark 3.4 that our method work for any $P_X$ with heavier tails than Gaussian. The choice of $h$ is critical when $P_X$ is a Gaussian distribution.
(see Proposition 3.5). We verify the statement by applying Algorithm 3 with different choices of $h > 0$ and $\sigma_\theta = h/\sqrt{20 + 10d}$ (see Remark 3.10 for the reasoning of such choice) to datasets generated from $P_X$ being either standard Gaussian or standard Cauchy. Under appropriate choices of $h$, our method works reasonably well for both cases.

**Choice of the Number of Partitions $m$ under Gaussian design** The number of partition $m$ affects the subspace estimation error rates in two ways. Firstly, when $m$ is not large enough, there is no guarantee that $\tilde{M}$ is exhaustive (Corollary 3.2). With a positive probability, one can only recover a proper subspace of the CMS $\mathcal{U}$, hence the subspace estimation error should be large. When $m$ gets larger, we expect the estimation error to grow in the order of $\sqrt{m}$ for a fixed $n$ according to the upper-bound in Theorem 3.6. We evaluate the performance of Algorithm 3 under different choices of $m$, with a fixed sampling budget $n = 100,000$ under a standard Gaussian design. The related results are presented in Figure 3.2.

![Figure 3.1](image)

**Figure 3.1:** The subspace estimation error $d(\tilde{U}, U)$ v.s. sampling budget $n$. Here, we fixed the number of partitions $m = 15$ and $\sigma_\theta = h/\sqrt{20 + 10d}$. We replicate 10 times for each pair of $(n, h)$ and plot the mean (the dots) and the standard error (the error bars) of the estimation error. When $P_X =$ standard Gaussian (left), the performance of Algorithm 3 is quite sensitive to the choice of $h$. The optimal choice of $h$ is around the data variance 1. When $h$ gets smaller (e.g., $h = 0.5$) or larger (e.g., $h = 1.2, 1.5$), we observe larger errors under the same budget level. This actually coincides approximately with the minimizer of $\mu_\rho$ in terms of $h$ (c.f. Proposition 3.5). When $P_X =$ standard Cauchy, the method is more robust to the choice of $h$. 

61
Figure 3.2: The subspace estimation error $d(\hat{U}, U)$ v.s. the choice of $m$. We replicate 10 times for each $m$, and plot the mean (the dots) and the standard error (the error bars) of the subspace estimation errors. When $m$ is small, the ASGOP $\tilde{M}$ is not guaranteed to be exhaustive, and only a proper subspace of $\mathcal{U}$ can be recovered. This results in a large subspace estimation error. When $m$ is larger than a certain threshold (c.f., Corollary 3.2), the ASGOP $\tilde{M}$ is exhaustive with high probability, and the subspace error has an upper-bound that grows at the rate of $O(\sqrt{m})$. The result in the figure matches the reasoning above, as the subspace estimation error drops sharply at the regime where $m$ is small, and grows gradually for large $m$.

3.6 Discussions

3.6.1 SAVE Is Not Exhaustive

In this section, we provide an example where the Sliced Average Variance Estimator (SAVE) proposed by [30] cannot extract all the relevant directions, and hence cannot recover the whole central mean subspace (CMS), let alone the central subspace (CS). In the setting of the multi-index Model 3.1, SAVE extracts the index space by estimating the SAVE matrix:

$$V = \mathbb{E} \left[ \left( \mathbb{E} \left[ XX^T \mid Y \right] - I \right)^2 \right].$$

One critical condition needed for exhaustiveness here is that $\text{col}(V)$ should contain all the relevant directions in the index space. This condition is not always satisfied.
Consider a simple noiseless single-index model:

$$Y = f(u^\top X), \quad (3.8)$$

where $X = (X_1, X_2)^\top \sim \mathcal{N}(0, I_2)$, and $u = (1, 0)^\top$. Next, we show that there is a class of link functions with which $V$ is a zero matrix. Let $\xi(z_1, z_2) := \mathbb{E}_{Z \sim \mathcal{N}(0, 1)}[Z^2 | Z \in [z_1, z_2]]$ for $0 \leq z_1 \leq 1 \leq z_2$. Note that $\xi$ is continuous, and monotonically increasing in each of $z_1$ and $z_2$, $\xi(1, 1) = \mathbb{E}[Z^2 | Z = 1] = 1$ and $\xi(0, \infty) = \mathbb{E}[Z^2 | Z \geq 0] = 1$. Therefore, there exists a continuous decreasing function $\nu : [1, \infty) \to (0, 1]$ such that $\nu(1) = 1, \lim_{z \to \infty} \nu(z) = 0$ and

$$\xi(\nu(z), z) = 1 \text{ for any } z \in [1, \infty).$$

Let $f_0 : [0, 1] \to \mathbb{R}_+$ be any continuous, non-negative, and monotone increasing function defined on $[0, 1]$ with $f_0 = 0$. Then, we define $f : \mathbb{R} \to \mathbb{R}_+$ in terms of $f_0$ as follows:

$$f(z) := \begin{cases} 0, & z < 0; \\ f_0(z), & 0 \leq z \leq 1; \\ f_0(\nu(z)), & 1 < z < \infty. \end{cases} \quad (3.9)$$

The following proposition shows the SAVE matrix $V$ is a zero matrix with link function $f$.

**Proposition 3.8.** Suppose $(X, Y)$ is an example generated from Model (3.8) with link function as defined in (3.9). Then, we have $V = 0$.

**Proof.** It suffices to show that for any $y \geq 0$, $\mathbb{E}[XX^\top \mathbb{1}(Y \geq y)] = I_2 \mathbb{P}(Y \geq y)$. For $y > f_0(1)$ or $y \leq 0$, this is vacuously true. For $0 < y \leq f_0(1)$, we can show from the definition of $f$ that:

$$\{Y \geq y\} = \{X_1 : y \leq f(X_1) \leq f_0(1)\} = \{f_0^{-1}(y) \leq X_1 \leq \nu^{-1}(f_0^{-1}(y))\}. \quad (3.10)$$
Figure 3.3: An example plot of the link function $f$ as defined in (3.9). Here, we have $\{Y > y\} = [z_1, z_2]$ from the plot $\{x : y \leq f(x)\}$, where $z_1 = f_0^{-1}(y)$ and $z_2 = \nu^{-1}(z_1) = \nu^{-1}(f_0^{-1}(y))$.

See also Figure 3.3 for a graphical illustration. Therefore,

$$
E \left[ XX^T \bigg| Y \geq y \right] = E \left[ \begin{bmatrix} X_2^2 & X_1X_2 \\ X_1X_2 & X_2^2 \end{bmatrix} \bigg| Y \geq y \right]
= \begin{bmatrix}
E \left[ X_1^2 \bigg| Y \geq y \right] & E \left[ X_1 \bigg| Y \geq y \right] E[X_2] \\
E \left[ X_1 \bigg| Y \geq y \right] E[X_2] & E[X_2^2]
\end{bmatrix}
= \begin{bmatrix}
\xi(f_0^{-1}(y), \nu^{-1}(f_0^{-1}(y))) & 0 \\
0 & 1
\end{bmatrix} = I,
$$

Finally, we have:

$$
E \left[ XX^T 1(Y \geq y) \right] = E \left[ XX^T \bigg| Y \geq y \right] P(Y \geq y) = I P(Y \geq y).
$$

The proof is completed.

3.6.2 Connection to Local Linear Regression

In this section, we will show that our estimator $\hat{\beta}_h(\theta) := \hat{\beta}_h(\theta; D)$, which is defined similarly as in Equation 3.5 but with a full sample $D = \{(X_i, Y_i)\}_{i=1}^n$, is asymptotically equivalent to the following local linear estimator (LLE) of $\nabla g(\theta)$ with a data-dependent kernel:

$$
\hat{\beta}_{h, \text{LLE}}(\theta) := \arg\min_{(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^d} \sum_{i=1}^n (Y_i - \alpha - \beta^T (X_i - \theta))^2 K_h(\theta^T X_i),
$$

64
where $K^\theta_h(x) := \rho_h(x; \theta)$.

**Proposition 3.9.** For each $\theta \in \mathbb{R}^d$,

$$\left\| \hat{\beta}_h(\theta) - \beta_{h}^{\text{LLE}}(\theta) \right\| \overset{p}{\to} 0,$$

where, the notation $\overset{p}{\to}$ represents convergence in probability.

**Proof.** By standard solution for the weighted least square error problems gives:

$$\hat{\beta}_{h}^{\text{LLE}}(\theta) = I_{d+1}^{-1}(X_\theta W_\theta X_\theta)^{-1}X_\theta^T W_\theta Y \tag{3.11}$$

where $I_{d+1}^{\ominus} \in \mathbb{R}^{d \times (d+1)}$ is obtained by removing the first row of a $(d + 1) \times (d + 1)$ identity matrix $I_{d+1}$.

$$X_\theta = \begin{bmatrix} 1 & (X_1 - \theta)^\top \\ \vdots & \vdots \\ 1 & (X_n - \theta)^\top \end{bmatrix}, \quad W_\theta = \text{diag}(K^\theta_h(X_1), \ldots, K^\theta_h(X_n)), \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

One can show from Equation 3.11 and the block matrix inversion formula (Lemma 3.23) that

$$\hat{\beta}_h(\theta) = h^{-2} \left( \hat{\Sigma}(\theta) \hat{\beta}_{h}^{\text{LLE}}(\theta) + \hat{\mu}_y(\theta) \hat{\mu}_x(\theta) \right),$$

where

$$\hat{\Sigma}(\theta) := n^{-1} \sum_{i=1}^{n} \rho_h(X_i; \theta)(X_i - \theta)(X_i - \theta)^\top;$$

$$\hat{\mu}_x(\theta) := n^{-1} \sum_{i=1}^{n} \rho_h(X_i; \theta)(X_i - \theta);$$

$$\hat{\mu}_y(\theta) := n^{-1} \sum_{i=1}^{n} \rho_h(X_i; \theta)Y_i.$$
By the weak law of large numbers, \( \hat{\mu}_x \overset{p}{\to} 0 \), \( \hat{\mu}_y \overset{p}{\to} \mathbb{E}_{X \sim N(0_d, I_d)}[f(U^T \tilde{X})] \). The proposition follows and the proof is completed.

### 3.6.3 Estimation Bias Under \( P_X \) with a Bounded Support

In this section, we discuss the performance of our estimator when Assumption 3.1(2) is not satisfied and the Lebesgue density \( p \) of \( P_X \) has a bounded support \( \{ x \in \mathbb{R}^d : \| x \| \leq K \} \) for some \( K < \infty \). In this case, we note that the smoothed gradient estimator \( \hat{\beta}_h(\theta_j; D_{j,l}) \) as defined in Equation 3.5 is biased for \( \beta_h(\theta_j) \). In particular, for a fixed \( \theta_j \),

\[
\mathbb{E}[\hat{\beta}_h(\theta_j; D_{j,l})] = h^{-2} \mathbb{E}[\rho_h(X; \theta) \cdot Y \cdot (X - \theta_j)] = h^{-2} \int_{\|X\| \leq K} \rho_h(X; \theta_j) \cdot g(X) \cdot (X - \theta_j) dP_X
\]

\[= h^{-2} \mathbb{E}_{Z \sim N(\theta_j, h^2 I_d)}[\mathbb{I}(\|Z\| \leq K) \cdot g(Z) \cdot (Z - \theta_j)].\]

Hence, the bias is given as:

\[
B_{h,j} := \mathbb{E}[\hat{\beta}_h(\theta_j; D_{j,l})] - \beta_h(\theta_j)
\]

\[= -h^{-2} \mathbb{E}_{Z \sim N(\theta_j, h^2 I_d)}[\mathbb{I}(\|Z\| > K) \cdot g(Z) \cdot (Z - \theta_j)]. \tag{3.12}\]

Intuitively, \( \|B_{h,j}\| \) should shrink to zero as \( h \to 0 \) when \( K \gg \|\theta_j\| \), which holds with high probability if we choose the variance \( \sigma^2_\theta \) of \( \theta_j \)'s is small compared to \( K \). On the other hand, it is clear that if \( \sup_{j \in [m]} B_{h,j} = o(n^{-1/2}) \), then the bias will not affect the rate in the upper-bound for subspace estimation error. The following proposition shows how small \( h \) needs to be to guarantee such small biases.

**Proposition 3.10.** Suppose that \( \|\theta_j\| \leq h \). Let \( h = K(\sqrt{d} + \sqrt{\log n})^{-1}/4 \), then we have the bias \( \|\mathbb{E}[\hat{\beta}_{j,\ell} - \beta_h(\theta_j)]\| = O(\log n \cdot n^{-2/3}) \).
Proof. By Cauchy-Schwarz Inequality and Equation 3.12, we have:

$$\| B_{h,j} \| \leq h^{-2} \left( \mathbb{E}_{\tilde{Z} \sim N(h^{-2}\theta, I_d)} [1(\| \tilde{Z} \| > Kh^{-1})] \cdot \mathbb{E}_{Z \sim N(\theta, h^2 I_d)} [g(Z)^3] \right) \cdot \mathbb{E}_{Z \sim N(\theta, h^2 I_d)} [\| Z - \theta \|^3]^{1/3}.$$ 

We will upper-bound each of the terms in the parentheses. By Assumption 3.4 and 3.5, it can be verified that there exists some $C > 0$, independent of $h$ and $n$, such that:

$$\mathbb{E}_{Z \sim N(\theta, h^2 I_d)} [g(Z)^3] \leq C \cdot \rho \cdot \mu.$$ \hspace{2cm} (3.13)

The condition $\| \theta \| \leq h$ implies that:

$$\mathbb{E}_{Z \sim N(\theta, h^2 I_d)} [\| Z - \theta \|^3] \leq C \cdot h^3.$$ \hspace{2cm} (3.14)

Finally, we have under the assumption $\| \theta \| < h < K/2$ that:

$$\mathbb{E}_{\tilde{Z} \sim N(h^{-2}\theta, I_d)} [1(\| \tilde{Z} \| > Kh^{-1})] \leq \mathbb{P} \left( \| N(0_d, I_d) \| > \frac{K h^{-1}}{2} \right) = \mathbb{P} \left( \chi^2_d > \frac{K^2 h^{-2}}{4} \right).$$ \hspace{2cm} (3.15)

Consider the following concentration inequality for the chi-squared distribution:

**Lemma 3.11** (Concentration for $\chi^2_d$, adapted from Lemma 1 of [52]). Let $W$ be a random variable follows the centered chi-squared distribution with degree of freedom $d$, i.e., $W \sim \chi^2_d$, then for any $t > 0$:

$$\mathbb{P}(W > d + 2\sqrt{dt} + 2t) \leq \exp(-t).$$

By Lemma 3.11, when $h = K(\sqrt{d} + \sqrt{\log n})^{-1}/4$, we have:

$$\mathbb{P} \left( \chi^2_d > \frac{K^2 h^{-2}}{4} \right) \leq n^{-2}.$$ \hspace{2cm} (3.16)
From Equation 3.13 to Equation 3.16, we have:

\[ \|B_{h,j}\| \leq Ch^{-2}(n^{-2} \cdot \mu_\rho \cdot \mu_f \cdot h^3)^{1/3} \leq \tilde{C} \cdot \log n \cdot n^{-2/3}. \]

where \( \tilde{C} \) is independent of \( n \).

**Remark 3.12.** The condition \( \|\theta_j\| \leq h \) in Proposition 3.10 holds uniformly for \( j \in [m] \), with probability at least \( 1 - \delta \) with the choice \( \sigma_\theta \lesssim h \cdot \log(m/\delta) \). This is compatible with our choice of \( \sigma_\theta \) in the main theorem.

### 3.6.4 Connection to Transfer Learning

The estimation of index space \( \text{col}(U) \) allows efficient transfer learning. Consider the following domain adaption setting:

- **Source domain:** the joint distribution is \( P_{X,Y}^{(S)} \) with \( \mathbb{E}_S[Y|X] = f_S(U^T X) \),
- **Target domain:** the joint distribution is \( P_{X,Y}^{(T)} \) with \( \mathbb{E}_T[Y|X] = f_T(U^T X) \);

where the lower-dimensional subspace \( \text{col}(U) \) is shared by both domain, while the link functions differs: \( f_S \neq f_T \). Let \( D^{(S)} := \{ X^{(S)}_i, Y^{(S)}_i \}_{i \in [n_S]} \) and \( D^{(T)} := \{ X^{(T)}_j, Y^{(T)}_j \}_{j \in [n_T]} \) be random samples in the two domains respectively. We propose to get an estimate of \( f_T \) using the following two-step approach: 1) get an estimate \( \hat{U} \) using Algorithm 3 with \( D^{(S)} \); 2) get an estimate \( \hat{E}_T[Y|X = x] \) in target domain through the Nadaraya-Watson regression estimator:

\[
\hat{E}_T[Y|X = x] := \hat{f}_T(U^T x) = \frac{\sum_{j=1}^{n_T} K_h(U^T (X^{(T)}_j - x)) \cdot Y^{(T)}_j}{\sum_{j=1}^{n_T} K_h(U^T (X^{(T)}_j - x))} \cdot \mathbb{I} \left( \sum_{j=1}^{n_T} K_h(U^T (X^{(T)}_j - x)) > 0 \right)
\]

with a box kernel \( K_h(\cdot) := \mathbb{I}(\|\cdot\| \leq h) \) with a proper choice of bandwidth \( h > 0 \).

**Proposition 3.12.** Suppose that the joint distribution \( P_{X,Y}^{(S)} \) satisfies Assumption 3.1-3.5, \( P_{X}^{(T)} \) has
a bounded support \( X \in \mathbb{R}^d \) and furthermore, \( f_T \) is \( L \)-Lipschitz, i.e.,

\[
|f_T(z) - f_T(z')| \leq L|z - z'|, \quad \forall z, z' \in \mathbb{R}^k,
\]

for some \( L > 0 \). Then, with probability at least \( 1 - \delta \), there exists some \( C > 0 \) independent of \( d, \delta, n_S, n_T \) and \( h \), such that

\[
\int_{\mathbb{R}^d} |\hat{f}_T(\hat{U}^T x) - f_T(U^T x)|^2 dP^{(T)}_X \leq C \cdot \left( L^2 \cdot \left( h^2 + \frac{C_d^2}{n_S} \right) + \frac{1}{n_T h^k} \right) \cdot \tilde{O} \left( 1/\sqrt{\delta} \right).
\]

Recall that \( C_d \) is the prefactor in Theorem 3.6.

**Remark 3.13.** Proposition 3.12 elucidates the gain from transferring the estimated subspace from another domain. In particular, when \( n_S \gtrsim C_d^2 \cdot (L^2 n_T)^2 / (2 + k) \), one can achieve the \( L_2^2 \)-squared prediction risk of order \( L_2^2 n_S / (2 + h) \cdot n^{-2/2 + k} \) in the target domain as if the index space is known, by choosing the bandwidth \( h \sim (L^2 n_T)^{-1/2 (2 + k)} \).

**Proof of Proposition 3.12.** By Theorem 3.6, we have that on a favorable event \( A \) with \( \mathbb{P}(A) \geq 1 - \delta \),

\[
d(\hat{U}, U) \leq C_d \cdot n_S^{-1/2} \cdot \tilde{O} \left( 1/\sqrt{\delta} \right) =: \varepsilon.
\]

We next establish the variance bound and bias bound on this favorable event \( A \).

**Variance bound** Let \( n_h(x) := \sum_{j=1}^{n_T} K_h(\hat{U}^T (X_j^{(T)} - x)) \), then

\[
v(x) := \text{Var} \left[ \hat{f}_T(\hat{U}^T x) \mid \{X_j^{(T)}\}_{j \in [n_T]}, \hat{U} \right] \leq \frac{\sigma_Y^2}{n_h(x)} \cdot 1(n_h(x) > 0)
\]

where \( \sigma_Y^2 := \sup_x \text{Var}[Y \mid X = x] \). We note that

\[
n_h(x) \sim \text{Binomial}(n_T, p_h(x)), \quad p_h(x) = \mathbb{P} \left( \|\hat{U}^T (X_1^{(T)} - x)\| \leq h \right),
\]

69
and by Lemma 4.1 of [25], we have
\[ \mathbb{E} \left[ \frac{1_{(n_h(x) > 0)}}{n_h(x)} \right] \leq \frac{2}{(n_T + 1)p_h(x)}. \]

Therefore, we have
\[ \mathbb{E}[v(x) | \hat{U}] \leq \frac{2\sigma_Y^2}{(n_T + 1)p_h(x)}. \quad (3.18) \]

**Bias bound** When \( n_h(x) > 0 \), we have
\[
\begin{align*}
b(x) &:= \mathbb{E} \left[ \hat{f}_T(\hat{U}^T x) | \{ X_j^{(T)} \}_{j \in [n_T], \hat{U}} - f_T(U^T x) \right] \\
&= \frac{1}{n_h(x)} \sum_{j : \hat{U}^T (X_j^{(T)} - x) \leq h} \left( f_T(U^T X_j^{(T)}) - f_T(U^T x) \right) \\
&\leq L(h + \varepsilon D_X).
\end{align*}
\]

Note that we use the Lipschitzness of \( f_T \) and the boundedness of \( \mathcal{X} \) (Here, \( D_X \) is the diameter of \( \mathcal{X} \)).

When \( n_h(x) = 0 \), the conditional bias
\[
b(x) = 0 - f_T(U^T x) = -f_T(U^T x).
\]

Combining the two cases above, we have
\[
\begin{align*}
\mathbb{E}[b(x)^2 | \hat{U}] &\leq L^2(h + \varepsilon D_X)^2 + f_T(U^T x)^2 \mathbb{P}(n_h(x) = 0) \\
&= L^2(h + \varepsilon D_X)^2 + f_T(U^T x)^2 \cdot (1 - p_h(x))^{n_T} \\
&\leq L^2(h + \varepsilon D_X)^2 + f_T(U^T x)^2 \cdot \frac{1}{n_T p_h(x)} \quad (3.19)
\end{align*}
\]
By combining (3.18), (3.19) and integrating over \( x \), we have

\[
\int_X |\hat{f}_T(\hat{U}^T x) - f_T(U^T x)|^2 dP_X(x) = \int_{\mathbb{R}^d} \mathbb{E}\{v(x) \mid \hat{U}U^T x \} + \mathbb{E}\{|h(x)|^2 \mid \hat{U}U^T x \} dP_X(x)
\]

\[
\leq L^2(\hat{n} + \varepsilon D_x)^2 + \frac{2\sigma_Y^2 + f_{\text{max}}^2}{n_T} \int_X \frac{1}{p_h(x)} dP_X(x)
\]

We can conclude the proof by bounding the second term.

Let \( Z \) be a \( h/2 \)-cover of \( \{\hat{U}^T x : x \in \mathcal{X}\} \subset \mathbb{R}^k \), it is clear that \( |Z| \simeq (2D_X/h)^k \). Let \( \{A_z\}_{z \in Z} \) be the Voronoi diagram of \( \{\hat{U}^T x : x \in \mathcal{X}\} \), where \( A_z := \{\hat{U}^T x : \min_{z' \in Z} \|\hat{U}^T x - z'\| = \|\hat{U}^T x - z\|\} \). Then,

\[
\int_X \frac{1}{p_h(x)} dP_X(x) \leq \sum_{z \in Z} \int_{A_z} \frac{1}{p_h(x)} dP_X(x)
\]

\[
= \sum_{z \in Z} \int_{A_z} \frac{1}{P_X(\{x' : U^T x' \leq h\})} dP_X(x)
\]

\[
\leq \sum_{z \in Z} \int_{A_z} \frac{1}{P_X(\{x' : U^T x' - z \leq h/2\})} dP_X(x)
\]

\[
= \sum_{z \in Z} \frac{P_X(A_z)}{P_X(\{x' : U^T x' - z \leq h/2\})}
\]

\[
\leq |Z| \leq c(D_X/h)^k.
\]

for some absolute constant \( c > 0 \). Consequently, we have:

\[
\int_X |\hat{f}_T(\hat{U}^T x) - f_T(U^T x)|^2 dP_X(x) \leq L^2(\hat{n} + \varepsilon D_x)^2 + c \left( \frac{f_{\text{max}}^2 + \sigma_Y^2}{n_T h^k} \right) D_X^k
\]

\[
\leq L^2(\hat{n} + \varepsilon D_x)^2 + c \left( \frac{f_{\text{max}}^2 + \sigma_Y^2}{n_T h^k} \right) D_X^k
\]

\[
\leq L^2(\hat{n} + C_d \cdot D_X^k \cdot n_s^{-1/2} \cdot \tilde{O}(1/\sqrt{\delta}))^2 + c \left( \frac{f_{\text{max}}^2 + \sigma_Y^2}{n_T h^k} \right) D_X^k.
\]

This concludes the proof.
3.7 Proofs of the Main Results

In this section, we provide the detailed proofs of results in Section 3.4.

3.7.1 Proof of Proposition 3.1 (Exhaustiveness of $M$)

Proof. We first consider the case $h = 1$. Using the Hermite expansion $f(z) = \sum_{\alpha \in \mathbb{N}^k} w_\alpha H_\alpha(z)$, we have

$$
E_{Z \sim \mathcal{N}(0,I_d)} \left[ \partial_h f(U^T(Z + \theta)) \right] = \sum_{\alpha \in \mathbb{N}^k, \alpha_i \geq 1} w_\alpha \left( \prod_{j \neq i} \frac{1}{\sqrt{\alpha_j}} (U_j^T \theta)^{\alpha_j} \right) \frac{\sqrt{\alpha_i}}{\sqrt{(\alpha_i - 1)!}} (U_i^T \theta)^{\alpha_i - 1} 
$$

$$
= \sum_{\alpha \in \mathbb{N}^k, \alpha_i = \alpha_i - 1} \sqrt{\alpha_i + 1} \cdot w_{\alpha(i)} \prod_{j \in [k]} (U_j^T \theta)^{\alpha_j} / \sqrt{\alpha_j!},
$$

$$
=: \sum_{\alpha \in \mathbb{N}^k} \lambda_{\alpha,i} S_\alpha(U^T \theta),
$$

(3.20)

where $\alpha(i) := (\alpha_1, \ldots, \alpha_i - 1, \alpha_i + 1, \alpha_{i+1}, \ldots, \alpha_k) \in \mathbb{R}^k$, $\lambda_{\alpha,i} := \sqrt{\alpha_i + 1} \cdot w_{\alpha(i)} \in \mathbb{R}$ and $S_\alpha(z) = \prod_{j \in [k]} (z_j^{\alpha_j} / \sqrt{\alpha_j!})$ is a scaled monomial basis indexed by $\alpha$.

According to the definition, $M = U \Lambda U^T$, where $\Lambda := E_{\theta}[\nabla_h f(\theta) \nabla_h f(\theta)^T] \in \mathbb{R}^{k \times k}$. To show $\text{col}(M) = \text{col}(U)$, we only need to show that the matrix $\Lambda$ is full-rank. Suppose not, then there exists a unit vector $\eta \in \mathbb{R}^k$ such that $\eta^T \Lambda \eta = 0$, which indicates $\eta^T \nabla_h f(\theta) = 0$, for almost every $\theta$ with respect to Gaussian measure. Therefore,

$$
\forall \alpha \in \mathbb{N}^k, \sum_{i \in [k]} \eta_i \lambda_{\alpha,i} = 0.
$$

(3.21)

Next, we show that Equation (3.21) implies that $\eta^T \nabla f(U^T x) = 0$ for almost every $x \in \mathbb{R}^d$. For
each \( i \in [k] \),

\[
\partial_i f(U^T x) = \sum_{\alpha \in \mathbb{N}^k} w_\alpha \cdot \partial_i H_\alpha(U^T x) \\
= \sum_{\alpha \in \mathbb{N}^k : \alpha_i \geq 1} w_\alpha \cdot (\sqrt{\alpha_i} H_{\alpha,-1}(U^T x)) \prod_{j \neq i} H_{\alpha_j}(U^T x) \\
= \sum_{\alpha \in \mathbb{N}^k : \alpha_i \geq 1} \lambda_{\alpha,i} \cdot H_\alpha(U^T x).
\]

Therefore, we have for almost every \( x \in \mathbb{R}^d \) with respect to the Gaussian measures, and hence \( P_X \)-a.s. (since \( P_X \) is dominated by Gaussian),

\[
\eta^T \nabla f(U^T x) = \sum_{\alpha \in \mathbb{N}^k} \left( \sum_{i \in [k]} \eta_i \lambda_{\alpha,i} \right) H_\alpha(U^T x) = 0.
\]

Then, it follows that

\[
(U \eta)^T \nabla g(x) = \eta^T U^T U \nabla f(U^T x) = \eta^T \nabla f(U^T x) = 0
\]

In other words, \( U \eta \in \mathcal{U} \) is a direction along which the directional derivative of \( g \) is zero almost surely. In other words, \( \mathcal{U} \) admits an irrelevant direction for the regression function \( g \), which contradicts the assumption that \( \mathcal{U} \) is the CMS and hence cannot be further reduced.

For the case \( h \neq 1 \), one can let \( f_h(z) := f(hz) \), then \( E_{Z \sim \mathcal{N}(0, I_d)}[f_h(Z)^2] = 1 \), and

\[
\nabla_h f(\theta) := E_{Z \sim \mathcal{N}(0, h^2 I_d)}[\nabla f(U^T(Z + \theta))] = h^{-1} E_{Z \sim \mathcal{N}(0, I_d)}[\nabla f_h(U^T(Z + \theta))] = h^{-1} \nabla_1 f_h(\theta).
\]

Clearly, \( E_\theta[\nabla_1 \tilde{f}(\theta) \nabla_1 \tilde{f}(\theta)^T] \) is full-rank by the reasoning for the case \( h = 1 \), and it follows that \( \overline{M} = h^{-2} U E_\theta[\nabla_1 \tilde{f}(\theta) \nabla_1 \tilde{f}(\theta)^T]U^T \) is exhaustive.
3.7.2 Proof of Corollary 3.2 (Exhaustiveness of $\tilde{M}$)

The following lemma establishes the concentration of the ASGOP $\tilde{M}$ around ESGOP $M$, and is the key ingredient for proving the exhaustiveness of $\tilde{M}$.

**Lemma 3.13.** Under the setting of Theorem 3.6, with probability $1 - \delta$,

$$\|\tilde{M} - M\|_{op} \leq \left(\sqrt{2\mu^2} \cdot \frac{1}{\sqrt{m}}\right) \cdot \frac{1}{\sqrt{\delta}} \cdot \log\left(\frac{4d}{\delta}\right),$$

for all $0 < \delta < 1$.

**Proof of Lemma 3.13.** Let $R > 0$, $\tilde{Z}_j := \beta(\theta_j)\beta(\theta_j)^\top - M$, and $\tilde{Z}_j' := \tilde{Z}_j \mathbb{1}(\|\tilde{Z}_j\|_{op} \leq R)$. Note that

$$\tilde{M} - M = \frac{1}{m} \sum_{j=1}^{m} \tilde{Z}_j,$$

and

$$\mathbb{E} \left[\|\tilde{Z}_j\|_{op}^2\right] = \mathbb{E} \left[\|\beta(\theta_j)\beta(\theta_j)^\top - M\|_{op}^2\right] = \mathbb{E} \left[\|U[\nabla_h f(\theta)\nabla_h f(\theta)^\top - \mathbb{E}[\nabla_h f(\theta)\nabla_h f(\theta)]U^\top\|_{op}^2\right]$$

$$\leq \mathbb{E} \left[\|\nabla_h f(\theta)\nabla_h f(\theta)^\top - \mathbb{E}[\nabla_h f(\theta)\nabla_h f(\theta)]\|_{op}^2\right] \leq 2\mu^4.$$

Next, we establish upper-bounds for three critical quantities to invoke Lemma 3.22:

$$\mathbb{P}\left(\|\tilde{Z}_j\|_{op} > R\right) \leq \frac{1}{R^2} \mathbb{E} \left[\|\tilde{Z}_j\|_{op}^2\right] = \frac{2\mu^4}{R^2};$$

$$\left\|\mathbb{E} \left[\tilde{Z}_j - \tilde{Z}_j'\right]\right\|_{op} \leq \mathbb{E} \left[\|\tilde{Z}_j\|_{op} \mathbb{1}(\|\tilde{Z}_j\|_{op} > R)\right] = \int_{R}^{\infty} \mathbb{P}\left(\|\tilde{Z}_j\|_{op} > t\right) dt$$

$$\leq \int_{R}^{\infty} \frac{1}{t^2} \cdot \mathbb{E} \left[\|\tilde{Z}_j\|_{op}^2\right] dt \leq \frac{2\mu^4}{R} =: \Delta;$$

$$\left\|\sum_{j=1}^{m} \mathbb{E}[\tilde{Z}_j' - \mathbb{E}\tilde{Z}_j']^2\right\|_{op} \leq m \mathbb{E} \left[\|\tilde{Z}_j'^2\|_{op}\right] \leq m \mathbb{E} \left[\|\tilde{Z}_j^2\|_{op}\right] \leq 2m\mu^4 =: \sigma^2.$$
Let \( \tilde{\delta} \in (0, 1/2) \), set the threshold levels \( R := \sigma/\sqrt{\delta} \). Then,

\[
\Delta = R\tilde{\delta}/m < R\log(2d/\tilde{\delta})/m,
\]

(3.23)

\[
m \mathbb{P}\left( \| \tilde{Z}_j \|_{op} \geq R \right) \leq \tilde{\delta}, \text{ and}
\]

(3.24)

\[
\sigma^2/R^2 = \tilde{\delta} < \log(2d/\tilde{\delta})/3,
\]

(3.25)

Then, we apply Lemma 3.22 with and \( t := R \log(2d/\tilde{\delta})/m + \Delta \):

\[
\mathbb{P}\left\{ \| \tilde{M} - M \|_{op} \geq \frac{2R\log(2d/\tilde{\delta})}{m} \right\} \leq \mathbb{P}\left\{ \| \tilde{M} - M \|_{op} \geq \frac{R\log(2d/\tilde{\delta})}{m} + \Delta \right\}
\]

\[
= \mathbb{P}\left\{ \| \tilde{M} - M \|_{op} \geq \frac{R\log(2d/\tilde{\delta})}{m} + \Delta \right\}
\]

\[
\leq 2d \exp\left( -\frac{\log^2(2d/\tilde{\delta})}{\sigma^2/R^2 + 2\log(2d/\tilde{\delta})/3} \right) + \tilde{\delta}
\]

\[
\leq 2\tilde{\delta}.
\]

Finally, by setting \( \delta = 2\tilde{\delta} \in (0, 1) \) and substituting \( R = \sigma/\sqrt{\delta} \), we have with probability at least \( 1 - \delta \),

\[
\| \tilde{M} - M \|_{op} \leq \left( \sqrt{2} \mu_k^2 \cdot \frac{1}{\sqrt{m}} \right) \cdot \frac{1}{\sqrt{\delta}} \cdot \log \left( \frac{4d}{\delta} \right).
\]

This concludes the proof.

Next, we proceed to the proof of the Corollary 3.2.

**Proof of Corollary 3.2 (Exhaustiveness of \( \tilde{M} \)).** By Lemma 3.13, we have with probability at least \( 1 - \delta \),

\[
\| \tilde{M} - M \|_{op} \leq \lambda_k(M) / 2,
\]

when \( m \geq 8\mu_k^4 \lambda_k(M)^{-2}\delta^{-1}(\log (4d/\delta))^2 \). Consequently, \( \lambda_k(\tilde{M}) \geq \lambda_k(M)/2 > 0 \) by Weyl’s inequality. Therefore, \( \text{col}(\tilde{M}) \) is of rank \( k \), and hence equals to \( \mathcal{U} \).
3.7.3 Proof of Proposition 3.5

**Proof.** By definition,
\[
\mu_\rho^5 := \mathbb{E}_{x \sim \mathcal{N}(0_d, \sigma^2 I_d)} \left[ \left( \frac{1}{(2\pi)^{d/2}} \exp \left( \frac{-x^T x}{2\sigma^2} \right) \right)^5 \right]
\]
\[
= \mathbb{E}_{x \sim \mathcal{N}(0_d, I_d)} \left[ \frac{(\sigma/h)^5 \exp \left( \frac{5h^2 - 5\sigma^2}{2h^2\sigma^2} x^T x \right)}{(2\pi)^{d/2}} \right]
\]
\[
= \frac{\sigma^{4d}}{h^{5d}} \cdot \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp \left( \frac{4h^2 - 5\sigma^2}{2h^2\sigma^2} \right) dx.
\]
(3.26)

When \( h \geq \sqrt{5\sigma}/2, (4h^2 - 5\sigma^2)/(2h^2\sigma^2) \geq 0 \) and thus \( \mu_\rho = \infty \). When \( 0 < h < \sqrt{5}/2 \):
\[
\mu_\rho = \left( \frac{\sigma^{4d}}{h^{5d}} \right)^{1/5} \cdot \left( \frac{5\sigma^2 - 4h^2}{h^2\sigma^2} \right)^{-d/10} = \left( \frac{5h^8 - 4h^{10}}{\sigma^8} \right)^{-d/10} < \infty.
\]

This completes the proof.

3.7.4 Proof of Theorem 3.6 (Main Theorem)

The following Lemma 3.14 establishes a probabilistic bound on \( \| \hat{M} - \tilde{M} \|_{op} \). It is a key step for the proof of the main theorem.

**Lemma 3.14.** Under the setting of Theorem 3.6, for all \( 0 < \delta < 1 \) and some absolute constant \( C > 0 \), with probability \( 1 - \delta \),
\[
\| \hat{M} - \tilde{M} \|_{op} \leq C \left( \tilde{C}_{h,\sigma,1} \cdot d(\mu_f + \sigma_Y)^2 \cdot \sqrt{m/n} + \tilde{C}_{h,\sigma,2} \cdot d^{1/2}(\mu_f + \sigma_Y) \cdot \frac{1}{\sqrt{n}} \right) \frac{\log(4d/\delta)}{\sqrt{\delta}}.
\]
(3.27)

where \( \tilde{C}_{h,\sigma,1}, \tilde{C}_{h,\sigma,2} \) are as defined as:
\[
\tilde{C}_{h,\sigma,1} := (1 - 20\sigma_\theta^2/h^2)^{-d/2} \cdot \mu_\rho^{5/3}; \quad \tilde{C}_{h,\sigma,2} := (1 - 10\sigma_\theta^2/h^2)^{-d/2} \cdot \mu^3 \cdot \mu_\rho^{5/6}.
\]

In what follows, we first show the proof of the main theorem using Lemma 3.14, and postpone
the proof of the till the end of the Section.

**Proof of Theorem 3.6 (Main Theorem).** Let $\tilde{\delta} = \delta / 2 \in (0, 1/2)$. We define two good events with desired properties for the concentration of the subspace estimator: Let $\xi_1$ denote the event where $U$ can be fully recovered from $\tilde{M}$, i.e., $\text{col}(\tilde{M}) = U$. Since $m \geq 8 \mu_4^2 \lambda_k(\tilde{M})^{-2} \tilde{\delta}^{-1} (\log(4d/\tilde{\delta}))^2$, we have the guarantee that $P(\xi_1) \geq 1 - \tilde{\delta}$ by Corollary 3.2. Let $\xi_2$ denote the good event in Lemma 3.14 with $\delta = \tilde{\delta}$. Next, we show that Equation (3.6) holds on $\xi_1 \cap \xi_2$, whose probability is at least $1 - \delta$. This follows from a standard application of the Davis-Kahan Theorem (Theorem V.3.6 from [53]). In particular, the eigengap can be calculated as:

$$
\lambda_k(\tilde{M}) - \lambda_{k+1}(\tilde{M}) = \lambda_k(\tilde{M}) - (\lambda_{k+1}(\tilde{M}) - \lambda_{k+1}(\tilde{M})) \\
\geq \lambda_k(\tilde{M}) - \|\tilde{M} - \tilde{M}\|_{\text{op}} \\
\geq \lambda_k(\tilde{M}) / 2,
$$

for large $n$ that ensures $\|\tilde{M} - \tilde{M}\|_{\text{op}} \leq \lambda_k(\tilde{M}) / 2$. Therefore,

$$
d(\hat{U}, U) \leq \frac{\sqrt{2}\|\tilde{M} - \tilde{M}\|_{\text{op}}}{\lambda_k(\tilde{M}) - \lambda_{k+1}(\tilde{M})} \leq \frac{2\sqrt{2}\|\tilde{M} - \tilde{M}\|_{\text{op}}}{\lambda_k(\tilde{M})} \leq \frac{4\sqrt{2}\|\tilde{M} - \tilde{M}\|_{\text{op}}}{\lambda_k(\tilde{M})}.
$$

By substituting Equation (3.27) into the last inequality, we conclude the proof.

We now prove Lemma 3.14 to conclude this section. In the remainder of this section, we use the simplified notation $\beta_j \equiv \beta_h(\theta_j)$, and define:

$$
\hat{M}_j := \frac{1}{2} \left( \hat{\beta}_{j,1} \hat{\beta}_{j,2}^\top + \hat{\beta}_{j,2} \hat{\beta}_{j,1}^\top \right), \quad \hat{Z}_j := \hat{M}_j - \tilde{M}_j,
$$

Then, $\hat{M} - \tilde{M} = m^{-1} \sum_{j=1}^m \hat{Z}_j$.

**Proof of Lemma 3.14 (Concentration of $\tilde{M}$).** Let $R > 0$ be the truncation level to be determined later, and $\hat{Z}_j' := \hat{Z}_j \mathbb{1}(\|\hat{Z}_j\|_{\text{op}} \leq R)$. By standard but cumbersome calculations (of which the
details we present after this proof), we can show that there exists some absolute constant \( c_1 > 0 \), such that for any \( R > 0 \),

(i) (Lemma 3.17)
\[
P \left\{ \| \hat{Z}_j \|_{op} \geq R \right\} \leq c_1 \left( \tilde{C}_{h,\sigma,1}^2 \cdot \frac{d^2(\mu_f + \sigma_Y)^4}{R^2} \cdot \frac{m^2}{n^2} + \tilde{C}_{h,\sigma,2}^2 \cdot \frac{d(\mu_f + \sigma_Y)^2}{R} \cdot \frac{m}{n} \right);
\]

(ii) (Lemma 3.18)
\[
\left\| \mathbb{E} \left[ \hat{Z}_j - \tilde{Z}_j' \right] \right\|_{op} \leq \mathbb{E} \left[ \left\| \hat{Z}_j \right\|_{op} 1(\| \hat{Z}_j \|_{op} > R) \right] \\
\leq c_1 \left( \tilde{C}_{h,\sigma,1}^2 \cdot \frac{d^2(\mu_f + \sigma_Y)^4}{R} \cdot \frac{m^2}{n^2} + \tilde{C}_{h,\sigma,2}^2 \cdot \frac{d(\mu_f + \sigma_Y)^2}{R} \cdot \frac{m}{n} \right) \\
=: \Delta ;
\]

(iii)
\[
\left\| \sum_{j=1}^{m} \mathbb{E} [(\hat{Z}_j' - \mathbb{E} \hat{Z}_j')^2] \right\|_{op} \leq m \mathbb{E} \left[ \left\| \hat{Z}_j^2 \right\|_{op} \right] \leq m \mathbb{E} \left[ \left\| \hat{Z}_j \right\|_{op}^2 \right] \\
\overset{\text{Lemma 3.16}}{\leq} c_1 \left( \tilde{C}_{h,\sigma,1}^2 \cdot \frac{d^2(\mu_f + \sigma_Y)^4}{n^2} \cdot \frac{m^3}{n^2} + \tilde{C}_{h,\sigma,2}^2 \cdot \frac{d(\mu_f + \sigma_Y)^2}{n} \cdot \frac{m^2}{n} \right) \\
=: \sigma^2 .
\]

Let \( \tilde{\delta} \in (0, 1/2) \), set the threshold levels \( R := \sigma / \sqrt{\tilde{\delta}} \). Then,

\[
\Delta = R\tilde{\delta}/m < R \log(2d/\tilde{\delta})/m , \quad \text{(3.28)}
\]

\[
m \mathbb{P} \left\{ \| \hat{Z}_j \|_{op} \geq R \right\} \leq \tilde{\delta} , \quad \text{(3.29)}
\]

and

\[
\sigma^2 / R^2 = \tilde{\delta} < \log(2d/\tilde{\delta})/3 , \quad \text{(3.30)}
\]
Then, we apply Lemma 3.22 with \( t := R \log(2d/\tilde{\delta})/m + \Delta \):

\[
\mathbb{P}\left\{ \| \hat{\mathbf{M}} - \mathbf{M} \|_{op} \geq \frac{2R \log(2d/\tilde{\delta})}{m} \right\} \stackrel{(3.28)}{\leq} \mathbb{P}\left\{ \| \hat{\mathbf{M}} - \mathbf{M} \|_{op} \geq \frac{R \log(2d/\tilde{\delta})}{m} + \Delta \right\} \leq 2d \exp\left( -\frac{\log^2(2d/\tilde{\delta})}{\sigma^2/R^2 + 2 \log(2d/\tilde{\delta})/3} \right) + \tilde{\delta} \leq 2\tilde{\delta}.
\]

Finally, by setting \( \delta = 2\tilde{\delta} \in (0,1) \) and substituting \( R = \sigma/\sqrt{\delta} \), we have with probability at least \( 1 - \delta \),

\[
\| \hat{\mathbf{M}} - \tilde{\mathbf{M}} \|_{op} \leq C \left( \tilde{C}_{h,\sigma_{\theta,1}} \cdot d(\mu_f + \sigma_Y)^2 \cdot \frac{\sqrt{m}}{n} + \tilde{C}_{h,\sigma_{\theta,2}} \cdot d^{1/2}(\mu_f + \sigma_Y) \cdot \frac{1}{\sqrt{n}} \right) \frac{\log(4d/\delta)}{\sqrt{\delta}}.
\]

This concludes the proof.

The following Lemmas 3.15-3.18 provides the detailed calculations to justify (i), (ii), and (iii) in the above proof of Lemma 3.14.

**Lemma 3.15.** Under Assumption 3 and 5, we have

\[
\mathbb{E}[Y^6] \leq \left( \mu_f + \sqrt{6}e^{1/e} \sigma_Y \right)^6.
\]

**Proof of Lemma 3.15.** Let \( \epsilon = Y - \mathbb{E}[Y \mid \mathbf{X}] = Y - f(\mathbf{U}^\top \mathbf{X}) \), then

\[
\mathbb{E}[Y^6] = \mathbb{E}\left[ (f(\mathbf{U}^\top \mathbf{X}) + \epsilon)^6 \right] \leq \left( \mathbb{E}[f^6(\mathbf{U}^\top \mathbf{X})] \right)^{1/6} + \left( \mathbb{E}[\epsilon^6] \right)^{1/6} \leq \left( \mathbb{E}[f^6(\mathbf{U}^\top \mathbf{X})] \right)^{1/6} + \sqrt{6}e^{1/e} \sigma_Y = \left( \mu_f + \sqrt{6}e^{1/e} \sigma_Y \right)^6.
\]
Lemma 3.16. Under Assumption 1-5, there exists some absolute constant $c_1 > 0$ such that,

$$
\mathbb{E}_{P_{X,Y}} \left[ \left\| \hat{Z}_j \right\|_{op}^2 \right] \leq c_1 \left( \tilde{C}_{h,\sigma_0,1}^2 \cdot d^2(\mu_f + \sigma_Y)^4 \cdot \frac{m^2}{n^2} + \tilde{C}_{h,\sigma_0,2}^2 \cdot d(\mu_f + \sigma_Y)^2 \cdot \frac{m}{n} \right).
$$

Proof of Lemma 3.16. For simplicity, we write $\beta_j = \beta_h(\theta_j)$. Note that for $\ell = 1, 2,$

$$
\mathbb{E} \left[ \left\| \hat{\beta}_{j,\ell} - \beta_j \right\|^2 \mid \theta_j \right] \leq \frac{2m\sigma^2}{n} \mathbb{E} \left[ \left\| \rho_h(X;\theta)(X - \theta)Y \right\|^2 \mid \theta = \theta_j \right]
$$

By Lemma 3.15, $\mathbb{E}[Y^6] \leq (\mu_f + \sqrt{6}e^{1/\sigma})^6 \leq c_2(\mu_f + \sigma)^6$. It suffices to upper-bound the term $\mathbb{E} \left[ \left\| \rho_h(X;\theta)(X - \theta) \right\|^3 \mid \theta = \theta_j \right]$.

Let $\tilde{\rho}_h(x;\theta) := \varphi_h(x - \theta)/\varphi_h(x)$ be the density ratio between $\mathcal{N}(\theta, h^2 I_d)$ and $\mathcal{N}(0_d, h^2 I_d)$, then $\rho_h(X;\theta) = (\varphi_h(X)/p(X)) \cdot \tilde{\rho}_h(X;\theta)$. Hence,

$$
\mathbb{E}_{X \sim P_X} \left[ \left\| \rho_h(X;\theta)(X - \theta) \right\|^3 \mid \theta = \theta_j \right]
$$

$$
= \mathbb{E} \left[ \left( \frac{\varphi_h(X)}{p(X)} \cdot \tilde{\rho}_h(X;\theta) \right)^{5/2} \cdot (\rho_h(X;\theta))^{1/2} \cdot \left\| (X - \theta) \right\|^3 \mid \theta = \theta_j \right]
$$

$$
\leq \mu_\rho^{5/2} \left\{ \int \varphi_h(z) \exp \left( \frac{5}{h^2} z^T \theta_j + \frac{5}{2h^2} \left\| \theta_j \right\|^2 \right) \left\| z \right\|^6 dz \right\}^{1/2}
$$

$$
\leq \mu_\rho^{5/2} \left\{ \exp \left( 15 \left\| \theta_j \right\|^2 \right) \mathbb{E}_{Z \sim \mathcal{N}(0,h^2 I_d)} \left[ \left\| Z + 5\theta_j \right\| \right] \right\}^{1/2}
$$

$$
\leq c_2 \mu_\rho^{5/2} \exp \left( \frac{15}{2h^2} \left\| \theta_j \right\|^2 \right) \left( h^3 d^{3/2} + \left\| \theta_j \right\|^3 \right)
$$

$$
\leq c_3 \mu_\rho^{5/2} h^3 d^{3/2} \exp \left( \frac{15}{2h^2} \left\| \theta_j \right\|^2 \right),
$$

for some absolute constant $c_2, c_3 > 0$, and thus

$$
\mathbb{E} \left[ \left\| \hat{\beta}_{j,\ell} - \beta_j \right\|^2 \mid \theta = \theta_j \right] \leq \frac{c_4 \mu_\rho^{5/3} d(\mu_f + \sigma_Y)^2 m}{n} \cdot \exp \left( \frac{5}{h^2} \left\| \theta_j \right\|^2 \right).
$$
for some $c_4 > 0$ and

$$
\mathbb{E} \left[ \| \hat{Z}_j \|_{\text{op}}^2 \right] = \mathbb{E} \left[ \| \hat{\beta}_{j,1} \hat{\beta}_{j,2}^\top - \beta_j \beta_j^\top \|_{\text{op}}^2 \right] \\
\leq 2 \mathbb{E} \left[ \mathbb{E} \left[ \| \hat{\beta}_{j,1} - \beta_j \| | \theta_j \right] \mathbb{E} \left[ \| \hat{\beta}_{j,2} - \beta_j \| | \theta_j \right] + 2 \| \beta_j \| \mathbb{E} \left[ \| \hat{\beta}_{j,1} - \beta_j \| | \theta_j \right] \right] \\
\leq c_1(\frac{\mu_p^{10/3} d^2(\mu_f + \sigma_Y)^4 m^2}{n^2} \cdot \left( 1 - \frac{20\sigma_\theta^2}{h^2} \right)^{-d/2} + \frac{\mu_p^2 \mu_p^{5/3} (\mu_f + \sigma_Y)^2 m}{n} \cdot \left( 1 - \frac{10\sigma_\theta^2}{h^2} \right)^{-d/2} ) \\
:= c_1 \left( \tilde{C}_{h,\sigma_\theta,1} \cdot d^2(\mu_f + \sigma_Y)^4 \cdot \frac{m^2}{n^2} + \tilde{C}_{h,\sigma_\theta,2} \cdot d(\mu_f + \sigma_Y)^2 \cdot \frac{m}{n} \right),
$$

for some $c_1 > 0$, where in the second last step we use the fact $\mathbb{E} \left[ \| \beta_j \| \right] \leq \sqrt{\mathbb{E} \left[ \| \beta_j \|^4 \right]} \leq \mu_\beta^2$, and $\mathbb{E}[t\| \beta_j \|^2] = (1 - 2\sigma_\theta t)^{-d/2}, \forall t < 1/(2\sigma_\theta)$. This concludes the proof.

**Remark 3.14.** The upper-bound in Lemma 3.16 and Lemma 3.14 is related implicitly to the largest eigenvalue of the ESGOP $\lambda_1(M)$. In particular, notice that we have the following two upper-bounds for $\lambda_1(M)$:

- $\lambda_1(M) \leq \mathbb{E}_{\theta \sim N(0,d,\sigma^2 I_d)} \left[ \| \beta(\theta) \beta(\theta)^\top \| \right] \leq \mu_\beta$;
- $\lambda_1(M) \leq \mathbb{E}_{\theta \sim N(0,d,\sigma^2 I_d)} \left[ \| \beta(\theta) \beta(\theta)^\top \| \right] \leq c_4 \mu_p^{5/3} d(\mu_f + \sigma_Y)(1 - 5\sigma_\theta^2/h^2)^{-d/2}$.

**Lemma 3.17.** For any $j \in [m]$ and $R > 0$,

$$
\mathbb{P} \left\{ \| \hat{Z}_j \|_{\text{op}} \geq R \right\} \leq c_1 \left( \tilde{C}_{h,\sigma_\theta,1} \cdot \frac{d^2(\mu_f + \sigma_Y)^4}{R^2} \cdot \frac{m^2}{n^2} + \tilde{C}_{h,\sigma_\theta,2} \cdot \frac{d(\mu_f + \sigma_Y)^2}{R} \cdot \frac{m}{n} \right).
$$

**Proof of Lemma 3.17.** It follows from Lemma 3.16 and Markov inequality.

**Lemma 3.18.** For any $j \in [m]$ and $R > 0$,

$$
\mathbb{E} \left[ \| \hat{Z}_j \|_{\text{op}} \mathbb{1} \left( \| \hat{Z}_j \|_{\text{op}} > R \right) \right] \leq c_1 \left( \tilde{C}_{h,\sigma_\theta,1} \cdot \frac{d^2(\mu_f + \sigma_Y)^4}{R} \cdot \frac{m^2}{n^2} + \tilde{C}_{h,\sigma_\theta,2} \cdot \frac{d(\mu_f + \sigma_Y)^2}{R} \cdot \frac{m}{n} \right).
$$

**Proof of Lemma 3.18.** It follows from Lemma 3.17 and the fact that for any nonnegative random
variable $\xi \geq 0$:

$$
E[\xi 1(\xi > R)] = \int_{\mathbb{R}} P\{\xi > t\} \, dt.
$$

### 3.7.5 Proof of Corollary 3.7

To invoke the main Theorem 3.6 for the special case of polynomial links, we rely on the following lemma that quantifies the various moment parameters and $\lambda_k(M)$ in terms of $d, r, \tau, h$ and $\sigma_\theta$.

**Lemma 3.19.** Under the conditions for Corollary 3.7, we have:

(i) $\mu_\rho = (5h^8 - 4h^{10})^{-d/10}$;

(ii) $\mu_f \leq 5^{r/2}$;

(iii) $\mu_\nabla \leq (r + 1)(6k + r + 2)/4 \cdot \max\{h^{r-1}, h^{-1}\}$;

(iv) $\lambda_k(M) \geq r^{-r-4k} \cdot 2^{-3(r-1)-k} \cdot ((r-1)!)^{-1} \cdot (\sigma_\theta^2/h^2)^{r-1} \cdot \min\{h^{2(r-1)}, 1\} \cdot \tau$.

Next, we proceed to prove Corollary 3.7, which merely substitutes the bounds from Lemma 3.19 into Theorem 3.6.

**Proof of Corollary 3.7.** By Equation (3.7) and Lemma 3.19, we have:

$$
C_{h,\sigma_\theta,1} \leq C_{k,r} \cdot \max\{h^{-2r+2}, 1\}(5h^8 - 4h^{10})^{-d/6} (\sigma_\theta^2/h^2)^{-(r-1)} (1 - 20\sigma_\theta^2/h^2)^{-d/2} \cdot \tau^{-1};
$$

$$
C_{h,\sigma_\theta,2} \leq C_{k,r} \cdot \max\{h^{-2r+1}, h^{r-1}\}(5h^8 - 4h^{10})^{-d/12} (\sigma_\theta^2/h^2)^{-(r-1)} (1 - 10\sigma_\theta^2/h^2)^{-d/2} \cdot \tau^{-1},
$$

where $C_{k,r}$ is a constant depends only on $k$ and $r$. Thus, we have $\max\{d \cdot C_{h,\sigma_\theta,1}, d^{1/2} \cdot C_{h,\sigma_\theta,2}\} \leq C_{k,r} \cdot A_{\sigma_\theta^2/h^2,d,r} \cdot B_{h,d,r} \cdot d$. Finally, we invoke Theorem 3.6 and get:

$$
d(\hat{U}, U) \leq C_2 \cdot A_{\sigma_\theta^2/h^2,d,r} \cdot B_{h,d,r} \cdot \frac{d}{\tau \sqrt{n}} \cdot \frac{1}{\tau} \cdot \frac{\log(8d/\delta)}{\sqrt{\delta}},
$$
where $C_2 := C_1 \cdot C_{k,r} \cdot \max\{\mu_f + \sigma_Y, (\mu_f + \sigma_Y)^2\}$.

We conclude this section by proving Lemma 3.19. In what follows, we use the short notation $\mathcal{I}_{r,k} := \{\alpha \in \mathbb{N}^k \text{ such that } \|\alpha\|_1 \leq r\}$. The following lemma gives an upper-bound for $|\mathcal{I}_{r,k}|$, or equivalently the dimension of the space spanned by polynomials of degree at most $r$.

**Lemma 3.20.** For any integers $r \geq 0$ and $k \geq 1$,

$$|\mathcal{I}_{r,k}| \leq (r + 1)^k.$$  

**Proof of Lemma 3.20.** The upper-bound can be achieved by:

$$|\mathcal{I}_{r,k}| = \|\{\alpha \in \mathbb{N}^k : \|\alpha\|_1 \leq r\}\| = \sum_{l=0}^{r} \|\{\alpha \in \mathbb{N}^k : \|\alpha\|_1 = l\}\|$$

$$\leq \sum_{l=0}^{r} (l + 1)^{(k - 1)} \leq (r + 1)^k.$$

The smoothed gradient $\nabla_h f(\theta)$ is defined as an expectation under Gaussian with variance $h^2 I$, which deviates from the normalization assumption $E[f(Z)^2] = 1$ in which the expectation is w.r.t. standard Gaussian. The next lemma shows at most how large these expectations can differ from each other for the polynomial class we are considering.

**Lemma 3.21.** Let $h > 0$, we have under Assumption 3.6 that

$$(r + 1)^{-2k-(r/2)} \cdot \min\{h^{2r}, 1\} \leq \mathbb{E}_{Z \sim \mathcal{N}(0, h^2 I_d)}[f(Z)^2] \leq (r + 1)^{2k+(r/2)} \cdot \max\{h^{2r}, 1\}.$$  

**Proof of Lemma 3.21.** We first show the upper-bound of $\mathbb{E}_{Z \sim \mathcal{N}(0, h^2 I_d)}[f(Z)^2]$. Since $f$ admits a
Hermite expansion \( f(z) = \sum_{\alpha \in I_{r,k}} w_\alpha H_\alpha(z) \), with \( \sum_{\alpha \in I_{r,k}} w_\alpha^2 = 1 \). Note that

\[
E_{Z \sim \mathcal{N}(0_k, h^2 I_k)} [f(Z)^2] = E_{Z \sim \mathcal{N}(0_k, I_k)} [f(hZ)^2] \]

\[
= E_{Z \sim \mathcal{N}(0_k, I_k)} \left[ \left( \sum_{\alpha \in I_{r,k}} w_\alpha H_\alpha(hZ) \right)^2 \right] \]

\[
\leq \left| I_{r,k} \right| \cdot \max_{\alpha \in I_{r,k}} E_{Z \sim \mathcal{N}(0_k, I_k)} [H_\alpha(hZ)^2] \quad (3.31) \]

Since \( |I_{r,k}| \leq (r + 1)^k \) by Lemma 3.20, it remains to find an upper-bound (uniformly in \( \alpha \)) for

\[
E_{Z \sim \mathcal{N}(0_k, I_k)} [H_\alpha(hZ)^2] :\]

\[
= E \left[ \prod_{j=1}^{k} H_{\alpha_j}(hZ_j)^2 \right] \]

\[
\overset{\text{Fact 3.1(iii)}}{=} \left( \sum_{i=0}^{\lfloor \alpha_j/2 \rfloor} \frac{\alpha_j!}{(i!)^2(\alpha_j - 2i)!} \right) \cdot h^{2\alpha_j} \prod_{j=1}^{k} \left( \frac{1 - h^{-2}}{2} \right)^{2i} \cdot \frac{\alpha_j!}{(i!)^2(\alpha_j - 2i)!} \leq 2\alpha_j^{\alpha_j/2} \leq 2r^{r/2}. \]

By Stirling’s approximation that for integer a: \( a! \sim (a/e)^a \cdot (2\pi a)^{1/2} \), one can show that for any \( i \in [\alpha_j/2 + 1] \):

\[
\frac{\alpha_j!}{(i!)^2(\alpha_j - 2i)!} \leq 2\alpha_j^{\alpha_j/2} \leq 2r^{r/2}. \]

Therefore, we have:

\[
E_{Z \sim \mathcal{N}(0_k, I_k)} [H_\alpha(hZ)^2] \leq r^{r/2} \cdot \prod_{j=1}^{k} h^{2\alpha_j} \cdot \left( \sum_{i=0}^{\lfloor \alpha_j/2 \rfloor} \left( \frac{1 - h^{-2}}{2} \right)^{2i} \right). \]
When \( h \geq \sqrt{3}/3 \), we have \((1 - h^{-2})/2 \in [-1, 1/2] \), thus
\[
\sum_{i=0}^{\lfloor \alpha_j/2 \rfloor} \left( \frac{1 - h^{-2}}{2} \right)^{2i} \leq \frac{1}{1 - ((1 - h^{-2})/2)^2} \leq 4/3,
\]
and hence:
\[
\mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} \left[ H_\alpha(hZ)^2 \right] \leq h^{2\|\alpha\|} \cdot r^{r/2} \cdot (4/3)^k.
\]

When \( h < \sqrt{3}/3 \), using the inequality \((a^{q+2} - 1)/(a^2 - 1) \leq (q + 2)a^q/2\) for any \( a > 1 \) and \( q \in \mathbb{N} \), we have:
\[
\sum_{i=0}^{\lfloor \alpha_j/2 \rfloor} \left( \frac{1 - h^{-2}}{2} \right)^{2i} \leq \frac{(h^2 - 1)/2^{\alpha_j+2} - 1}{(h^2 - 1)/2} \leq \frac{\alpha_j + 2}{2} \cdot \left( \frac{h^2 - 1}{2} \right)^{\alpha_j},
\]
and hence:
\[
\mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} \left[ H_\alpha(hZ)^2 \right] \leq r^{r/2} \cdot \left( \frac{r + 2}{2} \right)^k \cdot \left( \frac{1 - h^2}{2} \right)^{\|\alpha\|} \leq r^{r/2} \cdot \left( \frac{r + 2}{2} \right)^k.
\]

By combining the two cases, we have:
\[
\mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} \left[ H_\alpha(hZ)^2 \right] \leq r^{r/2} \cdot \left( \frac{r + 2}{2} \right)^k \cdot \max\{h^{2r}, 1\} \leq (r + 1)^{k+(r/2)} \cdot \max\{h^{2r}, 1\}.
\]

and the upper-bound for \( \mathbb{E}_{Z \sim \mathcal{N}(0_k, h^2 I_k)} [f(Z)^2] \) follows.

To get the lower-bound, we let \( f_h(z) = f(hz) \), and hence
\[
\mathbb{E}_{Z \sim \mathcal{N}(0_k, h^{-2} I_k)} [f_h(Z)] = \mathbb{E}_{Z \sim \mathcal{N}(0_k, h^{-2} I_k)} [f(Z)] = 1.
\]

On the other hand, we have by applying the upper-bound with \( h \) replaced by \( h^{-1} \) that:
\[
1 = \mathbb{E}_{Z \sim \mathcal{N}(0_k, h^{-2} I_k)} [f_h(Z)^2] \leq (r + 1)^{2k+(r/2)} \cdot \max\{h^{-2r}, 1\} \cdot \mathbb{E}_{Z \sim \mathcal{N}(0_k, h^2 I_k)} [(f_h(Z))^2]
\]
\[
= (r + 1)^{2k+(r/2)} \cdot \max\{h^{-2r}, 1\} \cdot \mathbb{E}_{Z \sim \mathcal{N}(0_k, h^2 I_k)} [(f(Z))^2].
\]
The lower-bound then follows.

Now, we are ready to prove Lemma 3.19.

**Proof of Lemma 3.19 (i).** This is essentially Proposition 3.11 with $\sigma = 1$.

**Proof of Lemma 3.19 (ii).**

$$
\mu_f = \left( \mathbb{E}_{Z \sim N(0_k, I_k)} [f(Z)^6] \right)^{1/6} \leq 5^{r/2} \cdot \left( \mathbb{E}_{Z \sim N(0_k, I_k)} [f(Z)^2] \right)^{1/2} = 5^{r/2}.
$$

**Proof of Lemma 3.19 (iii).** Let $f_h(z) = f(hz)$, then:

$$
\nabla_h f(\theta) = \mathbb{E}_{Z \sim N(0_k, I_k)} [\nabla f(hZ + U^T \theta)] = h^{-1} \mathbb{E}_{Z \sim N(0_k, I_k)} [\nabla f_h(Z + U^T \theta/h)]. \tag{3.32}
$$

For simplicity, define $S_j := U_j^T \theta/h$ and note that $S_j \overset{\text{iid}}{\sim} N(0, \sigma_0^2/h^2)$ for $j \in [k]$. Furthermore, we use $w_{h, \alpha} := \langle f_h, H_\alpha \rangle_{N(0_k, I_k)}$ to denote the Hermite coefficient of $f_h$. By Lemma 3.21, we have

$$
\sum_{\alpha \in I_{r,k}} w_{h, \alpha}^2 \leq (r + 1)^{2k+(r/2)} \cdot \max\{h^{2r}, 1\}.
$$

We then proceed to calculate $\|\nabla_h f(\theta)\|^4$ by Equation (3.32) and (vii) from Fact 3.1:

$$
\left\| \mathbb{E}_{Z \sim N(0_k, I_k)} \nabla f_h(Z + U^T \theta/h) \right\|^2
\leq \sum_{i=1}^k \left( \sum_{\alpha \in I_{r,k} : \alpha_i \geq 1} w_{h, \alpha}^2 \prod_{j \neq i} \frac{1}{\alpha_j!} \frac{S_{j}^{\alpha_j} \sqrt{\alpha_i}}{\sqrt{(\alpha_i - 1)!}} S_{j}^{\alpha_i - 1} \right)^2
\leq r \cdot \left( \sum_{\alpha \in I_{r,k} : \alpha_i \geq 1} w_{h, \alpha}^2 \right) \left( \sum_{\alpha \in I_{r-1,k} : \alpha_j \in [k]} \prod_{j \neq i} \frac{S_{j}^{\alpha_j} \alpha_i!}{\alpha_j!} \right)
\leq r \cdot (r + 1)^{2k+(r/2)} \cdot \max\{h^{2r}, 1\} \cdot \left( \sum_{\alpha \in I_{r-1,k} : \alpha_j \in [k]} \prod_{j \neq i} \frac{S_{j}^{\alpha_j} \alpha_i!}{\alpha_j!} \right).
$$

86
By taking square on both sides,

\[
\left\| \mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} \nabla f_h(Z + U^\top \theta / h) \right\|^4 
\leq (r + 1)^{4k + r + 2} \cdot \max\{h^{4r}, 1\} \cdot \left( \sum_{\alpha \in \mathcal{I}_{r-1,k}} \prod_{j \in [k]} \frac{S_{2 \alpha_j}^{\alpha_j}}{\alpha_j!} \right)^2 
\leq (r + 1)^{4k + r + 2} \cdot \max\{h^{4r}, 1\} \cdot \left( \sum_{\alpha \in \mathcal{I}_{r-1,k}} \prod_{j \in [k]} \frac{S_{4 \alpha_j}^{\alpha_j}}{(\alpha_j!)^2} \right) 
\leq (r + 1)^{5k + r + 2} \cdot \max\{h^{4r}, 1\} \cdot \left( \sum_{\alpha \in \mathcal{I}_{r-1,k}} \prod_{j \in [k]} \frac{S_{4 \alpha_j}^{\alpha_j}}{(\alpha_j!)^2} \right) 
\]

(3.33)

It remains upper-bound the expectation. By Fact 3.3, we have:

\[
\left( \frac{h}{\sigma_\theta} \right)^{4 \alpha_j} \mathbb{E} \left[ \frac{S_{\alpha_j}^{4 \alpha_j}}{(\alpha_j!)^2} \right] = \frac{(4 \alpha_j)!}{2^{2 \alpha_j}(2 \alpha_j)!(\alpha_j!)^2} \quad \text{Stirling Approximation} \quad \leq 2^{4 \alpha_j},
\]

and hence by our choice \(0 < \sigma_\theta < h/\sqrt{20}\):

\[
\mathbb{E} \left[ \frac{S_{\alpha_j}^{4 \alpha_j}}{(\alpha_j!)^2} \right] \leq \left( \frac{2}{\sqrt{20}} \right)^{4 \alpha_j} \leq 1.
\]

Finally, we have by Equation (3.32):

\[
\mathbb{E}_{\theta \sim \mathcal{N}(0_d, \sigma_\theta^2 I_d)} \left[ \left\| \nabla_h f(\theta) \right\|_2^4 \right] = h^{-4} \mathbb{E}_{\theta \sim \mathcal{N}(0_d, \sigma_\theta^2 I_d)} \left[ \left\| \mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} \nabla f_h(Z + U^\top \theta / h) \right\|^4 \right] 
\leq (r + 1)^{5k + r + 2} \cdot \max\{h^{4r-4}, h^{-4}\} \cdot \mathbb{E} \left[ \sum_{\alpha \in \mathcal{I}_{r-1,k}} \prod_{j \in [k]} \frac{S_{4 \alpha_j}^{\alpha_j}}{(\alpha_j!)^2} \right] 
\leq (r + 1)^{5k + r + 2} \cdot \max\{h^{4r-4}, h^{-4}\} \cdot |\mathcal{I}_{r-1,k}| 
\leq (r + 1)^{6k + r + 2} \cdot \max\{h^{4r-4}, h^{-4}\}.
\]
Proof of Lemma 3.19 (iv). Let $S = (S_1, \ldots, S_k)$ with $S_i = U_i^\top \theta / h$. Clearly, we have $S \sim \mathcal{N}(0_k, (\sigma_\theta / h)^2 I_k)$. By Equation (3.32), recalling that $f_h(z) := f(hz)$, we have

\[
\nabla_h f(\theta) = \mathbb{E}_{z \sim \mathcal{N}(0, h^2 I_d)} \left[ \nabla f(U^\top(Z + \theta)) \right] = h^{-1} \mathbb{E}_{\tilde{z} \sim \mathcal{N}(0, I_k)} \left[ \nabla f_h(\tilde{Z} + S) \right] = h^{-1} \left( \sum_{\alpha \in I_r, k : \alpha_i \geq 1} w_{h, \alpha} \prod_{j \neq i} \frac{1}{\sqrt{\alpha_j}!} \frac{\sqrt{\alpha_i}}{\sqrt{(\alpha_i - 1)!}} S_j^{\alpha_j} \right)_{i \in [k]} =: h^{-1} g(S).
\]

We aim to find a lower-bound for

\[
\lambda_k(M) = \min_{\eta \in \mathbb{R}^k: \|\eta\| = 1} \eta^\top \mathbb{E}_{\theta \sim \mathcal{N}(0, \sigma_\theta^2 I_k)} \left[ \nabla_h f(\theta) \nabla_h f(\theta)^\top \right] \eta
\]

\[
= h^{-2} \min_{\eta \in \mathbb{R}^k: \|\eta\| = 1} \mathbb{E}_{S \sim \mathcal{N}(0, (\sigma_\theta / h)^2 I_k)} \left[ (\eta^\top g(S))^2 \right]
\]

Here, we note that $\eta^\top g$ is a $k$-variate polynomial with a degree at most $p - 1$:

\[
\eta^\top g(s) = \sum_{i \in [k]} \eta_i \left( \sum_{\alpha \in I_{r-1,k} : \alpha_i \geq 1} w_{h, \alpha} \prod_{j \neq i} \frac{1}{\sqrt{\alpha_j}!} \frac{\sqrt{\alpha_i}}{\sqrt{(\alpha_i - 1)!}} s_j^{\alpha_j} \right) = \sum_{\alpha \in I_{r-1,k}} \left( \sum_{i \in [k]} \eta_i \cdot w_{h, \alpha(i)} \cdot \sqrt{\alpha_i + 1} \right) \cdot \prod_{j \in [k]} \frac{s_j^{\alpha_j}}{\sqrt{\alpha_j}!}.
\]  

(3.34)

where $\alpha(i) := (\alpha_1, \ldots, \alpha_{i-1}, \alpha_i + 1, \alpha_{i+1}, \ldots, \alpha_k)$ for $i \in [k]$. Thus, we can apply Lemma 3.21 on $\eta^\top g$:

\[
\mathbb{E}_{S \sim \mathcal{N}(0, (\sigma_\theta / h)^2 I_k)} \left[ (\eta^\top g(S))^2 \right] \geq r^{-2k-(r-1)/2} \cdot (\sigma_\theta / h)^{2r-2} \cdot \mathbb{E}_{S \sim \mathcal{N}(0, I_k)} \left[ (\eta^\top g(S))^2 \right].
\]  

(3.35)

Next, we will establish a lower-bound for $\mathbb{E}_{S \sim \mathcal{N}(0, I_k)} \left[ (\eta^\top g(S))^2 \right]$. Let $D_1 \in \mathbb{R}^{|[I_{r-1,k}]|$ and $D_2 \in \mathbb{R}^{|[I_{r-1,k}]|$ be the coefficient vectors of $\eta^\top g(s)$ with respect to the Hermite basis and monomial basis,
and $B_{r-1,k} \in \mathbb{R}^{[I_{r-1,k} \times [I_{r-1,k}]}$ be the invertible linear map that converts the monomial basis to the Hermite basis of the corresponding polynomial spaces. Then,

$$
\mathbb{E}_{S \sim \mathcal{N}(0_k, I_k)} [(\eta^T g(S))^2] = \|D_1\|^2 = \|B_{r-1,k} D_2\|^2
\geq \lambda_{\min} \left( B_{r-1,k}^T B_{r-1,k} \right) \cdot \|D_2\|^2 \overset{Lemma 3.2.2}{\geq} 2^{-3(r-1)-k} \cdot \|D_2\|^2. \quad (3.36)
$$

Next, we establish a lower-bound of $\|D_2\|^2$.

$$
\|D_2\|^2 = \sum_{\alpha \in I_{r-1,k}} \left( \prod_{j \in [k]} \alpha_j! \right)^{-1} \cdot \left( \sum_{i=1}^k \eta_i \cdot w_{h,\alpha(i)} \cdot \sqrt{\alpha_i+1} \right)^2
= (r!)^{-1} \cdot \sum_{\alpha \in I_{r-1,k}} \left( \sum_{i=1}^k \eta_i \cdot w_{h,\alpha(i)} \cdot \sqrt{\alpha_i+1} \right)^2. \quad (3.37)
$$

For any unit $\eta^T \in \mathbb{R}^k$, we note that $\eta^T \nabla f$ is a polynomial of $k$ variables with degree at most $r - 1$, we have by Lemma 3.21 that:

$$
\mathbb{E}_{Z \sim \mathcal{N}(0_k, h^2 I_k)} [(\eta^T \nabla f(Z))^2] \geq r^{-2k-(r-1)/2} \cdot \min\{h^{2r-2}, 1\} \cdot \mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} [(\eta^T \nabla f(Z))^2]
\geq r^{-2k-(r-1)/2} \cdot \min\{h^{2r-2}, 1\} \cdot \tau. \quad (3.38)
$$

On the other hand, we have:

$$
\mathbb{E}_{Z \sim \mathcal{N}(0_k, h^2 I_k)} [(\eta^T \nabla f(Z))^2] = h^{-2} \mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} [(\eta^T \nabla f_h(Z))^2]
= h^{-2} \sum_{\alpha \in I_{r-1,k}} \left( \sum_{i=1}^k \eta_i \cdot w_{h,\alpha(i)} \cdot \sqrt{\alpha_i+1} \right)^2, \quad (3.39)
$$

where the last step we use multiple properties from Fact 3.1. From Equation (3.34)-Equation (3.39), we conclude that:

$$
\lambda_k(M) \geq r^{-r-4k} \cdot ((r-1)!)^{-1} \cdot 2^{-3r-k+3} (\sigma_\theta/h)^{2r-2} \cdot \min\{h^{2r-2}, 1\} \cdot \tau. \quad (3.40)
$$
3.7.6 Auxiliary Results

In this section, we provide some auxiliary results that were used throughout the proofs in this chapter.

**Fact 3.1** (Facts about Hermite Polynomials, [37, 54]). Let \( \{H_\ell\}_{\ell \in \mathbb{N}} \) be the normalized Probabilist’s Hermite polynomials over \( \mathbb{R} \), given by:

\[
H_\ell(z) = \sum_{i=1}^{\lfloor \ell/2 \rfloor} \frac{(-1)^i \sqrt{\ell!}}{2^i i! (\ell - 2i)!} \cdot z^{\ell - 2i}.
\]

Then,

(i) \( \{H_\ell\}_{\ell \in \mathbb{N}} \) is an orthonormal basis for \( L_2(\mathcal{N}(0, 1)) \), i.e.,

\[
\langle H_\ell, H_{\ell'} \rangle_{\mathcal{N}(0, 1)} = \begin{cases} 1 & \text{if } \ell = \ell' \\ 0 & \text{otherwise} \end{cases}.
\]

(ii) \( \forall \ell \in \mathbb{N} \) and \( \forall s \in \mathbb{R} \),

\[
\mathbb{E}_{Z \sim \mathcal{N}}[H_\ell(Z + s)] = s^\ell / \sqrt{\ell!}.
\]

(iii) \( \forall \ell \in \mathbb{N} \) and \( \forall h \in \mathbb{R} \),

\[
H_\ell(hz) = \sqrt{\ell!} \sum_{i=0}^{\lfloor \ell/2 \rfloor} \frac{h^{\ell - 2i} (h^2 - 1)^i}{2^i i! \sqrt{(\ell - 2i)!}} H_{\ell - 2i}(z).
\]

(iv) \( \forall \ell \geq 1 \),

\[
\frac{d}{dz} H_\ell(z) = \sqrt{\ell} H_{\ell-1}(z).
\]
Furthermore, for $\alpha \in \mathbb{N}^k$, let

$$H_\alpha (z) := \prod_{j=1}^{k} H_{\alpha_j} (z_j).$$

Then, the properties of univariate Hermite Polynomials can be generalized as follows:

(v) $\{ H_\alpha \}_{\alpha \in \mathbb{N}^k}$ is an orthonormal basis for $L_2(\mathcal{N}(0_k, I_k))$, i.e.,

$$\langle H_\alpha, H_{\alpha'} \rangle_{\mathcal{N}(0_k, I_k)} = \begin{cases} 1 & \text{if } \alpha = \alpha' \\ 0 & \text{otherwise} \end{cases}.$$

(vi) $\forall f \in L_2(\mathcal{N}(0_k, I_k))$ can be written in terms of Hermite polynomials as

$$f(z) = \sum_{\alpha : \alpha \in \mathbb{N}^k} w_\alpha H_\alpha (z),$$

where $w_\alpha = \langle f, H_\alpha \rangle_{\mathcal{N}(0_k, I_k)}$. Furthermore, we have

$$\| f \|_{\mathcal{N}(0_k, I_k)}^2 = \sum_{\alpha \in \mathbb{N}^k} w_\alpha^2.$$

(vii) $\forall s \in \mathbb{R}^k$,

$$\mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} [\partial_i f (Z + s)] = \sum_{\alpha \in \mathbb{N}^k : \alpha_i \geq 1} w_\alpha \prod_{j \neq i} \frac{1}{\sqrt{\alpha_j}} s_j^{\alpha_j} \frac{\sqrt{\alpha_i}}{\sqrt{(\alpha_i - 1)!}} S_i^{\alpha_i - 1}.$$

**Fact 3.2** (Lemma 36 from [37]). Let $B_{r,k} \in \mathbb{R}^{||I_r,k|| \times ||I_r,k||}$ denote the invertible linear map that converts the monomial basis to the Hermite basis for the space of the polynomial of $k$ variables with degree at most $p$, then we have

$$\lambda_{\min} (B_{r,k}^T B_{r,k}) \geq 2^{-3r-k}.$$
Fact 3.3 (Gaussian Moments). For $X \sim \mathcal{N}(0, 1)$, the $m^{th}$ moment is

$$
\mathbb{E}[X^m] = \begin{cases} 
0 & m \text{ odd} \\
2^{-(m/2)} \frac{m!}{(m/2)!} & m \text{ even}
\end{cases}.
$$

Fact 3.4. Suppose that $\epsilon$ is subgaussian with parameter $\sigma^2$, then $\mathbb{E}[|\epsilon|^q] \leq \sqrt{2\pi} \sigma$, and for any positive integer $q \geq 2$, the $q^{th}$ absolute moment satisfies the bound

$$(\mathbb{E}[|\epsilon|^q])^{1/q} \leq \sigma e^{1/e} q^{1/2}.$$  

Fact 3.5 (Gaussian Hypercontractivity, Theorem 11.23 from [54]). Suppose $f$ is a polynomial with degree $p$ in $k$ variables. Given any integer $q \geq 2$, we have

$$
\left( \mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} \left[ f(Z)^q \right] \right)^{1/q} \leq (q-1)^{r/2} \left( \mathbb{E}_{Z \sim \mathcal{N}(0_k, I_k)} \left[ f(Z)^2 \right] \right)^{1/2}.
$$

Lemma 3.22 (Truncated matrix Bernstein, Lemma 31 from [10]). Consider a truncation level $R > 0$. Let $\{Z_i\}_{1 \leq i \leq m}$ be a sequence of symmetric independent real random matrices with dimension $d \times d$ and $Z'_i = Z_i 1(\|Z_i\|_{\text{op}} \leq R)$, then

$$
\mathbb{P} \left\{ \left\| \frac{1}{m} \sum_{i=1}^m Z_i - \mathbb{E}[Z_i] \right\|_{\text{op}} \geq t \right\} \leq \mathbb{P} \left\{ \left\| \frac{1}{m} \sum_{i=1}^m Z'_i - \mathbb{E}[Z'_i] \right\|_{\text{op}} \geq t - \Delta \right\} + \sum_{i=1}^m \mathbb{P} \{\|Z_i\|_{\text{op}} > R\},
$$

where $\Delta \geq \|\mathbb{E}[Z_i - Z'_i]\|_{\text{op}}$. For $t \geq \Delta$, we further have

$$
\mathbb{P} \left\{ \left\| \frac{1}{m} \sum_{i=1}^m Z'_i - \mathbb{E}[Z'_i] \right\|_{\text{op}} \geq t - \Delta \right\} \leq 2d \exp \left( -\frac{m^2(t-\Delta)^2}{\sigma^2 + 2Rm(t-\Delta)/3} \right),
$$

where $\sigma^2 \geq \|\sum_{i=1}^m \mathbb{E}[(Z'_i - \mathbb{E}[Z'_i])^2]\|_{\text{op}}$.

Lemma 3.23 (Block Matrix Inversion, 14.11. from [55]). Let $A$ be an invertible matrix with the
following partition:

\[
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix},
\]

where \(A_{11}, A_{22}\) are invertible. Then,

\[
A^{-1} = \begin{bmatrix}
(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & 0 \\
0 & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}
\end{bmatrix}
\cdot
\begin{bmatrix}
I & -A_{12}A_{22}^{-1} \\
-A_{21}A_{11}^{-1} & I
\end{bmatrix}.
\]
Chapter 4: Individual Tree Crown Delineation in Tropical Rain Forest

4.1 Introduction

The rapid development of remote sensing technologies has a deep influence on tropical forest research in the past decade. Compared to traditional methods of gathering information about individual trees from the field, the collection of remote sensing data is much more efficient and accurate. This opens possibilities for studying forests at unprecedented levels of detail and over-large areas [56]. Airborne laser scanning (ALS) point clouds and aerial RGB imagery are two important types of data that have been extensively used to segment individual tree crowns (ITC). For a variety of ecological applications, ITC segmentation results with ALS point clouds usually provide information: they provide useful ITC characteristics such as crown volume and shape that one cannot obtain from RGB imagery. For this reason, there is an increasing interest in developing ITC algorithms using point clouds [16, 15, 57].

Despite its advantages, ALS-based ITC segmentation is usually considered as a more challenging task than image-based ITC segmentation, especially in dense, complex-structured forest areas. One important source of difficulty is the lack of ground truth segmentation. Identifying clusters as crowns from point clouds is impracticable for large areas of forest. For that reason, the existing methods are typically unsupervised: the algorithms do not rely on ground truth labels, but exploit certain data patterns in point clouds to identify individual crowns. For example, the 3D Adaptive Mean Shift (AMS3D) algorithm [16] locates tree tops by detecting the local maxima in point density and height; and the Multi-Class Graph Cut (MCGC) algorithm [15] transforms the segmentation task to a graph cut problem based on a density-based similarity score. Like most of the unsupervised algorithms, these methods are unstable and highly sensitive to the choice of hyperparameters.
To that end, we aim to improve over these existing methods by incorporating the information one can obtain from multiple sources of weak supervision. Although it is impractically difficult to manually annotate crowns from 3D point clouds, we are able to construct a 2D individual crowns on with the help of RGB or hyperspectral images. These 2D manual crowns provide partial information about the correctness of the segmentation of 3D crowns and can be used to refine ALS-based segmentation. In the language of weakly supervised learning, the manually delineation on the image can be regarded as weak labels. Most recently, similar weak labels has been used to improve the performance of ITC segmentation algorithm [58, 59]. For instance, in [58], the image-level manual labels are used to train a deep learning model called DeepForest [60] that can capture bounding boxes that contains complete tree crowns on unseen images. These bounding boxes are then used as the tool to detect potential mis-segmentations. In particular, they notice that AMS3D tends to over-segment the large crowns, which are usually visible from the top and can be well-segmented by DeepForest. Their strategy is to merge the point clusters that are corresponding to a single crown according to the DeepForest prediction. They show that the ratio of over-segmentation can be well reduced at the price of a slight increase of under-segmentation.

In this chapter, we propose a novel 3D crown refinement algorithm that address both the over-segmentation and under-segmentation issue, with the aid of manually delineated crowns on RGB image. That is, our method is able to detect both 1) small point clusters that belongs to a same crown; and 2) overly large crowns that need to be further splitted. As a comparison, [58] only detect and refine over-segmentation. The other significant advantage of our method over the earlier works is that our method avoid the over-complication of training of any deep learning model that identifies crowns on the image or their bounding boxes, and only uses a pre-trained ResNet50 Model as a visual feature extractor. Unlike [58], we only use the visual features to identify wrong segmentation directly via a binary XGBoost classifier. We note that object detection is much harder task: it is a pixel-level multi-class prediction problem but what we really need for the refinement is essentially a binary classification problem, i.e., whether a 3D crown is over-segmented or under-segmented. Our method is evaluated using a congruence analysis and a paired-tree analysis. The
remainder of the chapter is organized as follows: Section 4.2 introduces the experiment site and the detailed dataset we work on. The main refinement algorithm is briefly explained in Section 4.3 with detailed implementation and performance evaluations in Section 4.4.

4.2 Experiment Site and Data sets

4.2.1 Luquillo Forest Dynamic Plot

Our study area, the Luquillo Forest Dynamics Plot (LFDP), is a 16-ha (320m × 500m) forest plot in El Yunque National Forest of Puerto Rico located at SW corner 18° 20’ N, 65° 49’ W. Figure 4.1 shows the map of the National Forest including the area LFDP. The forest is classified as a subtropical wet forest in the Holdridge life zone system and has an average annual rainfall of 3,500 mm per year [61]. The LFDP was established to record long-term plant community dynamics and the effect of hurricane damage on a tropical forest. Multiple types of data are available for the vegetation in the plot area. The plot is censused regularly in the LFDP area, with all the stems with diameter at least 1cm at 1.30m height (diameter at breast height, DBH) being surveyed and recorded. For our purpose of study, we use the sixth census in conducted in 2016, which is the most recent one before Hurricane Maria. All the stems are georeferenced, with additional information (e.g. DBH, species) available. See [62] for inventory data collection details.

4.2.2 Remote Sensing Data

The remote sensing data for the study were from NASA Goddard’s LiDAR, Hyperspectral, and Thermal (G-LiHT) Airborne Imager campaign in 2017\(^1\). The data were collected from a nominal flying altitude of 335m AGL, using the G-LiHT Airborne Imager ([63]). In particular, we used high-resolution ($10000 \times 10000$ pixels with $3cm \times 3cm$ per pixel) imagery and Airborne LiDAR point cloud ($\sim 36.75$ points/m\(^2\)) of the interested area. The raw data were pro-processed (shadow removing, ground point removing, normalization, etc.) before passed to further analysis.

\(^1\)https://glihtdata.gsfc.nasa.gov/puertorico/index.html
Figure 4.1: The Map of the El Yunque rainforest in Puerto Rico, the yellow bounding box shows the location of LFDP, where ground observations are available.
4.2.3 Manual Delineation of Crowns

In this section, we present a fully annotated dataset of individual crowns on image for the LFDP area, which we name *LFDP-Crowns*. The dataset includes:

- a georeferenced aerial RGB image of the LFDP;
- a polygon shapefile [64] containing all of the manually delineated crowns in the LFDP.

Our dataset contains all of the required information to readily serve as a training set for CNN-based ITC delineation models. We note that some earlier works [65, 59] also include manual segmentation of ITC as an ad-hoc procedure, as a step for model training and evaluation. Our hope is that our *LFDP-Crowns* dataset can serve as a prototype for the future release of similar types of human annotations.

**Image preprocessing**  The high-resolution (3cm × 3cm) georeferenced aerial RGB images of the LFDP can be downloaded from G-LiHT Data Center (https://glihtdata.gsfc.nasa.gov/). A large portion of the target area is covered by cloud shadows, adding difficulties to manual delineation and crown prediction. Therefore, we first removed shadows and balanced colors in the raw images using a similar strategy in an earlier work [66]. Briefly speaking, we partitioned the whole image into small patches, and aligned the RGB histograms of each patch with that of a benchmark image, in which there is no shadow. Such shadow-removal process yield an image that is more homogeneous and ready for further analysis, see Figure 4.2 for a comparison of the original and processed image.

**Auxiliary datasets**  Manual delineation solely based on the RGB imagery is a difficult task and prone to human errors. In many circumstances, auxiliary data sources of different types can provide additional information about individual crowns and thus are extremely helpful. For instance, if two species of trees have visually similar crowns (e.g., in color), it can be nearly impossible to delineate them individually from the image. However, if these species differ substantially in tree heights, the
delineation of crown boundaries would be much easier with the help of a LiDAR-derived canopy height model (CHM) [67]. The CHM is also available in the G-LiHT Data Center.

Another useful data source is field measurement. As mentioned in the introduction, all stems in the LFDP with a DBH greater than 1 cm have been mapped and georeferenced, identified to species, and measured following a detailed protocol [68]. Approximately every five years since 1990, stems have been re-measured and their status has been assessed, and new stems have been added as they appear. These stem locations and associated species labels have been used as weak supervision in a novel framework for a pixel-level mapping of tree species using aerial images of rainforest [66]. Our analyses used data from the last census of the LFDP completed in 2016. The census data is available at EDI Data Portal².

**Manual Delineation**  To carry out the manual delineations, we visually inspected the color, shape and texture of the vegetation from the RGB image, with the assistance of the auxiliary datasets (i.e. CHM and field inventories). Delineations were conducted using QGIS, with the toggle editing tool.

Figure 4.3: Illustration of the use of field measurement for aiding manual segmentation. In the left panel, it is hard to tell, solely based on the RGB image, whether the circular area annotated by the solid red curve is a single canopy, or two canopies separated by the dashed curve. In the right panel, each dot represents a tree stem from field measurements. The two highlighted stems are of same species (Guarea guidonia) and comparable diameter(DBH). This step supports the dashed curve as the boundary between the two crowns.

In total, we delineated 3,855 over-story individual tree crowns that are visible from the aerial image in the LFDP area, which takes around 100 working hours. This number matches approximately with the total number of stems with DBH ≥ 20 cm (3,791) (Figure 4.4). For quality assurance, the preliminary manual delineations were carefully examined by a group of forest ecologists with knowledge of the site.
Data Records  The complete dataset is publicly available at Mendeley Data \(^3\). The repository has the following structure:

```
./
  LFDP.tif
  manual_labels.cpg
  manual_labels.dbf
  manual_labels.prj
  manual_labels.qpj
  manual_labels.shp
  manual_labels.shx
```

Among these files, **LFDP.tif** is the denoised georeferenced RGB image; **manual_labels.shp** is the ESRI shapefile that contains 3,855 polygons, representing manually delineated individual crowns; and other auxiliary files that contain useful information (e.g. geographic projection, character encoding) about the shapefile.

\(^3\)http://doi.org/10.17632/gfpwnpw4jk
4.3 Methods

The idea of data fusion has been widely exploited in the fields of remote sensing and crown delineation ([69, 58]). The main idea behind data fusion is that data from multiple sources provide different views of the target objects. Combining these different views Despite the recent success in ITC detection using airborne LiDAR, e.g., AMS3D by [16] and MCGC by [15], none of these segmentation methods are perfect. Many of these algorithms have satisfactory overall performance on the plot level, however the segmentation is not perfect in certain locations. Some of these imperfect crowns can be visually seen by overlaying the delineation result with the RGB image. This motivates the use of RGB image and manual segmentation to correct the raw results from 3D LiDAR delineation algorithms.

Our refinement algorithm is built on the state-of-the-art machine learning models. The workflow is summarized as in Figure 4.5.
4.3.1 Incorrect segmentation detection

We aim to detect under- and over-segmentation in the raw segmentation results. With manual segmentation as our ground truth, it is straightforward to mark both under- and over-segmentation following a similar procedure as described in [70]. Specifically, the overlap coefficients (OC) between all the pairs of manual crowns and algorithm-detected crowns (after being projected to the plane of RGB images) are calculated:

$$\text{OC}(C, C') = \frac{|C \cap C'|}{\min(|C|, |C'|)}.$$  

where $|\cdot|$ is the area of a crown. Here, we choose the overlap coefficient over the Jaccard index (JI, as in [70]) for the detection. To see its superiority, consider a large algorithm-detected crown with several non-overlapping small manual crowns residing in it. The JIs of the large algorithm-detected crown and each of the small manual crown is small, indicating the existence of mis-segmentation. However, nothing can be inferred about whether it is over-segmentation or under-segmentation. This could be easily solved by using OC. If an algorithm-detected crowns has OCs at least 0.5 with more than one manual crown, then it is labeled as under-segmented. Similarly, if a manual crown has OCs at least 0.5 with a pair of algorithm-detected crowns, then the pair of crowns is labeled as over-segmented.

The above procedure relies on manual segmentation, which is extremely expensive to get for a study of a larger scale. To this end, we take the manually delineated dataset we created in Section 4.2.3 (with both RGB images and manual labels available) as the training set and train predictive models that detect incorrect segmentation in a larger scale of the forest without manual labels. The training procedures for under- and over-segmentation detection models are as follows:

**Under-segmentaion detection model** For each algorithm-detected crown in the LFDP, we crop out the corresponding part of RGB image and feed it to the a pre-trained deep CNN model for feature extraction. Existing choices of the feature extraction models include AlexNet [71], ResNet
Finally, we train a XGBoost classifier using the visual features and the binary label that indicates whether the crowns are under-segmented.

**Over-segmentation detection model** The procedure is similar to under-segmentation detection, except that the basic observation unit is a pair of algorithm-detected crowns. The visual features are now the concatenation of both crowns in a pair and the binary label is the indicator of whether the pair should be merged.

4.3.2 Crown merging and splitting

The operation of crown merging is straightforward. Once a pair of crowns are detected as over-segmented and should be merged, we can simply merge the two crowns by concatenate the corresponding LiDAR clusters. The crown splitting depends on the choice of 3D segmentation algorithm. In particular, these segmentation algorithms usually contain hyperparameters that control the size of the resulting segmentation. Such examples include the height to crown size parameters in AMS3D and the number of clusters $k$ in MCGC. The crown splitting is achieved through rerunning the 3D segmentation algorithm locally and tuning the hyperparameters to get finer segmentation.

4.4 Implementation Details and Results

4.4.1 Baseline 3D LiDAR segmentation

After comparing a couple with existing methods, we choose the AMS3D algorithm as our baseline 3D LiDAR segmentation algorithm, since the methods is designed for tropical forest area with complex geographical structure and high vegetation density. In a nutshell, the AMS3D algorithm apply the traditional mean shift algorithm throughout the point cloud, while choose the bandwidth in kernel density estimation automatically according to the local allometries over different space. Generally speaking, larger bandwidth were using in higher locations, as larger crowns (clusters) are more likely to appear.
With proper choice of hyper-parameters \( H2CW = 0.42, H2CL = 0.35 \), a total number of 3899 crowns were detected after eliminating small (i.e. those contain < 5 points) and under-story (i.e. those are not showing up in CHM) clusters. This roughly matches the number of manual delineated crowns. Nonetheless, we can spot some over- and under-segmentation in certain locations.

4.4.2 Refinement algorithm

Now, we introduce the detailed implementation of the refinement algorithm, by listing some key steps of the refinement algorithm.

**Training-test split**  To show that the predictive models have robust out-of-sample performances, we split the whole plot of size 320m ×500m into the two parts. The crowns in the northernmost 320m ×400m of the Plot is used to train the refinement model, and the remaining are used as the test set.

**Visual feature extraction**  We choose the ResNet50 Model ([72]), pre-trained on ImageNet, as the feature extractor. The image corresponding to each algorithm-detected crown is first filled with white pixels to a smallest square image patch, and then resized to 100 pixels × 100 pixels before feeding into the deep CNN model. A 2048-dimensional vector feature will be extracted from each image patch.

**Under-segmentation detection**  As shown in the Figure 4.5, we first detect under-segmented crowns and split them. In particular, we fit an XGBoost classifier \( n\text{-estimator} = 100 \), selected by 10-fold cross-validation) with the visual features and the manual segmentation. A decision threshold of 0.2 on the predicted probability is used to adjust for the imbalanced sample. The detection results sites are summarized in Table 4.1 and the ROC curve for the prediction is presented in Figure 4.6. The LiDAR clouds corresponding to the under-segmented crowns are then further segmented by running the AMS3D algorithm with smaller allometric parameters \( H2CL = H2CW \).
= 0.20). After splitting, the number of crowns increases to 5453 (from 3899).

**Over-segmentation detection**  Before fitting the over-segmentation detection model, we pair the split crowns up. Specifically, the pairs of crowns whose centroid distances are within 10m are considered as the basic observation unit. Similar to the under-segmentation detection, we fit an XGBoost classifier \((n\text{-estimator} = 35, \text{selected by 10-fold cross-validation})\) with the concatenated visual features and the manual segmentation. A decision threshold of 0.3 is used. The detection results are summarized in Table 4.1 and the ROC curve for the prediction is presented in Figure 4.6. After the detection, the over-segmented crowns, including those that are overly split after under-segmentation detection, are merged.

<table>
<thead>
<tr>
<th></th>
<th>detected as should split</th>
<th>detected as should not split</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>should split</td>
<td>719</td>
<td>0</td>
</tr>
<tr>
<td>should not split</td>
<td>0</td>
<td>3142</td>
</tr>
<tr>
<td><strong>Testing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>should split</td>
<td>94</td>
<td>46</td>
</tr>
<tr>
<td>should not split</td>
<td>59</td>
<td>597</td>
</tr>
</tbody>
</table>

**Table 4.1:** Under- and over-segmentation detection results

<table>
<thead>
<tr>
<th></th>
<th>detected as should merge</th>
<th>detected as should not merge</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>should merge</td>
<td>4132</td>
<td>0</td>
</tr>
<tr>
<td>should not merge</td>
<td>0</td>
<td>13220</td>
</tr>
<tr>
<td><strong>Testing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>should merge</td>
<td>775</td>
<td>163</td>
</tr>
<tr>
<td>should not merge</td>
<td>2</td>
<td>3016</td>
</tr>
</tbody>
</table>

Figure 4.6: ROC curves for under-segmentation detection (left) and over-segmentation detection (right)
4.4.3 Performance Evaluation

To evaluate the performance of refinement algorithm, we conduct both the _congruence analysis_ and _paired-tree analysis_ on raw segmentation (by AMS3D) and the refined segmentation with the whole LFDP area. These analysis compared the machine segmentation with either the manual segmentation or the tree inventories and give us a comprehensive evaluation on the segmentation performance. An earlier work by [70] used similar analysis to compare different segmentation methods. Here, we aim to show the performance boost from refinement.

**Congruence analysis** The congruence analysis summarize the similarity of the machine segmentation and the manual segmentation. In particular, we consider two popular types of similarity metric:

- Jaccard index (JI): the JI of two crowns \( C \) and \( C' \) is defined by

\[
JI(C, C') = \frac{|C \cap C'|}{|C \cup C'|}
\]

- Boundary similarity (BS): the BS of two crowns \( C \) and \( C' \) is defined by the proportion of the shared boundary of these two crowns. To illustrate the detailed calculation, we consider the simple example in figure 4.7.

![Similarity metrics: boundary similarity (left), Jarccard index (right)](image)

Figure 4.7: Similarity metrics: boundary similarity (left), Jarccard index (right)

Just like any object detection problem, two fundamental aspects should be studied:
<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Refined</th>
<th>Refined</th>
<th>Refined</th>
</tr>
</thead>
<tbody>
<tr>
<td>JI &gt; 0.3</td>
<td>81.97</td>
<td>87.57</td>
<td>84.29</td>
<td>89.66</td>
<td>89.66</td>
<td>89.66</td>
</tr>
<tr>
<td>JI &gt; 0.5</td>
<td>67.08</td>
<td>73.54</td>
<td>62.41</td>
<td>74.07</td>
<td>74.07</td>
<td>74.07</td>
</tr>
<tr>
<td>JI &gt; 0.8</td>
<td>36.21</td>
<td>40.70</td>
<td>34.19</td>
<td>40.34</td>
<td>40.34</td>
<td>40.34</td>
</tr>
<tr>
<td>BS &gt; 0.3</td>
<td>68.02</td>
<td>79.51</td>
<td>65.88</td>
<td>76.41</td>
<td>76.41</td>
<td>76.41</td>
</tr>
<tr>
<td>BS &gt; 0.5</td>
<td>25.88</td>
<td>40.44</td>
<td>25.64</td>
<td>38.85</td>
<td>38.85</td>
<td>38.85</td>
</tr>
<tr>
<td>BS &gt; 0.8</td>
<td>4.12</td>
<td>7.91</td>
<td>4.07</td>
<td>7.57</td>
<td>7.57</td>
<td>7.57</td>
</tr>
</tbody>
</table>

Table 4.2: Percentage of predicted crowns with a matched manual crown

<table>
<thead>
<tr>
<th></th>
<th>IoU &gt; 0.3</th>
<th>IoU &gt; 0.5</th>
<th>IoU &gt; 0.8</th>
<th>BS &gt; 0.3</th>
<th>BS &gt; 0.5</th>
<th>BS &gt; 0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>84.29</td>
<td>62.41</td>
<td>34.19</td>
<td>65.88</td>
<td>25.64</td>
<td>4.07</td>
</tr>
<tr>
<td>Refined</td>
<td><strong>89.66</strong></td>
<td><strong>74.07</strong></td>
<td><strong>40.34</strong></td>
<td><strong>76.41</strong></td>
<td><strong>38.85</strong></td>
<td><strong>7.57</strong></td>
</tr>
</tbody>
</table>

Table 4.3: Percentage of manual crowns with a matched predicted crown

i) **Precision:** what percentage of the predicted crowns are true crowns? That is, how correct the crown detection is. To answer this question, we refer to Table 4.4.3, where the proportion of the predicted crowns with a matched manual crown is reported. Here, two crowns match if the similarity metric (either JI or BS) surpass a certain threshold (0.3, 0.5 or 0.8).

ii) **Recall:** what percentage of the true crowns are detection? That is, what is the detection rate. To answer this question, we refer to Table 4.4.3, where the proportion of the true crowns with a matched predicted crown is reported.

**Paired-tree analysis** In this section, we perform a paired-tree analysis to evaluate the segmentation using the tree inventory. The evaluation relies on a matching algorithm that connects manually annotated crowns with the stems from field measurements. For each crown-stem pair, a matching score that is calculated as:

\[ S(C_i, S_j) = D(C_i, S_j) * L(C_i, S_j) \]  \hspace{1cm} (4.1)

where \( D(C_i, S_j) \) is the distance score between the crown \( C_i \) and the stem \( S_j \); and \( L(C_i, S_j) \) is the likelihood of observing the pair \( (C_i, S_j) \) based on species-specific allometric models. A high matching score indicates 1) the crown centroid and the stem are geographically close to each other; and 2) the sizes of crown and stem reflect the species allometries. Similar matching mechanism have been adopted in previous comparative studies, e.g., [58], [74].
**Distance Score**  The distance score uses a relative distance that based on distance rankings of crown-stem pairs, rather than absolute distances, since the latter are proportional to tree sizes, and thus not comparable between trees of different sizes. We first calculate the pairwise distances of all the crown-stem pairs, specifically the distances between the crown centroids and stems. Then, $R(C_i, S_j)$ is defined as the distance ranking of $S_j$ to $C_i$ among all the stems in the field inventory. Finally, the distance score is defined as $D(C_i, S_j) = \alpha \exp(-\alpha R(C_i, S_j))$, where $\alpha > 0$ is a tuning parameter.

**Likelihood Score**  The other metric that goes into the matching score is how well the sizes of the crowns and the stems match according to the allometric relationship of the stem species. These allometric models are built based on inventory data, which contains complete information about each tree, including species information, stem size, height and maximum crown diameter. We split the dataset into nine groups by the tree species (Table 4.4), and fitted the following allometric models separately for each species:

$$
\log \text{DBH}_{sp,k} = \beta_{sp,0} + \beta_{sp,1} \log \text{Height}_{sp,k} + \beta_{sp,2} \log \text{CD}_{sp,k} + \varepsilon_{sp,k}; \quad (4.2)
$$

where $sp$ is the species, and $\varepsilon_{sp,i} \sim N(0, \sigma^2_{sp})$ are i.i.d. errors. Once the allometric models were fitted, we calculated the likelihood score of a crown-stem pair as:

$$
L(C_i, S_j) = \frac{1}{\sqrt{2\pi \hat{\sigma}_{sp_j}}} \exp \left[ -\frac{(\log \text{DBH}_i - \log \hat{\text{DBH}}_j)^2}{2} \right],
$$

where $sp_j$ is species of $S_j$ and $\log \hat{\text{DBH}}_j = \hat{\beta}_{sp,j,0} + \hat{\beta}_{sp,j,1} \log \text{Height}_j + \hat{\beta}_{sp,j,2} \log \text{CD}_j$.

<table>
<thead>
<tr>
<th>Species</th>
<th>CECSCH</th>
<th>CASARB</th>
<th>DACEXC</th>
<th>GUAGUI</th>
<th>MANBID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts</td>
<td>25</td>
<td>23</td>
<td>43</td>
<td>48</td>
<td>31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species</th>
<th>PREMON</th>
<th>SCHMOR</th>
<th>SLOBER</th>
<th>OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts</td>
<td>29</td>
<td>24</td>
<td>29</td>
<td>113</td>
</tr>
</tbody>
</table>

Table 4.4: Number of stems used to derive species allometries for the target species
Figure 4.8: Histogram of absolute crown-stem distances (left, in meters) and distance rankings (right) for the 3,329 matched crown-stem pairs. The majority of the stems are paired with their closest or second closest neighboring crowns. The range of the crown-stem distances is 0-20m, with mean distance of 6m.

Figure 4.9: Allometric plots of (A) DBH(cm) v.s. height(m)), and (B) DBH(cm) v.s. crown depth(m) for all of the matched crown-stem pairs, by species.

To avoid matching conflicts, we ordered the stem by decreasing DBHs and paired them to the crowns one by one. Each stem was matched to the crown with the largest matching score, i.e. $S_j$
will be matched to $C_i$ such that $S(C_i, S_j)$ is maximized over all available crowns. Once paired with a stem, the crown was removed from the list so that it would not be matched with multiple stems. Matched pairs have small crown-stem distances (Figure 4.8) and there is a clear positive correlation between 1) stem size and tree height (Figure 4.9A); 2) stem size and crown size (Figure 4.9B).
References


[41] Z. Yang, K. Balasubramanian, Z. Wang, and H. Liu, “Estimating high-dimensional non-
gaussian multiple index models via stein’s lemma,” in *Advances in Neural Information
Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish-


[43] S. Trivedi, J. Wang, S. Kpotufe, and G. Shakhnarovich, “A consistent estimator of the ex-
pected gradient outerproduct,” in *Proceedings of the Thirtieth Conference on Uncertainty
in Artificial Intelligence*, ser. UAI’14, Quebec City, Quebec, Canada: AUAI Press, 2014,

works and kernel machines that recursively learn features,” *arXiv preprint arXiv:2212.13881*,
2022.


Heidelberg: Springer, 2011.


[50] M. Kearns, “Efficient noise-tolerant learning from statistical queries,” *J. ACM*, vol. 45, no. 6,

[51] I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis, “Algorithms and sq lower bounds
for pac learning one-hidden-layer relu networks,” in *Proceedings of Thirty Third Confer-


