Nonlinear Inverse Problems: Efficient and Guaranteed Algorithms

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Seonho Kim, M.S.

Graduate Program in Electrical and Computer Engineering

The Ohio State University

2024

Dissertation Committee:

Kiryung Lee, Advisor

Philip Schniter

Yoonkyung Lee

Abstract

Nonlinear inverse problems arise in fields such as engineering, statistics, and machine learning. Unlike linear inverse problems, which can be formulated as convex programs, the main challenge in nonlinear inverse problems is the non-convex nature of the optimization involved. Solving non-convex optimization problems is NP-hard, susceptible to local minima, and often computationally intractable, making it essential to design practical algorithms with guaranteed performance.

This thesis addresses two specific nonlinear inverse problems. The first problem is robust phase retrieval, which has applications in areas including X-ray crystallography, diffraction and array imaging, and optics. In this problem, the forward model is the magnitude of linear measurements, and the observations are corrupted by sparse outliers. We employ a least absolute deviation (LAD) approach to robust phase retrieval, which aims to recover a signal from its absolute measurements contaminated by sparse noise. To tackle the resulting non-convex optimization problem, we propose a robust alternating minimization (Robust-AM) approach, derived as an unconstrained Gauss-Newton method. For solving the inner optimization in each step of Robust-AM, we adopt two computationally efficient methods. We provide a non-asymptotic convergence analysis of these practical algorithms for Robust-AM under the standard Gaussian measurement assumption. With suitable initialization, these algorithms are guaranteed to converge linearly to the ground truth at an order-optimal sample

complexity with high probability, assuming the noise support is arbitrarily fixed and the sparsity level does not exceed 1/4. Furthermore, comprehensive numerical experiments on synthetic and image datasets demonstrate that Robust-AM outperforms existing methods for robust phase retrieval, while offering comparable theoretical guarantees.

The second problem is max-affine regression, where the forward model is a convex piecewise-linear function, also known as the *max-affine model*, which combines k affine models using a max function. This model is advantageous for approximating the data relationship in a way that is both interpretable and effective for fitting shape-restricted data which often arises in economic, financial, and engineering applications.

In the first part of this study, we focus on the max-linear model, a simplified version of the max-affine model without the bias term, with observations corrupted by deterministic noise. For this scenario, we propose a scalable convex estimator. Under the assumption of Gaussian covariates, we establish a non-asymptotic performance guarantee, demonstrating that the convex estimator recovers the parameters with high probability. When the k linear components are equally likely to achieve the maximum, our results show that the number of noise-free observations required for exact recovery scales as $\mathcal{O}(k^4 p)$ up to a logarithmic factor, matching the sample complexity of alternating minimization (Ghosh et al., 2019). This sample complexity also holds when observations are corrupted by arbitrary deterministic noise. Empirical results further demonstrate that our method performs in line with our theoretical predictions and competes favorably with the alternating minimization algorithm, especially in the presence of multiplicative Bernoulli noise. Additionally, we show that recursively applying the estimator can significantly improve estimation accuracy.

The second part focuses on max-affine regression under sub-Gaussian noise. We present a non-asymptotic convergence analysis of gradient descent (GD) and mini-batch stochastic gradient descent (SGD) for max-affine regression when the model is observed at random locations, assuming sub-Gaussianity and anti-concentration with additive sub-Gaussian noise. Under these conditions, suitably initialized GD and SGD converge linearly to a neighborhood of the ground truth, with the error bound specified accordingly. Numerical results support our theoretical findings, showing that SGD not only converges faster in runtime with fewer observations than both alternating minimization and GD in the noiseless case, but also outperforms these methods in low-sample scenarios with noise.

Dedicated to my family.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Prof. Kiryung Lee, for his invaluable support throughout my Ph.D. studies. Under his guidance, I learned how to formulate problems mathematically, pose meaningful theoretical questions, and communicate effectively through writing and presentations. These lessons have been instrumental in my development into a mathematically mature and independent researcher. I am deeply thankful for his mentorship, which has shaped me into the kind of researcher I aspired to become at the start of my Ph.D. journey.

I am sincerely grateful to my thesis committee members, Prof. Philip Schniter and Prof. Yoonkyung Lee, for their insightful comments and invaluable suggestions throughout my research proposal and thesis. Their feedback helped me contextualize my work within a broader framework, including connections to state-of-the-art algorithms in statistics and machine learning. I also deeply appreciate Prof. Yingbin Liang for her thoughtful feedback during my qualifying and candidacy exams, which significantly shaped and improved my work.

I would also like to thank my collaborators, Dr. Sohail Bahmani and Dr. Rakshith Sharma Srinivasa, for their valuable insights, which greatly enhanced the quality of our work. Sohail consistently provided insightful comments and demonstrated how to solve problems with elegance and simplicity.

My sincere thanks go to Prof. Lixin Ye for serving as the graduate faculty representative for my defense and for offering insightful perspectives on how my work can be connected to and understood within the field of economics.

I am profoundly grateful to my undergraduate and master's advisor, Prof. Songnam Hong, who guided my first steps into research in machine learning and statistics. He helped me understand the importance of mathematical theory in solving engineering problems and inspired me to pursue a Ph.D. abroad. His advice and support during my master's degree, especially during challenging times, were instrumental in shaping my academic path.

I am forever thankful to my family—Dae Young Kim, Sook Ja Kang, Jungmi Kim, Bangmi Kim, and Sunmi Kim—for their unwavering support. Their love, trust, and encouragement made it possible for me to complete this journey successfully.

I am also fortunate to have had the support of my lab mates: Meghna Kalra, Haitham Kanj, Noah Levine, Charles Berdanier, and Alex Thieken. It has been a privilege to work alongside them and share so many meaningful experiences.

My heartfelt thanks go to my Korean friends at OSU, including Kyeong Joo Jung, Chae Eun Jang, Jong Hoon Shin, Ju-seung Byun, Won Yong Chung, Seungbin Park, Dr. Jonghoo Lee, Ji Yoon Kim, Eugine Caroline Shin, Jihyun Lee, Prof. Min Ho Cho, Hyeran Cho, Dr. Hyeong Jun Kim, Prof. Seunghyun Lee, Dr. Hyunsoo Lee, Dr. Hyemin Jung, Dr. Taeyoung Kim, and Dr. Yunsik Hahn. From the beginning of my Ph.D. to the challenging times during COVID, their friendship and support have been invaluable. I will always treasure the beautiful memories we shared, whether in moments of happiness or hardship.

I also wish to thank my long-time Korean friends, Jaewon Jung, Joonmin Lee, Dong-chan Son, Daechul Ahn, HoKi Seo, Kiwook Kwon, Seongyeol Park, and Jinwoo Jang. Although we now follow different paths, our connections from high school or university remain strong. Whenever I visited Korea, it was a joy to catch up and share our life stories.

Lastly, I am deeply grateful to my girlfriend, Ye Jin Choi, for her endless understanding, support, and love. She has been my best friend and constant source of motivation. Her unwavering trust and encouragement helped me overcome every challenge during this journey. I feel truly lucky to have her by my side, now and in the future.

Finally, I acknowledge the support provided by the ECE department and the NSF CAREER Award CCF-1943201, which made this work possible.

Vita

2017	B.S., Electrical and Computer Engineering, Ajou University, Suwon, South Korea
2019	M.S., Electrical and Computer Engineering, Ajou University, Suwon, South Korea
2019-present	Ph.D., Electrical and Computer Engineering The Ohio State University, Columbus, USA

Publications

H. Kanj, S. Kim, and K. Lee, "Variable Selection in Convex Piecewise Linear Regression," *submitted to SIAM Journal on Mathematics of Data Science*, (arXiv:2411.02225).

H. Kanj, S. Kim, and K. Lee, "Variable Selection for Max-Affine Regression via Sparse Gradient Descent," *in Proceedings of the 2024 IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop (SAM)*, vol. 13, no. 1, Corvallis, OR, USA, pp. 1-5, 2024.

S. Kim and K. Lee, "Robust Phase Retrieval by Alternating Minimization," *Under minor revision in IEEE Transactions on Signal Processing.*

S. Kim and K. Lee, "Max-Affine Regression via First-Order Methods," *SIAM Journal on Mathematics of Data Science*, vol. 6, no. 2, pp. 534–552, 2024.

S. Kim, S. Bahmani, and K. Lee, "Max-Linear Regression by Convex Programming," *IEEE Transactions on Information Theory*, vol. 70, no. 3, pp. 1897–1912, 2024.

S. Kim and K. Lee, "Sequence of Linear Program for Robust Phase Retrieval," *in Proceedings of ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2024, no. 1, Seoul, Korea, Republic of, pp. 9791-9795, 2024.

S. Kim and K. Lee, "Fast Max-Affine Regression via Stochastic Gradient Descent," *in Proceedings of the 59th Annual Allerton Conference on Communication, Control, and Computing*, vol. 59, no. 1, Monticello, IL, USA, pp. 1-5, 2023.

R. S. Srinivasa, S. Kim, and K. Lee, "Sketching Low-Rank Matrices with a Shared Column Space by Convex Programming," *IEEE Journal on Selected Areas in Information Theory*, vol. 4, pp. 54–60, 2023.

Fields of Study

Major Field: Electrical and Computer Engineering
Minor Field: Mathematics

Table of Contents

Page

List of Tables

Table **Page**

List of Figures

Figure **Page**

Chapter 1: Introduction

In this dissertation, we examine nonlinear inverse problems, which are essential yet challenging in the fields of engineering, statistics, and machine learning. An inverse problem involves recovering the true signal $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ from a forward model $\{f_i\}_{i=1}^n$, where $f_i : \mathbb{R}^d \to \mathbb{R}$, and observations $\{y_i\}_{i=1}^n$ generated by

$$y_i = f_i(\boldsymbol{\theta}_\star) + z_i, \quad i = 1, \dots, n,$$

where $z_1, \dots, z_n$ represent noise. Inverse problems are prevalent in various applications, including biomedical imaging [9], computer vision [67], and scientific research [73].

To emphasize the complexities inherent in nonlinear inverse problems, we first consider the simpler case of linear inverse problems. In this scenario, the observation model is given by

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle + z_i, \quad i = 1, \dots, n, \tag{1.0.1}$$

where $\{\boldsymbol{x}_i\}_{i=1}^n$ are measurement vectors (often called sensing vectors in signal processing or covariates in statistical regression).

The choice of method to recover $\boldsymbol{\theta}_\star$ from $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ in (1.0.1) depends on the nature of the noise $\{z_i\}_{i=1}^n$. If we assume Gaussian noise, i.e., $\{z_i\}_{i=1}^n \overset{i.i.d}{\sim} N(0, \sigma^2)$, commonly encountered in linear regression or as sensing noise in signal processing,

the maximum likelihood estimation (MLE) approach is often used:

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \prod_{i=1}^{n} P(y_i; \boldsymbol{x}_i, \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\mathbb{R}^d} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle)^2}{2\sigma^2}\right). \tag{1.0.2}$$

Maximizing the likelihood in (1.0.2) is equivalent to minimizing the negative log-likelihood, leading to the least squares estimator (LSE):

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle)^2. \tag{1.0.3}$$

Alternatively, in scenarios with sparse outliers, where the noise $\{z_i\}_{i=1}^n$ is sparse but can take arbitrary values, the least squares approach (1.0.3) is sensitive to such outliers. In these cases, robust estimation methods, such as the least absolute deviation (LAD) estimator, are more suitable:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} |y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle|. \tag{1.0.4}$$

It is well known that LAD is robust to outliers [10]. Notably, the optimization formulations in linear inverse problems, such as (1.0.3) and (1.0.4), are convex, ensuring that any minimizer is a global solution.

We now turn our attention to nonlinear inverse problems. Consider the following magnitude-based model as a simple example:

$$y_i = |\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star\rangle| + z_i, \quad i = 1, \ldots, n. \tag{1.0.5}$$

The forward model in (1.0.5) is nonlinear due to the absolute value function, and such models arise in phase retrieval problems, as we will discuss later. Following optimization approaches similar to those in linear inverse problems, one can consider various formulations. Under Gaussian noise, for instance, the least squares estimator (LSE) can be used:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} (y_i - |\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle|)^2. \tag{1.0.6}$$

In the presence of outliers, the LAD estimator can be applied:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} \left| y_i - |\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle| \right|. \tag{1.0.7}$$

Compared to the convex optimizations in (1.0.3) and (1.0.4), the formulations in (1.0.6) and (1.0.7) are non-convex due to the nonlinear forward model. Consequently, there may be local minima and saddle points in the optimization landscape, making these problems NP-hard and computationally intractable [54]. Thus, designing practical and guaranteed algorithms for non-convex optimizations is critical.

Common approaches for tackling non-convex optimizations include first-order methods and convex relaxation techniques. Key considerations in designing an effective algorithm are:

- Computational Efficiency: This is particularly vital for large-scale problems. In iterative algorithms, convergence rate is crucial; even if each iteration is computationally inexpensive, slow convergence can result in high overall computational cost.

- Sample Efficiency: A desirable algorithm should perform well with fewer samples. Theoretical study of sample complexity, the number of samples required for reliable performance, is also important.

This thesis focuses on theoretical aspects of algorithm performance, particularly convergence and sample complexity.

We study two specific nonlinear inverse problems. The first is robust phase retrieval, where observations are corrupted by outliers. This problem has numerous applications in engineering, which we will discuss in detail later. The second problem is max-affine regression, relevant to either prediction or optimization problems in economic,

financial, and engineering. For the max-affine model, we consider deterministic or sub-Gaussian noise scenarios. We propose three algorithms tailored to these nonlinear inverse problems, accounting for both the forward model and noise conditions.

In the following subsections, we summarize the problem formulation, background, and our main contribution for each non-linear inverse problem.

1.1 Robust Phase Retrieval

Phase retrieval refers to the recovery of unknown signals $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ (or $\mathbb{C}^d$) from the magnitudes of its linear measurements, which are formulated as

$$y_i = |\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle|, \quad i = 1, \dots, n, \tag{1.1.1}$$

where $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in \mathbb{R}^d$ (or $\mathbb{C}^d$) are known measurement vectors. Solving the set of nonlinear equations in (1.1.1) arises in numerous applications including X-ray crystallography, diffraction and array imaging, and optics (e.g. [15, 20, 78, 96]). We consider the robust phase retrieval from the amplitude measurements in (1.1.1) corrupted with sparse noise, i.e.

$$y_i = \begin{cases} \xi_i & \text{if } i \in I_{\text{out}} \\ |\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle| & \text{if } i \in I_{\text{in}} \end{cases} \tag{1.1.2}$$

where $I_{\text{out}} \subset [n]$ and $I_{\text{in}} = [n] \setminus I_{\text{out}}$ collect the unknown indices of outliers and inliers respectively, and $\{\xi_i\}_{i \in I_{\text{out}}}$ is an arbitrary sequence in $\mathbb{R}$. For example, such a scenario arises in phase retrieval imaging applications [101] due to various reasons including detection failures and recording errors.

A suite of methods designed for the plain phase retrieval [30] has been adapted to address the outliers. These methods provide not only empirically successful performances but also theoretical analyses under random measurement models. For

instance, anchored regression [3] and PhaseMax [39] formulate phase retrieval given an initial estimate as a linear program. RobustPhaseMax [46] modifies these methods to offer robust estimation by introducing auxiliary variables to describe the outliers. In another example, reshaped wirtinger flow (RWF) [105] and Amplitude Flow [97] follow a generalized gradient descent approach for a least squares estimator (LSE). Median-RWF [104] is a variant of these methods tailored to robust phase retrieval. Specifically, Median-RWF uses a truncation type of regularization that identifies and excludes outliers in each iteration by median-based thresholding on the consistency of the current estimate to the measurements. Median-RWF significantly improves the empirical performance of RobustPhaseMax by tolerating a higher fraction of outliers. However, the regularization of Median-RWF involves algorithmic parameters that have been tuned specifically for the Gaussian measurement model. It has not been discussed how to generalize the tuning parameters to other measurement models.

A recent work proposed an approach to robust phase retrieval in the classical robust regression framework in statistics [31]. Instead of the least squares, they adopted the *least absolute deviation* (LAD) [10] to enforce the consistency to the squared amplitude measurements with outliers (i.e. $y_i = |\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle|^2$ for $i \in I_{\text{in}}$). The parameter estimation is then cast as a nonconvex optimization problem. They proposed a prox-linear method that updates the estimate iteratively through local linearization of the forward model. This algorithm can be viewed as a variant of the Gauss-Newton method that regularizes the updates with the proximity to the previous iterate. The prox-linear algorithm iteratively refines the estimate through a sequence of quadratic programs. Importantly, the prox-linear method provides comparable performance to Median-RWF without involving any tuning parameter. Furthermore, to accelerate

iterative methods for large-scale applications such as astronomical or medical imaging, they adopted the proximal operator graph splitting (POGS) solver.

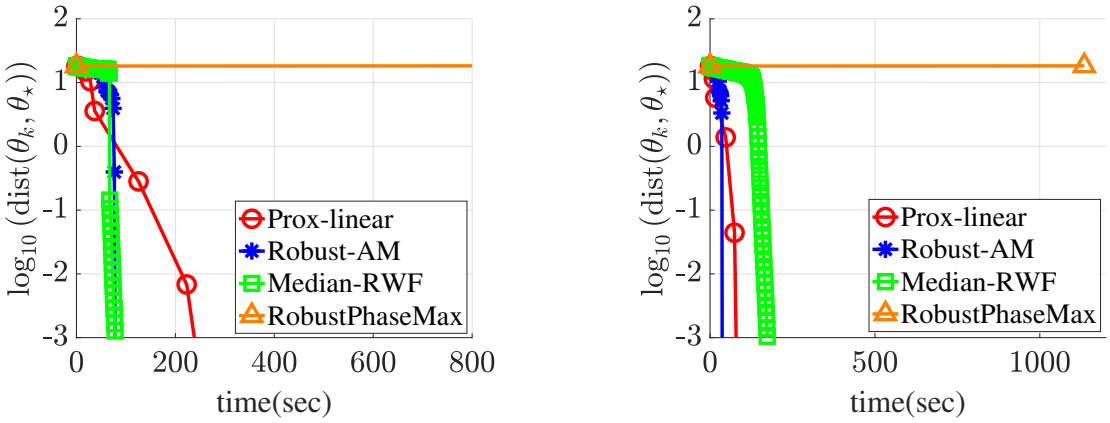
In this chapter, we propose a novel optimization approach to robust phase retrieval that shares strong theoretical guarantees (high tolerance of outlier ratio and no tuning parameters) with the prox-linear algorithm and further improves its computational cost. The objective is achieved by a simple unconstrained Gauss-Newton method for LAD under the amplitude measurements in (1.1.2). The resulting algorithm is equivalent to an alternating minimization for LAD. Since LAD is robust in the presence of outliers, we refer to this optimization as *Robust-AM*. Our main theoretical result demonstrates that a suitably initialized Robust-AM linearly converges to the ground-truth signal from $n \gtrsim d$ random amplitude-only measurements including up to 25% outliers. The desired initialization can be obtained by the existing robust spectral estimators [31, 104].

Table 1.1: Comparison of RobustPhaseMax [46], Median-RWF [104], Prox-linear [31], and Robust-AM for robust phase retrieval in terms of computational cost to obtain an ϵ-accurate solution and sparse noise assumptions for the performance guarantees.

Method	Computational cost	Algorithm type	Measurement	Support model	Sparsity
RobustPhaseMax	$\mathcal{O}(n^3 + (n+d)^2 \log(1/\epsilon))$ [99] $\widetilde{\mathcal{O}}((n+d)^{2.38} \log(1/\epsilon))$ [87]	Linear program	Amplitude	Adversarial	Unspecified
Median-RWF	$\mathcal{O}(nd \log(1/\epsilon))$	Truncated gradient descent	Amplitude	Arbitrary fixed	Unspecified
Prox-linear	$\mathcal{O}\left(nd \log\log(1/\epsilon)(d + \log(1/\epsilon))\right)$[1]	Regularized Gauss-Newton	Squared	Arbitrary fixed	1/4
Robust-AM (Theorem 4)	$\mathcal{O}\left(n^3 + (n+d)^2 \log^2(1/\epsilon)\right)$ [99] $\widetilde{\mathcal{O}}\left((n+d)^{2.38} \log^2(1/\epsilon)\right)$ [87]	**Unconstrained Gauss-Newton**	Amplitude	Arbitrary fixed	1/4

[1]We establish this computational cost under the assumption that POGS linearly converges to the solution for the inner optimization of prox-linear. However, to the

best of our knowledge, the convergence rate of POGS has not been shown. Thus, this computational cost is a conjecture.



(a) zero (b) Cauchy distribution

Figure 1.1: Convergence of Robust-AM by ADMM [14], prox-linear by POGS, Median-RWF, and RobustPhaseMax in run time ($d = 1,000$, $n = 10,000$, and $\eta = 0.3$).

We explicitly compare Robust-AM to the aforementioned methods providing a performance guarantee, as summarized in Table 1.1. These methods consider their own optimization approaches to robust phase retrieval. RobustPhaseMax introduced explicit variables for sparse noise to the original constrained optimization [3, 39]. Median-RWF employed a truncation by median to convert gradient descent to minimize the ℓ_2 fidelity into a robust algorithm. Robust-AM is most similar to that of the prox-linear which considered solving LAD by a regularized Gauss-Newton method. By adopting the amplitude measurement model without a regularizer, unlike the squared measurement model with a regularizer used in the prox-linear method, Robust-AM admits a computationally efficient ADMM algorithm that runs faster than POGS, as shown in Figure 1.1. In this experiment, the fraction of outliers $\eta := |I_{\mathrm{out}}|/n$ is

set to 0.3 where $|I_{\text{out}}|$ represents the cardinality of the set I_{out}. Outlier entries are either set to zero or generated following the Cauchy distribution with median 0 and mean-absolute-deviation 1. The convergence is measured by the metric $\text{dist}(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) := \min_{\alpha\in\{\pm 1\}} \|\boldsymbol{\theta} - \alpha\boldsymbol{\theta}_\star\|_2$ for $\boldsymbol{\theta}, \boldsymbol{\theta}_\star \in \mathbb{R}^d$. Figure 1.1 shows that Robust-AM, without any explicit control over the proximity to previous iterates, converges to the ground truth signal $\boldsymbol{\theta}_\star$ without overshooting. More importantly, Robust-AM empirically outperforms the existing methods for robust phase retrieval. We will verify through comprehensive numerical simulations that Robust-AM can tolerate a higher fraction of outliers and provide exact recovery with fewer observations.

1.2 Max-Affine regression

The *max-affine* model combines k affine models in the form of

$$y = \max_{j\in[k]} \left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j^\star\rangle + b_j^\star\right) \tag{1.2.1}$$

to produce a piecewise-linear mutivariate functions, where $\boldsymbol{x}$ and y respectively denote the covariate and the response, and $[k]$ denotes the set $\{1, \ldots, k\}$. The max-affine model frequently appears in applications across statistics, machine learning, economics, and signal processing. Specifically, the max-affine model has been used for simple auction problems [69, 77] and multiclass classification problems. In the multiclass SVM formulation [23, 26], the hypothesis takes the form:

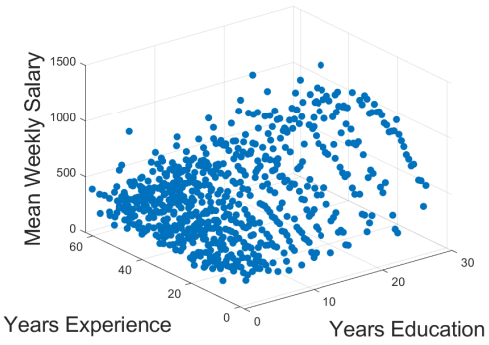
$$H_{\{(\boldsymbol{\theta}_j, b_j)\}_{j=1}^k}(\boldsymbol{x}) = \arg\max_{j\in[k]} \left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j\rangle + b_j\right), \tag{1.2.2}$$

where $\langle \boldsymbol{x}, \boldsymbol{\theta}_j\rangle + b_j$ is referred to as the *confidence* or *similarity* score for the j-th class. The hypothesis in (1.2.2) generalizes the linear separator (halfspaces) used in binary classifiers, which take the form $H_{\boldsymbol{\theta},b}(\boldsymbol{x}) = \text{sign}\left(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle + b\right)$. According to the

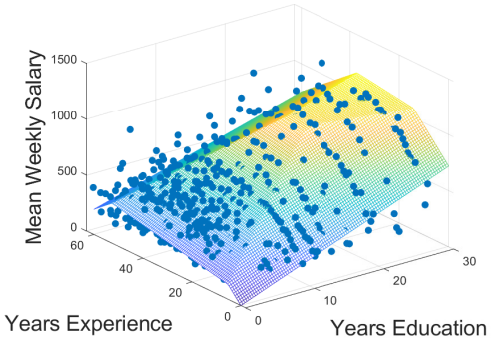
framework (1.2.2), the predicted label is determined by the class with the highest similarity score with $\boldsymbol{x}$. Compared to learning a max-affine function in (1.2.1), which involves learning k separate linear functions and determining the response by taking their maximum, the multiclass classifier in (1.2.2) learns k linear functions for the classes and selects the class with the highest score using arg max.

Moreover, it has been adopted as a predictive model for shape-restricted data, which often arises in utility functions [1, 91, 92]. The piecewise linear approximation simplify the relationship in these types of data in an interpretable way [7, 47, 79]. Figure 1.2 visualizes fitted regression functions estimated by the max-affine model for the mean weekly wages dataset [76] and the Boston housing dataset [48], both of which are examples of shape-restricted data. [1]

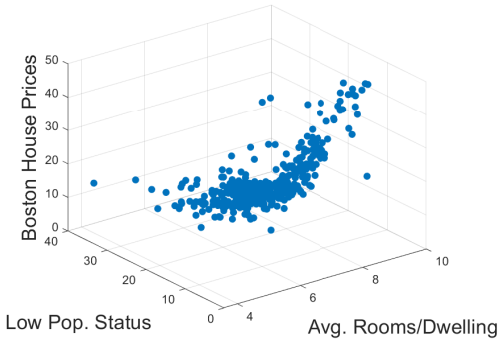
[1]For shape-restricted data with concavity, we can alternatively use the min-affine function by substituting the max function in (1.2.1) with the min function. However, since the results for the max-affine function apply directly to the min-affine function, we only study the max-affine model in this thesis.



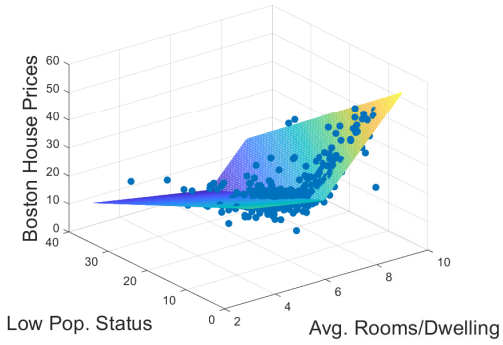
Mean weekly wages data



Max-affine model fitted for mean weekly wages data



Boston housing data



Max-affine model fitted for boston housing data

Figure 1.2: Visualizations of max-affine fitted regression functions for shape-restricted datasets.

We consider a regression of the max-affine model in (1.2.1) via least squares

$$\min_{\{\boldsymbol{\theta}_j, b_j\}_{j=1}^k} \frac{1}{2n} \sum_{i=1}^n \left(y_i - \max_{j \in [k]} (\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j \rangle + b_j) \right)^2 \tag{1.2.3}$$

from statistical observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ potentially corrupted with noise.

A suite of numerical methods has been proposed to solve the nonconvex optimization in (1.2.3) (e.g., [7, 47, 65, 86]). The fact that (1.2.1) is a special case of piecewise linear function allows us to divide $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ into k partitions based on their

membership in the polyhedral cones

$$\mathcal{C}_j := \{\boldsymbol{w} \in \mathbb{R}^d \ : \ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j - \boldsymbol{\beta}_l \rangle > 0,\ \forall l < j,\ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j - \boldsymbol{\beta}_l \rangle \geq 0,\ \forall l > j\}. \tag{1.2.4}$$

The set $\mathcal{C}_j$ contains all inputs maximizing the jth linear model.[2] Note that each $\mathcal{C}_j$ is determined by $k-1$ half spaces given by the pairwise difference of the jth linear model and the others. If this oracle information is known a priori, then the estimation is divided into k decoupled linear least squares given by

$$(\widehat{\boldsymbol{\theta}}_j, \widehat{b}_j) = \underset{\{(\boldsymbol{\theta}_j, b_j)\}_{j=1}^k}{\operatorname{argmin}} \sum_{i \in \mathcal{C}_j} \left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j \rangle + b_j - y_i\right)^2, \quad j \in [k]. \tag{1.2.5}$$

However, since the oracle partition information is not available in practice, various adaptive partitioning methods have been studied. The *least-squares partition algorithm* [65] iteratively refines the parameter estimate by alternating between the partition and the least-squares steps when the number of affine models k is known a priori. The partitioning step classifies the inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with respect to the maximizing affine models given estimated model parameters. The least-squares step updates the parameters for each affine model by using the corresponding observations. Later variations of the alternating minimization algorithm used an adaptive search for unknown k [7, 47]. The consistency of these estimators has been derived. In more recent works, Ghosh et al. [36–38] established finite-sample analysis of the *alternating minimization* (AM) estimator [65] for the special case when the observations are generated from a ground-truth model. One can interpret their analysis through the lens of the popular *teacher-student framework* [64]. This framework has been widely

[2]In case of a tie when multiple linear models attain the maximum for a given sample, we assign the sample to the smallest maximizing index. Since the event of duplicate maximizing indices will happen with probability 0 for any absolutely continuous probability measure on $\boldsymbol{x}_i$s, the choice of a tie-break rule does not affect the analysis.

adopted in statistical mechanics [33, 64] and machine learning [40, 52, 106, 107]. It provides a theoretical understanding of how a specific model is trained and generalized through a ground-truth generative model [52]. In this framework, a max-affine model (student) is trained by data generated from a ground-truth max-affine model (teacher) from k fixed affine models. By using the provided data, the student model recovers parameters that produce the ground-truth model via AM. Since the max affine model is invariant under the permutation of the component affine models, the minimizer to (1.2.3) is determined only up to the corresponding equivalence class. Ghosh et al. [38] established a finite-sample analysis of AM under the standard Gaussian covariate assumption with independent stochastic noise. They showed that a suitably initialized alternating minimization converges linearly to a consistent estimate of the ground-truth parameters along with a non-asymptotic error bound. Moreover, they proposed and analyzed a spectral method that provides the desired initialization. They also further extended the theory to a generalized scenario with relaxed assumptions on the covariate model [36, 37].

Related Work

Relation to phase retrieval and ReLU regression: The max-affine model includes well-known models in signal processing and machine learning as special cases. The instance of (1.2.1) for $k = 2$ with $b_1^\star = b_2^\star = 0$ and $\boldsymbol{\theta}_1^\star = -\boldsymbol{\theta}_2^\star = \boldsymbol{\theta}^\star$ reduces to $y = |\langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle|$, which corresponds to a measurement model in phase retrieval. Similarly, the rectified linear unit (ReLU) $y = \max(\langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle, 0)$ is written in the form of (1.2.1) for $k = 2$ with $\boldsymbol{\theta}_1^\star = \mathbf{0}$ and $\boldsymbol{\theta}_2^\star = \boldsymbol{\theta}^\star$. A series of studies in [59, 82, 84, 85, 90, 98, 103, 105] has developed a statistical analysis of GD and SGD for phase retrieval and ReLU regression. It has been shown that for the noiseless case, GD and SGD converge

linearly to a near-optimal estimate of the ground-truth parameters when the number of observations grows linearly with the ambient dimension d. In the context of bounded noise, GD converges to the ground truth within a radius determined by the noise level [98, 105]. However, it remained an open question whether GD is consistent under stochastic noise assumptions. Additionally, SGD in the presence of noise has not been thoroughly investigated yet. The main results of this chapter address these questions on phase retrieval as a special case of max-affine regression.

Relation to convex regression: The max-affine model has also been adopted in parametric approaches to convex regression [5–7, 45, 47, 65, 79–81]. Let $f_\star : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary convex function. The observations are given by $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where $y_i = f_\star(\boldsymbol{x}_i)$ for all i in $[n]$. The nonparametric convex regression problem aims to estimate $f_\star$ by solving

$$\min_{f \in \mathcal{F}_{\text{cvx}}} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2, \tag{1.2.6}$$

where $\mathcal{F}_{\text{cvx}}$ denotes the set of convex functions. Since f exists in the space of continuous real-valued functions on $\mathbb{R}^d$, the optimization problem in (1.2.6) is infinite-dimensional. A line of research [6, 13, 81] investigated the interpolation approach with a max-affine model in the form of

$$\widehat{f}(\boldsymbol{x}) = \max_{i \in [n]} \left(y_i + \boldsymbol{g}_i^{\mathsf{T}} (\boldsymbol{x} - \boldsymbol{x}_i) \right). \tag{1.2.7}$$

It provides a perfect interpolation of data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with zero training error. For example, the interpolation is achieved by choosing $\boldsymbol{g}_i \in \partial f_\star(\boldsymbol{x}_i)$ for all $i \in [n]$. It has been show that the least squares estimator provides near-optimal generalization bounds relative to a matching minimax bound [7, 42, 45, 61, 62]. However, the minimax bound for the parametric model in (1.2.7) decays slowly due to the curse of dimensionality

for a set of max affine with n segments. The least squares for the model in (1.2.7) is formulated as a quadratic program (QP) [13, Section 6.5.5]. However, off-the-shelf interior-point methods do not scale to large instances of this QP due to the high computational cost $O(d^4n^5)$ [47, 65].

The k-max-affine model in (1.2.1) is considered as an alternative compact parametrization to approximate convex regression. The worst-case error in approximating d-variate Lipschtiz convex functions on a bounded domain by a k-max-affine model decays as $O(k^{-2/d})$ [7, Lemma 5.2]. However, data in practical applications such as aircraft wing design, wage prediction, and pricing stock options are often well approximated by the k-max-affine model with small k (e.g., [47, Section 6], [7, Section 7]). Unlike the interpolation approach to convex regression, if the compact model fits data in applications, the estimation error decays much faster in n.

We will examine two different noise scenarios. The first scenario is deterministic noise, where we make no assumptions regarding any stochastic distribution of the noise. Under this scenario, we focus on the max-linear model, a special case of the max-affine model obtained by removing the bias term $\{b_j^\star\}_{j=1}^k$ in (1.2.1). Subsequently, we study the max-affine model under sub-Gaussian noise.

1.2.1 Max-linear regression under the deterministic noise

We consider the problem of estimating the parameters $\boldsymbol{\theta}_{\star,1}, \ldots, \boldsymbol{\theta}_{\star,k} \in \mathbb{R}^d$ that determine the *max-linear* function

$$\boldsymbol{x} \in \mathbb{R}^d \mapsto \max_{j \in [k]} \langle \boldsymbol{\theta}_{\star,j}, \boldsymbol{x} \rangle , \tag{1.2.8}$$

from independent and identically distributed (i.i.d.) observations, where $[k]$ denotes the set $\{1, \ldots, k\}$. Specifically, given the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, and denoting

the value of a max-linear function, with parameter $\boldsymbol{\theta}$, at these points by

$$f_i(\boldsymbol{\theta}) := \max_{j\in[k]} \langle \boldsymbol{x}_i, \boldsymbol{\theta}_j \rangle \,, \tag{1.2.9}$$

we observe the nonlinear observation

$$y_i = f_i(\boldsymbol{\theta}_\star) + z_i \,,$$

of the parameter vector $\boldsymbol{\theta}_\star = [\boldsymbol{\theta}_{\star,1}; \,\ldots\, ; \boldsymbol{\theta}_{\star,k}] \in \mathbb{R}^{kd}$ where z_i denotes noise for $i \in [n]$.

The most relevant prior work studied an *alternating minimization* (AM) algorithm to solve a slightly more general problem of max-affine regression [38]. Each iteration consists of a step to identify the maximizing linear models followed by least-squares update of model parameters. However, we observed that their empirical performance significantly degrades with outliers, mainly due to the sensitivity of the "maximizer identification" step. Leveraging recent theory for convexifying nonlinear inverse problems in the original domain [2–4], we propose an alternative approach by convex programming. Due to the inherent geometry of the formulation, the convex estimator provides stable performance in the presence of adversarial noise. It is worth mentioning that Ghosh et al. [38] considered a random noise model, whereas we consider a deterministic "gross error" model. Nevertheless, in the noiseless case, both results achieve exact parameter recovery at comparable sample complexities.

Convex estimator

The common estimators for $\boldsymbol{\theta}_\star$ such as the *least absolute deviation* (LAD), i.e.,

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \;\; \frac{1}{n}\sum_{i=1}^{n} |f_i(\boldsymbol{\theta}) - y_i| \,, \tag{1.2.10}$$

are generally hard to compute as they involve nonconvex optimization. Given an "anchor vector" $\boldsymbol{a}$, we study the estimation of $\boldsymbol{\beta}_\star$ through *anchored regression* (AR)

that formulates the estimation by the convex program

$$\begin{aligned} \underset{\boldsymbol{\theta}}{\text{maximize}} \quad & \langle \boldsymbol{a}, \boldsymbol{\theta} \rangle \\ \text{subject to} \quad & \frac{1}{n} \sum_{i=1}^{n} (f_i(\boldsymbol{\theta}) - y_i)_+ \leq \eta, \end{aligned} \tag{1.2.11}$$

where $(\cdot)_+$ denotes the positive-part function. The parameter η should be chosen so that the feasible set of (1.2.11) is not empty. The anchored regression can be interpreted as a convexification of the LAD estimator. Since the observation functions (1.2.9) are convex, the LAD is nonconvex mainly due to the effect of the absolute value operator in (1.2.10). This source of nonconvexity is removed in anchored regression by relaxing the absolute deviation to the positive part of the error. The linear objective that is determined by the anchor vector $\boldsymbol{a}$ acts as a "regularizer" to prevent degenerate solutions and guarantees exact recovery of the true parameter $\boldsymbol{\theta}_\star$ under certain conditions on the measurement model in the noiseless scenario.

Anchored regression has been originally developed as a scalable convex program to solve the phase retrieval problem [3,39] with provable guarantees. Anchored regression is highly scalable compared to other convex relaxations in this context [18,95] that rely on semidefinite programming. The idea of anchored regression is further studied in a broader class of nonlinear parametric regression problems with convex observations [4] and *difference of convex* functions [2].

Contributions

We provide a scalable convex estimator for the max-linear regression problem that is formulated as a linear program and is backed by statistical guarantees. Under the standard Gaussian covariate model, the convex estimator (1.2.11) is guaranteed to recover the regression parameters exactly with high probability if the number of

observations n scales as $\pi_{\min}^{-4}d$ up to some logarithmic factors where $\pi_{\min}$ is defined as $\min_{j\in[k]} \mathbb{P}\left(\boldsymbol{g} \in \mathcal{C}_j\right)$ for $\boldsymbol{g} \in \text{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$. This sample complexity implicitly depends on k (i.e., the number of components) through $\pi_{\min}$. Particularly, when the k linear components form a "well-balanced partition" in the sense that they are equally likely to achieve the maximum, the smallest probability $\pi_{\min}$ is close to $1/k$ and the derived sample complexity reduces to $k^4 d$ up to the logarithmic factors. This is comparable to the sufficient condition for exact recovery $n = \mathcal{O}(kd\pi_{\min}^{-3})$ of alternating minimization algorithm [38] in the noise-free scenario. Monte Carlo simulations show that our proposed convex estimator, as a convexification of the LAD estimator, exhibits robustness against outliers, whereas AM appears to be fragile in the presence of impulsive noise. Furthermore, the repetition of AR significantly improves the accuracy of the estimation.

1.2.2 Max-affine regression under the subGaussian noise

We consider the max-affine regression problem, where the observations are given by

$$y_i = \max_{j\in[k]} \left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j^\star \rangle + b_j^\star\right) + z_i, \quad i = 1, \ldots, n,$$

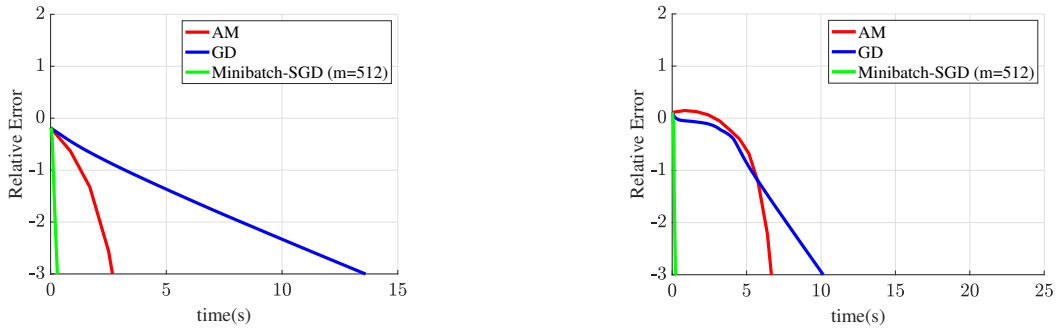
where we assume that $\{z_i\}_{i=1}^n$ are independently generated from a σ-sub-Gaussian distribution.

We present theoretical and numerical results on max-affine regression by first-order methods including *gradient descent* (GD) and *stochastic gradient descent* (SGD). The first-order methods have been widely used to solve various nonlinear least squares problems in machine learning [34, 41, 58, 83]. We observe that GD and SGD also perform competitively on max-affine regression compared to AM. In particular, SGD

converges significantly faster (in run time) than AM in a noise-free scenario. Figure 1.3 compares AM, GD, and a mini-batch SGD on random 50 trials of max-affine regression where the ground-truth parameter vectors $\{\boldsymbol{\beta}_j^\star\}_{j=1}^k$ are selected randomly from the unit sphere. Covariates are independently generated from either Normal$(\mathbf{0}, \boldsymbol{I}_{500})$ or Unif$[-\sqrt{3}, \sqrt{3}]^{\otimes 500}$. We plot the median of relative errors versus the average run time where the relative error is calculated as

$$\min_{\pi \in \mathrm{Perm}([k])} \log_{10} \left(\sum_{j=1}^k \|\widehat{\boldsymbol{\beta}}_{\pi(j)} - \boldsymbol{\beta}_j^\star\|_2^2 / \sum_{j=1}^k \|\boldsymbol{\beta}_j^\star\|_2^2 \right)$$

with Perm$([k])$ and $\{\widehat{\boldsymbol{\beta}}_j\}_{j=1}^k$ denoting the set of all possible permutations over $[k]$ and the estimated parameters, respectively. Our main result provides a theoretical analysis of SGD that explains this empirical observation.



(a) Gaussian covariates (b) Uniform covariates

Figure 1.3: Convergence of estimators for noise-free max-affine regression ($k = 3$, $d = 500$, and $n = 8,000$).

Main results

We derive convergence analyses of GD and mini-batch SGD under the same covariate and noise assumptions in the previous work on AM by Ghosh et al. [37].

They assumed that covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent copies of a random vector $\boldsymbol{x}$ that satisfies the sub-Gaussianity and anti-concentration defined below.

Assumption 1 (Sub-Gaussianity) *The covariate distribution satisfies*

$$\|\langle \boldsymbol{v}, \boldsymbol{x} \rangle\|_{\psi_2} \leq \eta, \quad \forall \boldsymbol{v} \in \mathbb{S}^{d-1},$$

where $\|\cdot\|_{\psi_2}$ *and* $\mathbb{S}^{d-1}$ *denote the sub-Gaussian norm (i.e., see [93, Equation 2.13]) and the unit sphere in* ℓ_2^d*, respectively.*

Assumption 2 (Anti-concentration) *The covariate distribution satisfies*

$$\sup_{w \in \mathbb{R}, \boldsymbol{v} \in \mathbb{S}^{d-1}} \mathbb{P}((\langle \boldsymbol{v}, \boldsymbol{x} \rangle + w)^2 \leq \epsilon) \leq (\gamma \epsilon)^{\zeta}, \quad \forall \epsilon > 0.$$

The class of covariate distributions by Assumptions 1 and 2 generalizes the standard independent and identically distributed Gaussian distribution. For example, the uniform and beta distributions satisfy Assumptions 1 and 2. Therefore, the theoretical result under this relaxed covariate model will apply to a wider range of applications. They also assumed that observations are corrupted with independent additive σ-sub-Gaussian noise.

This work establishes the first theoretical analysis of GD and mini-batch SGD for max-affine regression. The following pseudo-theorem demonstrates that GD shows a local linear convergence under the above assumptions.

Theorem 1 (Informal) *Let* $\boldsymbol{\beta}^\star \in \mathbb{R}^{k(d+1)}$ *denote the column vector that collects all ground-truth parameters* $(\boldsymbol{\theta}_j^\star, b_j^\star)_{j \in [k]}$*. Given* $\widetilde{O}(C_{\boldsymbol{\beta}^\star} kd(k^3 \vee \sigma^2))$ *observations, a suitably initialized GD for max-affine regression converges linearly to an estimate of* $\boldsymbol{\beta}^\star$ *with* ℓ_2*-error scaling as* $\widetilde{O}(\sigma k^2 \sqrt{d/n})$*, where* $C_{\boldsymbol{\beta}^\star}$ *is a constant that implicitly depends on* k *through* $\boldsymbol{\beta}^\star$ *but is independent of* d*.*

The error bound by this theorem improves upon the best-known result on max-affine regression achieved by AM [37, Theorem 2]. The error bound for AM is larger by a factor that grows at least as $k^{-1+2\zeta^{-1}}$. We also present an analogous analysis for SGD. A specification for the noise-free observation scenario is stated as follows.

Theorem 2 (Informal) *A suitably initialized mini-batch SGD for max-affine regression with $\widetilde{O}(C_{\boldsymbol{\beta}^\star} k^9 d)$ noise-free observations converges linearly to the ground truth $\boldsymbol{\beta}^\star$ for any batch size.*

The per-iteration cost of a mini-batch SGD with batch size m is $O(kmd)$, which is significantly lower than those for GD $O(knd)$ and of AM $O(knd^2)$. This implies the faster convergence of SGD in run time shown in Figure 1.3. We also observe that SGD empirically recovers the ground-truth parameters from fewer observations (see Figures 3.2 and 3.4).

1.2.3 Notation and Organization of the Thesis

Boldface lowercase letters denote column vectors (e.g., $\boldsymbol{a}$), and boldface capital letters denote matrices (e.g., $\boldsymbol{A}$). The concatenation of two column vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by $[\boldsymbol{a};\ \boldsymbol{b}]$. The subvector of $\boldsymbol{a} \in \mathbb{R}^{d+1}$ containing the first d entries is denoted by $(\boldsymbol{a})_{1:d}$. Various norms are used throughout this thesis. We use $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_{\mathrm{F}}$, and $\|\cdot\|_{\psi_2}$ to denote the ℓ_1 norm, Euclidean norm, Frobenius norm, and sub-Gaussian norm, respectively, while the spectral norm of a matrix is denoted by $\|\cdot\|$. The d-dimensional Euclidean unit ball is denoted by B_2^d, and the unit sphere in d dimensions is denoted by $\mathbb{S}^{d-1}$. We use big-O notation to describe asymptotic bounds. For two scalars q and p, we write $q \lesssim p$ or $q = \mathcal{O}(p)$ if there exists an absolute constant $C > 0$ such that $q \leq Cp$. The notation $\widetilde{\mathcal{O}}$ is used to ignore logarithmic factors.

Absolute constants that may vary from line to line are denoted by $C, C_1, C_2, \ldots$ and $c, c_1, c_2, \ldots$. For brevity, the shorthand notation $[n]$ denotes the set $\{1, \ldots, n\}$ for $n \in \mathbb{N}$. Additionally, $a \vee b$ and $a \wedge b$ denote $\max(a, b)$ and $\min(a, b)$ for $a, b \in \mathbb{R}$.

We organize the rest of the thesis as follows. In Section 1.2.3, we present the robust alternating minimization algorithm for the robust phase retrieval problem and provide theoretical results regarding its convergence and sample complexity. In Chapter 2, we establish the theoretical guarantees of the convex program (1.2.11) for max-linear regression under deterministic noise. In Chapter 3, we analyze the convergence guarantees and sample complexity of first-order methods for max-affine regression under sub-Gaussian noise.

Each of these chapters includes a comprehensive discussion of the algorithms in terms of computational complexity and hyperparameter tuning (e.g., η in the convex problem (1.2.11) and the step size in first-order methods). Furthermore, we show that numerical experiments corroborate our theoretical findings and that the proposed algorithms are comparable to or outperform existing methods in terms of computational efficiency and sample efficiency. We conclude the thesis by discussing future research directions and ongoing work in Chapter 4. The proofs for the theoretical results presented in this thesis are provided in Appendices.

In this chapter, we present the results for alternating minimization approach for the robust phase retrieval problem, with the problem formulation described in Section 1.1.

1.3 Robust Alternating Minimization

We consider the minimization of the composite function $\ell = h \circ F$ where $h : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $F : \mathbb{R}^d \to \mathbb{R}^n$ is a nonlinear mapping. In the special case when F is differentiable, Burke and Ferris [16] proposed a constrained Gauss-Newton method where the amount of the update is upper-bounded by a threshold. Duchi and Ruan [31] considered a variant where the constraint on the proximity on consecutive iterates is substituted by regularization with an additive penalty. We consider a more challenging case where F is non-differentiable and propose an unconstrained Gauss-Newton method where the variable sequence $(\boldsymbol{\theta}_k)_{k \in \mathbb{N} \cup \{0\}}$ is iteratively updated by

$$\boldsymbol{\theta}_{k+1} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \; h(F(\boldsymbol{\theta}_k) + F'(\boldsymbol{\theta}_k)(\boldsymbol{\theta} - \boldsymbol{\theta}_k)) \tag{1.3.1}$$

where $F'(\boldsymbol{\theta}_k) \in \mathbb{R}^{n \times d}$ denotes the Clarke's generalized Jacobian matrix at $\boldsymbol{\theta}_k$ [21]. Due to the local linear approximation of F at $\boldsymbol{\theta}_k$ in (1.3.1), $\boldsymbol{\theta}_{k+1}$ is obtained as a solution to a convex program. In a special case where $h : \mathbb{R}^n \to \mathbb{R}$ and $F : \mathbb{R}^d \to \mathbb{R}^n$ are respectively given by

$$h(\boldsymbol{z}) = \|\boldsymbol{z}\|_1 \tag{1.3.2}$$

and

$$F(\boldsymbol{\theta}) = (|\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle| - b_i)_{i=1}^n \,, \tag{1.3.3}$$

their composition reduces to

$$\ell(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} ||\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle| - b_i| \,. \tag{1.3.4}$$

Then the minimization of ℓ corresponds to the LAD approach to robust phase retrieval with the amplitude measurement model. Furthermore, given h and F as in (1.3.2) and (1.3.3), the update rule in (1.3.1) is explicitly written as

$$\boldsymbol{\theta}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} \left|\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \operatorname{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k\rangle)\cdot b_i\right|. \tag{1.3.5}$$

The resulting algorithm (1.3.5), derived from an unconstrained Gauss-Newton method of robust phase retrieval, is equivalent to an alternating minimization to the LAD formulation of robust phase retrieval when noisy measurements with a negative sign are discarded. The alternating minimization iteratively updates two variables $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\phi} := (\phi_1, \ldots, \phi_n) \in \{\pm 1\}^n$ to recover the ground-truth $\boldsymbol{\theta}_\star$ and true phase $\operatorname{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star\rangle)$, respectively, by alternatively solving the following optimization:

$$\phi_i^k = \operatorname*{argmin}_{\phi_i\in\{\pm 1\}} \left|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k\rangle - \phi_i \cdot b_i\right|, \quad \forall i \in [n], \tag{1.3.6a}$$

$$\boldsymbol{\theta}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} \left|\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \phi_i^k \cdot b_i\right|, \tag{1.3.6b}$$

where k denotes the iteration index. Since $b_i \geq 0$, (1.3.6a) yields the closed form expression $\phi_i^k = \operatorname{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k\rangle)$ for all $i \in [n]$. Therefore, plugging this into (1.3.6b) results in (1.3.5). An analogous alternating minimization for least-squares phase retrieval has been studied in the literature [35, 72, 94].

Remark 3 *Besides the magnitude loss in* (1.3.4)*, there exist more sophisticated robust loss functions such as Huber [53], Cauchy [66], Welsch [27], and HOW [100] (see [8] for more examples). Despite its simplest form, Robust-AM derived with the magnitude loss empirically performed significantly better than competing methods providing a performance guarantee. It would be interesting to study whether one can further improve empirical and theoretical performances with other loss functions.*

1.4 Optimization Algorithms

This section discusses numerical algorithms for Robust-AM. First, we note that the optimization in (1.3.5) is equivalent to a linear program

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\mathbb{R}^d,(t_i)_{i=1}^n}{\text{minimize}} \ & \langle \boldsymbol{t}, \mathbf{1}_n \rangle \\ \text{subject to } & t_i \geq \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle - \text{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle) \cdot b_i, \\ & t_i \geq -\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle + \text{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle) \cdot b_i, \quad \forall i \in [n] \end{aligned} \tag{1.4.1}$$

where $\mathbf{1}_n = [1, \ldots, 1]^\mathsf{T} \in \mathbb{R}^n$. There exist various computationally efficient numerical methods to solve linear programs. For example, the derandomized algorithm by van den Brand [87] finds an exact solution to a linear program with d variables and n constraints at the cost of $\widetilde{\mathcal{O}}\left((n+d)^c\right)$ multiplications where $c \approx 2.38$.

To further accelerate the convergence of Robust-AM, we also adopt iterative numerical algorithms that provide an approximate solution to the inner optimization in (1.3.5). In particular, we consider two alternating direction methods of multipliers (ADMM) algorithms for inner optimization. We refer to the Robust-AM with approximate solutions to the inner optimization by these ADMM algorithms as *fast Robust-AM* since they provide a significantly lower computational cost for the entire convergence of Robust-AM to an ϵ-accurate estimate.

1.4.1 ADMM for LAD

Given $\boldsymbol{\theta}_k$, the optimization in (1.3.5) is viewed as LAD for linear regression and one can use an ADMM algorithm for LAD [14, Chapter 6.1]. To describe the update rule of the ADMM algorithm, we introduce shorthand notations for the sake of brevity. Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be a matrix whose i-th row is $\boldsymbol{x}_i^\mathsf{T}$ for $i \in [n]$, $\boldsymbol{b} := (b_1, \ldots, b_n) \in \mathbb{R}^n$, and $\boldsymbol{\Lambda}_k = \text{diag}(\text{sign}(\langle \boldsymbol{x}_1, \boldsymbol{\theta}_k \rangle), \ldots, \text{sign}(\langle \boldsymbol{x}_n, \boldsymbol{\theta}_k \rangle))$. By introducing an auxiliary variable

$\boldsymbol{y} \in \mathbb{R}^n$, (1.3.5) is equivalently rewritten as

$$\begin{array}{ll} \underset{\boldsymbol{\theta}\in\mathbb{R}^d, \boldsymbol{y}\in\mathbb{R}^n}{\text{minimize}} & \|\boldsymbol{y} - \boldsymbol{\Lambda}_k \boldsymbol{b}\|_1 \\ \text{subject to} & \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}. \end{array} \tag{1.4.2}$$

The augmented Lagrangian function of (1.4.2) is written as

$$\mathcal{L}_\rho(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{\phi}) = \|\boldsymbol{y} - \boldsymbol{\Lambda}_k \boldsymbol{b}\|_1 + \boldsymbol{\phi}^\intercal (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}) + \frac{\rho}{2}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2,$$

from which the update rules are derived as follows:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{X}^+ \left(\boldsymbol{y}^t - \frac{\boldsymbol{\phi}^t}{\rho} \right), \tag{1.4.3a}$$

$$\begin{aligned} \boldsymbol{y}^{t+1} &= \boldsymbol{\Lambda}_k \boldsymbol{b} \\ &+ \operatorname{sign}\left(\boldsymbol{X}\boldsymbol{\theta}^t + \frac{\boldsymbol{\phi}^t}{\rho} - \boldsymbol{\Lambda}_k \boldsymbol{b} \right) \odot \left[\left| \boldsymbol{X}\boldsymbol{\theta}^t + \frac{\boldsymbol{\phi}^t}{\rho} - \boldsymbol{\Lambda}_k \boldsymbol{b} \right| - \frac{1}{\rho} \right]_+, \end{aligned} \tag{1.4.3b}$$

$$\boldsymbol{\phi}^{t+1} = \boldsymbol{\phi}^t + \rho(\boldsymbol{X}\boldsymbol{\theta}^{t+1} - \boldsymbol{y}^{t+1}), \tag{1.4.3c}$$

where $t, k \in \{0\} \cup \mathbb{N}$ denote the indices respectively for the inner iteration in (1.4.3) and the outer iteration (1.3.5), $\odot$ denotes the Hadamard product, and $[\cdot]_+$ takes the positive part of each entry of the input vector. The most expensive step in (1.4.3) is the least squares problem in (1.4.3a). Since it repeats with the same $\boldsymbol{X}$, the pseudo inverse $\boldsymbol{X}^+$ of $\boldsymbol{X}$ can be pre-computed as $\boldsymbol{X}^+ = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$ with cost $\mathcal{O}(d^3 + d^2 n)$ and be used on memory over iterations. For faster convergence, we adopt the varying step size strategy for ρ [14, Section 3.4.1]. The prox-linear with the POGS algorithm [31, Section 5] involves a similar matrix inversion. However, since their matrix evolves over the outer iteration, unlike the fast Robust-AM with ADMM, it is necessary for POGS to repeat the matrix inversion. Recall that we wanted to adopt ADMM for the inner iteration of Robust-AM to accelerate the convergence with approximate solutions. Therefore, the convergence rate in the inner optimization is crucial. However, to the best of our knowledge, the convergence rate has not been

shown for the above ADMM algorithm and the POGS algorithm. Below we will present another ADMM algorithm for (1.3.5) with proven linear convergence in the next section.

1.4.2 ADMM for linear program with linear convergence

Wang and Shroff [99] proposed the ADMM approach for a linear program and showed that their ADMM approach solves a linear program significantly faster than standard software such as CPLEX [51] and Gurobi [43]. Moreover, they showed the linear convergence result for their ADMM approach. To apply their approach to our linear program (1.4.1), we reformulate it into the standard form of a linear program (only with equality constraints) [99, Equation 1] by introducing $2n$ auxiliary variables $\boldsymbol{u}, \boldsymbol{s} \in \mathbb{R}^n$ as

$$\begin{aligned} &\underset{\boldsymbol{w}\in\mathbb{R}^{d+3n}}{\text{minimize}} \quad \langle \boldsymbol{c}, \boldsymbol{w} \rangle \\ &\text{subject to} \quad \boldsymbol{B}\boldsymbol{w} = \boldsymbol{p}_k, \quad \boldsymbol{u}, \boldsymbol{s} \geq \boldsymbol{0}_n, \end{aligned} \tag{1.4.4}$$

with

$$\begin{aligned} \boldsymbol{c} &:= [\boldsymbol{0}_d;\, \boldsymbol{1}_n;\, \boldsymbol{0}_n;\, \boldsymbol{0}_n] \in \mathbb{R}^{d+3n} \\ \boldsymbol{w} &:= [\boldsymbol{\theta};\, \boldsymbol{t};\, \boldsymbol{u};\, \boldsymbol{s}] \in \mathbb{R}^{d+3n} \\ \boldsymbol{p}_k &:= [\boldsymbol{\Lambda}_k \boldsymbol{b};\, \boldsymbol{\Lambda}_k \boldsymbol{b}] \in \mathbb{R}^{2n} \\ \boldsymbol{B} &:= \begin{bmatrix} \boldsymbol{X} & -\boldsymbol{I}_n & \boldsymbol{0}_{n,n} & \boldsymbol{I}_n \\ \boldsymbol{X} & \boldsymbol{I}_n & -\boldsymbol{I}_n & \boldsymbol{0}_{n,n} \end{bmatrix} \in \mathbb{R}^{2n\times(d+3n)}, \end{aligned}$$

where $\boldsymbol{0}_n \in \mathbb{R}^n$ and $\boldsymbol{0}_{n,d} \in \mathbb{R}^{n\times d}$ denote the column vector and the matrix with zero entries. By following [99, Algorithm 1], introducing auxiliary variable $\boldsymbol{y} = [\boldsymbol{y}_1;\, \boldsymbol{y}_2] \in \mathbb{R}^{d+3n}$ and dual variable $\boldsymbol{z}^k = [\boldsymbol{z}_1;\, \boldsymbol{z}_2] \in \mathbb{R}^{d+5n}$ for $\boldsymbol{y}_1 \in \mathbb{R}^{d+n}$, $\boldsymbol{y}_2, \boldsymbol{z}_1 \in \mathbb{R}^{2n}$, and $\boldsymbol{z}_2 \in \mathbb{R}^{d+3n}$ provides the augmented Lagrangian function of (1.4.4):

$$\begin{aligned} \mathcal{L}_\rho(\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{c}^\mathsf{T}\boldsymbol{w} + g(\boldsymbol{y}_2) + \boldsymbol{z}^\mathsf{T}(\boldsymbol{B}_1\boldsymbol{\theta} + \boldsymbol{B}_2\boldsymbol{y} - \bar{\boldsymbol{p}}_k) \\ + \frac{\rho}{2}\|\boldsymbol{B}_1\boldsymbol{\theta} + \boldsymbol{B}_2\boldsymbol{y} - \bar{\boldsymbol{p}}_k\|_2^2, \end{aligned} \tag{1.4.5}$$

where

$$g(\boldsymbol{y}_2) := \begin{cases} 0 & \text{if } \boldsymbol{y}_2 \geq \mathbf{0}_{2n}, \\ \infty & \text{otherwise,} \end{cases}$$

and

$$\boldsymbol{B}_1 := \begin{bmatrix} \boldsymbol{B} \\ \boldsymbol{I}_{d+3n} \end{bmatrix}, \quad \boldsymbol{B}_2 := \begin{bmatrix} \mathbf{0}_{d+2n,d+3n} \\ -\boldsymbol{I}_{d+3n} \end{bmatrix}, \quad \bar{\boldsymbol{p}}_k := \begin{bmatrix} \boldsymbol{p}_k \\ \mathbf{0}_{3n} \end{bmatrix}.$$

The update rule by (1.4.5) is then given in a closed form as

$$\boldsymbol{w}^{t+1} = \frac{1}{\rho} \left(\boldsymbol{I} + \boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \left(\boldsymbol{B}_1^\top \left(\boldsymbol{z}^t + \rho(\boldsymbol{B}_2 \boldsymbol{y}^t - \bar{\boldsymbol{p}}_k)\right) + \boldsymbol{c}\right), \tag{1.4.6a}$$

$$\boldsymbol{y}^{t+1} = \boldsymbol{w}^{t+1} + \frac{\boldsymbol{z}_y^t}{\rho}, \quad \boldsymbol{y}_2^{t+1} = [\boldsymbol{y}_2^{t+1}]_+, \tag{1.4.6b}$$

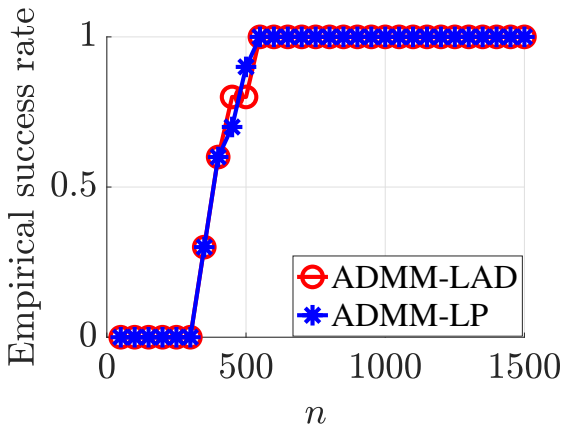
$$\boldsymbol{z}_1^{t+1} = \boldsymbol{z}_1^t + \rho \left(\boldsymbol{B}\boldsymbol{\theta}^{t+1} - \boldsymbol{p}\right), \ \boldsymbol{z}_2^{t+1} = \boldsymbol{z}_2^t + \rho(\boldsymbol{w}^{t+1} - \boldsymbol{y}^{t+1}). \tag{1.4.6c}$$

The most expensive step is the matrix inversion given in (1.4.6a). It is calculated via the matrix-inversion lemma as

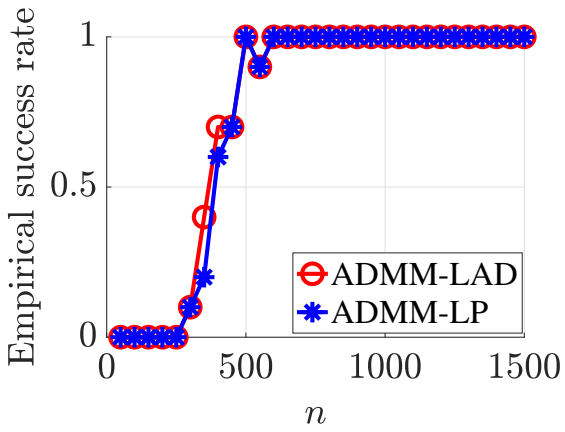
$$\left(\boldsymbol{I}_{d+3n} + \boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} = \boldsymbol{I}_{d+3n} - \boldsymbol{B}^\top (\boldsymbol{I}_{2n} + \boldsymbol{B}\boldsymbol{B}^\top)^{-1} \boldsymbol{B}$$

with cost $\mathcal{O}(n^3)$. Since this step does not depend on previous outer iterations, one can use a pre-computed result on memory over the inner and outer iterations. Hence, by the linear convergence result [99, Theorem 1], the cost for an ϵ_k-accurate solution to (1.4.4) is $\mathcal{O}\left(n^3 + (n+d)^2 \log(1/\epsilon_k)\right)$. However, due to more auxiliary variables in (1.4.4) compared to (1.3.5), in our numerical studies, the ADMM algorithm by (1.4.6) showed slower convergence in the run time relative to the algorithm by (1.4.3).

Figure 1.4 compares the empirical success rate of Robust-AM by two ADMM algorithms, where the success is declared if the estimate $\widehat{\boldsymbol{\theta}}$ satisfies $\text{dist}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_\star) \leq 10^{-3}$. The table in (c) of Figure 1.4 compares the average run time for the experiments in (a) and (b) of Figure 1.4. The empirical study illustrates that the two Robust-AM algorithms show comparable performances but Robust-AM with the first ADMM



(a) zero



(b) Cauchy distribution

(c) Run time comparison

	zero	Cauchy distribution
ADMM-LAD	1.75s	4.75s
ADMM-LP	767.19s	155.64s

Figure 1.4: (Top) Empirical success rate per number of measurements n for Robust-AM by ADMM algorithms ($d = 100$ and $\eta = 0.25$) under the outlier settings in Figure 1.1. (Bottom) Run time comparison of Robust-AM by ADMM algorithms.

(ADMM-LAD) is much faster than Robust-AM with the second ADMM (ADMM-LP). In the analysis of computation cost, we considered the slower ADMM-LP for the inner optimization. However, since Robust-AM with ADMM-LAD is significantly faster and the two algorithms show similar success rates, we adopted Robust-AM with ADMM-LAD for numerical experiments.

1.5 Theoretical results

In this section, we present the convergence analysis of the Robust-AM algorithms under the following assumptions. First, we adopt the standard random linear measurements and outliers with arbitrary support and adversarial values [31].

Assumption 3 *The measurement vectors* $(\boldsymbol{x}_i)_{i=1}^n$ *are independent copies of* $\boldsymbol{x} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$.

Assumption 4 *The outliers are supported on an arbitrarily fixed set* I_{out} *with* $|I_{\text{out}}| = \eta n$ *for* $\eta \in [0, 1/4]$ *and their magnitudes* $|\xi_i|$ *can be adversarial.*

Additionally, to provide the convergence analysis of the fast Robust-AM, we introduce an extra assumption that quantifies the suboptimality of solving (1.3.5) by ADMM.

Assumption 5 *There exists a bounded sequence* $(\epsilon_k)_{k\in\mathbb{N}}$ *such that* $\boldsymbol{\theta}_k$ *is an inexact minimizer up to the sub-optimality level* ϵ_k *for all* $k \in \mathbb{N}$, *i.e.*

$$\begin{aligned} &\sum_{i=1}^n |\text{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle)\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} \rangle - b_i| \\ &\leq \epsilon_k + \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^n |\text{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle)\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle - b_i| \,. \end{aligned} \tag{1.5.1}$$

We denote the highest sub-optimality level as $\epsilon_{\max}$, *i.e.*

$$\epsilon_{\max} := \max_{k\in\mathbb{N}} \epsilon_k.$$

Theorem 4 *Suppose that Assumptions 3, 4, and 5 hold. Then there exist absolute constants* $C, c > 0$ *and constants* $\nu_\eta \in (0, 1), \lambda_\eta > 0$ *depending only on* η, *for which*

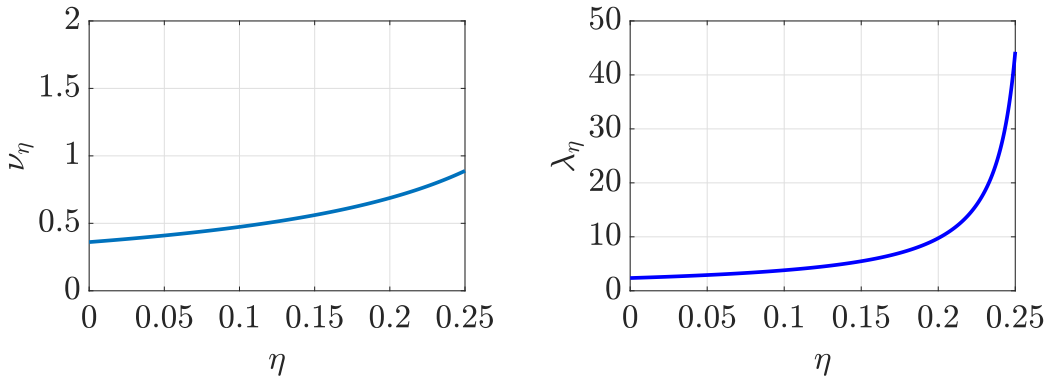
the following statement holds for all $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ *with probability at least* $1 - \exp(-cd)$*: If* $n \geq Cd$ *and*

$$\max\left(\operatorname{dist}\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\star\right), \lambda_\eta \epsilon_{\max}\right) \leq \sin(1/20)\|\boldsymbol{\theta}_\star\|_2, \tag{1.5.2}$$

then the sequence $(\boldsymbol{\theta}_k)_{k\in\mathbb{N}\cup\{0\}}$ *by the fast Robust-AM algorithm satisfies*

$$\operatorname{dist}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_\star\right) \leq \nu_\eta^k \cdot \operatorname{dist}\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\star\right) + \lambda_\eta \epsilon_{\max} \tag{1.5.3}$$

for all $k \in \mathbb{N}$*, where* $\operatorname{dist}(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) := \min_{\alpha\in\{\pm 1\}} \|\boldsymbol{\theta} - \alpha\boldsymbol{\theta}_\star\|_2$.



(a) ν_η (b) λ_η

Figure 1.5: The dependence of parameters η_n and λ_n in Theorem 4 on the outlier fraction η.

Theorem 4 establishes a local linear convergence of the Robust-AM with $n \gtrsim d$ Gaussian random measurements when the support of outliers is arbitrarily fixed and the fraction of outliers is no larger than 1/4.

For a more detailed illustration of the linear convergence result in (3.1.13), we discuss how the parameters ν_η and λ_η depend on the outlier ratio η. The linear convergence parameter ν_η in (3.1.13) is explicitly specified as an increasing function of

η shown in (a) of Figure 1.5 in the proof of Theorem 4. Therefore, smaller η implies faster convergence. The final error bound by (3.1.13) with k going to infinity is given as the amplification of the sub-optimality parameter $\epsilon_{\max}$ in the inner optimization by a factor of λ_η. On one hand, the parameter λ_η is also explicitly given as an increasing function of η and is bounded for $\eta \in [0, 1/4]$ (see (b) of Figure 1.5). On the other hand, $\epsilon_{\max}$ can be small in practice as one uses the linear program packages in readily available software such as CPLEX and Gurobi, and the default accuracy parameter is sufficiently low (less than 10^{-4}). Furthermore, for ADMM algorithms in Section 1.4, the accuracy is determined by the employed stop condition. A sufficiently small $\epsilon_{\max}$ and bounded λ_η make the final error bound sufficiently small.

Now we discuss two conditions in (2.1.4). A condition $\lambda_\eta \epsilon_{\max} \leq \sin(1/20)\|\boldsymbol{\theta}_\star\|_2$ is easily satisfied in practice because $\lambda_\eta \epsilon_{\max}$ can be small, as discussed in the final error bound. The other condition on the initial estimate is easily satisfied by existing initialization methods studied in [31, 104]. Specifically, [104, Proposition 2] and [31, Theorem 3] guarantee that $\boldsymbol{\theta}_0$ provided by the initialization methods obeys the initial condition with the same order of sample complexity of Theorem 4 with high probability.

Next, we compare the specification of Theorem 4 to this scenario to the analogous results for competing methods: RobustPhaseMax [46], Median-RWF [104], and prox-linear [31]. Theorem 4 as well as the previous results achieve the exact recovery when the number of observations n exceeds a multiple of the signal dimension d. For the threshold of η, earlier theoretical results on RobustPhaseMax and Median-RWF showed that there exists an unspecified numerical constant so that the algorithms provide the exact recovery if the outlier fraction is below this constant. In contrast,

the analyses of the prox-linear [31] and Robust-AM (Theorem 4) demonstrate that these methods can tolerate outliers up to 1/4 of the total observations. Furthermore, these theoretical guarantees consider different degrees of adversary for their outlier models. The performance guarantee of RobustPhaseMax by Hand [46] assumed the highest adversary so that both the support and values of sparse noise are adversarial. The performance guarantees of Median-RWF by Zhang et al. [104] considered the same outlier model as in Assumption 4, but they also introduced additive noise of a bounded norm in addition to sparse noise. Duchi and Ruan [31] used the lowest adversary so that the support of sparse noise is random but the nonzero values of sparse noise can depend on the measurements. Despite providing performance guarantees under the highest adversary, as shown in Section 2.2, RobustPhaseMax showed significantly inferior empirical performance relative to the other methods in terms of the tolerable outlier ratio.

As discussed in Section 1.3, Robust-AM has no explicit control over the amount of the update in each iteration unlike the constrained or regularized versions of the Gauss-Newton method [16,31]. However, despite its simple form, Robust-AM provides the monotone decrease of the estimation error toward zero without any overshooting for robust phase retrieval in the setting of Theorem 4. All convergence analyses by Theorem 4 and previous work [31,104] require an initialization within a neighborhood of the ground truth. The size of the basin of convergence was determined with an explicit numerical constant only in [46] and Theorem 4.

Lastly, we compare the computational costs of the robust estimators. First, RobustPhaseMax is formulated as a linear program and thus it can be exactly solved with $\widetilde{\mathcal{O}}((n+d)^{2.38}\log(1/\epsilon))$ multiplications by derandomized algorithm [87].

Furthermore, as we discussed in Section 1.4.2, there exists an ADMM algorithm for the linear program that costs $\mathcal{O}(n^3 + (n+d)^2 \log(1/\epsilon))$ for an ϵ-accurate solution. Due to the term $\log(1/\epsilon)$, if the desired accuracy decreases in proportion to the size of the problem, it is preferable to use ADMM. Otherwise, the derandomized algorithm will be computationally efficient. The other estimators are given as an iterative algorithm with a proven convergence rate. Therefore, we compare their computational costs to obtain an ϵ-accurate solution. Median-RWF is a truncated gradient descent with the per-iteration cost of $\mathcal{O}(nd)$. Since the linear convergence of Median-RWF has been established, the total cost is $\mathcal{O}(nd \log(1/\epsilon))$. Unlike Median-RWF, the updates in prox-linear and Robust-AM involve a nontrivial inner optimization, respectively cast as a quadratic program and a linear program. One may use an exact solver for these subproblems. For example, there exists an interior point method for quadratic programs with the cost $\mathcal{O}((n+d)^4)$ [102]. Since it has been shown that prox-linear converges quadratically, the total cost with this exact inner solver is $\mathcal{O}((n+d)^4) \log\log(1/\epsilon)$. The inner optimization in Robust-AM can be exactly solved at the cost $\widetilde{\mathcal{O}}((n+d)^{2.38} \log(1/\epsilon))$ by the derandomized algorithm [87]. Due to its linear convergence, the total cost of Robust-AM is $\widetilde{\mathcal{O}}((n+d)^{2.38} \log(1/\epsilon))$. However, as shown in Theorem 4, the linear convergence of Robust-AM remains valid when the inner optimization problems are solved only approximately. The fast Robust-AM with the ADMM solver for linear programs has the per-iteration cost of $\mathcal{O}(n^3 + (n+d)^2 \log(1/\epsilon_{\max}))$ as shown in Section 1.4. Due to its linear convergence in Theorem 4, the total cost to obtain the $\epsilon + \lambda_\eta \epsilon_{\max}$ accuracy is $\mathcal{O}(n^3 + (n+d)^2 \log(1/\epsilon_{\max}) \log(1/\epsilon))$. In contrast, the convergence rate of POGS for the inner optimization in prox-linear has not been

established. We summarize the comparison for the computational costs of algorithms in Table 1.1.

1.6 Numerical Results

This section compares the empirical performances of Robust-AM to its theoretical analysis in Theorem 4. Robust-AM is also compared against the competing methods, which include RobustPhaseMax, Median-RWF, and the prox-linear. Recall that all these methods require an initial estimate. For this purpose, we adopt the spectral method by Zhang et al. [104].

1.6.1 Synthetic data experiments

First, through experiments on synthetic data, we show that the numerical results corroborate our theoretical findings in Theorem 4 and Robust-AM outperforms the competing methods. In this experiment, the measurement vectors are generated so that $\{\boldsymbol{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \text{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$ by following the assumptions in Theorem 4 and analogous theoretical analyses of the other methods. The ground-truth signal is generated as $\boldsymbol{\theta}_\star \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$ independently from the measurement vectors. The outlier support is randomly selected following the uniform distribution on all possible subsets $I_{\text{out}} \subset [n]$ of size ηn.

Figure 1.6 shows the phase transition of the empirical success rate by Robust-AM through Monte Carlo simulations, where the outlier values are i.i.d. following the Cauchy distribution with median 0 and mean-absolute-deviation 1. The fraction of outliers is fixed to $\eta = 0.25$. Recall that the performance guarantee in Theorem 4 applies uniformly to all ground-truth signals. To observe the empirical performance in an analogous setting, we design the experiment as follows: 1) Generate 20 sets

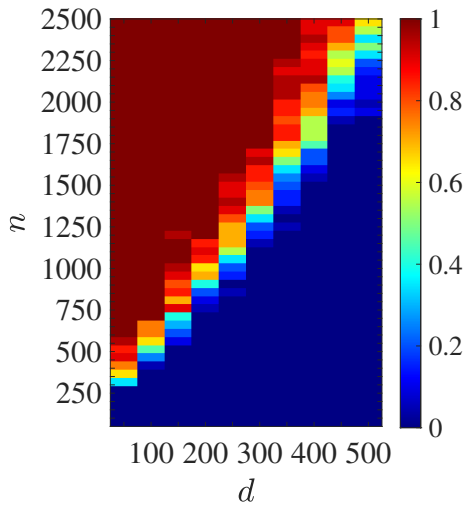


Figure 1.6: Phase transition of empirical success rate by Robust-AM per the number of measurements n and the dimension d.

of random measurement vectors $\{\boldsymbol{x}_i\}_{i=1}^n$. Generate 30 sets of random ground-truth $\boldsymbol{\theta}_\star$; 2) For each fixed $\{\boldsymbol{x}_i\}_{i=1}^n$, success is declared if the estimator recovers all 30 ground-truth signals by satisfying $\text{dist}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_\star) \leq 10^{-3}$ where $\widehat{\boldsymbol{\theta}}$ denotes the estimate; 3) The empirical success rate is calculated on the outcomes from 20 distinct sets of measurement vectors. The transition occurs at the boundary where the number of measurements is proportional to the ambient dimension (signal length). This empirical result corroborates our theoretical finding in Theorem 4. Next, we repeat the same experiment on RobustPhaseMax, Median-RWF, and the prox-linear. Figure 1.7(a) compares the empirical performance of Robust-AM against RobustPhaseMax, Median-RWF, and the prox-linear by displaying the phase transition of these methods for a range of the outlier fraction η in this setting. The ambient dimension is set to $d = 100$. Figure 1.7(a) shows that Robust-AM outperforms all the other methods with a significantly lower threshold for the phase transition. We further expand

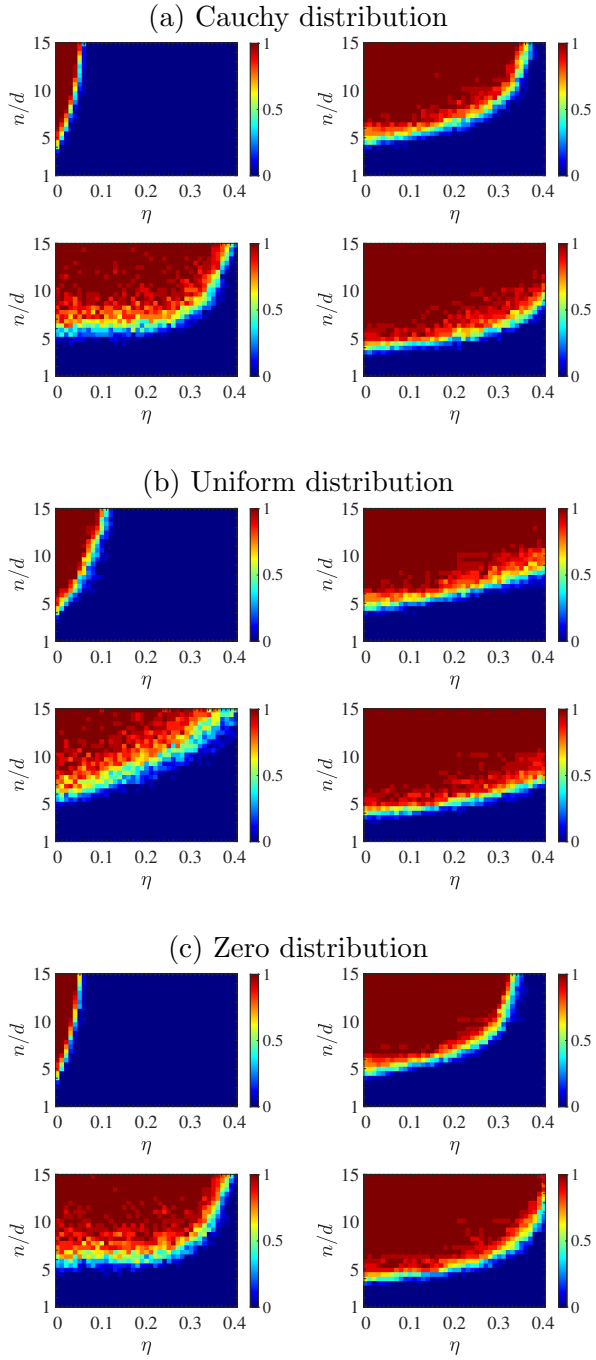
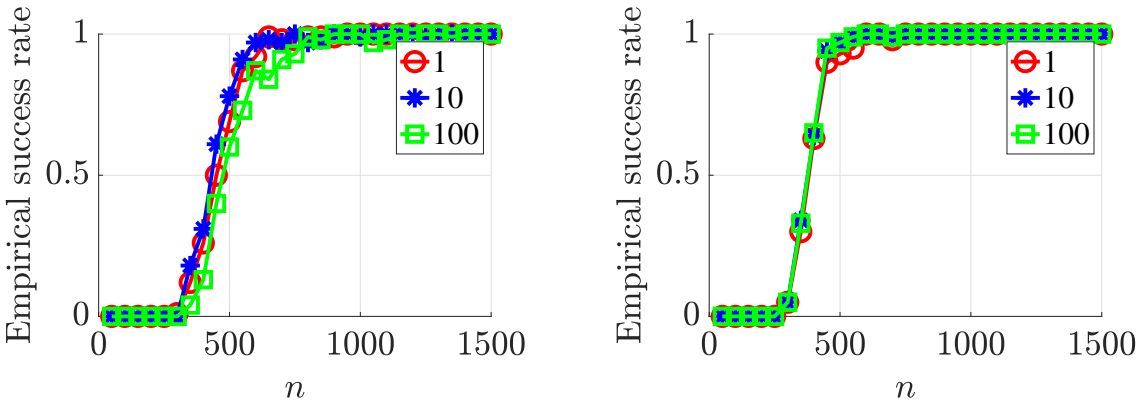


Figure 1.7: Phase transition of success rate per measurement ratio n/d and fraction of outliers η for various outlier magnitude models. Arranged as: (top-left) RobustPhaseMax, (top-right) Median-RWF, (bottom-left) prox-linear method, (bottom-right) Robust-AM.

the comparison to other models for outlier values. The second scenario draws ξ_i from the uniform distribution on $(-d\|\boldsymbol{\theta}_\star\|_2^2/2, d\|\boldsymbol{\theta}_\star\|_2^2/2)$. The third scenario sets ξ_i to 0. As observed in Figure 1.7(b) and Figure 1.7(c), similar trends appear in the other outlier models. RobustPhaseMax, while providing the strongest theoretical performance guarantee, shows the worst empirical performance in the comparison. There is no consistent dominance between Median-RWF and the prox-linear algorithm. Median-RWF outperforms the prox-linear in the second scenario, but the other way around in the other scenarios.



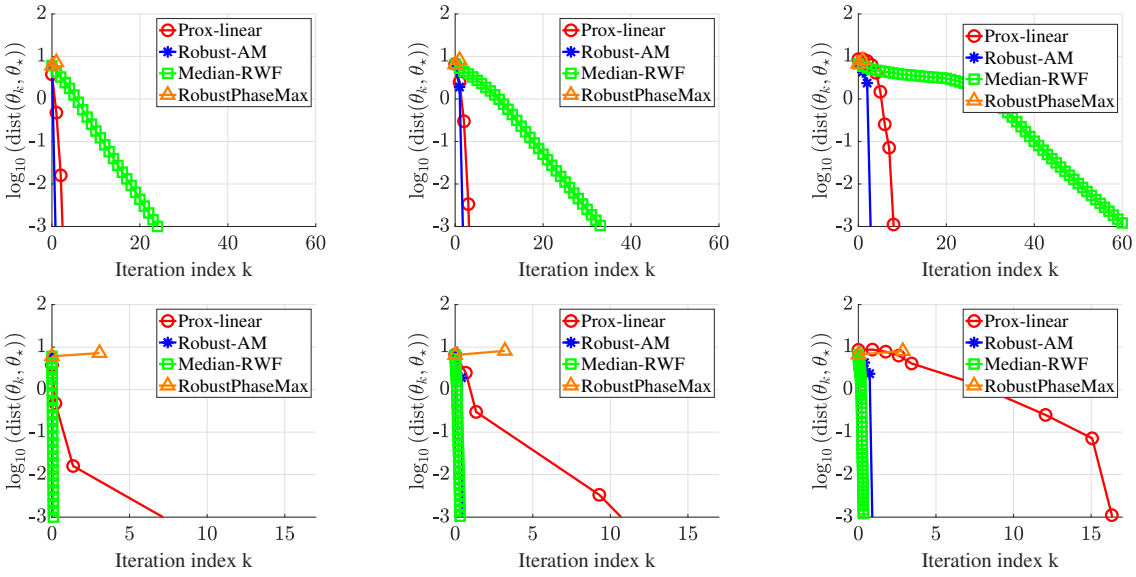
(a) Cauchy distribution

(b) Uniform distribution

Figure 1.8: Empirical success rate per number of measurements n by Robust-AM with the amplification of outliers ($d = 100$ and $\eta = 0.25$). The outlier values are generated following the Cauchy/uniform distribution and then amplified by a factor from {1, 10, 100}.

We also investigated whether the performance of Robust-AM is affected by outlier magnitudes. We repeat the experiment in Figure 1.7 with the outliers amplified by constant factors of 1, 10, and 100. As shown in Figure 1.8, the amplification of the outlier magnitudes does not affect the empirical success rate significantly. This

phenomenon is expected since the least absolute deviation formulation is a median estimator [10].



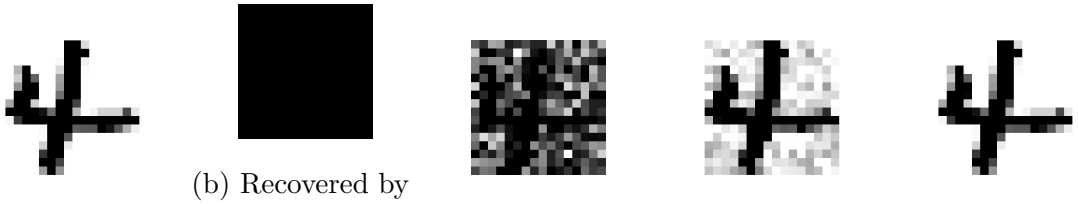
$\eta = 0.1$ $\eta = 0.2$ $\eta = 0.3$

Figure 1.9: Convergence of RobustPhaseMax, Median-RWF, prox-linear method, and Robust-AM in the iteration count (first row) and the run time (second row).

Next, we compare the convergence speed of Robust-AM and the prox-linear algorithm. In this experiment, the dimension parameters are set to $n = 1,500$ and $d = 200$ where the values of outliers are zero. The outlier ratio varies over $\eta \in \{0.1, 0.2, 0.3\}$. Figure 1.9 illustrates how the log of $\mathrm{dist}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_\star)$ decays over the iteration index k. The median over 10 trials is plotted. In their theoretical analyses, the prox-linear algorithm converges faster at a quadratic rate than the linear convergence of Robust-AM in Theorem 4. However, as shown in Figure 1.9, Robust-AM empirically

converges faster than the prox-linear algorithm in both the iteration count and run time for all considered η. Moreover, Figure 1.9 illustrates that the number of iterations for Robust-AM increases as η increases. This implies that for each iteration, the convergence rate of Robust-AM is proportional to η. This supports our theoretical finding that the convergence parameter ν_η in Theorem 4 is an increasing function of η as shown in Figure 1.5(a).

1.6.2 Real image experiments



(a) Ground-truth (b) Recovered by RobustPhase-Max (c) Recovered by Median-RWF (d) Recovered by Prox-linear (e) Recovered by Our Method

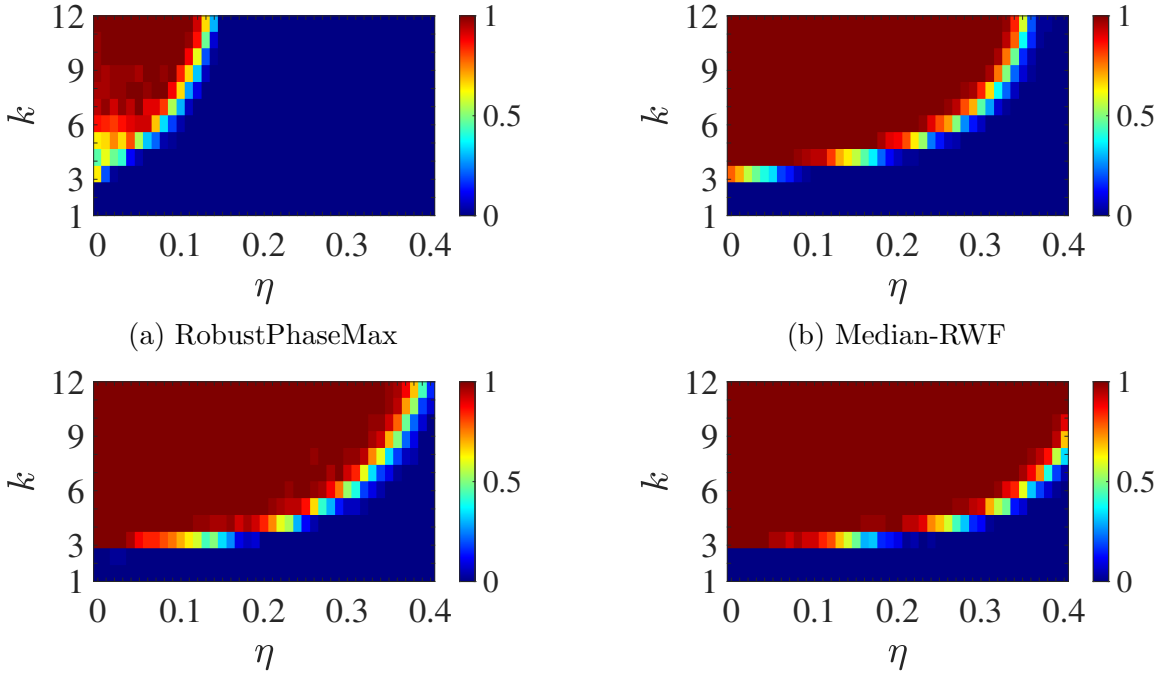
Figure 1.10: Example of recovery for an image data.

We further apply Robust-AM to a set of image data to show that Robust-AM continues outperforming the other competing methods for non-Gaussian measurement models. We adopt the structured random measurement model in the experimental setting in [31, Section 6.3] given by

$$\boldsymbol{X}_{\mathrm{H}} = (\boldsymbol{I}_k \otimes \boldsymbol{H}_d)[\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_k]^{\mathsf{T}} \in \mathbb{R}^{kd \times d}, \tag{1.6.1}$$

where $\boldsymbol{H}_n \in \mathbb{R}^{d\times d}$ denotes the normalized Hadamard matrix and $\boldsymbol{S}_1, \ldots \boldsymbol{S}_k \in \mathbb{R}^{d\times d}$ are diagonal matrices whose diagonal entries are independently drawn uniformly random from $\{\pm 1\}$. The measurement vector $\boldsymbol{x}_i$ is the i-th column of $\boldsymbol{X}_{\mathrm{H}}^{\mathsf{T}}$ for $i \in [n]$, where

$n = kn$. The linear measurement operator in (1.6.1) applies to the vectorized version of a 2D input image $\boldsymbol{\theta}_\star \in \mathbb{R}^{d_1 \times d_2}$ denoted by $\boldsymbol{\theta}_\star := \mathrm{Vec}(\boldsymbol{\theta}_\star) \in \mathbb{R}^d$ with $d = d_1 \times d_2$. The measurements corresponding to outliers are substituted by zero in the experiment.



(a) RobustPhaseMax (b) Median-RWF

(c) Prox-linear method (d) Robust-AM

Figure 1.11: Phase transition of success rate per k and the fraction of outliers η for zero outlier magnitude models. Subfigures are displayed according to (a) RobustPhaseMax (top-left), (b) Median-RWF (top-right), (c) Prox-linear method (bottom-left), and (d) Robust-AM (bottom-right).

Robust-AM and the competing algorithms are tested on the collection of 50 images of handwritten digits[3]. Figure 1.11 compares the two methods in the empirical success rate over 50 images, where the number of random modulations k and the outlier fraction η respectively vary over $k \in \{1, \ldots, 12\}$ and $\eta \in [0, 0.4]$. Similar to the

[3]`https://hastie.su.domains/ElemStatLearn/datasets/zip.digits`.

previous experiments on synthetic data, Figure 1.11 demonstrates that Robust-AM outperforms the competing algorithms by providing recovery with smaller k for each observed η. Since the algorithmic parameters of Median-RWF were specifically selected for Gaussian measurements in [104], we heuristically tuned the step size to 0.2 for the non-Gaussian measurement model (1.6.1).

1.7 Discussion on a tolerable fraction of outliers in robust phase retrieval

Although Theorem 4 provides a local convergence guarantee when the outlier fraction satisfies $\eta \leq 1/4$, Robust-AM empirically continues to provide exact recovery for η above $1/4$ for larger n as shown in Figure 1.7(a), Figure 1.7(b), and Figure 1.7(c). We discuss these phenomena in this section. First, we present an analysis of tolerable η for LAD in (1.3.4). Next, we elaborate on the analysis of Robust-AM to elucidate the dependence of tolerable η on n and the error in the initialization. The resulting condition on η is compared to that for LAD.

1.7.1 Tolerable η by LAD

We consider the population-level analysis of LAD in (1.3.4) under the Gaussian measurement assumption. Furthermore, since linear regression can be considered as a special case of phase retrieval with the oracle phase information, the upper bound on tolerable η for phase retrieval is no larger than that for linear regression. Therefore, we analyze η for the population-level linear regression for LAD given by

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \ \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} |\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle - \beta_i| \right), \tag{1.7.1}$$

where

$$\beta_i := \begin{cases} \xi_i & \text{if } i \in I_{\text{out}} \\ \langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle & \text{if } i \in I_{\text{in}}. \end{cases} \tag{1.7.2}$$

The following lemma characterizes the condition for the exact recovery of $\boldsymbol{\theta}_\star$ by (1.7.1).

Lemma 5 *Suppose that Assumptions 3, 4 and 5 hold with modifications so that* $\eta \in [0, 1)$ *and* $\epsilon_{\max} = 0$. *The solution to* (1.7.1) *coincides with* $\boldsymbol{\theta}_\star$ *for any instances of* $(\xi_i)_{i \in I_{\text{out}}}$ *if and only if* $\eta < 1/2$.

Theorem 5 provides a sharp characterization of tolerable η for the population-level LAD. The original LAD in (1.3.4) that minimizes the empirical loss will behave similarly when n is sufficiently large. In the subsequent analysis, we compare tolerable η for Robust-AM to the condition $\eta < 1/2$.

1.7.2 Tolerable η by Robust-AM

In the proof of Theorem 4, for the sake of simple presentation, we did not attempt to expand the tolerable range of η beyond $[0, 1/4]$. However, with a slight modification of the proof, we obtain the following lemma that illustrates how tolerable η is determined by n and the initialization error.

Lemma 6 *Fix* $\delta \in (0, 1), \psi_0 \in (0, 0.12)$ *arbitrarily and define*

$$u(x) := \frac{1}{2} - \frac{3x}{\pi} \left(1 + \sqrt{\pi \log \left(\frac{e\pi}{2x} \right)} \right), \quad \forall x \in \mathbb{R}_+.$$

There exists an absolute constant C_1 *for which the following statement holds: Suppose that Assumptions 3,4 and 5 hold with modifications so that*

$$0 \le \eta < u(\psi_0) - C_1 \sqrt{\frac{d \vee \log(1/\delta)}{n}} \tag{1.7.3}$$

and $\epsilon_{\max} = 0$. *Furthermore, suppose that*

$$\operatorname{dist}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\star) \leq \sin(\psi_0)\|\boldsymbol{\theta}_\star\|_2, \tag{1.7.4}$$

$$n \geq C_1^2 \cdot \min(1 - 2\eta, \psi_0, u(\psi_0))^{-2} \cdot (d \vee \log(1/\delta)) \tag{1.7.5}$$

Then Robust-AM linearly converges to $\boldsymbol{\theta}_\star$ *with probability at least* $1 - \delta$.

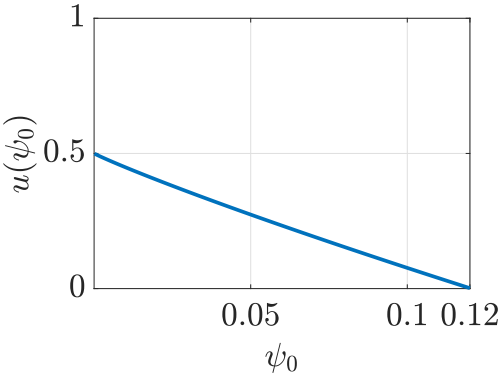


Figure 1.12: Plot of $u(\psi_0)$ with respect to $\psi_0 \in [0, 0.12]$.

Theorem 6 shows that Robust-AM tolerates a higher value of η as the initialization error decreases and/or n increases. In particular, as ψ_0 and n approach 0 and ∞, respectively, the highest tolerable level of η by Robust-AM converges 1/2, which matches the corresponding condition on η by LAD Theorem 5.

Next we explain why we choose $\psi_0 = 1/20$ and $\eta \leq 1/4$ in Theorem 4. On one hand, the requirement on the initialization in (1.7.4) and the sample complexity in (1.7.5) become more stringent as ψ_0 decreases toward 0. On the other hand, as shown in Figure 1.12, with ψ_0 increasing toward 0.12, $u(\psi_0)$ converges to 0 and the sample complexity in (1.7.5) blows up to infinity. The choice of $\psi_0 = 1/20$ compromises these

conflicting conditions. Then choosing C_1 in (1.7.3) sufficiently large yields the range $0 \leq \eta \leq 1/4$. The factor $\min(1 - 2\eta, \psi_0, u(\psi_0))$ in (1.7.5) is then bounded from below by 0.12 and Theorem 6 coincides with a special case of Theorem 4 with $\epsilon_{\max} = 0$.

1.8 Summary

The least absolute deviation (LAD) has been a popular statistical method for regression in the presence of outliers. We consider the LAD approach to robust phase retrieval with the magnitude-only measurement model. To solve the resulting non-convex optimization, we derive a robust alternating minimization method (Robust-AM) as an unconstrained Gauss-Newton method. Furthermore, we propose fast Robust-AM by exploiting efficient solvers and show that Robust-AM by ADMM converges faster than a similar approach known as the prox-linear by its efficient solver POGS [31].

We established a local convergence analysis of Robust-AM under the standard Gaussian measurement model when the support of sparse noise is arbitrarily fixed but magnitudes can be adversarial. A suitably initialized Robust-AM converges linearly to the ground truth uniformly over all ground-truth signals when the number of measurements n is proportional to the signal length d and the outlier fraction is up to $1/4$. This theoretical result is comparable to existing prior art in the literature. Furthermore, the numerical results show that Robust-AM outperforms the existing guaranteed methods for various outlier models in both synthetic and real-image data.

Chapter 2: Max-linear regression by convex program

In this chapter, we present the results for convex program for max-linear regression under the deterministic noise, with the problem formulation described in Section 1.2.1.

2.1 Accuracy of the Convex Estimator

In this section, we provide our main results on the estimation error of the convex program in (1.2.11). We consider the anchor vector $\boldsymbol{a}$ constructed from a given initial estimate $\widetilde{\boldsymbol{\theta}} = [\widetilde{\boldsymbol{\theta}}_1; \ldots; \widetilde{\boldsymbol{\theta}}_k] \in \mathbb{R}^{kd}$ as

$$\boldsymbol{a} = \frac{1}{2n} \sum_{i=1}^{n} \nabla f_i(\widetilde{\boldsymbol{\theta}}) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}_{\{\boldsymbol{x}_i \in \widetilde{\mathcal{C}}_j\}} \boldsymbol{e}_j \otimes \boldsymbol{x}_i, \tag{2.1.1}$$

where

$$\widetilde{\mathcal{C}}_j := \left\{ \boldsymbol{z} \in \mathbb{R}^d \ : \ \langle \boldsymbol{z}, \widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_l \rangle \geq 0, \ \forall l \neq j \right\}, \quad j \in [k] \tag{2.1.2}$$

and $\boldsymbol{e}_j \in \mathbb{R}^k$ denotes the jth column of the k-by-k identity matrix $\boldsymbol{I}_k$ for $j \in [k]$. Since f_i is differentiable except on a set of measure zero, with a slight abuse of terminology, ∇f_i in (2.1.1) is referred to as the "gradient". In (2.1.1), the choice of anchor vector follows from the geometry of convex equations [4, Section 1.4]. In particular, in the noiseless case, $\boldsymbol{\theta}_\star$ would be a solution to

$$\begin{aligned} \underset{\boldsymbol{\theta}}{\text{maximize}} \quad & \langle \boldsymbol{a}, \boldsymbol{\theta} \rangle \\ \text{subject to} \quad & f_i(\boldsymbol{\theta}) \leq y_i, \quad \forall i \in [n]. \end{aligned}$$

if it satisfies the Karush–Kuhn–Tucker condition

$$-\boldsymbol{a} + \sum_{i=1}^{n} \lambda_i \nabla f_i(\boldsymbol{\theta}_\star) = \mathbf{0}$$

for some $\lambda_1, \ldots, \lambda_n \geq 0$. In other words, the anchor vector $\boldsymbol{a}$ needs to be in the cone $(\{\nabla f_i(\boldsymbol{\theta})\}_{i=1}^n)$. The choice of $\boldsymbol{a}$ in (2.1.1) is inspired by this condition.

The following theorem illustrates the sample complexity and the corresponding estimation error achieved by the estimator in (1.2.11). The estimation error is measured as the sum of the ℓ_2 norms of the difference between the corresponding components of the ground truth $\boldsymbol{\theta}_\star$ and the estimate $\widehat{\boldsymbol{\theta}}$.

Theorem 7 *Let $\{\mathcal{C}_j\}_{j=1}^k$ be polyhedral cones constructed by $(\boldsymbol{\theta}_{\star,j})_{j=1}^k$*

$$\mathcal{C}_j := \left\{ \boldsymbol{z} \in \mathbb{R}^d \ : \ \langle \boldsymbol{z}, \boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,l} \rangle \geq 0, \ \forall l \neq j \right\}, \quad j \in [k], \tag{2.1.3}$$

and $\{\widetilde{\mathcal{C}}_j\}_{j=1}^k$ be and (2.1.2). Let $\boldsymbol{a}$ be as in (2.1.1) and $\{\boldsymbol{x}_i\}_{i=1}^n$ be independent copies of $\boldsymbol{g} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$. Then there exist absolute constants $c, C > 0$, for which the following statement holds for all $\boldsymbol{z} \in \mathbb{R}^n$ with probability at least $1 - \delta$: Suppose that $\widetilde{\boldsymbol{\theta}}$ is independent of $\{\boldsymbol{x}_i\}_{i=1}^n$ satisfies

$$\begin{aligned} &\frac{\|(\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_{j'}) - (\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'})\|_2}{\|\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'}\|_2} \leq \\ &\min\left(0.1, \frac{c\pi_{\min}^4}{2k} \log^{-1/2}\left(\frac{k}{c\pi_{\min}^4}\right)\right), \ \ \forall j, j' \in [k] : j \neq j'. \end{aligned} \tag{2.1.4}$$

If the feasible set of the optimization problem in (1.2.11) is not empty and the number of observations satisfies

$$n \geq C\, \zeta^{-2} \left(4d \log^3 d \log^5 k + 4 \log(1/\delta) \log k\right), \tag{2.1.5}$$

where

$$\zeta := \min_{j \in [k]} \sqrt{\frac{\pi}{32}}\, \mathbb{P}^2\{\boldsymbol{g} \in \mathcal{C}_j\} - 2 \max_{j \in [k]} \sqrt{\mathbb{P}\{\boldsymbol{g} \in \widetilde{\mathcal{C}}_j \triangle \mathcal{C}_j\}},$$

then the solution $\widehat{\boldsymbol{\theta}}$ *to* (1.2.11) *obeys*

$$\sum_{j=1}^{k} \|\boldsymbol{\theta}_{\star,j} - \widehat{\boldsymbol{\theta}}_j\|_2 \leq \frac{2}{\zeta}\left(\eta + \frac{1}{n}\sum_{i=1}^{n}(z_i)_+\right). \tag{2.1.6}$$

To make the optimization problem in (1.2.11) feasible, it suffices to include the ground-truth $\boldsymbol{\theta}_\star$ in the feasible set, i.e.

$$\eta \geq \frac{1}{n}\sum_{i=1}^{n}(-z_i)_+, \tag{2.1.7}$$

The error bound in (2.1.6) reduces to $\frac{2}{\zeta n}\sum_{i=1}^{n}|z_i|$ when the parameter η is chosen so that the equality in (2.1.7) is achieved. In practice, the noise entries are unknown and this error cannot be achieved. If η, as a parameter that determines the power of the adversary, is chosen so that $\eta \geq \|\boldsymbol{z}\|_1/n$, then the resulting error bound becomes $\frac{4\|\boldsymbol{z}\|_1}{n\zeta}$. In particular, if η satisfies $\eta \geq \|\boldsymbol{z}\|_\infty$, then the resulting error bound will be $\frac{4\|\boldsymbol{z}\|_\infty}{\zeta}$. The latter condition will be readily satisfied in practical applications. Furthermore, as shown in the empirical sensitivity analysis in Section 2.2, the estimation error does not crucially depend on the choice of η.

2.1.1 Comparison with an oracle estimator

Assuming that the additive noise is i.i.d. sub-Gaussian with zero mean and variance σ^2, the error bound in (2.1.6) becomes $\widetilde{O}(\sigma/\zeta)$, which implies that our estimator is not consistent. However, in the adversarial noise setting which is our focus, we can compare the performance case of our estimator with an oracle-assisted estimator, similar to the analysis carried out in [17] for the matrix completion problem. In this scenario, the error bound by the convex estimator nearly matches the performance of an oracle-assisted estimator (up to a factor determined by $\boldsymbol{\theta}_\star$).

Lemma 1 *Consider the same regression problem as in Theorem 7 with* $\{\boldsymbol{x}_i\}_{i=1}^n$ *being independent copies of* $\boldsymbol{g} \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$. *Suppose that* $\{\mathcal{C}_j\}_{j=1}^k$ *in* (2.1.3) *is given as the oracle information. Then there exists an absolute constant* $C > 0$ *such that if*

$$n \geq C\pi_{\min}^{-2} \max(kd \log(n/d), \log(1/\delta)), \tag{2.1.8}$$

then the estimates $\{\widehat{\boldsymbol{\theta}}_j\}_{j=1}^k$ *obtained through the decoupled least-squares* (1.2.5) *with* $b_j = 0$ *for all* $j \in [k]$ *satisfy*

$$\sup_{\|\boldsymbol{z}\|_\infty \leq \eta'} \sum_{j=1}^k \|\boldsymbol{\theta}_{\star,j} - \widehat{\boldsymbol{\theta}}_j\|_2 \gtrsim \frac{\pi_{\min}^{3/2} \eta'}{\pi_{\max}} \tag{2.1.9}$$

with probability at least $1 - \delta$, *where* $\pi_{\max} := \max_{j \in [k]} \mathbb{P}(\boldsymbol{g} \in \mathcal{C}_j)$.

Proof 1 *See Appendix B.2.1.*

One expects that the oracle estimator nearly achieves the optimal performance. However, since the lower bound by Lemma 1 does not vanish as n increases to infinity, the oracle estimator is also biased in the presence of adversarial noise. Note that the lower bound in (2.1.9) remains the same with the feasible set substituted by $\|\boldsymbol{z}\|_1 \leq n\eta'$. Furthermore, if η achieves the equality in (2.1.7), then the error bound in (2.1.6) implies

$$\sup_{\|\boldsymbol{z}\|_1 \leq n\eta'} \sum_{j=1}^k \|\boldsymbol{\theta}_{\star,j} - \widehat{\boldsymbol{\theta}}_j\|_2 \leq \frac{2\eta'}{\zeta}. \tag{2.1.10}$$

Therefore, in this scenario, the error bound in (2.1.10) matches that by the oracle estimator up to an extra factor $O(\pi_{\max}/\zeta\pi_{\min}^{3/2})$. In particular, if $\pi_{\max} \approx \pi_{\min} \approx 1/k$, then the error by the convex estimator is sub-optimal up to a factor $k^{5/2}$ relative to the oracle estimator.

2.1.2 Initialization

Theorem 7 provides an error bound by the convex estimator given an initial estimate satisfying (2.1.4). Finding such an initial estimate is not a trivial task. Ghosh et al. [38] proposed an initialization scheme that consists of dimensionality reduction by a spectral method [38, Algorithm 2], followed by a low-dimensional random search [38, Algorithm 3]. It has been shown that if the observations are corrupted with independent sub-Gaussian noise, then the initialization scheme provides an estimate within a certain neighborhood of the ground-truth in a polynomial time when $k = O(1)$. Their proof only uses the fact that the maximum magnitude of sub-Gaussian noise entries is bounded with high probability. Below, we extend the analysis of their initialization scheme to the scenario where the noise vector $\boldsymbol{z}$ is a fixed *deterministic vector* under the only condition that $\|\boldsymbol{z}\|_\infty \leq \eta'$.

To this end, we first recall the first stage in their initialization scheme that extracts the eigenvectors corresponding to the k dominant eigenvalues of the following matrix:

$$\widehat{\boldsymbol{M}} = \frac{2}{n}\left(\sum_{i=1}^{n/2} y_i \boldsymbol{x}_i\right)\left(\sum_{i=1}^{n/2} y_i \boldsymbol{x}_i\right)^\top + \frac{2}{n}\sum_{i=1}^{n/2} y_i\left(\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T} - \boldsymbol{I}_d\right). \tag{2.1.11}$$

Let $\widetilde{\boldsymbol{M}}$ denote the noise-free version of $\widehat{\boldsymbol{M}}$, i.e.,

$$\widetilde{\boldsymbol{M}} = \frac{2}{n}\sum_{i=1}^{n/2}\left(\max_{j\in[k]}\langle\boldsymbol{\theta}_{\star,j}, \boldsymbol{x}_i\rangle\right)\left(\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T} - \boldsymbol{I}_d\right) + \frac{2}{n}\left(\sum_{i=1}^{n/2}\left(\max_{j\in[k]}\langle\boldsymbol{\theta}_{\star,j}, \boldsymbol{x}_i\rangle\right)\boldsymbol{x}_i\right)\left(\sum_{i=1}^{n/2}\left(\max_{j\in[k]}\langle\boldsymbol{\theta}_{\star,j}, \boldsymbol{x}_i\rangle\right)\boldsymbol{x}_i\right)^\top.$$

Then the ground-truth parameter vectors $\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_k^\star$ are in the columns space of $\mathbb{E}\widetilde{\boldsymbol{M}}$. Ghosh et al. [38] derived a tail bound on the perturbation of those eigenvectors due to sub-Gaussian noise. We provide an analogous perturbation analysis in the

deterministic noise setting. The following lemma provides upper bounds on the contributions of the noise to the two summands in the right-hand side of (2.1.11).

Lemma 2 *Suppose that* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \text{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$ *and* $\boldsymbol{z} := (z_1, \ldots, z_n) \in \mathbb{R}^n$ *are arbitrary fixed. Then the following inequalities hold with probability at least* $1 - \delta$*:*

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^{n} z_i \boldsymbol{x}_i \right\|_2 \lesssim \|\boldsymbol{z}\|_\infty \cdot \sqrt{\frac{d + \log(1/\delta)}{n}}, \\ &\left\| \frac{1}{n} \sum_{i=1}^{n} z_i \left(\boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{I}_d \right) \right\| \lesssim \\ &\qquad \|\boldsymbol{z}\|_\infty \cdot \max\left(\sqrt{\frac{d + \log(1/\delta)}{n}}, \frac{d + \log(1/\delta)}{n} \right). \end{aligned} \tag{2.1.12}$$

Proof 2 *See Appendix B.2.3.*

Let $\widehat{\boldsymbol{U}}$ be a matrix whose columns are the k dominant eigenvectors of $\widehat{\boldsymbol{M}}$. Furthermore, let the columns of $\boldsymbol{U}^\star$ be the eigenvectors of the noise-free component of $\mathbb{E}\widetilde{\boldsymbol{M}}$. Then, plugging the results in Lemma 2 into the proof of [38, Lemma 8] yields that

$$\begin{aligned} &\left\| \widehat{\boldsymbol{U}} \widehat{\boldsymbol{U}}^\top - \boldsymbol{U}^\star \left(\boldsymbol{U}^\star \right)^\top \right\|_{\text{F}}^2 \lesssim \\ &\left(\frac{\|\boldsymbol{z}\|_\infty^2 + \max_{j \in [k]} \|\boldsymbol{\theta}_{\star,j}\|_1^2}{\lambda_k^2(\mathbb{E}\widetilde{\boldsymbol{M}})} \right) \frac{kd \log^3(dk/\delta)}{n} \end{aligned} \tag{2.1.13}$$

holds with probability at least $1 - \delta$. This is analogous to [38, Theorem 2] which addresses the case of the sub-Gaussian noise. The remainder of their initialization scheme does not depend on any assumption on the noise model. Therefore, the resulting initial estimate satisfies (2.1.4) if

$$\begin{aligned} &n \gtrsim \frac{k^6 \log(k/\pi_{\min})}{\pi_{\min}^{13}} \cdot \\ &\max\left\{ \|\boldsymbol{z}\|_\infty^2 \log\left(1 + \frac{\max_{j \in [k]} \|\boldsymbol{\theta}_{\star,j}\|_2 k^4 \log^{1/2}(k/\pi_{\min})}{\pi_{\min}^{5.5}} \right), \right. \\ &\left. \left(\|\boldsymbol{z}\|_\infty^2 + \max_{j \in [k]} \|\boldsymbol{\theta}_{\star,j}\|_1^2 \right) \frac{k^3 d \log^3(n/k) \cdot \max_{j \in [k]} \|\boldsymbol{\theta}_{\star,j}\|_2}{\lambda_k^2(\mathbb{E}\widetilde{\boldsymbol{M}})} \right\}. \end{aligned} \tag{2.1.14}$$

The condition in (2.1.14) is obtained by applying the initialization condition (2.1.4) and substituting σ by $\|\boldsymbol{z}\|_\infty$ in [38, Equation 20]. The requirement for the initial point of anchored regression (2.1.4) is more relaxed in terms of the dependence on $\pi_{\min}$, compared to the similar requirement for the alternating minimization method [38, Theorem 1]. Furthermore, for both anchored regression and alternating minimization, the sample complexity of the initialization dominates that of the subsequent stages of the algorithms.

In the above paragraphs, we have shown that the anchored regression combined with the spectral initialization provides a stable estimate in the presence of an arbitrarily fixed deterministic noise of bounded magnitudes. However, this result does not extend to the adversarial noise setting in Theorem 7 and Lemma 1. Maximization over $\boldsymbol{z}$s that obey $\|\boldsymbol{z}\|_\infty \leq \eta'$ in (2.1.12), can be addressed effectively by taking the union bound over extreme points of ℓ_∞^n ball with the radius η' and choosing $\delta = 2^{-n}\bar{\delta}$ with $\bar{\delta} \in [0, 1]$ denoting overall error probability. Therefore, the terms $\frac{d+\log(1/\delta)}{n}$ in (2.1.12) are equal to $\frac{d+n\log(2)+\log(1/\bar{\delta})}{n}$, which are clearly bounded from below by $\log 2$. Consequently, in the adversarial setting, the error in the spectral method does not vanish as n grows, and the desired accuracy for the initialization scheme cannot be established. Considering a relaxed condition $\|\boldsymbol{z}\|_1 \leq n\eta'$ exacerbates the situation and the error bound in the spectral method becomes even larger.

2.1.3 Compariosn with alternating minimization in computational cost

This section compares AR and AM in their computational costs. First, AR is implemented via an equivalent formulation with auxiliary variables $\boldsymbol{t} := [t_1; \ldots; t_n] \in \mathbb{R}^n$ as

$$\begin{aligned}
&\underset{(\boldsymbol{\theta}_j)_{j=1}^k,(t_i)_{i=1}^n}{\text{maximize}} \ \langle \boldsymbol{a}, [\boldsymbol{\theta}_1; \ldots; \boldsymbol{\theta}_k] \rangle \\
&\text{subject to } t_i \geq 0,\ \langle \boldsymbol{x}_i, \boldsymbol{\theta}_j \rangle - y_i \leq t_i,\ \frac{1}{n}\sum_{i=1}^n t_i \leq \eta, \\
&\qquad\qquad \forall i \in [n],\ \forall j \in [k]\,.
\end{aligned} \tag{2.1.15}$$

To compute the computational costs for (4.2.2), we further reformulate it into the form of a linear program $\min_{\boldsymbol{As}=\boldsymbol{b},\boldsymbol{s}\geq\boldsymbol{0}}\langle \boldsymbol{c}, \boldsymbol{s}\rangle$ by introducing an additional $nk+1$ auxiliary variables to convert the second and third inequality constraints into equality constraints. Then, we have $nk+1$ equality constraints and $2dk+nk+n+1$ variables. By [87], finding its exact solution costs $\widetilde{O}\left(((n+d)k)^c\right)$ with $c \approx 2.38$. In contrast, with finitely many operations, AM can find only an approximate solution. The per-iteration cost of AM is $O(nkd^2)$. In the noiseless case, due to the linear convergence of AM, the total cost to obtain an ϵ-accurate solution is $O(nkd^2 \log(1/\epsilon))$.

In a special case where the observations are almost equally distributed over the linear components of the max-linear model, we have $\pi_{\min} \approx \pi_{\max} \approx 1/k$. Consequently, the sample complexity for both estimators is $\widetilde{O}(dk^4)$. Thus, the computational costs for AR and AM become $\widetilde{O}(d^{2.38}k^{12})$ and $\widetilde{O}(d^3k^5)$, respectively. When d is much larger than k (specifically, $d > k^{14}$, the computational cost of AR is significantly lower than that of AM. However, in the opposite scenario, AM is more cost-effective. We summarize the comparison with respect to the computational cost, sample complexity and model assumption in Table 2.1.

[4]The spectral initialization is not included in this comparison. To incorporate the initialization into the analysis, it is necessary to modify the noise model from an adversarial noise model to a gross error model as discussed in Section 2.1.2.

Table 2.1: Comparison of local convergence of AR and AM.[4]

	AR	AM [38]
Cost for ϵ-accuracy	$\widetilde{O}\left(((n+d)k)^{2.38}\right)$	$O(nkd^2\log(1/\epsilon))$
Cost for an ideal instance	$\widetilde{O}(d^{2.38}k^{12})$	$\widetilde{O}(d^3k^5)$
Sample complexity	$\widetilde{O}\left(\pi_{\min}^{-4}d\right)$	$O(\pi_{\min}^{-3}kd)$
Covariate model	Gaussian	Gaussian
Noise model	Adversarial	Sub-Gaussian

2.2 Numerical results

We present a set of Monte Carlo simulations to evaluate the performance of the estimator by anchored regression numerically. The experiments were designed to illustrate the following perspectives on the estimation performance: i) The empirical phase transition on exact recovery without noise corroborates Theorem 7; ii) Further iterations of AR with updated anchor vectors significantly reduce the estimation error; iii) AR provides a competitive empirical performance with additive Gaussian noise to AM; iv) AR provides a stable estimation in the presence of sparse noise, where the performance of AM significantly deteriorates. We implement AR by the linear program given in (4.2.2). Since (4.2.2) is in the standard form of a linear program, it can be solved efficiently by readily available software such as CPLEX and Gurobi [44]. AR is compared to the version of AM by Ghosh et al. [38]. For a fair comparison, we let both methods start from the same initial estimate, which will be specified later.

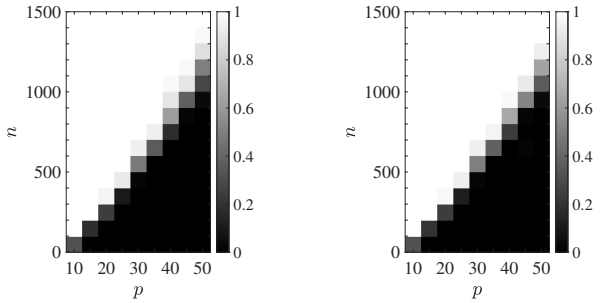
In the Monte Carlo simulations, the regressors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are generated as independent copies of a random vector following Normal$(\mathbf{0}, \boldsymbol{I}_d)$, as assumed in Theorem 7. For each run, the estimation error is measured up to permutation ambiguity, that is, the error is calculated as the minimum of $\sum_{j=1}^k \|\widehat{\boldsymbol{\theta}}_{\pi(j)} - \boldsymbol{\theta}_{\star,j}\|_2 / \sum_{j=1}^k \|\boldsymbol{\theta}_{\star,j}\|_2$ over all

possible permutation π over segment indices, where $(\boldsymbol{\theta}_{\star,j})_{j=1}^{k}$ and $(\widehat{\boldsymbol{\theta}}_j)_{j=1}^{k}$ denote the ground-truth parameters and their estimates, respectively.

Since both AR and AM algorithms operate provided suitably initialized parameter, it is crucial to obtain an initial estimate, which lands near the ground-truth parameter. To this end, throughout the simulations, we apply the heuristic known as the *AM with repeated random initialization* in [7], summarized as follows: One repeats the following procedure for $q \in [m]$: i) Randomly generate parameters $\boldsymbol{\theta}_{q,1}^{r}, \ldots, \boldsymbol{\theta}_{q,k}^{r} \in \mathbb{R}^d$. ii) Run the AM algorithm from given initial estimates for I_{init} iterations and obtain estimates $\boldsymbol{\theta}_{q,1}^{o}, \ldots, \boldsymbol{\theta}_{q,k}^{o}$. Then choose the set of parameters $\boldsymbol{\theta}_{q',1}^{o}, \ldots \boldsymbol{\theta}_{q',k}^{o}$, which achieves the least empirical loss in (1.2.3), i.e.

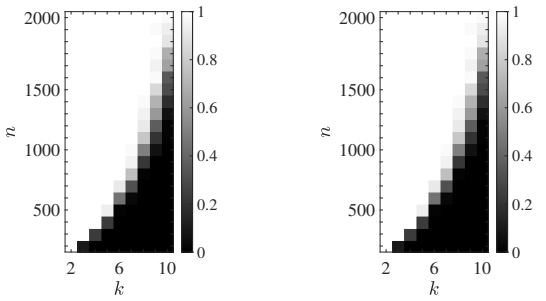
$$q' = \operatorname*{argmin}_{q \in [m]} \sum_{i=1}^{n} \left(\max_{1 \le j \le k} \langle \boldsymbol{x}_i, \boldsymbol{\theta}_{q,j}^{o} \rangle - y_i \right)^2 .$$

Throughout all simulations, the initialization parameters are set to $m = 200$ and $I_{\text{init}} = 10$. Moreover, the maximum iteration number for the AM algorithm, denoted by I_{AM}, is set to $I_{\text{AM}} = 120$.



(a) AR (b) AM

Figure 2.1: Phase transition of recovery rate for varying n and d in the noiseless case ($k = 5$).



(a) AR (b) AM

Figure 2.2: Phase transition of recovery rate for varying n and k in the noiseless case ($p = 20$).

Figures 2.1 and 2.2 illustrate the empirical phase transition of exact recovery in the noise-free scenario as a function of the sample size n per varying dimension parameters, which are the ambient dimension p and the number of segments k. The reconstruction is determined as success if the normalized estimation error is below 10^{-5}. The recovery rate is calculated as the ratio of success out of 50 trials. In this simulation, we assume that $k \leq d$. To satisfy the "well-balance partition" condition, we generate the ground-truth parameter vectors so that they are mutually orthogonal one another.

Figure 2.1 shows that for both AR and AM, the phase transition occurs when n grows linearly with p while k is fixed to 5. This observation qualitatively coincides with the sample complexity by Theorem 7. A complementary view is provided by Figure 2.2 for varying k while p is fixed to 20. Here, the phase transition occurs when n is proportional to k^t for some constant $t \in (1, 2)$. The order of this polynomial is smaller than the corresponding result by Theorem 7, where n is proportional to k^4. A similar gap between theoretical sufficient condition and empirical phase transition

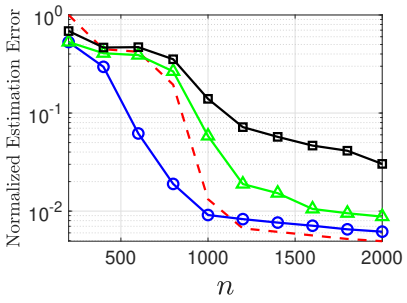
Algorithm 1: Iterative Anchored Regression (IAR)

Input: data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$; initialized parameter $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{kd}$; fidelity upper bound η; max. number of iterations I_{IAR}
Output: estimated parameter $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^{dk}$
for $i = 1$ **to** I_{IAR} **do**
    Compute anchor vector $\boldsymbol{a}$ from $\widetilde{\boldsymbol{\theta}}$ by (2.1.1)
    Estimate $\widehat{\boldsymbol{\theta}}$ by anchored regression in (4.2.2)
    $\widetilde{\boldsymbol{\theta}} \leftarrow \widehat{\boldsymbol{\theta}}$
end for

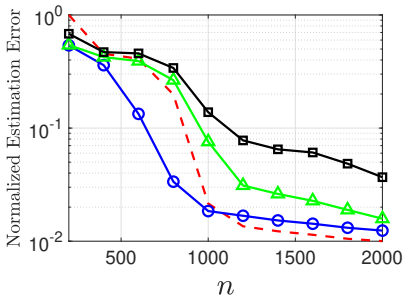
was observed for AM in the noise-free setting [37, Appendix L]. Overall, as shown in these figures, AR and AM provide similar empirical performance in the noiseless scenario.

In practice, observations are often corrupted with noise. Next, we study the estimation under two noise models. In these experiments, the ground-truth parameter vectors are i.i.d Normal$(\mathbf{0}, \boldsymbol{I}_{kd})$. Furthermore, to deduce statistical performance, the median of the estimation error in 50 trials is observed.

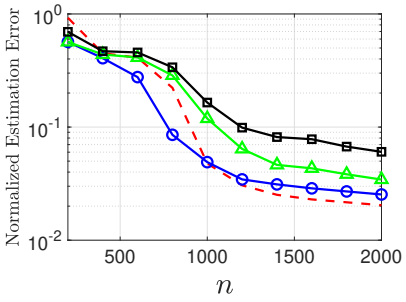
First, we consider the i.i.d. Gaussian noise model, i.e. $y_i = f_i(\boldsymbol{\theta}_\star) + z_i$, where $f_i(\boldsymbol{\theta}_\star)$ is defined in (1.2.9) and $\{z_i\}_{i=1}^n$ are i.i.d following Normal$(0, \sigma^2)$. To track the change of the estimation performance as a function of the noise strength, the dimension parameters are fixed as $d = 30$ and $k = 6$. AM has shown to be consistent, with an error rate that vanishes as n grows [38]. Its empirical estimation error decays similarly in the experiment. However, we observe that AR has a larger estimation error compared to AM, which remains nontrivial even for large n. We conjecture that this bias term is due to the regularizer with an imperfect anchor vector. In fact, as the anchor vector is obtained from a more accurate initial estimate, the result estimation error decays accordingly. Motivated by this observation, we consider a



(a) $\sigma = 0.05$



(b) $\sigma = 0.1$



(c) $\sigma = 0.2$

Figure 2.3: Estimation error versus the number of observations n under Gaussian noise of variance σ^2 ($k = 6$ and $d = 30$): repeated random initialization (black line with square markers), AR (green line with triangle markers), iterative AR (blue line and circle markers), and AM (red dashed line). All methods start from the repeated random initialization.

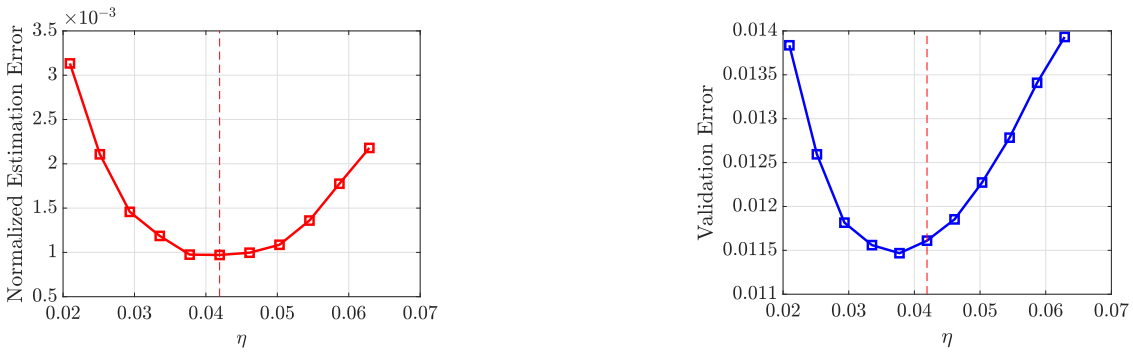
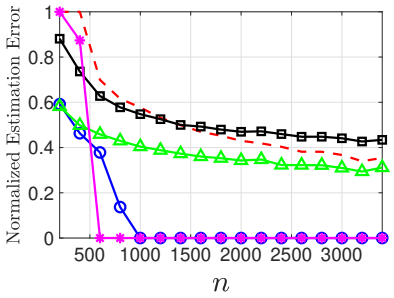
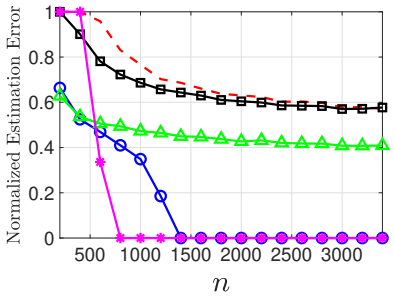


Figure 2.4: Estimation error and validation error via cross-validation by AR for varying η ($k = 3, d = 30$, and $n = 1,500$): The dotted vertical line indicates the location of $\eta_\star$ that achieves the equality in (2.1.7).

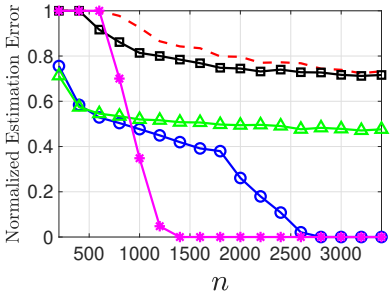
modification of AR with further iterative refinements, which we call the *iterative anchored regression* (IAR). The first iteration of IAR is equivalent to AR, but in the subsequent iterations, the anchor vector is refined by using the estimate from the previous iteration. The entire IAR algorithm is summarized in Algorithm 1. The number of iterations in IAR is set to $I_{\mathrm{IAR}} = 40$. Figure 2.3 shows that with more iterations the performance of iterative AR becomes as good as that of AM. Moreover, for small n (e.g. $n \leq 1,000$), IAR provides a smaller estimation error than AM. Moreover, we also study the sensitivity to the choice of the parameter η in (1.2.11). The need to tune this parameter can be a weakness of AR since AM does not involve any such parameter. As shown in Figure 2.4, the estimation error by AR does not critically depend on η. In this experiment, we vary η around $\eta_\star$ that achieves the equality in (2.1.7) with $\pm 50\%$ margin. Within this range, the estimation error remains small. Also, note that the minimum estimation error is achieved when η is slightly smaller than $\eta_\star$. It still remains to set the value of η within this range. Since the observations are corrupted with i.i.d. noise in this experiment, we applied a 5-fold



(a) $\varphi = 0.1$



(b) $\varphi = 0.2$



(c) $\varphi = 0.3$

Figure 2.5: Estimation error versus the number of observations n under multiplicative Bernoulli noise model with probability φ ($k = 6$ and $d = 30$): repeated random initialization (black line with square markers), AR (green line with triangle markers), IAR (blue line with circle markers), AM (red dashed line), and AM-LAD (magenta line with asterisk markers). All methods start from repeated random initialization.

cross-validation to estimate the validation error. Figure 2.4 suggests that choosing an η value that yields the smallest prediction error will likely result in the smallest estimation error.

Next, we study the empirical performance of the estimators under a gross error model. In Section 2.1, we have shown that the theoretical analysis of AR combined with the initialization by Ghosh et al. [38] applies to this model. Specifically, each observation is corrupted by a sparse noise according to the multiplicative Bernoulli

model with probability φ, that is, $\mathbb{P}\{y_i = -f_i(\boldsymbol{\theta}_\star)\} = \varphi$ and $\mathbb{P}\{y_i = f_i(\boldsymbol{\theta}_\star)\} = 1 - \varphi$ for $i \in [n]$. The multiplicative Bernoulli noise model has a similarity with the Massart noise [28, Definition 1.1]. Similar to the previous experiment, we compare AR to IAR and AM. Furthermore, we also study the performance of a variation of AM in which the least squares update is substituted by LAD. It will be denoted by AM-LAD. Figure 2.5 illustrates the estimation error in this setting where $d = 30$ and $k = 6$. Unlike the case of Gaussian noise, AR outperforms AM in the presence of multiplicative Bernoulli noise. Furthermore, IAR and AM-LAD achieve exact recovery over the range of φ in this experiment.

2.3 Discussion

As discussed in Section 2.2, the proposed convex estimator provides a comparable error bound relative to an oracle estimator in the adversarial noise case. However, it does not provide a consistent estimator with random noise. This inconsistency arises due to the maximization of the correlation with the anchor vector $\boldsymbol{a}$. Since the direction of the anchor vector does not coincide with the ground truth, the convex estimator introduces a bias. As a way to mitigate the bias in the convex estimator, we propose the iterative anchored regression that recursively refines the anchor vector to better align its direction with that of the ground truth. We have demonstrated that the iterative anchored regression empirically provides an exact recovery of the ground-truth parameters in the presence of outliers. Hence, it would be fruitful to pursue the theoretical analysis of the iterative anchored regression, particularly in terms of its behavior in the presence of outliers and random noise. Each iteration solves a linear program, which costs $\widetilde{O}\left(((n+d)k)^c\right)$ with $c \approx 2.38$ as discussed in

Section 2.1.3. Therefore, the per-iteration cost of the iterative anchored regression might be higher than that of the alternating minimization, which is $O(nkd^2)$. To further alleviate the computational cost of the iterative version, one might consider warm-start strategies in interior-point methods for linear programming (e.g. [56]).

Chapter 3: Max-affine regression by first-order methods

In this chapter, we present the results for the first-order methods for max-affine regression, with the problem formulation described in Section 1.2.2.

3.1 Convergence analysis of gradient descent

We first formulate the least squares estimator for max-affine regression and derive the gradient descent algorithm. For brevity, let $\boldsymbol{\xi} := [\boldsymbol{x};\ 1] \in \mathbb{R}^{d+1}$ and $\boldsymbol{\beta}_j := [\boldsymbol{\theta}_j;\ b_j] \in \mathbb{R}^{d+1}$. Then the model in (1.2.1) with an additive noise is rewritten as

$$y = \max_{j \in [k]} \langle \boldsymbol{\xi}, \boldsymbol{\beta}_j^\star \rangle + \text{noise}. \tag{3.1.1}$$

The least squares estimator minimizes the quadratic loss function given by

$$\ell(\boldsymbol{\beta}) := \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle \right)^2, \tag{3.1.2}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1;\ \ldots;\ \boldsymbol{\beta}_k] \in \mathbb{R}^{k(d+1)}$.

The gradient descent algorithm iteratively updates the estimate by

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \mu \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}^t),$$

where $\mu > 0$ denotes a step size. A generalized gradient [50] of the cost function in (3.1.2) with respect to the jth block $\boldsymbol{\beta}_j$ is written as

$$\nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \left(\max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle - y_i \right) \boldsymbol{\xi}_i, \tag{3.1.3}$$

where $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are defined in (1.2.4).

We show that the expression in (3.1.3) provides a valid generalized gradient of $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_\ell$. We apply the chain rule on the generalized gradient [50]. The cost function in (3.1.2) is the composition $\varrho \circ F$ where

$$\varrho((t_i)_{i=1}^n) = \frac{1}{2n} \sum_{i=1}^n t_i^2$$

and $\boldsymbol{\beta} \mapsto F(\boldsymbol{\beta}) = (f_i(\boldsymbol{\beta}))_{i=1}^n$ with

$$f_i(\boldsymbol{\beta}) = \left| \max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle - y_i \right|, \quad i \in [n].$$

Since each max-affine function f_i is regular at each point of the domain, the equality in [50, Eq. (5.7)] holds and it characterizes the generalized gradient of ℓ as

$$\nabla_{\boldsymbol{\beta}_\ell} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \left(\max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle - y_i \right) \cdot \nabla_{\boldsymbol{\beta}_\ell} \left(\max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \right).$$

Since a sub-gradient of a convex function is a generalized gradient [22], it suffices to show that $\mathbb{1}_{\{\boldsymbol{x}_i \in C_\ell\}} \boldsymbol{\xi}_i$ is a sub-gradient of the convex function $\nabla_{\boldsymbol{\beta}_\ell} \left(\max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \right)$. To this end, we verify that the following inequality holds for all $i \in [n]$:

$$\max \left(\langle \boldsymbol{\beta}_\ell + \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \max_{j \neq \ell \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \right) - \max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \geq \mathbb{1}_{\{\boldsymbol{x}_i \in C_\ell\}} \langle \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \quad \forall \boldsymbol{h} \in \mathbb{R}^{d+1}. \tag{3.1.4}$$

Let $i \in [n]$ be arbitrarily fixed. First, we consider the case when ℓ is a maximizer in the max-affine function in (3.1.1) at $\boldsymbol{\xi}_i$. Then we have $\langle \boldsymbol{\beta}_\ell, \boldsymbol{\xi}_i \rangle = \max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle$ and $\mathbb{1}_{\{\boldsymbol{x}_i \in C_\ell\}} = 1$. Therefore, (3.1.4) holds since

$$\max \left(\langle \boldsymbol{\beta}_\ell + \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \max_{j \neq \ell \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle \right) \geq \langle \boldsymbol{\beta}_\ell + \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \quad \forall \boldsymbol{h} \in \mathbb{R}^{d+1}.$$

Next, we assume that ℓ is not a maximizer. Then $\mathbb{1}_{\{\boldsymbol{x}_i \in C_\ell\}} = 0$ and there exists $\ell' \in [k] \setminus \{\ell\}$ such that $\langle \boldsymbol{\beta}_{\ell'}, \boldsymbol{\xi}_i \rangle = \max_{j \in [k]} \langle \boldsymbol{\beta}_j, \boldsymbol{\xi}_i \rangle > \langle \boldsymbol{\beta}_\ell, \boldsymbol{\xi}_i \rangle$. Therefore, (3.1.4) is also

satisfied since

$$\max\left(\langle \boldsymbol{\beta}_\ell + \boldsymbol{h}, \boldsymbol{\xi}_i \rangle, \langle \boldsymbol{\beta}_{\ell'}, \boldsymbol{\xi}_i \rangle\right) \geq \langle \boldsymbol{\beta}_{\ell'}, \boldsymbol{\xi}_i \rangle, \quad \forall \boldsymbol{h} \in \mathbb{R}^{d+1}.$$

Then the generalized gradient $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ is obtained by concatenating $\{\nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta})\}_{j=1}^k$ by

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{j=1}^k \boldsymbol{e}_j \otimes \nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}),$$

where $\boldsymbol{e}_j \in \mathbb{R}^k$ denotes the jth column of the k-by-k identity matrix $\boldsymbol{I}_k$ for $j \in [k]$. Moreover, $\ell(\boldsymbol{\beta})$ is differentiable except on a set of measure zero, with a slight abuse of terminology, $\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ is referred to as the "gradient".

Next, we present a convergence analysis of the gradient descent estimator. The analysis depends on a set of geometric parameters of the ground-truth model. The first parameter $\pi_{\min}$ describes the minimum portion of observations corresponding to the linear model which achieved the maximum least frequently. It is formally defined as a lower bound on the probability measure on the smallest partition set, i.e.

$$\min_{j \in [k]} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star) \geq \pi_{\min}, \tag{3.1.5}$$

where $\mathcal{C}_1^\star, \ldots, \mathcal{C}_k^\star$ are polytopes determined by

$$\mathcal{C}_j^\star := \{\boldsymbol{w} \in \mathbb{R}^d \ : \ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle > 0,\ \forall l < j,\ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle \geq 0,\ \forall l > j\}. \tag{3.1.6}$$

The next parameter κ quantifies the separation between all pairs of distinct linear models in (1.2.1) so that the pairwise distance on two distinct linear models satisfy

$$\min_{j' \neq j} \|(\boldsymbol{\beta}_j^\star)_{1:d} - (\boldsymbol{\beta}_{j'}^\star)_{1:d}\|_2 \geq \kappa. \tag{3.1.7}$$

Next, we present a convergence analysis of the gradient descent estimator. The analysis depends on a set of geometric parameters of the ground-truth model. The

first parameter $\pi_{\min}$ describes the minimum portion of observations corresponding to the linear model which achieved the maximum least frequently. It is formally defined as a lower bound on the probability measure on the smallest partition set, i.e.

$$\min_{j\in[k]} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star) \geq \pi_{\min}, \tag{3.1.8}$$

where $\mathcal{C}_1^\star, \ldots, \mathcal{C}_k^\star$ are polytopes determined by

$$\mathcal{C}_j^\star := \{\boldsymbol{w} \in \mathbb{R}^d \ : \ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle > 0,\ \forall l < j,\ \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle \geq 0,\ \forall l > j\}. \tag{3.1.9}$$

The next parameter κ quantifies the separation between all pairs of distinct linear models in (1.2.1) so that the pairwise distance on two distinct linear models satisfy

$$\min_{j'\neq j} \|(\boldsymbol{\beta}_j^\star)_{1:d} - (\boldsymbol{\beta}_{j'}^\star)_{1:d}\|_2 \geq \kappa. \tag{3.1.10}$$

Our main result in the following theorem presents a local linear convergence of the gradient descent estimator uniformly over all $\boldsymbol{\beta}^\star$ satisfying (3.1.9) and (3.1.10).

Theorem 8 *Let $\delta \in (0, 1/e)$, $y_i = \max_{j\in[k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star \rangle + z_i$ for $i \in [n]$ with $\boldsymbol{\xi}_i = [\boldsymbol{x}_i;\ 1]$, and $\{z_i\}_{i=1}^n$ being additive σ-sub-Gaussian noise independent from everything else. Suppose that Assumptions 1 and 2 hold.*[5] *Then there exist absolute constants $C, C', R > 0$, and $\nu \in (0, 1)$, for which the following statement holds with probability at least $1 - \delta$: If the initial estimate $\boldsymbol{\beta}^0$ belongs to a neighborhood of $\boldsymbol{\beta}^\star$ given by*

$$\mathcal{N}(\boldsymbol{\beta}^\star) := \left\{ \boldsymbol{\beta} \in \mathbb{R}^{k(d+1)} \ : \ \max_{j\in[k]} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\|_2 \leq \kappa\rho \right\} \tag{3.1.11}$$

with

$$\rho := \frac{R\pi_{\min}^{\zeta^{-1}(1+\zeta^{-1})}}{4k^{\zeta^{-1}}} \cdot \log^{-1/2}\left(\frac{k^{\zeta^{-1}}}{R\pi_{\min}^{\zeta^{-1}(1+\zeta^{-1})}} \right) \wedge \frac{1}{4}, \tag{3.1.12}$$

[5]To simplify the presentation, we assume that the parameters η, ζ, γ in Assumptions 1 and 2 are fixed numerical constants in the statement and proof of Theorem 8. Therefore, any constant determined only by η, ζ, γ will be treated as a numerical constant.

then for all $\boldsymbol{\beta}^\star$ *satisfying* (3.1.8) *and* (3.1.10)*, the sequence* $(\boldsymbol{\beta}^t)_{t\in\mathbb{N}}$ *by the gradient descent method with a constant step size satisfies*

$$\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\right\|_2 \leq \nu^t \left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\right\|_2 + C'\sigma k \frac{\sqrt{k\left(kd\log(n/d) + \log(k/\delta)\right)}}{\sqrt{n}}, \quad \forall t \in \mathbb{N}, \tag{3.1.13}$$

provided that

$$n \geq C\pi_{\min}^{-2(1+\zeta^{-1})} \cdot \left(k^{1.5}\pi_{\min}^{-(1+\zeta^{-1})} \vee \frac{\sigma}{\kappa\rho}\right)^2 \cdot \left(kd\log(n/d) + \log(k/\delta)\right). \tag{3.1.14}$$

Proof 3 *See Appendix C.3.*

Theorem 8 demonstrates that the GD estimator with a constant step size converges linearly to a neighborhood of the ground-truth parameter of radius $\widetilde{O}\left(\sigma^2 k^4 d/n\right)$. The number of sufficient observations to invoke this convergence result scales linearly in d and is proportional to a polynomial in $\pi_{\min}^{-1}$ and k. This result implies the consistency of the gradient descent estimator. To compare Theorem 8 to the analogous result for AM under the same covariate and noise models [36, Theorem 1], we have the following remarks in order.

- First, the final estimation error by (3.1.13) with $t \to \infty$ is smaller than that by [36, Theorem 1] by being independent of $\pi_{\min}^{-1}$, which grows at least proportional to k. A larger estimation error bound in their result is due to the analysis of the least squares update, wherein the smallest singular value of the design matrix of each linear model is utilized. These quantities do not appear in the analysis of the gradient descent update.

- Second, the convergence parameter ν in (3.1.13) is smaller than 3/4 for AM[6], which might result in a slower convergence of GD in iteration count. The convergence speed issue becomes significant for large k and $\pi_{\min}^{-1}$. For example, in the illustration by Figure 1.3, GD shows a slower convergence in run time despite the lower per-iteration cost $O(knd)$, which is lower than that of AM $O(knd^2)$ by a factor of d. However, as discussed in Section 3.2, the slow convergence of GD can be improved by modifying the algorithm into a (mini-batch) SGD.

- Third, the sample complexity results by Theorem 8 and [36, Theorem 1] are qualitatively comparable. There were mistakes in the proof of [36, Theorem 1]. We think that their result could be corrected with an increased order of dependence in their sample complexity on k and $\pi_{\min}$ (see Appendix C.5 for a detailed discussion).

- Lastly, regarding the proof technique, we adapt and improve the strategy by Ghosh et al. [36, 37]. Note that the subgradient of the loss function in (3.1.3) involves clustering of covariates with respect to maximizing linear models such as (1.2.4), which also arises in alternating minimization. Due to this similarity, key quantities in the analysis have been estimated in [36, 37]. We provide sharpened estimates via different techniques. For example, Theorem 26 provides a tighter bound than [37, Lemma 7] by a factor of $\alpha^{\zeta^{-1}}$ for a scalar $\alpha \in (0, 1)$.

[6]As shown in the proof in Appendix C.3, the parameter ν is given as $\nu = (1 - \mu\lambda)$ by (C.3.19). The quantity $\mu\lambda$ is determined by (C.3.8) and (C.3.29) as a function of $\pi_{\min}$, $\pi_{\max}$, and ζ so that it decreases in k and $\pi_{\min}^{-1}$.

Theorem 8 also provides an auxiliary result. As a direct consequence of Theorem 8, we obtain an upper bound on the prediction error, which is defined by

$$\mathcal{E}(\widehat{\boldsymbol{\beta}}) := \mathbb{E}\left(\max_{j\in[k]}\langle \boldsymbol{\xi}, \widehat{\boldsymbol{\beta}}_j\rangle - \max_{j\in[k]}\langle \boldsymbol{\xi}, \boldsymbol{\beta}_j^\star\rangle\right)^2,$$

where $\widehat{\boldsymbol{\beta}} = [\widehat{\boldsymbol{\beta}}_1;\ \ldots;\ \widehat{\boldsymbol{\beta}}_k]$ denotes the estimated parameter vector by GD. Since the quadratic cost function in (1.2.3) is 1-Lipschitz with respect to the ℓ_2 norm, it follows that the prediction error $\mathcal{E}(\widehat{\boldsymbol{\beta}})$ is also bounded by $\widetilde{O}(\sigma^2 k^3 d/n)$ as in (3.1.13) with $t \to \infty$.

A limitation of Theorem 8 is that its local convergence analysis requires an initialization within a specific neighborhood of the ground-truth parameter. To obtain the desired initial estimate, one may use spectral initialization by [38, Algorithm 2, 3], which consists of dimensionality reduction followed by a grid search. They provided a performance guarantee of a spectral initialization scheme under the standard Gaussian covariate assumption [38, Theorems 2 and 3]. Therefore, the reduction of Theorem 8 to the Gaussian covariate case combined with [38, Theorems 2 and 3] provides a global convergence analysis of GD, which is comparable to that for alternating minimization [38]. Even in this case, the number of sufficient samples for the success of spectral initialization overwhelms that for the subsequent gradient descent step. Since multiple steps of their analysis critically depend on the Gaussianity, it remains an open question whether the result on the spectral initialization generalizes to the setting by Assumptions 1 and 2.

3.2 Convergence analysis of mini-batch SGD

SGD is an optimization method that updates parameters using a single or a small batch of randomly selected data point(s) instead of the entire dataset. SGD

converges faster in run time than GD due to its significantly lower per-iteration cost. In particular, when applied to max-affine regression, SGD empirically outperforms GD and AM in both sample complexity and convergence speed (see Figures 1.3, 3.2 and 3.4). In this section, we present an accompanying theoretical convergence analysis of mini-batch SGD for max-affine regression. The update rule of a mini-batch SGD with batch size m for max-affine regression is described as follows. For each iteration index $t \in \mathbb{N}$, let I_t be a multiset of m randomly selected indices with replacement so that the entries of I_t are independent copies of a uniform random variable in $[n]$. A mini-batch SGD iteratively updates the estimate by

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \mu \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}^t),$$

where

$$\ell_i(\boldsymbol{\beta}) := \frac{1}{2} \left(y_i - \max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle \right)^2, \quad i \in [n].$$

Then the following theorem presents a local linear convergence of SGD.

Theorem 9 *Under the hypothesis of Theorem 8, there exist absolute constants $C, C' > 0$ and $c, \nu \in (0,1)$, for which the following statement holds with probability at least $1 - \delta$: For all $\boldsymbol{\beta}^\star$ satisfying* (3.1.9) *and* (3.1.10)*, if the initial estimate $\boldsymbol{\beta}^0$ belongs to $\mathcal{N}(\boldsymbol{\beta}^\star)$ defined in* (3.1.11)*, n satisfies* (3.1.14)*, and m satisfies*

$$m \geq C \cdot \left(\frac{\sigma}{\kappa \rho} \right)^2 \cdot (d + \log(k/\delta)), \tag{3.2.1}$$

then the sequence $(\boldsymbol{\beta}^t)_{t\in\mathbb{N}}$ *by the mini-batch SGD with batch size* m *and step size* $\mu = c\,(1 \wedge m/(d+\log(n/\delta)))$ *satisfies*

$$\mathbb{E}_{I_t}\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\right\|_2 \leq \left(1 - \left(1 \wedge \frac{m}{d+\log(n/\delta)}\right) c\nu\right)^t \left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\right\|_2 + C'\sigma k\sqrt{\left(\frac{d+\log(n/\delta)}{m} \vee \frac{kd\log(n/d)+\log(1/\delta)}{n}\right)}, \quad \forall t \in \mathbb{N}. \tag{3.2.2}$$

Proof 4 *See Appendix C.4.*

Theorem 9 establishes linear convergence of mini-batch SGD in expectation to the ground-truth parameters within error $\widetilde{O}(\sigma^2 k^2\,(d/m \vee kd/n))$. The local linear convergence applies uniformly over all $\boldsymbol{\beta}^\star$ satisfying (3.1.9) and (3.1.10). In general, the convergence rate of SGD is much slower even with strong convexity [12, 49, 71]. However, in a special case where the cost function is in the form of $\sum_{i=1}^n \ell_i(\boldsymbol{\beta})$, smooth, and strongly convex, if $\boldsymbol{\beta}^\star$ is the minimizer of all summands $\{\ell_i(\boldsymbol{\beta})\}_{i=1}^n$, then SGD converges linearly to $\boldsymbol{\beta}^\star$ [70, Theorem 2.1]. The convergence analysis in Theorem 9 can be considered along with this result. The cost function in (3.1.2) in the noiseless case satisfies the desired properties locally near the ground truth, whence establishes the local linear convergence of SGD.

Theorem 9 also explains how the batch size m affects the final estimation error by (3.2.2) with $t \to \infty$. Let n and m satisfy (3.1.14) and (3.2.1) so that Theorem 9 is invoked. Under this condition, one can still choose m and n so that $m \lesssim n/k$. Then the $\widetilde{O}(\sigma^2 k^2 d/m)$ term determined by the batch size m dominates the final estimation error. In this regime, the SGD estimator is not consistent since the estimation error $\widetilde{O}(\sigma^2 k^2 d/m)$ does not vanish with increasing n. This result implies the trade-off

between the convergence speed and the final estimation error determined by the batch size.

Furthermore, since the condition on m in (3.2.1) becomes trivial when $\sigma = 0$, we obtain a stronger result in the noiseless case given by the following corollary.

Corollary 10 *Let $\delta, \delta' \in (0,1)$, and $\epsilon > 0$ fixed. Suppose that the hypothesis of Theorem 9 holds. If $t \geq (\log(1/\epsilon) + \log(1/\delta)) \left(1 \vee \frac{d+\log(n/\delta)}{m}\right) 1/\nu$, then*

$$\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\right\|_2 \leq \epsilon \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2$$

holds with probability at least $1 - \delta - \delta'$.

Proof 5 *By Theorem 9,* (3.2.2) *holds with probability at least $1 - \delta$. By applying Markov's inequality, we have*

$$\mathbb{P}\left(\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\right\|_2 \geq \epsilon \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2\right) \leq \frac{\mathbb{E}_{I_t}\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|_2}{\epsilon \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2} \leq \frac{\left(1 - \left(1 \wedge \frac{m}{d+\log(n/\delta)}\right)\nu\right)^t}{\epsilon} \leq \delta',$$

where the second and third inequalities hold by (3.2.2) *and assumption on t respectively.*

Theorem 10 presents the convergence of SGD with high probability, which is stronger than the convergence in expectation. Furthermore, there is no requirement on the batch size in invoking Theorem 10. This result is analogous to the recent theoretical analysis of phase retrieval by randomized Kaczmarz [85] and SGD [84].

3.3 Numerical results

We study the empirical performance of GD and mini-batch SGD for max-affine regression. The performance of these first-order methods is compared to AM [38]. We use a constant step size 0.5 for GD. The step size for SGD is set to $\frac{1 \wedge (m/d)}{2}$

adaptive to the batch size. In the synthetic data experiment, according to our covariate assumptions in Assumption 1 and Assumption 2, we consider the following two scenarios; The first scenario involves Gaussian covariates, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are generated as independent samples from a random vector following Normal$(\mathbf{0}, \mathbf{I}_d)$. The other scenario involves a uniform distribution, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are generated as independent samples from a random vector following Unif$[-\sqrt{3}, \sqrt{3}]^{\otimes d}$, which is also considered in the numerical setting in [37]. We use spectral initialization for the Gaussian covariate model [37], while for the uniform distribution case, we apply the multiple-restart random initialization method [7]. Next, in the real data experiment, we use the mean weekly wages and Boston housing pricing datasets, as presented in Figure 1.2. We apply random initialization, which is commonly used in practice.

3.3.1 Synthetic data experiments

First, we observe the performance of the three estimators for the exact parameter recovery in the noiseless case. In this experiment, the ground-truth parameters $\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_k^\star$ are generated as k random pairwise orthogonal vectors with $k < d$, and the offset terms are set to 0, i.e., $b_j^\star = 0$ for all $j \in [k]$. By the construction, the probability assigned to the maximizer set of each linear model will be approximately $\frac{1}{k}$. In other words, the parameters $\pi_{\max}$ and $\pi_{\min}$ of the ground truth concentrate around $\frac{1}{k}$ where $\pi_{\min}$ is defined in (3.1.8) and $\pi_{\max} := \max_{j\in[k]} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)$. Furthermore, due to the orthogonality, the pairwise distance satisfies $\|\boldsymbol{\theta}_j^\star - \boldsymbol{\theta}_{j'}^\star\|_2 = \sqrt{2}$ for all $j \neq j' \in [k]$. Consequently, the sample complexity results for GD and SGD by Theorem 8 and Theorem 9 simplify to an easy-to-interpret expression $\widetilde{O}(k^{16}d)$ that involves only k

and d for both Gaussian and uniform distribution scenarios. The sample complexity result on AM [37] simplifies similarly.

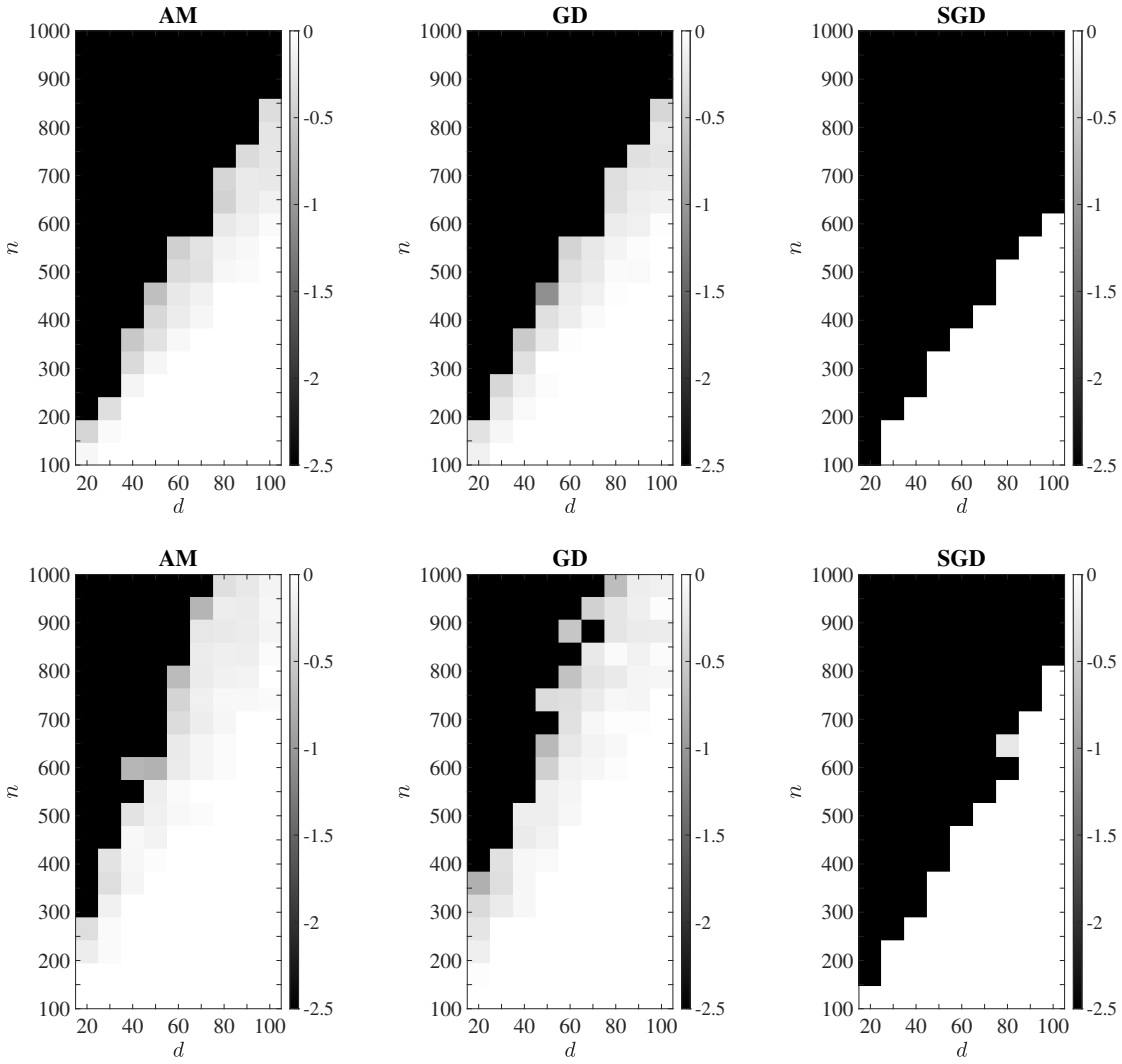


Figure 3.1: Gaussian covariate

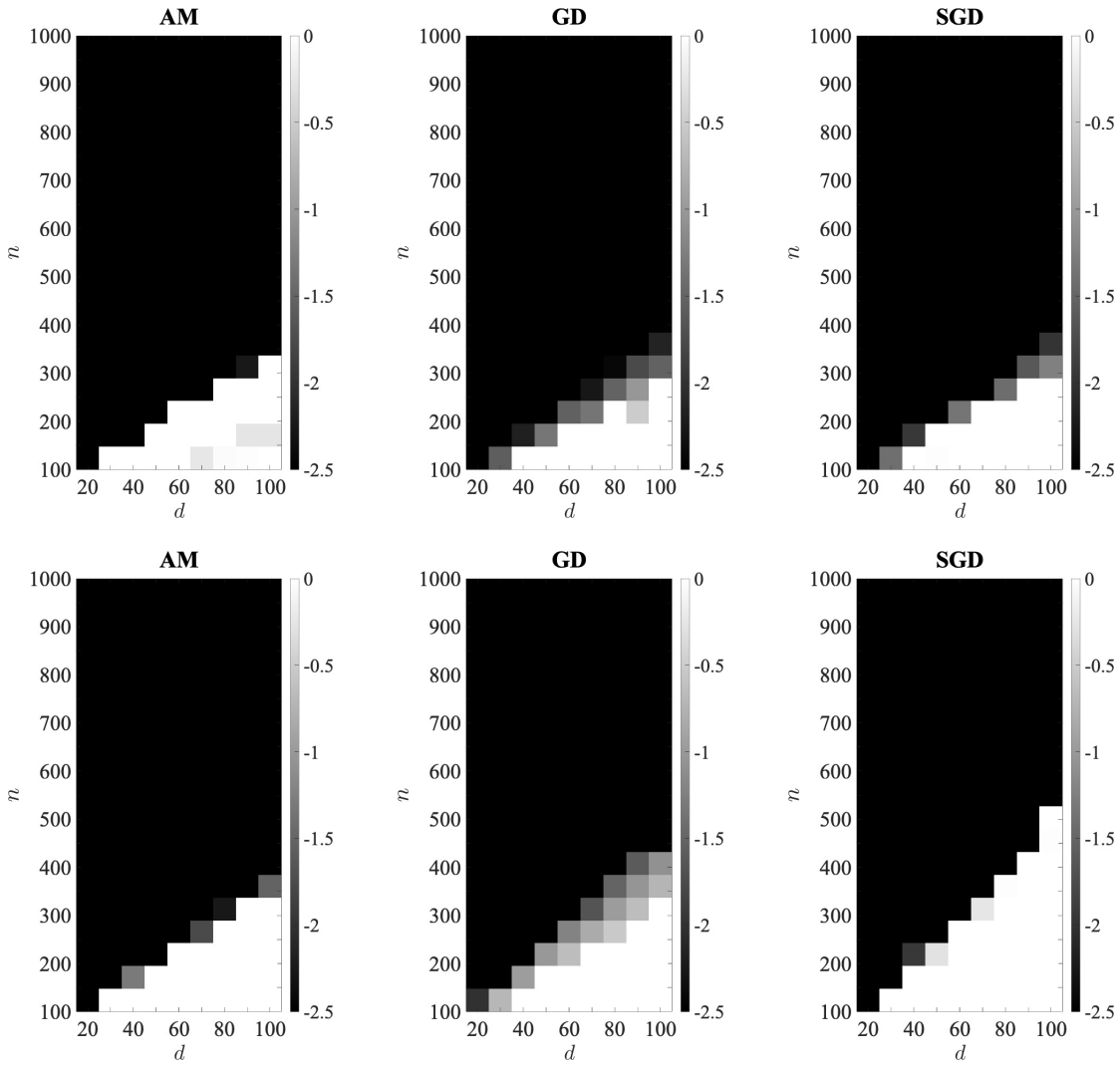


Figure 3.1: Uniform covariate

Figure 3.2: Phase transition of estimation error per the number of observations n and the ambient dimension d in the noiseless case (The number of linear models k and the batch size m are set to 3 and 64, respectively). The first row and the second row respectively show the median and the 90th percentile of estimation errors in 50 trials.

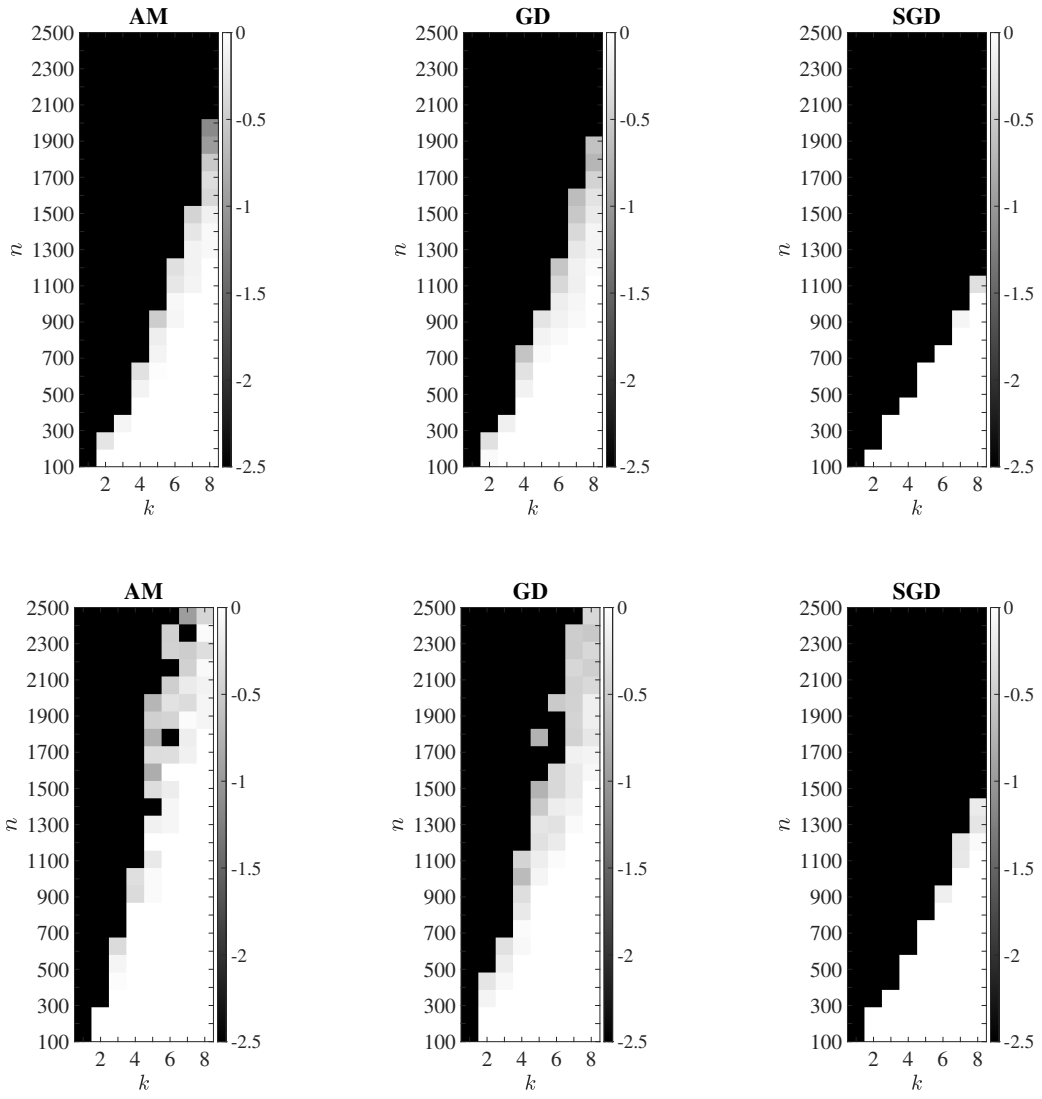


Figure 3.3: Gaussian covariate

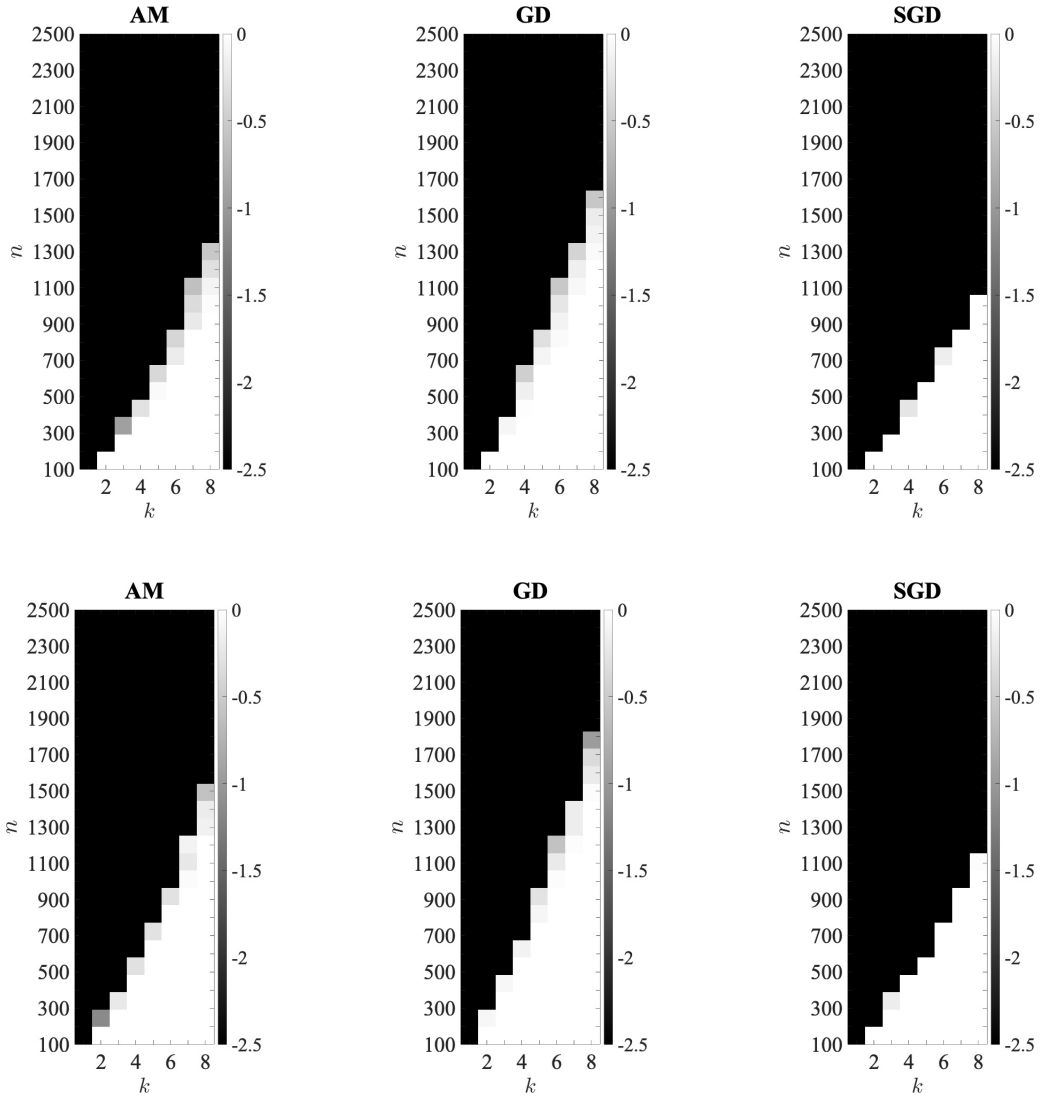


Figure 3.3: Uniform covariate

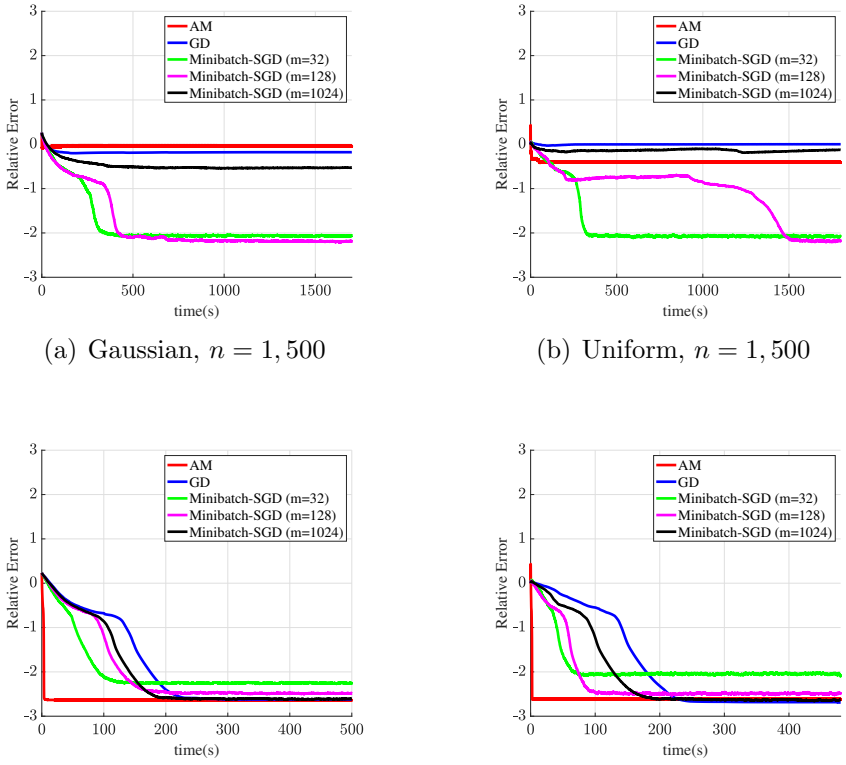
Figure 3.4: Phase transition of estimation error per number of observations n and number of linear models k in the noiseless case (The ambient dimension d and mini-batch size m are set to 50 and 64 respectively). The first row and the second row respectively show the median and the 90th percentile of estimation errors in 50 trials.

Figures 3.1 and 3.3 illustrate the empirical phase transition by the three estimators through Monte Carlo simulations under the Gaussian covariate model. The median

and the 90th percentile of 50 random trials are displayed. In these figures, the transition occurs when the sample size n becomes larger than a threshold that depends on the ambient dimension d and the number of linear models k. Figure 3.1 shows that the threshold for both estimators increases linearly with d for fixed k. This observation is consistent with the sample complexity by Theorem 8 and Theorem 9. A complementary view is presented in Figure 3.3 for varying k and fixed d. The thresholds in Figure 3.3 for GD and SGD are almost linear in k when d is fixed to 50, which scales slower than the corresponding sample complexity results in Theorem 8 and Theorem 9. A similar discrepancy between theoretical and empirical phase transitions has been observed for AM [37, Appendix L]. We also observe that mini-batch SGD outperforms GD and AM with a lower threshold for phase transition. It has been shown that the inherent random noise in the gradient helps the estimator to escape saddle points or local minima [25, 55]. This explains why SGD recovers the parameters with fewer samples than GD. We also note that the relative performance among the three estimators remains similar in both the median and the 90th percentile. This shows that SGD for noiseless max-affine regression does not suffer from a large variance, which corroborates the result in Corollary 10.

The phase transition boundaries in Figures 3.1 and 3.3 are higher with a larger success regime relative to the corresponding results in Figures 3.1 and 3.3. Recall that GD/SGD with the multiple-restart random initialization involves multiple runs of GD/SGD. The performance improvement is obtained at the cost of higher computational cost proportional to the number of repetitions.

Figures 3.5 and 3.6 study the estimation error by mini-batch SGD under zero-mean Gaussian noise with standard deviation $\sigma = 0.1$ in three different scenarios.

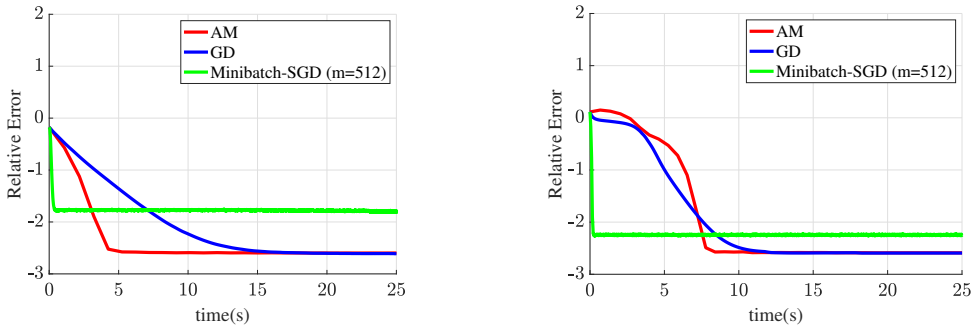


(a) Gaussian, $n = 1,500$ (b) Uniform, $n = 1,500$

(c) Gaussian, $n = 3,000$ (d) Uniform, $n = 3,000$

Figure 3.5: Convergence of estimators for max-affine regression under additive white Gaussian noise of variance $\sigma^2 = 0.01$ ($k = 8$ and $d = 50$). Comparison between Gaussian and Uniform covariates.

In Figure 3.5, we focus on observing how the batch size m affects the convergence speed and the estimation error. Figure 3.5(a) and Figure 3.5(b) consider the scenario where the spectral method provides a poor initialization due to a small number of observations. Consequently, GD and AM fail to provide a low estimation error. In contrast, mini-batch SGD with a small batch size ($m = 32$ or $m = 128$) relative to the total number of samples ($n = 1,500$) converges to a small estimation error ($< 10^{-2}$). In other words, there exists a trade-off between the convergence speed and the estimation error determined by the batch size m. SGD with $m = 128$ converges



(a) Gaussian distribution (b) Uniform distribution

Figure 3.6: Convergence of estimators for max-affine regression under additive white Gaussian noise of variance $\sigma^2 = 0.01$ ($k = 3$, $d = 500$, and $n = 8,000$).

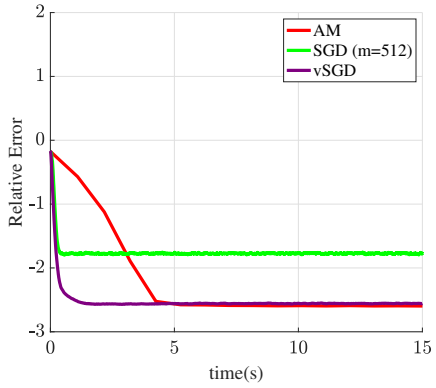


Figure 3.7: Comparison of vSGD, SGD, and AM for max-affine regression with Gaussian covariates under additive white Gaussian noise with variance $\sigma^2 = 0.01$ ($k = 3$, $d = 500$, and $n = 8,000$). vSGD starts with $m = 16$ and doubles m every 50 epochs.

slower to a smaller error than SGD with $m = 32$. This corroborates the theoretical result in Theorem 9. However, as the batch size m further increases to $m = 1,024$ close to $n = 1,500$, SGD starts to fail like GD and AM. Again, this phenomenon is

explained by the fact that the noisy gradient in SGD avoids saddle points and local minima efficiently [25, 55].

For the Gaussian and uniform covariates, Figure 3.5(c) and Figure 3.5(d) illustrate the comparison in a high-sample regime, where the number of samples is twice larger than that for Figure 3.5(a) and Figure 3.5(b), respectively. In this case, both GD and AM converge to a smaller error than SGD. Moreover, AM converges faster than the other algorithms in the run time, which is explained by the following two reasons. First, as discussed in Section 3.1, AM converges faster than GD and SGD in the iteration count with a smaller constant for linear convergence. Second, due to the small ambient dimension ($d = 50$), the gain in the per-iteration cost of SGD $O(kmd)$ over that of AM $O(knd^2)$ is not significant.

Lastly, Figure 3.6, compares the convergence of the estimators in the presence of noise when d, k, and n are set as in Figure 1.3. On one hand, SGD converges faster than AM with a significantly lower per-iteration cost $O(kmd)$ than $O(knd^2)$ due to the large ambient dimension ($d = 500$) and small batch size ($m = 512$ compared to $n = 8,000$). On the other hand, SGD yields a larger error than the other two estimators. The estimation error bound of SGD, as described in Theorem 9, is affected by m and behaves similarly in this case. To address the large error in SGD caused by the mini-batch size while maintaining fast convergence, we empirically propose SGD with a variable step size (vSGD). vSGD begins with a small m and gradually increases m to improve the estimator's accuracy. Figure 3.6 compares vSGD with both SGD and AM, showing that vSGD not only achieves the accuracy of AM but also converges at a speed comparable to SGD.

3.3.2 Real economic data experiments

The real data experiments are prediction tasks for mean weekly wages based on years of education and experience and Boston housing price with respect to lower population status and average of rooms per dwelling. Mean weakly wages dataset, provided in [76, Chapter 10, Exercise 29], contains $25,631$ records of weekly wages for adult males aged between 18 and 70 who worked full-time in the US, along with years of experience and education. The Boston housing dataset [48] includes various features related to housing in the Boston area. For this experiment, we focus on two features: the percentage of the population with lower socioeconomic status and the average number of rooms per dwelling. As shown in Figure 1.2, these data sets show shape restrictions, and the max-affine model is well-suited to fit this data. After normalizing the dataset for both covariates and responses, we applied SGD and AM with $k = 6$ and evaluated performance using RMSE in 5-fold cross-validation across 100 random initializations. The box plot results in Figures 3.8 and 3.9 show the performance across 100 runs. SGD provides more stable estimates that are less affected by initialization, whereas the performance of AM varies with initialization, consistent with observations in [65]. This may come from the fact that SGD escapes local minima well [25, 55], while AM is likely to get stuck in local minima depending on the initialization.

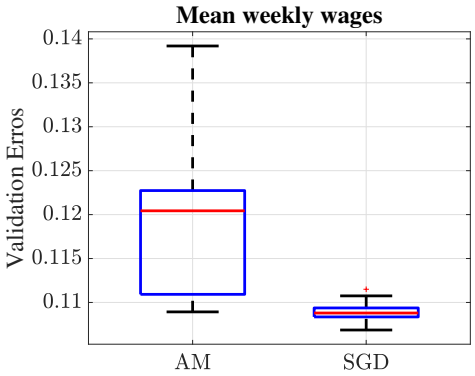


Figure 3.8: Box plot of RMSEs for mean weekly wages across 100 initializations.

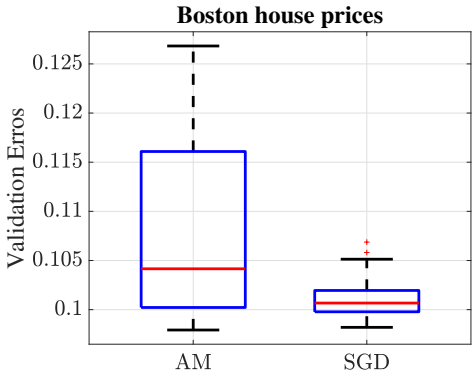


Figure 3.9: Box plot of RMSEs for Boston housing prices across 100 initializations.

3.4 Summary

We have established a local convergence analysis of GD and SGD for max-affine regression under a relaxed covariate model with σ-sub-Gaussian noise. The covariate distribution, characterized by sub-Gaussianity and anti-concentration, extends beyond the standard Gaussian model. It has been shown that suitably initialized GD and SGD converge linearly to within a non-asymptotic error bound, comparable to the analogous result for AM. Notably, when applied to noiseless max-affine regression, SGD empirically outperforms both GD and AM in terms of sample complexity and convergence speed. Furthermore, in the presence of noise, we show that variable batch size strategies for SGD converge faster than AM while achieving the same accuracy. In the special case of the Gaussian covariate model, the spectral method proposed by Ghosh et al. [38] can provide the desired initial estimate. Extending their theoretical results on the spectral method to the relaxed covariate model would be of great interest.

On the practical side, we demonstrate that SGD provides more stable estimates than AM in real data experiments when the algorithms are initialized randomly.

Chapter 4: Future Work

In this chapter, we outline a potential research direction, motivated by the empirical results in this Thesis.

4.1 Motivation

Figure 2.5 shows that in the presence of outliers, alternating minimization with LAD empirically outperforms competing methods, including the iterative version of the convex program in Algorithm 1. While the hyperparameter η for the convex program in (1.2.10) can be tuned via cross-validation in the presence of stochastic noise (as shown in Figure 2.4), this may not be feasible with outliers. In contrast, alternating minimization with LAD has no tuning parameter. These observations suggest that alternating minimization with LAD is a promising method for max-affine regression in the presence of outliers. Consequently, our future goal is to study alternating minimization with LAD, which we refer to as RobustPA, as it can be seen as a robust version of LSPA [65].

4.2 Algorithm for RobustPA and preliminary real data experiment results

We describe the RobustPA algorithm in more detail. Analogous to LSPA (equivalent to the AM algorithm), Robust-AM consists of two main steps. Using the notation

from (3.1.1), the first step partitions the covariates $(\boldsymbol{\xi}_i)_{i=1}^n$ into regions based on the current estimates. Fix $t \in \mathbb{N}$ arbitrarily. Suppose we are at the t-th iteration and have an estimator $\boldsymbol{\beta}^t$ at this step. We partition the covariates $(\boldsymbol{\xi}_i)_{i=1}^n$ into disjoint polyhedral cones $\mathcal{C}_1^t, \ldots, \mathcal{C}_k^t$, defined by substituting $\boldsymbol{\beta}$ in (1.2.4) with $\boldsymbol{\beta}^t$. The second step is to estimate the parameters $(\boldsymbol{\beta}_j^{t+1})_{j=1}^k$ based on the partitions $(\mathcal{C}_j^t)_{j=1}^k$:

$$\boldsymbol{\beta}_j^{t+1} \in \operatorname*{argmin}_{\boldsymbol{\beta}_j \in \mathbb{R}^{d+1}} \sum_{i=1}^n \mathbf{1}_{\left\{\boldsymbol{x}_i \in \mathcal{C}_j^t\right\}} \left| y_i - \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle \right|, \quad \forall j \in [k]. \tag{4.2.1}$$

The alternating minimization with LAD for max-affine regression can be derived by applying the Gauss-Newton method. This procedure parallels the approach in Section 1.3 but substitutes the linear measurement model (1.1.1) with the max-affine model (1.2.1). Solving (4.2.1) is equivalent to linear regression for LAD. Hence, one can use a linear program, alternating direction method of multipliers (ADMM), or iteratively reweighted least squares (IRLS). Here, we focus on the linear programming approach. By introducing auxiliary variables $\boldsymbol{t} := [t_1; \ldots; t_n] \in \mathbb{R}^n$, the optimization in (4.2.1) can be reformulated as:

$$\begin{aligned} &\underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d+1}, (t_i)_{i=1}^n}{\text{minimize}} \quad \langle \mathbf{1}_n, \boldsymbol{t} \rangle \\ &\text{subject to} \quad \begin{cases} t_i \geq y_i - \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle, & \text{if } \boldsymbol{x}_i \in \mathcal{C}_j^t, \\ t_i \geq -y_i + \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle, & \text{if } \boldsymbol{x}_i \in \mathcal{C}_j^t, \\ t_i = 0, & \text{otherwise} \end{cases} \quad \forall i \in [n],\, j \in [k], \end{aligned} \tag{4.2.2}$$

where $\mathbf{1}_n = [1, \ldots, 1]^\mathsf{T} \in \mathbb{R}^n$. The complete RobustPA algorithm is detailed in Algorithm 2.

To further motivate our goal, we compare the performance of RobustPA to AM [37] and SGD Section 3.2 in real data experiments using the datasets from Section 3.3.2. To contaminate the datasets, we amplify the response values by a factor of 15 in 10% of the data points selected randomly. Figure 4.1 visualizes the contaminated mean

Algorithm 2: RobustPA for Max-Affine Regression

Input: dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$; initial parameter estimates $\boldsymbol{\beta}^0$;
while *stop condition is not satisfied* **do**
  for $j = 1; j \le k; j = j + 1$ **do**
    Update $\boldsymbol{\beta}_j^{t+1}$ using (4.2.2):

$$\boldsymbol{\beta}_j^{t+1} \in \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \mathbf{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^t\}} \left| y_i - \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle \right|.$$

  end
  $t \leftarrow t + 1$
end
Output: final estimate $\widehat{\boldsymbol{\beta}} \leftarrow [\boldsymbol{\beta}_1^t; \ldots; \boldsymbol{\beta}_k^t]$

weekly wages data points, where the red points represent the amplified outliers. As shown in Figure 4.2, AM, which is designed for least squares (1.2.3), is sensitive to outliers, leading to poor model fitting. However, as seen in Figure 4.3, RobustPA provides a model that fits the majority of the data and is less affected by outliers.

We applied SGD, AM, and RobustPA with $k = 6$ and evaluated performance using median absolute error in 5-fold cross-validation across 100 random initializations. Figures 4.4 and 4.5 show that RobustPA outperforms the competing algorithms in this setting.

4.3 Problem setting for max-affine regression in the presence of outliers

Having observed the empirical success of RobustPA, we now seek to analyze the algorithm theoretically. For a formulation with measurements corrupted by sparse outliers, as in robust phase retrieval in (1.1.2), we consider robust max-affine regression

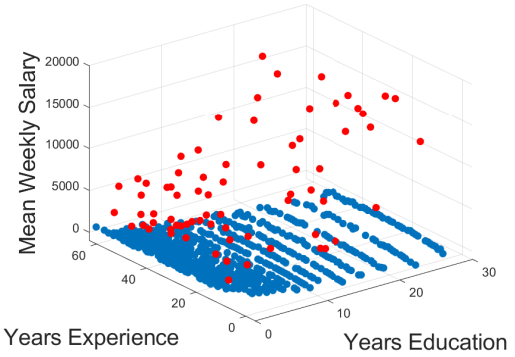


Figure 4.1: Mean weekly wages data with 10% outliers (in red)

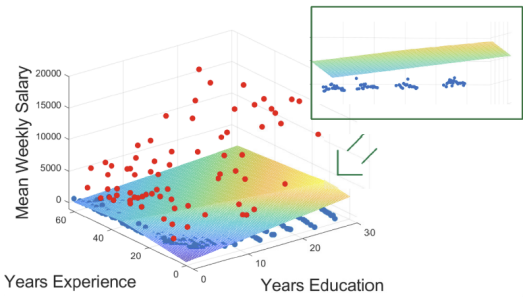


Figure 4.2: Fitted max-affine model by AM for mean weekly wages data with 10% outliers

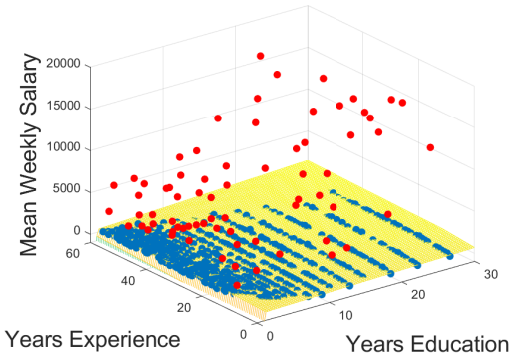


Figure 4.3: Fitted max-affine model by RobustPA for mean weekly wages data with 10% outliers

where the observations are corrupted by sparse outliers:

$$y_i = \begin{cases} \chi_i, & \text{if } i \in I_{\text{out}}, \\ \max_{j\in[k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star \rangle, & \text{if } i \in I_{\text{in}}, \end{cases} \tag{4.3.1}$$

where $I_{\text{out}} \subset [n]$ and $I_{\text{in}} := [n] \setminus I_{\text{out}}$ represent the unknown indices of outliers and inliers, respectively. The values of the outliers $(\chi_i)_{i\in I_{\text{out}}}$ are arbitrary in $\mathbb{R}$. The goal is to estimate the ground-truth parameters $(\boldsymbol{\beta}_j^\star)_{j=1}^k$ from the observations $(\boldsymbol{\xi}_i, y_i)_{i=1}^n$.

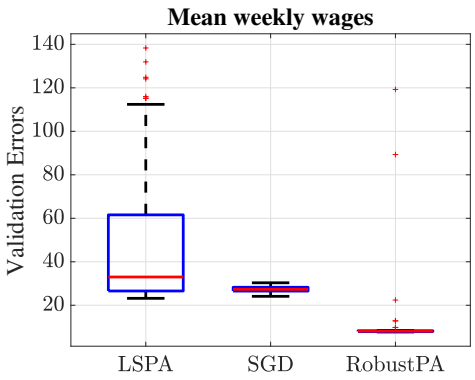


Figure 4.4: Performance of AM, SGD, and AM with LAD on mean weekly wages data

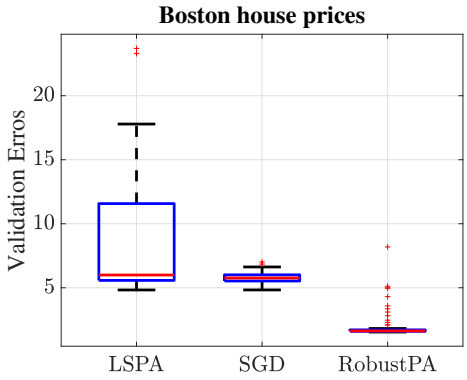


Figure 4.5: Performance of AM, SGD, and AM with LAD on Boston housing data

In the formulation (4.3.1), it is important to analyze the theoretical performance of RobustPA, particularly with respect to local convergence and sample complexity. Our first goal is to establish local analyses that may demonstrate convergence and sample complexity similar to the theoretical results studied in this thesis. Preliminary numerical results for local convergence in Figure 4.6 indicate that RobustPA converges to the ground truth at a rate faster than linear. Furthermore, Figure 4.7 shows that RobustPA also achieves global convergence from random initialization. The phase transition in Figure 4.8 shows that the sample complexity depends on the dimension linearly, resulting in (near) optimal sample complexity.

We summarize our future goals as follows:

- The first goal is to establish the local analysis of RobustPA. Based on preliminary numerical results, we expect that RobustPA locally converges to the ground truth at a rate faster than linear with (near) optimal sample complexity.

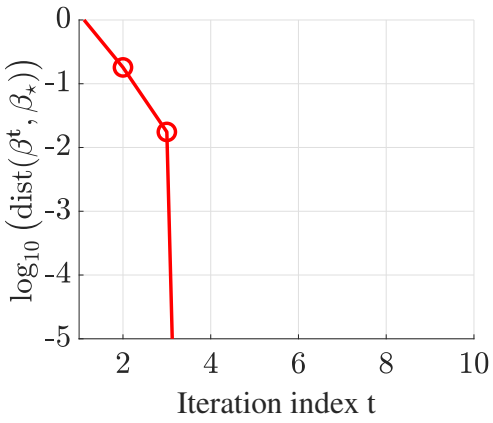


Figure 4.6: Convergence of RobustPA from suitable intializations in the iteration count.

- To complete the local analysis, it is necessary to study robust initial estimation methods that can provide suitable initial estimates.

- The final goal is to establish global convergence. This may involve two steps: (1) demonstrating that the dynamics from random initialization lead to the basin of the local convergence region, and (2) applying the local convergence result once the estimates lie within this region. A similar approach has been studied in the phase retrieval problem using SGD [84].

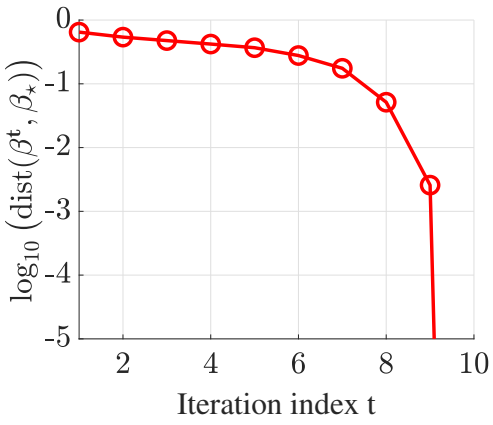


Figure 4.7: Convergence of RobustPA from random initializations in the iteration count.

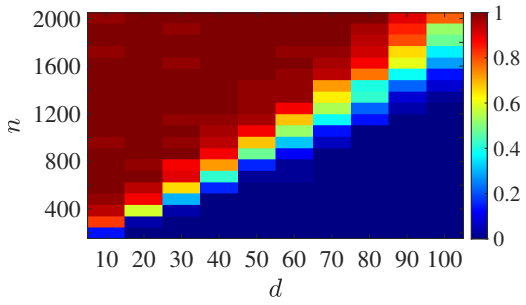


Figure 4.8: The phase transition for empirical success rate over 50 trials. We generate Gaussian measurements under $k = 5$, $p_{\text{fail}} = 0.04$ with the values of outliers $\xi_i = -y_i$ for $i \in I_{\text{out}}$.

Appendix A: Proofs for Section 1.2.3

A.1 Proof of Theorem 4

We first prove by induction on the iteration index j that

$$\operatorname{dist}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_\star) \leq \nu_\eta \cdot \operatorname{dist}(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_\star) + \frac{\epsilon_{j-1}}{C_\eta} \tag{A.1.1}$$

holds for all $j \in \mathbb{N}$ for some numerical constant $\nu_\eta \in (0,1)$ and $C_\eta > 0$ depending only on η. Let $k \in \mathbb{N}$ be arbitrarily fixed. Suppose that $\boldsymbol{\theta}_j$ satisfies (A.1.1) for all $j \leq k$. Note that the distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_\star$ is written as

$$\operatorname{dist}(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) = \|\boldsymbol{\theta} - \varphi(\boldsymbol{\theta})\boldsymbol{\theta}_\star\|_2, \tag{A.1.2}$$

where

$$\varphi(\boldsymbol{\theta}) := \operatorname*{argmin}_{\alpha \in \{\pm 1\}} \|\boldsymbol{\theta} - \alpha\boldsymbol{\theta}_\star\|_2 .$$

Then we have $\operatorname{dist}(\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_\star) \leq \|\boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2$ and $\operatorname{dist}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_\star) = \|\boldsymbol{\theta}_k - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2$. Therefore, it follows that

$$\|\boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2 \leq \nu_\eta \|\boldsymbol{\theta}_k - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2 + \frac{\epsilon_k}{C_\eta} \tag{A.1.3}$$

implies (A.1.1) for $j = k+1$. This completes the induction argument.

Therefore, it suffices to show that the hypothesis of the theorem implies (A.1.3). For the sake of brevity, we denote the objective function of the optimization formulation in (1.3.5) by

$$f_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \left|\operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k\rangle\right)\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - b_i\right|. \tag{A.1.4}$$

Then (1.5.1) provides

$$\underbrace{f_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_{k+1})}_{\text{(A)}} \le \underbrace{f_{\boldsymbol{\theta}_k}(\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star)}_{\text{(B)}} + \epsilon_k. \tag{A.1.5}$$

Next, we derive a lower bound (resp. an upper bound) on (A) (resp. (B)) of (A.1.5). From the definition of b_i in (1.1.2), (A) is written as

$$\begin{aligned}
\text{(A)} &= \frac{1}{n}\sum_{i=1}^{n} \left|\operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k\rangle\right)\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1}\rangle - b_i\right| \\
&= \underbrace{\frac{1}{n}\sum_{i\in I_{\text{in}}} \left|\operatorname{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k\rangle)\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1}\rangle - |\langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\rangle|\right|}_{\text{(a)}} \\
&\quad + \frac{1}{n}\sum_{i\in I_{\text{out}}} \left|\operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k\rangle\right)\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1}\rangle - \xi_i\right|.
\end{aligned} \tag{A.1.6}$$

To simplify the partial summation over I_{in}, we introduce the spherical wedge [85] defined by

$$W_{\boldsymbol{\theta},\boldsymbol{z}} := \{\boldsymbol{v} \in \mathbb{S}^{d-1} \,|\, \operatorname{sign}(\langle \boldsymbol{v}, \boldsymbol{\theta}\rangle) \neq \operatorname{sign}(\langle \boldsymbol{v}, \boldsymbol{z}\rangle)\}. \tag{A.1.7}$$

Then it follows that $\langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\rangle$ and $\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k\rangle$ have the opposite sign if and only if $\boldsymbol{a}_i \in W_{\boldsymbol{\theta}_k,\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}$. Therefore, the summand in (a) is rewritten as

$$\begin{aligned}
\text{(a)} &= \frac{1}{n}\sum_{i\in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \in W_{\boldsymbol{\theta}_k,\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}\}} \left|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} + \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\rangle\right| \\
&\quad + \frac{1}{n}\sum_{i\in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \notin W_{\boldsymbol{\theta}_k,\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}\}} \left|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\rangle\right|.
\end{aligned}$$

The second summand on the right-hand side provides a valid lower bound on (a) since the other summand is nonnegative. Combining the above results, we obtain that

$$\begin{aligned}(\mathrm{A}) \geq \frac{1}{n} \sum_{i \in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \notin W_{\boldsymbol{\theta}_k, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}\}} \left| \langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star \rangle \right| \\ + \frac{1}{n} \sum_{i \in I_{\text{out}}} \left| \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle \right) \langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} \rangle - \xi_i \right|.\end{aligned} \tag{A.1.8}$$

Similarly, (B) is written as

$$\begin{aligned}(\mathrm{B}) = & \frac{1}{n} \sum_{i \in I_{\text{in}}} \underbrace{\left| \operatorname{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle) \langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star \rangle - \left| \langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star \rangle \right| \right|}_{\text{(b)}} \\ & + \frac{1}{n} \sum_{i \in I_{\text{out}}} \left| \operatorname{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle) \langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star \rangle - \xi_i \right|.\end{aligned}$$

If $\boldsymbol{a}_i \in W_{\boldsymbol{\theta}_k, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}$, then $\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle$ and $\langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star \rangle$ have the opposite sign and hence (b) satisfies

$$(\mathrm{b}) = 2 \left| \langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star \rangle \right| \leq 2 \left| \langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star - \boldsymbol{\theta}_k \rangle \right|.$$

Otherwise, if $\boldsymbol{a}_i \notin W_{\boldsymbol{\theta}_k, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}$, then (b) $= 0$. Therefore, we have

$$\begin{aligned}(\mathrm{B}) \leq \frac{2}{n} \sum_{i \in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \in W_{\boldsymbol{\theta}_k, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}\}} \left| \langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star - \boldsymbol{\theta}_k \rangle \right| \\ + \frac{1}{n} \sum_{i \in I_{\text{out}}} \left| \operatorname{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle) \langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star \rangle - \xi_i \right|.\end{aligned} \tag{A.1.9}$$

By plugging in the bounds of (A.1.8) and (A.1.9) into (A.1.5), we obtain that (A.1.5) implies

$$\begin{aligned}& \frac{1}{n} \sum_{i \in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \notin W_{\boldsymbol{\theta}_k, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}\}} \left| \langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star \rangle \right| \\ & \quad + \underbrace{\frac{1}{n} \sum_{i \in I_{\text{out}}} \left| \operatorname{sign}\left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle \right) \langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} \rangle - \xi_i \right|}_{(*)} \\ & \quad - \underbrace{\frac{1}{n} \sum_{i \in I_{\text{out}}} \left| \operatorname{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_k \rangle) \langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star \rangle - \xi_i \right|}_{(**)} \\ & \leq \frac{2}{n} \sum_{i \in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \in W_{\boldsymbol{\theta}_k, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}\}} \left| \langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star - \boldsymbol{\theta}_k \rangle \right| + \epsilon_k.\end{aligned} \tag{A.1.10}$$

By applying the triangle inequality to the summands in $(*)$ and $(**)$, we obtain a necessary condition of (A.1.10) given by

$$\begin{aligned}
&\underbrace{\frac{1}{n}\sum_{i\in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \notin W_{\boldsymbol{\theta}_k,\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}\}} \left|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\rangle\right|}_{\text{(c)}} \\
&\qquad\qquad - \underbrace{\frac{1}{n}\sum_{i\in I_{\text{out}}} \left|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\rangle\right|}_{\text{(d)}} \\
&\leq \underbrace{\frac{2}{n}\sum_{i\in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \in W_{\boldsymbol{\theta}_k,\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}\}} \left|\langle \boldsymbol{x}_i, \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star - \boldsymbol{\theta}_k\rangle\right|}_{\text{(e)}} + \epsilon_k.
\end{aligned} \tag{A.1.11}$$

We have shown that (A.1.5) implies (A.1.11). In the remainder of the proof, we demonstrate that if (A.1.11) is satisfied, then (A.1.3) holds with high probability. This is achieved by applying a probabilistic lower bound on (c) and probabilistic upper bounds on (d) and (e), using concentration inequalities.

To this end, note that the measurement vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ depend not only on the current iterate $\boldsymbol{\theta}_k$ and the next iterate $\boldsymbol{\theta}_{k+1}$, but also on the indicator functions within the spherical wedge in (c) and (e). Therefore, we consider the uniform bounds for all iterates and the collection of spherical wedges with the largest angle less than $\psi \in (0, \pi)$. We introduce the corresponding lemmas below.

Lemma 11 *Let $\psi \in (0, \pi), \eta \in (0, 1/2)$ and $\delta > 0$. Suppose that $\{\boldsymbol{x}_i\}_{i=1}^n$ are independent copies of $\boldsymbol{g} \sim \text{Normal}(\mathbf{0}, \boldsymbol{I}_d)$. Let*

$$\mathcal{W}_\psi := \left\{W_{\boldsymbol{\theta},\boldsymbol{z}} : \boldsymbol{\theta}, \boldsymbol{z} \in \mathbb{R}^d, \angle(\boldsymbol{\theta}, \boldsymbol{z}) \leq \psi\right\}, \tag{A.1.12}$$

where $W_{\boldsymbol{\theta},\boldsymbol{z}}$ is defined in (A.1.7)*. Then there exists an absolute constant C such that*

$$\inf_{\substack{W\in\mathcal{W}_\psi\\ \boldsymbol{z}\in\mathbb{S}^{d-1}}} \frac{1}{n}\sum_{i\in I_{\mathrm{in}}} \mathbb{1}_{\{\boldsymbol{x}_i\notin W\}} \left|\langle \boldsymbol{x}_i,\boldsymbol{z}\rangle\right| \geq (1-\eta)\sqrt{\frac{2}{\pi}}$$
$$- \frac{2\psi}{\pi}\left(\sqrt{\frac{2}{\pi}} + \sqrt{2\log\left(\frac{e\pi}{2\psi}\right)}\right) - \frac{\psi}{20}\left(\sqrt{\frac{2\psi}{\pi}}+1\right), \tag{A.1.13}$$

$$\sup_{\boldsymbol{z}\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i\in I_{\mathrm{out}}} \left|\langle \boldsymbol{x}_i,\boldsymbol{z}\rangle\right| \leq \eta\sqrt{\frac{2}{\pi}} + \sqrt{\eta}\frac{\psi}{20}, \tag{A.1.14}$$

and

$$\sup_{\substack{W\in\mathcal{W}_\psi\\ \boldsymbol{z}\in\mathbb{S}^{d-1}}} \frac{1}{n}\sum_{i\in I_{\mathrm{in}}} \mathbb{1}_{\{\boldsymbol{x}_i\in W\}} \left|\langle \boldsymbol{x}_i,\boldsymbol{z}\rangle\right|$$
$$\leq \frac{2\psi}{\pi}\left(\sqrt{\frac{2}{\pi}} + \sqrt{2\log\left(\frac{e\pi}{2\psi}\right)}\right) + \sqrt{\frac{2\psi}{\pi}}\cdot\frac{\psi}{20} \tag{A.1.15}$$

hold with probability at least $1-\delta$ provided that

$$n \geq C\cdot\psi^{-2}\left(d\log(n/d)\vee\log(1/\delta)\right). \tag{A.1.16}$$

Proof 6 *See Appendix A.3.*

Now we derive the largest angle for the spherical wedge $W_{\boldsymbol{\theta}_k,\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}$. Since the angle between $\boldsymbol{\theta}_k$ and $\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star$ is always acute, we have

$$\begin{aligned}
\sin\left(\angle\left(\boldsymbol{\theta}_k,\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\right)\right) &= \left\|\left(\boldsymbol{I}_d - \frac{\boldsymbol{\theta}_k\boldsymbol{\theta}_k^{\mathrm{T}}}{\|\boldsymbol{\theta}_k\|_2^2}\right)\frac{\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star}{\|\boldsymbol{\theta}_\star\|_2}\right\| \\
&\leq \left\|\left(\boldsymbol{I}_d - \frac{\boldsymbol{\theta}_k\boldsymbol{\theta}_k^{\mathrm{T}}}{\|\boldsymbol{\theta}_k\|_2^2}\right)\frac{\varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star - \boldsymbol{\theta}_k}{\|\boldsymbol{\theta}_\star\|_2}\right\| \\
&\overset{\text{(i)}}{\leq} \frac{\|\boldsymbol{\theta}_k - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2}{\|\ theta_\star\|_2} = \frac{\mathrm{dist}\left(\boldsymbol{\theta}_k,\boldsymbol{\theta}_\star\right)}{\|\boldsymbol{\theta}_\star\|_2} \\
&\overset{\text{(ii)}}{\leq} \sin\left(\frac{1}{20}\right),
\end{aligned} \tag{A.1.17}$$

where (i) holds since the projection operator is non-expansive; (ii) follows since the induction hypothesis implies

$$\begin{aligned}
&\operatorname{dist}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_\star) \\
&\leq \nu_\eta^k \cdot \operatorname{dist}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\star) + \frac{\max_{i\in[0:k-1]} \epsilon_i}{C_\eta} \sum_{t=0}^{k-1} \nu_\eta^t \\
&\leq \nu_\eta^k \cdot \operatorname{dist}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\star) + (1-\nu_\eta) \sin\left(\frac{1}{20}\right) \|\boldsymbol{\theta}_\star\|_2 \sum_{t=0}^{k-1} \nu_\eta^t \\
&\leq \sin\left(\frac{1}{20}\right) \|\boldsymbol{\theta}_\star\|_2,
\end{aligned} \tag{A.1.18}$$

where the second and the last inequalities follow from (2.1.4).

Hence, in Theorem 11, we plug in $\psi = 1/20$. Then the sample complexity in Theorem 4 invokes Theorem 11, (A.1.13), (A.1.14), and (A.1.15) hold with probability at least $1-\delta$ simultaneously. The remainder of the proof is conditioned on the events that (A.1.13), (A.1.14), and (A.1.15) hold.

By applying (A.1.13) and (A.1.14) to (c) and (d) of (A.1.11) and (A.1.15) to (e) of (A.1.11) with the choice of $\psi = 1/20$, we obtain

$$\|\boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2 \leq \nu_\eta \|\boldsymbol{\theta}_k - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2 + \frac{\epsilon_k}{C_\eta} \tag{A.1.19}$$

for

$$\nu_\eta := \frac{2c_0}{C_\eta} \quad \text{and} \quad C_\eta := (1-2\eta)\sqrt{\frac{2}{\pi}} - c_0 - \frac{1}{400}(1+\sqrt{\eta}), \tag{A.1.20}$$

where

$$c_0 := \frac{1}{10\pi}\left(\sqrt{\frac{2}{\pi}} + \sqrt{2\log(5e\pi)}\right) + \frac{1}{100\sqrt{5\pi}}.$$

Since ν_η satisfies

$$\frac{d\nu_\eta}{d\eta} = \frac{c_0\left(2\sqrt{\frac{2}{\pi}} + \frac{1}{800\sqrt{\eta}}\right)}{\left((1-2\eta)\sqrt{\frac{2}{\pi}} - c_0 - \frac{1}{400}(1+\sqrt{\eta})\right)^2} > 0$$

for all $\eta \in [0, 1/4]$, it is monotonically increasing in η and upper-bounded as $\nu_\eta \leq \nu_{1/4} < 9/10$. This implies $\nu_\eta < 1$ uniformly over $\eta \in [0, 1/4]$. This completes the proof of (A.1.3).

A.2 Supporting Lemmas

Lemma 12 *Let $\boldsymbol{g} \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$ and $\psi \in (0, \pi)$. Let $\mathcal{W}_\psi$ be defined as in* (A.1.12). *Then we have*

$$\sup_{W \in \mathcal{W}_\psi} \mathbb{P}(\boldsymbol{g} \in W) \leq \frac{\psi}{\pi}.$$

Proof 7 *Let $W \in \mathcal{W}_\psi$ be arbitrarily fixed. It follows from the definitions in* (A.1.12) *and* (A.1.7) *that W is a cone. Therefore, $\boldsymbol{g} \in W$ if and only if $\boldsymbol{g}/\|\boldsymbol{g}\|_2 \in W$. Furthermore, note that $\boldsymbol{g}/\|\boldsymbol{g}\|_2$ is uniformly distributed in $\mathbb{S}^{d-1}$. Then we have*

$$\mathbb{P}\left(\boldsymbol{g} \in W\right) = \mathbb{P}\left(\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in W\right) \leq \frac{\psi}{\pi}. \tag{A.2.1}$$

The assertion follows since W was arbitrary.

Lemma 13 ([75, Lemma 2.1]) *Let $\delta \in (0, 1)$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ be independent copies of $\boldsymbol{g} \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$. Then it holds with probability at least $1 - \delta$ that*

$$\sup_{\boldsymbol{z} \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n |\langle \boldsymbol{x}_i, \boldsymbol{z} \rangle| - \sqrt{\frac{2}{\pi}} \right| \leq 4\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}. \tag{A.2.2}$$

Lemma 14 ([74, Lemma 6.4]) *Let $\delta \in (0, 1)$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ be independent copies of $\boldsymbol{g} \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$. Let $s \in \mathbb{N}$ satisfy $s < n$. Then it holds with probability at least $1 - \delta$ that*

$$\begin{aligned} &\sup_{\substack{\boldsymbol{z} \in \mathbb{S}^{d-1} \\ T : |T| \leq s}} \frac{1}{s} \sum_{i \in T} |\langle \boldsymbol{x}_i, \boldsymbol{z} \rangle| \\ &\quad \leq \sqrt{\frac{2}{\pi}} + 4\sqrt{\frac{d}{s}} + \sqrt{2\log\left(\frac{en}{s}\right)} + \sqrt{\frac{2}{s} \cdot \log\left(\frac{2}{\delta}\right)}. \end{aligned} \tag{A.2.3}$$

Lemma 15 ([85, Lemma 5.1]) *Let $\delta \in (0,1)$ and an acute angle $\psi > 0$. Suppose $\{\boldsymbol{x}_i\}_{i=1}^n$ be independent copies of a random variable $\boldsymbol{x} \in \mathbb{R}^d$ and we consider the set $\mathcal{W}_\psi$ given by* (A.1.12)*. Then, if*

$$n \geq (4\pi/\psi)^2(2d\log(2en/d) + \log(2/\delta)),$$

we have

$$\sup_{W \in \mathcal{W}_\psi} \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in W\}} \leq \frac{2\psi}{\pi}. \tag{A.2.4}$$

holds with probability at least $1-\delta$.

A.3 Proof of Theorem 11

We proceed with the proof under the following four events, each of which holds with probability at least $1-\delta/4$. The first event is defined as

$$\begin{aligned}\sup_{\boldsymbol{z}\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i\in I_{\text{in}}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle| - (1-\eta)\sqrt{\frac{2}{\pi}} \right| \\ \leq 4\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log(8/\delta)}{n}},\end{aligned} \tag{A.3.1}$$

which holds with probability at least $1-\delta/4$. Since by the assumption on outliers, we have a set $|I_{\text{in}}|$ with $|I_{\text{in}}| = (1-\eta)n$ and the outliers are independent of $\{\boldsymbol{x}_i\}_{i=1}^n$. Hence, (A.3.1) is a direct result of (A.2.2) in Theorem 13. By following the same argument, we also have that

$$\sup_{\boldsymbol{z}\in\mathbb{S}^{d-1}} \left| \frac{1}{n}\sum_{i\in I_{\text{out}}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle| - \eta\sqrt{\frac{2}{\pi}} \right| \leq 4\sqrt{\frac{\eta d}{n}} + \sqrt{\frac{2\eta\log(8/\delta)}{n}} \tag{A.3.2}$$

holds with probability at least $1-\delta/4$.

Next, we describe the following event: for an arbitrary fixed $\alpha \in (0,1)$, it holds with probability at least $1-\delta/4$ that

$$\begin{aligned}&\sup_{\substack{T:|T|\le \alpha n\\ \boldsymbol{z}\in\mathbb{S}^{d-1}}} \frac{1}{n}\sum_{i\in T\cap I_{\text{in}}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle| \le \\ &\alpha\sqrt{\frac{2}{\pi}} + 4\sqrt{\frac{\alpha d}{n}} + \alpha\sqrt{2\log\left(\frac{e}{\alpha}\right)} + \sqrt{\frac{2\alpha\log(8/\delta)}{n}}.\end{aligned} \tag{A.3.3}$$

Again, since by the Assumption 1, we have a fixed set $|I_{\text{in}}|$ with $|I_{\text{in}}| = (1-\eta)n$ and the outliers are independent of $\{\boldsymbol{x}_i\}_{i=1}^n$, (A.3.3) holds by (A.2.3) in Theorem 14.

Since (A.1.16) invokes Theorem 15 with probability at least $1-\delta/4$, it holds with probability at least $1-\delta/4$ that

$$\sup_{W\in\mathcal{W}_\psi} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in W\}} \le \frac{2\psi n}{\pi}. \tag{A.3.4}$$

Since we have shown that (A.3.1), (A.3.2), (A.3.3), and (A.3.4) hold with probability at least $1-\delta$, we will move forward with the remainder of the proof by assuming those conditions are satisfied.

We first show (A.1.13). We observe that for an arbitrary $W \in \mathcal{W}_\psi$ and $\boldsymbol{z} \in \mathbb{S}^{d-1}$, it holds deterministically that

$$\begin{aligned}&\frac{1}{n}\sum_{i\in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i\notin W\}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle| = \\ &\quad \frac{1}{n}\sum_{i\in I_{\text{in}}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle| - \frac{1}{n}\sum_{i\in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i\in W\}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle|.\end{aligned}$$

Hence, by taking infimum on both sides over sets $\mathcal{W}_\psi$ and $\mathbb{S}^{d-1}$, we have

$$\begin{aligned}&\inf_{\substack{W\in\mathcal{W}_\psi\\ \boldsymbol{z}\in\mathbb{S}^{d-1}}} \frac{1}{n}\sum_{i\in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i\notin W\}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle| \\ &\ge \underbrace{\inf_{\boldsymbol{z}\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i\in I_{\text{in}}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle|}_{\text{(A)}} - \underbrace{\sup_{\substack{W\in\mathcal{W}_\psi\\ \boldsymbol{z}\in\mathbb{S}^{d-1}}} \frac{1}{n}\sum_{i\in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i\in W\}} |\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle|}_{\text{(B)}}.\end{aligned} \tag{A.3.5}$$

We first obtain a lower bound on (A) and an upper bound on (B). We have a lower bound on (A) by (A.3.1):

$$\text{(A)} \geq (1-\eta)\sqrt{\frac{2}{\pi}} - 4\sqrt{\frac{d}{n}} - \sqrt{\frac{2\log(8/\delta)}{n}}. \tag{A.3.6}$$

By taking n according to (A.1.16) for a sufficiently large $C > 0$, we have

$$\text{(A)} \geq (1-\eta)\sqrt{\frac{2}{\pi}} - \frac{\psi}{20}. \tag{A.3.7}$$

It remains to show an upper bound on (B). Under the event (A.3.4), we have

$$\text{(B)} \leq \sup_{\substack{T:|T|\leq 2\psi n/\pi \\ \boldsymbol{z}\in\mathbb{S}^{d-1}}} \frac{1}{n} \sum_{i\in T\cap I_{\text{in}}} \left|\langle \boldsymbol{x}_i, \boldsymbol{z}\rangle\right|.$$

Therefore, by letting $\alpha = 2\psi/\pi$ in (A.3.3), (A.3.3) gives an upper bound on (B):

$$\text{(B)} \leq \frac{2\psi}{\pi}\sqrt{\frac{2}{\pi}} + 4\sqrt{\frac{2\psi d}{\pi n}} + \frac{2\psi}{\pi}\sqrt{2\log\left(\frac{e\pi}{2\psi}\right)} + \sqrt{\frac{4\psi\log(8/\delta)}{\pi n}}. \tag{A.3.8}$$

Taking n according to (A.1.16) yields

$$\text{(B)} \leq \frac{2\psi}{\pi}\left(\sqrt{\frac{2}{\pi}} + \sqrt{2\log\left(\frac{e\pi}{2\psi}\right)}\right) + \frac{\psi}{20}\sqrt{\frac{2\psi}{\pi}}. \tag{A.3.9}$$

Hence, putting the results (A.3.7) and (A.3.9) into (A.3.5) completes the proof of the statement (A.1.13).

For the proofs of remaining statements in (A.1.14) and (A.1.15), the upper bound in (A.1.14) is a direct consequence of (A.3.2) with choosing n according to (A.1.16). Lastly, (A.1.15) is the result of the upper bound of (B) in (A.3.9). These complete the proof of (A.1.14) and (A.1.15).

A.4 Proof of Theorem 5

Define the empirical loss function $\ell_{\text{linear}} : \mathbb{R}^d \to \mathbb{R}_+$ for the LAD linear regression as

$$\ell_{\text{linear}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \left|\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \beta_i\right|, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

Then it holds for any $\boldsymbol{\theta} \in \mathbb{R}^d$ that

$$\begin{aligned}\ell_{\text{linear}}(\boldsymbol{\theta}) - \ell_{\text{linear}}(\boldsymbol{\theta}_\star) = \frac{1}{n}\sum_{i\in I_{\text{in}}} |\langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_\star\rangle| \\ + \underbrace{\frac{1}{n}\sum_{i\in I_{\text{out}}} \left(|\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - \xi_i| - |\langle \boldsymbol{x}_i, \boldsymbol{\theta}_\star\rangle - \xi_i|\right)}_{(*)}.\end{aligned} \tag{A.4.1}$$

By the triangle inequality, the magnitude of $(*)$ is upper-bounded by

$$|(*)| \le \frac{1}{n}\sum_{i\in I_{\text{out}}} |\langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_\star\rangle|. \tag{A.4.2}$$

Then (A.4.1) and (A.4.2) yield the following inequality:

$$\begin{aligned}&\frac{1}{n}\sum_{i\in I_{\text{in}}} |\langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_\star\rangle| - \frac{1}{n}\sum_{i\in I_{\text{out}}} |\langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_\star\rangle| \\ &\le \ell_{\text{linear}}(\boldsymbol{\theta}) - \ell_{\text{linear}}(\boldsymbol{\theta}_\star).\end{aligned} \tag{A.4.3}$$

For brevity, let $\ell(\boldsymbol{\theta}) = \mathbb{E}\ell_{\text{linear}}(\boldsymbol{\theta})$, which is the cost function in (1.7.1). Since the outlier fraction is fixed to η, by taking the expectation in (A.4.3), we obtain

$$\sqrt{\frac{2}{\pi}}(1-2\eta)\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_2 \le \ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_\star). \tag{A.4.4}$$

It is obvious from (A.4.4) that $\eta < 1/2$ is a sufficient condition for $\boldsymbol{\theta}_\star$ to be the unique minimizer of ℓ regardless of the values of $(\xi_i)_{i\in I_{\text{out}}}$. We prove that $\eta < 1/2$ is also a necessary condition by contradiction. Suppose $\eta \ge 1/2$ and $\boldsymbol{\theta}_\star$ is the unique minimizer of ℓ. Since $(\xi_i)_{i\in I_{\text{out}}}$ can be arbitrary, we set the values by

$$\xi_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}_1\rangle, \quad \forall i \in I_{\text{out}}$$

for some $\boldsymbol{\theta}_1$. Then it follows that

$$\begin{aligned}\ell_{\text{linear}}(\boldsymbol{\theta}_1) &= \frac{1}{n}\sum_{i\in I_{\text{in}}} |\langle \boldsymbol{x}_i, \boldsymbol{\theta}_1 - \boldsymbol{\theta}_\star\rangle|, \\ \ell_{\text{linear}}(\boldsymbol{\theta}_\star) &= \frac{1}{n}\sum_{i\in I_{\text{out}}} |\langle \boldsymbol{x}_i, \boldsymbol{\theta}_1 - \boldsymbol{\theta}_\star\rangle|.\end{aligned}$$

Thus, by taking expectation in $\ell_{\text{linear}}(\boldsymbol{\theta}_1) - \ell_{\text{linear}}(\boldsymbol{\theta}_\star)$, we obtain

$$\ell(\boldsymbol{\theta}_1) - \ell(\boldsymbol{\theta}_\star) = (1 - 2\eta)\sqrt{\frac{2}{\pi}}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_\star\|_2 \leq 0,$$

where the last inequality holds by the assumption $\eta \geq 1/2$. This contradicts the assumption that $\boldsymbol{\theta}_\star$ is the unique minimizer of ℓ.

A.5 Proof of Theorem 6

The proof is almost similar to that of Theorem 4. We present only their differences below. First, to focus on the dependence of η on the other parameters, we consider a simple case with the exact inner iterations. Therefore, the corresponding error term will not appear in this proof. More importantly, we rewrite the concentration inequalities Theorem 11 to explicitly show the dependence of the deviation terms on dimension parameters d and n, and the probability parameter δ. This modification is straightforward and the corresponding results are stated in the following corollary.

Corollary 16 *Instate the assumptions of Theorem 11. Then there exists an absolute value $C_2 > 0$ such that*

$$\begin{aligned}
&\inf_{\substack{W \in \mathcal{W}_\psi \\ \boldsymbol{z} \in \mathbb{S}^{d-1}}} \frac{1}{n} \sum_{i \in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \notin W\}} \left| \langle \boldsymbol{x}_i, \boldsymbol{z} \rangle \right| \geq (1 - \eta) \sqrt{\frac{2}{\pi}}, \\
&- \frac{2\psi}{\pi} \left(\sqrt{\frac{2}{\pi}} + \sqrt{2 \log \left(\frac{e\pi}{2\psi} \right)} \right) - 2C_2 \sqrt{\frac{d \vee \log(1/\delta)}{n}} \\
&\sup_{\boldsymbol{z} \in \mathbb{S}^{d-1}} \frac{1}{n} \sum_{i \in I_{\text{out}}} \left| \langle \boldsymbol{x}_i, \boldsymbol{z} \rangle \right| \leq \eta \sqrt{\frac{2}{\pi}} + C_2 \sqrt{\frac{d \vee \log(1/\delta)}{n}},
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{\substack{W \in \mathcal{W}_\psi \\ \boldsymbol{z} \in \mathbb{S}^{d-1}}} \frac{1}{n} \sum_{i \in I_{\text{in}}} \mathbb{1}_{\{\boldsymbol{x}_i \in W\}} \left| \langle \boldsymbol{x}_i, \boldsymbol{z} \rangle \right| \\
&\leq \frac{2\psi}{\pi} \left(\sqrt{\frac{2}{\pi}} + \sqrt{2 \log \left(\frac{e\pi}{2\psi} \right)} \right) + C_2 \sqrt{\frac{\psi(d \vee \log(1/\delta))}{n}}
\end{aligned}$$

hold with probability at least $1 - \delta$.

Since the sample complexity in (1.7.5) implies (A.1.16), it invokes Theorem 16 with ψ substituted by ψ_0. Due to the modifications from Theorem 11 to Theorem 16, under the exact inner iterations, the inequality in (A.1.19) is rewritten as

$$\|\boldsymbol{\theta}_{k+1} - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2 \leq \nu_{\eta,\psi_0,n} \cdot \|\boldsymbol{\theta}_k - \varphi(\boldsymbol{\theta}_k)\boldsymbol{\theta}_\star\|_2 \tag{A.5.1}$$

with

$$\begin{aligned}
&\nu_{\eta,\psi_0,n} := \frac{2c_{\psi_0,n}}{C_{\eta,\psi_0,n}}, \\
&C_{\eta,\psi_0,n} := (1 - 2\eta)\sqrt{\frac{2}{\pi}} - 2C_2 \sqrt{\frac{d \vee \log(1/\delta)}{n}} - c_{\psi_0,n}, \\
&c_{\psi_0,n} := \\
&\frac{2\psi_0}{\pi} \left(\sqrt{\frac{2}{\pi}} + \sqrt{2 \log \left(\frac{e\pi}{2\psi_0} \right)} \right) + C_2 \sqrt{\frac{\psi_0(d \vee \log(1/\delta))}{n}}.
\end{aligned} \tag{A.5.2}$$

We show that the assumptions in Theorem 6 imply $\nu_{\eta,\psi_0,n} < 1$. By the definitions of $C_{\eta,\psi_0,n}$ and $c_{\psi_0,n}$, we have

$$\begin{aligned}
&C_{\eta,\psi_0,n} - 2c_{\psi_0,n} \\
&\geq (1-2\eta)\sqrt{\frac{2}{\pi}} - 5C_2\sqrt{\frac{d \vee \log(1/\delta)}{n}} \\
&\quad - \frac{6\psi_0}{\pi}\left(\sqrt{\frac{2}{\pi}} + \sqrt{2\log\left(\frac{e\pi}{2\psi_0}\right)}\right) \\
&\geq \sqrt{\frac{2}{\pi}}\left(\frac{6\psi_0}{\pi}\left(1 + \sqrt{\pi\log\left(\frac{e\pi}{2\psi_0}\right)}\right) - C_1\sqrt{\frac{d \vee \log(1/\delta)}{n}}\right) \\
&\quad - 5C_2\sqrt{\frac{d \vee \log(1/\delta)}{n}} - \frac{6\psi_0}{\pi}\left(\sqrt{\frac{2}{\pi}} + \sqrt{2\log\left(\frac{e\pi}{2\psi_0}\right)}\right) \\
&= C_1\sqrt{\frac{2}{\pi}}\sqrt{\frac{d \vee \log(1/\delta)}{n}} - 5C_2\sqrt{\frac{d \vee \log(1/\delta)}{n}} \\
&\geq \left(2\sqrt{\frac{2}{\pi}} - \frac{5C_2}{C_1}\right) \cdot \min(1-2\eta, \psi_0, u(\psi_0)),
\end{aligned} \tag{A.5.3}$$

where i) the first inequality holds by $\psi_0 \leq 1$; ii) the second inequality is obtained by substituting η by its upper bound by (1.7.3); iii) the last inequality follows from the condition in (1.7.5). It remains to show that the lower bound in (A.5.3) is strictly positive. First, one may choose C_1 sufficiently large so that $5C_2/C_1 < 2\sqrt{2/\pi}$. Next, since $u(x)$ obeys

$$\frac{du(x)}{dx} = -\frac{3\left(1 + \sqrt{\pi\log\left(\frac{e\pi}{2x}\right)}\right)}{\pi} + \frac{3}{2\sqrt{\pi\log\left(\frac{e\pi}{2x}\right)}} < 0$$

for all $x \in (0,1)$, it is monotonically decreasing on $(0,1)$. It then follows that $u(\psi_0) \geq u(0.12) > 0$ for all $\psi_0 \in (0, 0.12]$. Therefore, we have $\min(1-2\eta, \psi_0, u(\psi_0)) > 0$ for all $\eta \in [0, 1/2)$ and all $\psi_0 \in (0, 0.12]$. This completes the proof.

Appendix B: Proofs for Chapter 2

B.1 Proof of Theorem 7

We prove Theorem 7 in two steps. First, in the following proposition, we present a sufficient condition for stable estimation by convex program in (1.2.11). Then we derive an upper bound on ϱ in the proposition, which provides the sample complexity condition along with the corresponding error bound in Theorem 7.

Proposition 1 *Under the hypothesis of Theorem 7, suppose that $\widetilde{\boldsymbol{\theta}}$ satisfies*

$$\begin{aligned}\varrho := \inf_{\substack{j\in[k]\\ \boldsymbol{w}\in\mathbb{S}^{d-1}}} \mathbb{E}\,\mathbb{1}_{\mathcal{C}_j}(\boldsymbol{g})\,|\langle \boldsymbol{g},\boldsymbol{w}\rangle| - \sup_{\substack{j\in[k]\\ \boldsymbol{w}\in\mathbb{S}^{d-1}}} \mathbb{E}\mathbb{1}_{\widetilde{\mathcal{C}}_j\setminus\mathcal{C}_j}(\boldsymbol{g})\langle \boldsymbol{g},\boldsymbol{w}\rangle_+ \\ - \sup_{\substack{j\in[k]\\ \boldsymbol{w}\in\mathbb{S}^{d-1}}} \mathbb{E}\mathbb{1}_{\mathcal{C}_j\setminus\widetilde{\mathcal{C}}_j}(\boldsymbol{g})\langle \boldsymbol{g},\boldsymbol{w}\rangle_+ > 0\,. \end{aligned} \tag{B.1.1}$$

Then there exists an absolute constant $c>0$ such that if

$$n \geq c\varrho^{-2}\left(4d\log^3 d\log^5 k + 4\log(\delta^{-1})\log k\right), \tag{B.1.2}$$

then the solution $\widehat{\boldsymbol{\theta}}$ to the optimization problem in (1.2.11) obeys

$$\sum_{j=1}^{k}\|\boldsymbol{\theta}_{\star,j}-\widehat{\boldsymbol{\theta}}_j\|_2 \leq \frac{2}{\varrho n}\sum_{i=1}^{n}|z_i| \tag{B.1.3}$$

with probability $1-\delta$.

Proof 8 *We first show that there exists a constant $c > 0$ such that the condition in* (2.1.4) *implies $\zeta > 0$. Hence, we consider*

$$\underbrace{\min_{j\in[k]} \sqrt{\frac{\pi}{32}}\, \mathbb{P}^2\{\boldsymbol{g} \in \mathcal{C}_j\}}_{\text{(i)}} - \underbrace{2 \max_{j\in[k]} \sqrt{\mathbb{P}\{\boldsymbol{g} \in \widetilde{\mathcal{C}}_j \triangle \mathcal{C}_j\}}}_{\text{(ii)}} > 0. \tag{B.1.4}$$

It follows from the definition of $\pi_{\min}$ that (i) in (B.1.4) *is bounded from below as*

$$\text{(i)} \geq \sqrt{\frac{\pi}{32}} \pi_{\min}^2. \tag{B.1.5}$$

It only remains to find an appropriate upper bound on (ii). Since $\{\mathcal{C}_j\}_{j=1}^k$ consists of disjoint sets (except their boundaries corresponding to sets of measure zero), for a fixed $j \in [k]$, the symmetric difference between $\widetilde{\mathcal{C}}_j$ and $\mathcal{C}_j$ is written as

$$\widetilde{\mathcal{C}}_j \triangle \mathcal{C}_j = \left(\cup_{j'\neq j} \widetilde{\mathcal{C}}_j \cap \mathcal{C}_{j'}\right) \cup \left(\cup_{j'\neq j} \mathcal{C}_j \cap \widetilde{\mathcal{C}}_{j'}\right).$$

Therefore, we obtain

$$\text{(ii)} \leq 2\sqrt{2k} \max_{j\in[k]} \max_{j'\in[k]\setminus\{j\}} \sqrt{\mathbb{P}\left(\boldsymbol{g} \in \widetilde{\mathcal{C}}_j \cap \mathcal{C}_{j'}\right)}. \tag{B.1.6}$$

Moreover, since

$$\begin{aligned} \boldsymbol{g} \in \widetilde{\mathcal{C}}_j \cap \mathcal{C}_{j'} &\implies \boldsymbol{g}^\mathsf{T}\widetilde{\boldsymbol{\theta}}_j \geq \boldsymbol{x}_i^\mathsf{T}\widetilde{\boldsymbol{\theta}}_{j'},\ \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}_{\star,j'} \geq \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}_{\star,j} \\ &\implies \boldsymbol{g}^\mathsf{T}(\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_{j'}) \geq 0,\ \boldsymbol{g}^\mathsf{T}(\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'}) \leq 0 \\ &\implies \boldsymbol{g}^\mathsf{T}(\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_{j'}) \cdot \boldsymbol{g}^\mathsf{T}(\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'}) \leq 0, \end{aligned} \tag{B.1.7}$$

with [38, Lemma 9], (ii) *in* (B.1.6) *is further upper-bounded by*

$$\begin{aligned} &\text{(ii)} \\ &\leq 2\sqrt{2k} \cdot \\ &\quad \max_{j\in[k]} \max_{j'\in[k]\setminus\{j\}} \sqrt{\mathbb{P}\left(\boldsymbol{g}^\mathsf{T}(\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_{j'}) \cdot \boldsymbol{g}^\mathsf{T}(\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'}) \leq 0\right)} \\ &\leq C\sqrt{k} \cdot \max_{j\in[k]} \max_{j'\in[k]\setminus\{j\}} \left(\sqrt{\frac{\|(\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_{j'}) - (\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'})\|_2}{\|\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'}\|_2}}\right. \\ &\quad \left. \cdot \log^{1/4}\left(\frac{2\|\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'}\|_2}{\|(\widetilde{\boldsymbol{\theta}}_j - \widetilde{\boldsymbol{\theta}}_{j'}) - (\boldsymbol{\theta}_{\star,j} - \boldsymbol{\theta}_{\star,j'})\|_2}\right)\right), \end{aligned} \tag{B.1.8}$$

for an absolute constant $C > 0$. Then, by plugging in (B.1.5) *and* (B.1.8) *to* (B.1.4)*, we obtain a sufficient condition for* (B.1.4) *as*

$$\begin{aligned} &C\sqrt{k}\max_{j\in[k]}\max_{j'\in[k]\setminus\{j\}}\sqrt{\frac{\|(\widetilde{\boldsymbol{\theta}}_j-\widetilde{\boldsymbol{\theta}}_{j'})-(\boldsymbol{\theta}_{\star,j}-\boldsymbol{\theta}_{\star,j'})\|_2}{\|\boldsymbol{\theta}_{\star,j}-\boldsymbol{\theta}_{\star,j'}\|_2}}\cdot \\ &\log^{1/4}\left(\frac{2\|\boldsymbol{\theta}_{\star,j}-\boldsymbol{\theta}_{\star,j'}\|_2}{\|(\widetilde{\boldsymbol{\theta}}_j-\widetilde{\boldsymbol{\theta}}_{j'})-(\boldsymbol{\theta}_{\star,j}-\boldsymbol{\theta}_{\star,j'})\|_2}\right)<\sqrt{\frac{\pi}{32}}\pi_{\min}^2. \end{aligned} \tag{B.1.9}$$

For a fixed $j' \in [k] \setminus \{j\}$, let

$$a=\frac{\|(\widetilde{\boldsymbol{\theta}}_j-\widetilde{\boldsymbol{\theta}}_{j'})-(\boldsymbol{\theta}_{\star,j}-\boldsymbol{\theta}_{\star,j'})\|_2}{\|\boldsymbol{\theta}_{\star,j}-\boldsymbol{\theta}_{\star,j'}\|_2}\quad \textit{and}\quad b=\frac{\pi_{\min}^4}{k}.$$

Since $a, b \in (0, 0.1]$ and $a \le \frac{b}{2}\log^{-1/2}(1/b)$ imply $a\log^{1/2}(2/a) \le b$, if one chooses c in (2.1.4) *so that $c < \frac{\pi}{32C^2}$, then* (2.1.4) *implies* (B.1.9) *for all distinct $j, j' \in [k]$. In the remainder of the proof, we will assume that* (B.1.4) *holds.*

We show that, for a sufficiently large $\rho > 0$, the following three conditions cannot hold simultaneously:

$$\frac{1}{n}\sum_{i=1}^{n}\left(f_i(\boldsymbol{\theta}_\star+\boldsymbol{h})-y_i\right)_+\le\eta\,, \tag{B.1.10}$$

$$\|\boldsymbol{h}\|_{1,2}>\rho\,, \tag{B.1.11}$$

$$\langle\boldsymbol{a},\boldsymbol{h}\rangle\ge 0\,. \tag{B.1.12}$$

Therefore, assuming (B.1.11) and (B.1.12) hold, it suffices to show

$$\mathcal{L}(\boldsymbol{h}):=\frac{1}{n}\sum_{i=1}^{n}\left(f_i(\boldsymbol{\theta}_\star+\boldsymbol{h})-y_i\right)_+>\eta\,. \tag{B.1.13}$$

To this end, we derive a lower bound on $\mathcal{L}(\boldsymbol{h})$ as follows:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{h}) &\geq \frac{1}{n}\sum_{i=1}^{n}\left(f_i(\boldsymbol{\theta}_\star+\boldsymbol{h})-f_i(\boldsymbol{\theta}_\star)\right)_+ - \frac{1}{n}\sum_{i=1}^{n}(z_i)_+ \\
&\overset{\text{(a)}}{\geq} \frac{1}{n}\sum_{i=1}^{n}\left(\langle\nabla f_i(\boldsymbol{\theta}_\star),\boldsymbol{h}\rangle\right)_+ - \frac{1}{n}\sum_{i=1}^{n}(z_i)_+ \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{|\langle\nabla f_i(\boldsymbol{\theta}_\star),\boldsymbol{h}\rangle|}{2} + \frac{1}{n}\sum_{i=1}^{n}\frac{\langle\nabla f_i(\boldsymbol{\theta}_\star),\boldsymbol{h}\rangle}{2} \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}(z_i)_+ \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{|\langle\nabla f_i(\boldsymbol{\theta}_\star),\boldsymbol{h}\rangle|}{2} + \frac{1}{n}\sum_{i=1}^{n}\frac{\langle\nabla f_i(\boldsymbol{\theta}_\star),\boldsymbol{h}\rangle}{2} \\
&\quad - \langle\boldsymbol{a},\boldsymbol{h}\rangle + \langle\boldsymbol{a},\boldsymbol{h}\rangle - \frac{1}{n}\sum_{i=1}^{n}(z_i)_+ \\
&\overset{\text{(b)}}{=} \langle\boldsymbol{a},\boldsymbol{h}\rangle - \frac{1}{n}\sum_{i=1}^{n}(z_i)_+ + \frac{1}{n}\sum_{i=1}^{n}\frac{|\langle\nabla f_i(\boldsymbol{\theta}_\star),\boldsymbol{h}\rangle|}{2} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{\langle\nabla f_i(\boldsymbol{\theta}_\star)-\nabla f_i(\widetilde{\boldsymbol{\theta}}),\boldsymbol{h}\rangle}{2} \\
&\overset{\text{(c)}}{=} \langle\boldsymbol{a},\boldsymbol{h}\rangle - \frac{1}{n}\sum_{i=1}^{n}(z_i)_+ + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i)|\langle\boldsymbol{x}_i,\boldsymbol{h}_j\rangle|}{2} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\{\mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i)-\mathbb{1}_{\widetilde{\mathcal{C}}_j}(\boldsymbol{x}_i)\}\langle\boldsymbol{x}_i,\boldsymbol{h}_j\rangle}{2}\,,
\end{aligned} \tag{B.1.14}$$

where (a) holds by the convexity of f_i, which implies

$$f_i(\boldsymbol{\theta}_\star+\boldsymbol{h}) \geq f_i(\boldsymbol{\theta}_\star) + \langle\nabla f_i(\boldsymbol{\theta}_\star),\boldsymbol{h}\rangle\,,$$

(b) follows from (2.1.1), and (c) is obtained by calculating $\nabla f_i(\boldsymbol{\theta})$ at $\boldsymbol{\theta}=\boldsymbol{\theta}_\star$ and $\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}$. We further proceed by obtaining lower bounds on the last two terms in (B.1.14) by the following lemmas, which are proved in Appendices B.2.4 and B.2.5.

Lemma 3 *Let $(V_{\boldsymbol{h}})_{\boldsymbol{h}\in\mathbb{R}^{kd}}$ be a random process defined by*

$$V_{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i)\,|\langle\boldsymbol{x}_i,\boldsymbol{h}_j\rangle|\,,$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *are i.i.d.* $\mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$. *Then, for* $\boldsymbol{g} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$ *and any* $\delta \in (0,1)$, *there exists an absolute constant* $c_1 > 0$ *such that*

$$\begin{aligned}\underline{V} := \inf_{\|\boldsymbol{h}\|_{1,2}=1} V_{\boldsymbol{h}} \geq & \min_{j\in[k], \boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\, \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{g})\, |\langle \boldsymbol{g}, \boldsymbol{w}\rangle| \\ & - c_1 \left(\frac{d \log^3 d \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2}\end{aligned}$$

holds with probability at least $1 - \delta/2$.

Proof 9 *See Appendix B.2.4.*

Lemma 4 *Let* $(Q_{\boldsymbol{h}})_{\boldsymbol{h}\in B_{1,2}}$ *be a random process defined by*

$$Q_{\boldsymbol{h}} := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \left\{ \mathbb{1}_{\widetilde{\mathcal{C}}_j}(\boldsymbol{x}_i) - \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i) \right\} \langle \boldsymbol{x}_i, \boldsymbol{h}_j \rangle \,,$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *are i.i.d.* $\mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$. *Then, for* $\boldsymbol{g} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$ *and any* $\delta \in (0,1)$, *there exists an absolute constant* $c_2 > 0$ *such that*

$$\begin{aligned}\overline{Q} := \sup_{\|\boldsymbol{h}\|_{1,2}=1} Q_{\boldsymbol{h}} \leq & \max_{j\in[k], \boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\mathbb{1}_{\widetilde{\mathcal{C}}_j \setminus \mathcal{C}_j}(\boldsymbol{g}) \langle \boldsymbol{g}, \boldsymbol{w}\rangle_+ \\ & + \max_{j\in[k], \boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\mathbb{1}_{\mathcal{C}_j \setminus \widetilde{\mathcal{C}}_j}(\boldsymbol{g}) \langle \boldsymbol{g}, \boldsymbol{w}\rangle_+ \\ & + c_2 \left(\frac{d \log^3 d \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2}\end{aligned}$$

holds with probability at least $1 - \delta/2$.

Proof 10 *See Appendix B.2.5.*

Since $V_{\boldsymbol{h}}$ are $Q_{\boldsymbol{h}}$ are homogeneous in $\boldsymbol{h}$, we obtain that the third term in the right-hand side of (B.1.14) is written as $V_{\boldsymbol{h}}$ and lower-bounded by

$$\frac{V_{\boldsymbol{h}}}{2} \geq \frac{\underline{V} \|\boldsymbol{h}\|_{1,2}}{2} \,. \tag{B.1.15}$$

Similarly, the last term in the right-hand side of (B.1.14) is written as $-Q_{\boldsymbol{h}}$ and lower-bounded by

$$-\frac{Q_{\boldsymbol{h}}}{2} \geq -\frac{\overline{Q}\|\boldsymbol{h}\|_{1,2}}{2}\,. \tag{B.1.16}$$

Furthermore, by Lemmas 3 and 4, the condition in (B.1.2) implies that

$$\underline{V} - \overline{Q} \geq c_3\varrho > 0 \tag{B.1.17}$$

holds with probability $1-\delta$ for an absolute constant $c_3 > 0$. Then we choose ρ so that it satisfies

$$\rho = \frac{2}{\underline{V} - \overline{Q}} \cdot \left(\eta + \frac{1}{n}\sum_{i=1}^{n}(z_i)_+\right)\,.$$

Next, by plugging in the above estimates to (B.1.14), we obtain that, under the event in (B.1.17), the conditions in (B.1.11) and (B.1.12) imply

$$\begin{aligned}
\mathcal{L}(\boldsymbol{h}) &\geq \langle \boldsymbol{a}, \boldsymbol{h}\rangle - \frac{1}{n}\sum_{i=1}^{n}(z_i)_+ + \frac{(\underline{V} - \overline{Q})\|\boldsymbol{h}\|_{1,2}}{2} \\
&> -\frac{1}{n}\sum_{i=1}^{n}(z_i)_+ + \frac{(\underline{V} - \overline{Q})\rho}{2} \\
&= -\frac{1}{n}\sum_{i=1}^{n}(z_i)_+ + \frac{1}{n}\sum_{i=1}^{n}(z_i)_+ + \eta \\
&= \eta\,.
\end{aligned}$$

This lower bound implies (B.1.13). Therefore we have shown that the three conditions in (B.1.10), (B.1.11), and (B.1.12) cannot hold simultaneously. It remains to apply the claim to a special case.

Let $\widehat{\boldsymbol{h}} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\star$. Recall that both $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_\star$ are feasible for the optimization problem in (1.2.11). Moreover, since $\widehat{\boldsymbol{\theta}}$ is the maximizer, it follows that $\langle \boldsymbol{a}, \widehat{\boldsymbol{\theta}}\rangle \geq \langle \boldsymbol{a}, \boldsymbol{\theta}_\star\rangle$, which implies $\langle \boldsymbol{a}, \widehat{\boldsymbol{h}}\rangle \geq 0$. Therefore the conditions in (B.1.10) and (B.1.12) are satisfied with $\boldsymbol{h}$ substituted by $\widehat{\boldsymbol{h}}$. Since the three conditions cannot be satisfied simultaneously,

the condition in (B.1.11) cannot hold, i.e. $\widehat{\boldsymbol{h}}$ satisfies

$$\left\|\widehat{\boldsymbol{h}}\right\|_{1,2} \leq \rho \leq \frac{2}{\varrho}\left(\eta + \frac{1}{n}\sum_{i=1}^{n}(z_i)_+\right). \tag{B.1.18}$$

Since the noise vector $\boldsymbol{w}$ was arbitrary, (B.1.18) holds for any $\boldsymbol{w}$. Furthermore, since the random processes in Lemma 3 and Lemma 4 do not depend on the noise $\boldsymbol{w}$, the conclusion of the theorem applies to an adversarial noise without amplifying the error probability.

Next, we use the following lemma to obtain a lower bound on ϱ in (B.1.1).

Lemma 5 *Let $\mathcal{A} \subset \mathbb{R}^d$ be of finite Gaussian measure and $\boldsymbol{g} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$. Then we have*

$$\inf_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\,\mathbb{1}_{\mathcal{A}}(\boldsymbol{g})\,|\langle \boldsymbol{g}, \boldsymbol{w}\rangle| \geq \sqrt{\frac{\pi}{32}}\,\mathbb{P}^2\{\boldsymbol{g}\in\mathcal{A}\}$$

and

$$\sup_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\mathbb{1}_{\mathcal{A}}(\boldsymbol{g})\langle \boldsymbol{g}, \boldsymbol{w}\rangle_+ \leq \sqrt{\mathbb{P}\{\boldsymbol{g}\in\mathcal{A}\}}\,.$$

Proof 11 *For an arbitrarily fixed $\epsilon > 0$, let $S_\epsilon \subset \mathbb{R}^d$ denote the set defined by*

$$S_\epsilon := \{\boldsymbol{x}\in\mathbb{R}^d : |\langle \boldsymbol{x}, \boldsymbol{w}\rangle| < \epsilon\}\,.$$

Then we have

$$\begin{aligned}
\mathbb{E}\mathbb{1}_{\mathcal{C}}(\boldsymbol{g})|\langle \boldsymbol{g}, \boldsymbol{w}\rangle| &\geq \epsilon\,\mathbb{E}\mathbb{1}_{\mathcal{C}}(\boldsymbol{g})\mathbb{1}_{\mathcal{S}_\epsilon^{\mathrm{c}}}(\boldsymbol{g}) \\
&= \epsilon\,\mathbb{E}\left(\mathbb{1}_{\mathcal{C}}(\boldsymbol{g}) - \mathbb{1}_{\mathcal{C}}(\boldsymbol{g})\mathbb{1}_{\mathcal{S}_\epsilon}(\boldsymbol{g})\right) \\
&\geq \epsilon\,\mathbb{E}\left(\mathbb{1}_{\mathcal{C}}(\boldsymbol{g}) - \mathbb{1}_{\mathcal{S}_\epsilon}(\boldsymbol{g})\right) \\
&= \epsilon\,\left(\mathbb{P}\{\boldsymbol{g}\in\mathcal{C}\} - \mathbb{P}\{\boldsymbol{g}\in\mathcal{S}_\epsilon\}\right)\,.
\end{aligned} \tag{B.1.19}$$

Moreover, since $\langle \boldsymbol{g}, \boldsymbol{w} \rangle \sim \text{Normal}(0,1)$, $\mathbb{P}\{\boldsymbol{g} \in S_\epsilon\}$ *is upper-bounded by*

$$\mathbb{P}\{\boldsymbol{g} \in S_\epsilon\} = \mathbb{P}\{|\langle \boldsymbol{g}, \boldsymbol{w} \rangle| < \epsilon\} = \int_{-\epsilon}^{\epsilon} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \leq \epsilon \sqrt{\frac{2}{\pi}}\,. \tag{B.1.20}$$

By plugging in (B.1.20) *to* (B.1.19)*, we obtain*

$$\mathbb{E}\mathbb{1}_{\mathcal{C}}(\boldsymbol{g}) |\langle \boldsymbol{g}, \boldsymbol{w} \rangle| \geq \epsilon \left(\mathbb{P}\{\boldsymbol{g} \in \mathcal{C}\} - \epsilon \sqrt{\frac{2}{\pi}} \right). \tag{B.1.21}$$

Since the parameter $\epsilon > 0$ *was arbitrary, one can we maximize the right-hand side of* (B.1.21) *with respect to* ϵ *to obtain the tightest lower bound. Note that the objective is a concave quadratic function and the maximum is attained at* $\epsilon = \sqrt{\pi/8}\, \mathbb{P}\{\boldsymbol{g} \in \mathcal{C}\}$*. This provides the lower bound in the first assertion. Next, by the Cauchy-Schwarz inequality, we obtain the upper bound in the second assertion as follows:*

$$\begin{aligned}
\mathbb{E}\mathbb{1}_{\mathcal{A}}(\boldsymbol{g}) \langle \boldsymbol{g}, \boldsymbol{w} \rangle_+ &\leq \sqrt{\mathbb{E}\left(\mathbb{1}_{\mathcal{A}}(\boldsymbol{g})\right)^2} \sqrt{\mathbb{E}\langle \boldsymbol{g}, \boldsymbol{w} \rangle_+^2} \\
&= \sqrt{\mathbb{E}\mathbb{1}_{\mathcal{A}}(\boldsymbol{g})} \sqrt{\frac{\mathbb{E}\langle \boldsymbol{g}, \boldsymbol{w} \rangle^2}{2}} \\
&= \sqrt{\frac{\mathbb{P}\{\boldsymbol{g} \in \mathcal{A}\}}{2}}\,.
\end{aligned}$$

Finally, by applying Lemma 5 to each of the expectation terms in ϱ, we obtain a lower bound on ϱ given by

$$\begin{aligned}
\varrho &\geq \min_{j \in [k]} \sqrt{\frac{\pi}{32}}\, \mathbb{P}^2\{\boldsymbol{g} \in \mathcal{C}_j\} - \max_{j \in [k]} \sqrt{\mathbb{P}\left\{\boldsymbol{g} \in \mathcal{C}_j \setminus \widetilde{\mathcal{C}}_j\right\}} \\
&\quad - \max_{j \in [k]} \sqrt{\mathbb{P}\{\boldsymbol{g} \in \widetilde{\mathcal{C}}_j \setminus \mathcal{C}_j\}} \\
&\geq \min_{j \in [k]} \sqrt{\frac{\pi}{32}}\, \mathbb{P}^2\{\boldsymbol{g} \in \mathcal{C}_j\} - 2 \max_{j \in [k]} \sqrt{\mathbb{P}\left\{\boldsymbol{g} \in \mathcal{C}_j \triangle \widetilde{\mathcal{C}}_j\right\}}\,,
\end{aligned} \tag{B.1.22}$$

where the second inequality holds since $\widetilde{\mathcal{C}}_j \triangle \mathcal{C}_j = (\widetilde{\mathcal{C}}_j \setminus \mathcal{C}_j) \cup (\mathcal{C}_j \setminus \widetilde{\mathcal{C}}_j)$ for all $j \in [k]$. This implies that (2.1.5) is a sufficient condition for (B.1.2). Moreover, substituting ϱ in (B.1.3) by the lower bound in (B.1.22) provides (2.1.6). This completes the proof of Theorem 7.

B.1.1 Tightness of the lower bound on ϱ

In (B.1.22), we obtain a lower bound on ϱ by Lemma 5. We show through the following example that the lower bound is tight in terms of its dependence on $\mathbb{P}\{\boldsymbol{g} \in \mathcal{C}_j\}$ for $j \in [k]$.

Example 1 *Let $d = 2$. Then $\widetilde{\mathcal{C}}_j \setminus \mathcal{C}_j$ and $\mathcal{C}_j \setminus \widetilde{\mathcal{C}}_j$ are Lorentz cones. Let $\theta_{\mathcal{C}_j}$, $\theta_{\mathcal{C}_j \setminus \widetilde{\mathcal{C}}_j}$ and $\theta_{\widetilde{\mathcal{C}}_j \setminus \mathcal{C}_j}$ denote the angular width of $\mathcal{C}_j$, $\mathcal{C}_j \setminus \widetilde{\mathcal{C}}_j$, and $\widetilde{\mathcal{C}}_j \setminus \mathcal{C}_j$ respectively. Furthermore, we assume that*

$$\min_{j \in [k]} \mathbb{P}\{\boldsymbol{g} \in \mathcal{C}_j\} \geq \max_{j \in [k]} \mathbb{P}\left\{\boldsymbol{g} \in \mathcal{C}_j \triangle \widetilde{\mathcal{C}}_j\right\}. \tag{B.1.23}$$

In this case, the parameter ϱ in Proposition 1 is expressed as

$$\varrho = \frac{\sqrt{2}\Gamma(3/2)}{\Gamma(1)} \left[\min_{j \in [k]} \frac{2}{\pi} \sin^2\left(\frac{\theta_{\mathcal{C}_j}}{4}\right) - \max_{j \in [k]} \frac{1}{\pi} \sin\left(\frac{\theta_{\widetilde{\mathcal{C}}_j \setminus \mathcal{C}_j}}{2}\right) - \max_{j \in [k]} \frac{1}{\pi} \sin\left(\frac{\theta_{\mathcal{C}_j \setminus \widetilde{\mathcal{C}}_j}}{2}\right) \right]. \tag{B.1.24}$$

When $\theta_{\mathcal{C}}$ is small enough, $\sin(\theta_{\mathcal{C}}) \approx \theta_{\mathcal{C}}$ holds by the Taylor series approximation. Hence, there exists absolute constants $c_1 > 0$ and $c_2 > 0$ such that

$$\varrho = c_1 \min_{j \in [k]} \mathbb{P}^2\{\boldsymbol{g} \in \mathcal{C}_j\} - c_2 \max_{j \in [k]} \mathbb{P}\{\boldsymbol{g} \in \widetilde{\mathcal{C}}_j \triangle \mathcal{C}_j\}.$$

This example shows that ζ in Theorem 7 is tight in the sense that the dominating term in both ϱ and ζ is proportional to the squared probability measure of the smallest $\mathcal{C}_j$.

Let $\theta_{\mathcal{C}_j}$ denote the angular width of $\mathcal{C}_j$. Without loss of generality, we may assume that $\min_{j \in [k]} \theta_{\mathcal{C}_j} \leq \pi$. Furthermore, the assumption in (B.1.23) implies that the angular width of $\mathcal{C}_j \triangle \widetilde{\mathcal{C}}_j$ is at most π for all $j \in [k]$. Therefore, the identity in (B.1.24) is obtained by applying the following lemma, proved in Appendix B.2.2, to the infimum/supremum of expectation terms in (B.1.1).

Lemma 6 *Let $\mathcal{C}$ be a polyhedral cone in $\mathbb{R}^2$ and $\boldsymbol{g} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_2)$. Suppose that the angular width of $\mathcal{C}$, denoted by $\theta_{\mathcal{C}}$ satisfies $0 \le \theta_{\mathcal{C}} \le \pi$. Then we have*

$$\inf_{\boldsymbol{w} \in \mathbb{S}^1} \mathbb{E}\, \mathbb{1}_{\mathcal{C}}(\boldsymbol{g}) \left| \langle \boldsymbol{g}, \boldsymbol{w} \rangle \right| = \frac{2\sqrt{2}\Gamma(3/2)}{\pi \Gamma(2)} \sin^2\left(\frac{\theta_{\mathcal{C}}}{4}\right)$$

and

$$\sup_{\boldsymbol{w} \in \mathbb{S}^1} \mathbb{E}\mathbb{1}_{\mathcal{C}}(\boldsymbol{g}) \langle \boldsymbol{g}, \boldsymbol{w} \rangle_+ = \frac{\sqrt{2}\Gamma(3/2)}{\pi \Gamma(2)} \sin\left(\frac{\theta_{\mathcal{C}}}{2}\right).$$

Proof 12 *See Appendix B.2.2.*

B.2 Supporting lemmas

B.2.1 Proof of Lemma 1

For brevity, we introduce the shorthand notations

$$\boldsymbol{A}_j = \sum_{i=1}^n \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}, \quad \text{and} \quad \boldsymbol{b}_j = \sum_{i=1}^n \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i) y_i \boldsymbol{x}_i.$$

Then, since each $\mathcal{C}_j$ is given by the intersection of $(k-1)$ half-planes in $\mathbb{R}^d$, by [89, Theorem 2], it holds with probability at least $1 - \delta/3$ that

$$\sup_{j \in [k]} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i) - \mathbb{P}(\boldsymbol{g} \in \mathcal{C}_j) \right| \le C_1 \sqrt{\frac{\log(3/\delta) + kd \log(n/d)}{n}}, \tag{B.2.1}$$

which implies

$$c_2 n \pi_{\min} \le \sum_{i=1}^n \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i) \le C_3 n \pi_{\max}, \quad \forall j \in [k]. \tag{B.2.2}$$

Moreover, by [85, Theorem 5.7], with probability at least $1 - \delta/3$, we have

$$\sup_{\mathcal{I}: |\mathcal{I}| \le \alpha n} \lambda_{\max}\left(\sum_{i \in \mathcal{I}} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}} \right) \le C_4 \sqrt{\alpha} n$$

provided

$$n \ge \max\left(d, \frac{\log(3/\delta)}{\alpha}\right). \tag{B.2.3}$$

We also use the following claim: If

$$n \geq C_5 \theta^{-2} \max(d \log(n/d), \log(3/\delta)), \tag{B.2.4}$$

then it holds with probability $1 - \delta/3$ that

$$\inf_{\mathcal{I} \subset [n]: |\mathcal{I}| \geq \theta n} \lambda_{\min} \left(\sum_{i \in \mathcal{I}} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \geq c_6 n \theta^3. \tag{B.2.5}$$

Proof 13 (Proof of Claim) *For an arbitrarily fixed $T > 0$, we have*

$$\frac{1}{n} \sum_{i \in \mathcal{I}} \langle \boldsymbol{\xi}_i, \boldsymbol{v} \rangle^2 \geq \frac{\theta T}{2}, \quad \forall \mathcal{I} \subset [n] : |\mathcal{I}| \geq \theta n \tag{B.2.6}$$

provided

$$N(\boldsymbol{v}) := \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}: \langle \boldsymbol{x}, \boldsymbol{v} \rangle^2 > T\}}(\boldsymbol{x}_i) > n - \frac{\theta n}{2}. \tag{B.2.7}$$

Since $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{v} \rangle^2 > T\}$ is consists of two half-spaces in $\mathbb{R}^d$, by [89, Theorem 2], there exists an absolute constant $C_7 > 0$, for which it holds with probability at least $1 - \delta/3$ that

$$\frac{1}{n} N(\boldsymbol{v}) \geq \frac{1}{n} \mathbb{E} N(\boldsymbol{v}) - C_7 \sqrt{\frac{d \log(n/d) + \log(3/\delta)}{n}}, \ \forall \boldsymbol{v} \in \mathbb{S}^{d-1}. \tag{B.2.8}$$

Moreover, due to [37, Lemma 15], we have

$$\frac{1}{n} \mathbb{E} N(\boldsymbol{v}) = \mathbb{P}\left(|\langle \boldsymbol{x}, \boldsymbol{v} \rangle|^2 > T \right) \geq 1 - \sqrt{eT}. \tag{B.2.9}$$

Plugging in (C.2.6) *into* (C.2.5) *yields*

$$\frac{1}{n} N(\boldsymbol{v}) \geq 1 - \sqrt{eT} - C_7 \sqrt{\frac{d \log(n/d) + \log(3/\delta)}{n}}, \ \forall \boldsymbol{v} \in \mathbb{S}^{d-1}.$$

Then (C.2.4) *is satisfied for all $\boldsymbol{v} \in \mathbb{S}^d$ by $T = \frac{\theta^2}{16e}$ and $C_5 = (4C_7)^2$.*

Since (2.1.8) implies (C.2.1) and (C.2.2), combining the above results provides that

$$c_8 n \pi_{\min}^3 \leq \lambda_{\min}(\boldsymbol{A}_j) \leq \lambda_{\max}(\boldsymbol{A}_j) \leq C_9 n \sqrt{\pi_{\max}}, \quad \forall j \in [k],$$

holds with probability $1-\delta$. Then the least squares solution in (1.2.5) with $b_j = 0$ for all $j \in [k]$ is written as $\widehat{\boldsymbol{\theta}}_j = \boldsymbol{A}_j^{-1}\boldsymbol{b}_j$ and satisfies

$$\begin{aligned}\|\boldsymbol{\theta}_{\star,j} - \widehat{\boldsymbol{\theta}}_j\|_2 &\geq \frac{\lambda_{\min}^{1/2}(\boldsymbol{A}_j)}{\lambda_{\max}(\boldsymbol{A}_j)} \left\|(z_i)_{i:\boldsymbol{x}_i\in\mathcal{C}_j}\right\|_2 \\ &\geq \frac{c_{10}\pi_{\min}^{3/2}}{\sqrt{n\pi_{\max}}} \left\|(z_i)_{i:\boldsymbol{x}_i\in\mathcal{C}_j}\right\|_2 \geq \frac{c_{11}\pi_{\min}^{3/2}}{\pi_{\max}} \left\|(z_i)_{i:\boldsymbol{x}_i\in\mathcal{C}_j}\right\|_\infty .\end{aligned} \tag{B.2.10}$$

Then taking a sum over $j \in [k]$ and maximizing over $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}\|_\infty \leq \eta'$, we obtain

$$\sup_{\|\boldsymbol{w}\|_\infty\leq\eta'} \sum_{j=1}^{k} \|\boldsymbol{\theta}_{\star,j} - \widehat{\boldsymbol{\theta}}_j\|_2 \geq \frac{c_{12}\pi_{\min}^{3/2}\eta'}{\pi_{\max}}.$$

This completes the proof.

B.2.2 Proof of Lemma 6

We first prove the first assertion. Since $\mathcal{C}$ is a cone, it follows that $\boldsymbol{g} \in \mathcal{C}$ if and only $\boldsymbol{g}/\|\boldsymbol{g}\|_2 \in \mathcal{C}$. Moreover, Bayes' rule implies

$$\mathbb{E}\,\mathbb{1}_{\mathcal{C}}(\boldsymbol{g})\,|\langle \boldsymbol{g}, \boldsymbol{w}\rangle| = \mathbb{P}\left\{\boldsymbol{g}\in\mathcal{C}\right\}\mathbb{E}\left[\,|\langle \boldsymbol{g}, \boldsymbol{w}\rangle| \mid \boldsymbol{g}\in\mathcal{C}\right].$$

Therefore we have

$$\begin{aligned}&\inf_{\boldsymbol{w}\in\mathbb{S}^1} \mathbb{E}\,\mathbb{1}_{\mathcal{C}}(\boldsymbol{g})\,|\langle \boldsymbol{g}, \boldsymbol{w}\rangle| \\ &= \inf_{\boldsymbol{w}\in\mathbb{S}^1} \mathbb{P}\left\{\boldsymbol{g}\in\mathcal{C}\right\}\mathbb{E}\left[\|\boldsymbol{g}\|_2 \left|\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\rangle\right| \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}\in\mathcal{C}\right] \\ &\overset{\text{(a)}}{=} \inf_{\boldsymbol{w}\in\mathbb{S}^1} \mathbb{P}\left\{\boldsymbol{g}\in\mathcal{C}\right\}\mathbb{E}\left[\|\boldsymbol{g}\|_2\right]\mathbb{E}\left[\left|\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle\right| \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}\in\mathcal{C}\right] \\ &\overset{\text{(b)}}{=} \frac{\sqrt{2}\Gamma(3/2)}{\Gamma(2)} \inf_{\boldsymbol{w}\in\mathbb{S}^1} \frac{\theta_{\mathcal{C}}}{2\pi}\mathbb{E}\left[\left|\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle\right| \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}\in\mathcal{C}\right],\end{aligned} \tag{B.2.11}$$

where (a) holds since $\|\boldsymbol{g}\|_2$ and $\boldsymbol{g}/\|\boldsymbol{g}\|_2$ are independent and (b) follows from $\mathbb{E}\|\boldsymbol{g}\|_2 = \sqrt{2}\Gamma(3/2)/\Gamma(2)$ and

$$\mathbb{P}\{\boldsymbol{g}\in\mathcal{C}\} = \mathbb{P}\left\{\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}\in\mathcal{C}\right\} = \frac{\theta_{\mathcal{C}}}{2\pi}.$$

Then it remains to compute the expectation in (B.2.11). Below we show that

$$\inf_{\boldsymbol{w}\in\mathbb{S}^2} \mathbb{E}\left[\left|\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle\right| \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right] = \frac{4}{\theta_{\mathcal{C}}} \sin^2\left(\frac{\theta_{\mathcal{C}}}{2}\right) \tag{B.2.12}$$

and

$$\sup_{\boldsymbol{w}\in\mathbb{S}^2} \mathbb{E}\left[\left|\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle\right| \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right] = \frac{2}{\theta_{\mathcal{C}}} \sin\left(\frac{\theta_{\mathcal{C}}}{2}\right) . \tag{B.2.13}$$

Let $\mathcal{T} = \{\boldsymbol{a}, \boldsymbol{b}\} \subset \mathbb{S}^1$ satisfy that $\mathcal{C}$ is the conic hull of $\mathcal{T}$. Then let $\boldsymbol{h}$ be the unit vector obtained by normalizing $(\boldsymbol{a}+\boldsymbol{b})/2$. Then we have $\angle(\boldsymbol{a}, \boldsymbol{h}) = \theta_{\mathcal{C}}/2$ and $\angle(\boldsymbol{b}, \boldsymbol{h}) = \theta_{\mathcal{C}}/2$. Let $\phi : \mathbb{S}^1 \to \mathbb{R}$ be defined by $\phi(\boldsymbol{w}) := \angle(\boldsymbol{h}, \boldsymbol{w})$. Since the conditional expectation applies to $|\langle \boldsymbol{g}/\|\boldsymbol{g}\|_2, \boldsymbol{w}\rangle|$, which is invariant under the global sign change in $\boldsymbol{w}$, it suffices to consider $\boldsymbol{w}$ that satisfies $0 \le \phi(\boldsymbol{w}) \le \pi$. Since $\boldsymbol{g}/\|\boldsymbol{g}\|_2$ is uniformly distributed on the unit sphere, the expectation term in (B.2.12) is written as

$$\mathbb{E}\left[\left|\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle\right| \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right] = \frac{1}{\theta_{\mathcal{C}}} \int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} |\cos\theta| d\theta \,. \tag{B.2.14}$$

It follows from the assumption on the range of $\theta_{\mathcal{C}}$ and $\phi(\boldsymbol{w})$ that $-\pi/2 \le \phi(\boldsymbol{w}) - \theta_{\mathcal{C}}/2 \le \pi$ and $0 \le \phi(\boldsymbol{w}) + \theta_{\mathcal{C}}/2 \le 3\pi/2$. We proceed by separately considering the complementary cases for $(\theta_{\mathcal{C}}, \phi(\boldsymbol{w}))$ given below.

Case 1: Suppose that

$$-\frac{\pi}{2} \le \phi(\boldsymbol{w}) - \frac{\theta_{\mathcal{C}}}{2} < \phi(\boldsymbol{w}) + \frac{\theta_{\mathcal{C}}}{2} \le \frac{\pi}{2} \,. \tag{B.2.15}$$

Then $\phi(\boldsymbol{w})$ is constrained by

$$0 \le \phi(\boldsymbol{w}) \le \pi/2 - \theta_{\mathcal{C}}/2 \,. \tag{B.2.16}$$

Furthermore, the integral in (B.2.14) is rewritten as

$$\begin{aligned}\int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} |\cos\theta| d\theta &= \int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} \cos\theta d\theta \\ &= \sin\left(\phi(\boldsymbol{w})+\frac{\theta_{\mathcal{C}}}{2}\right) - \sin\left(\phi(\boldsymbol{w})-\frac{\theta_{\mathcal{C}}}{2}\right) \\ &= 2\cos\left(\phi(\boldsymbol{w})\right)\sin\left(\frac{\theta_{\mathcal{C}}}{2}\right) . \end{aligned} \tag{B.2.17}$$

Since $\sin(\theta_{\mathcal{C}}/2) \geq 0$, the expression in (B.2.17) monotonically decreases in $\phi(\boldsymbol{w})$ for the interval given in (B.2.16). Thus the maximum (resp. minimum) is attained as $2\sin(\theta_{\mathcal{C}}/2)$ at $\phi(\boldsymbol{w}) = 0$ (resp. $2\sin^2(\theta_{\mathcal{C}}/2)$ at $\phi(\boldsymbol{w}) = \pi/2 - \theta_{\mathcal{C}}/2$).

Case 2: Suppose that

$$-\frac{\pi}{2} \leq \phi(\boldsymbol{w}) - \frac{\theta_{\mathcal{C}}}{2} < \frac{\pi}{2} < \phi(\boldsymbol{w}) + \frac{\theta_{\mathcal{C}}}{2} \leq \frac{3\pi}{2} . \tag{B.2.18}$$

Then $\phi(\boldsymbol{w})$ satisfies

$$\frac{\pi}{2} - \frac{\theta_{\mathcal{C}}}{2} \leq \phi(\boldsymbol{w}) \leq \frac{\pi}{2} + \frac{\theta_{\mathcal{C}}}{2} \tag{B.2.19}$$

and the integral in (B.2.14) reduces to

$$\begin{aligned}&\int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} |\cos\theta| d\theta \\ &\qquad = \int_{\phi(\boldsymbol{w})-\frac{\theta_{\mathcal{C}}}{2}}^{\frac{\pi}{2}} \cos\theta d\theta - \int_{\frac{\pi}{2}}^{\phi(\boldsymbol{w})+\frac{\theta_{\mathcal{C}}}{2}} \cos\theta d\theta \\ &\qquad = 2 - \sin\left(\phi(\boldsymbol{w})-\frac{\theta_{\mathcal{C}}}{2}\right) - \sin\left(\phi(\boldsymbol{w})+\frac{\theta_{\mathcal{C}}}{2}\right) \\ &\qquad = 2 - 2\sin\left(\phi(\boldsymbol{w})\right)\cos\left(\frac{\theta_{\mathcal{C}}}{2}\right) . \end{aligned} \tag{B.2.20}$$

Since $\cos(\theta_{\mathcal{C}}/2) \geq 0$ for all $\theta_{\mathcal{C}} \in [0, \pi]$, the maximum (resp. minimum) is attained as $2\sin^2(\theta_{\mathcal{C}}/2)$ at $\phi(\boldsymbol{w}) = \pi/2 - \theta_{\mathcal{C}}/2$ (resp. $4\sin^2(\theta_{\mathcal{C}}/4)$ at $\phi(\boldsymbol{w}) = \pi/2$).

Case 3: Suppose that

$$\frac{\pi}{2} \le \phi(\boldsymbol{w}) - \frac{\theta_{\mathcal{C}}}{2} < \phi(\boldsymbol{w}) + \frac{\theta_{\mathcal{C}}}{2} \le \frac{3\pi}{2}\,. \tag{B.2.21}$$

Then we have

$$\frac{\pi}{2} + \frac{\theta_{\mathcal{C}}}{2} \le \phi(\boldsymbol{w}) \le \pi \tag{B.2.22}$$

and

$$\begin{aligned}
&\int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} |\cos\theta| d\theta \\
&\qquad = \int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} (-\cos\theta) d\theta \\
&\qquad = \sin\left(\phi(\boldsymbol{w}) - \frac{\theta_{\mathcal{C}}}{2}\right) - \sin\left(\phi(\boldsymbol{w}) + \frac{\theta_{\mathcal{C}}}{2}\right) \\
&\qquad = -2\cos\phi(\boldsymbol{w}) \sin\frac{\theta_{\mathcal{C}}}{2}\,.
\end{aligned} \tag{B.2.23}$$

The maximum (resp. minimum) of (B.2.23) is attained as $2\sin(\theta_{\mathcal{C}}/2)$ at $\phi(\boldsymbol{w}) = \pi$ (resp. $2\sin^2(\theta_{\mathcal{C}}/2)$ at $\phi(\boldsymbol{w}) = \pi/2 + \theta_{\mathcal{C}}/2$).

By combining the results in the above three cases, we obtain (B.2.12) and (B.2.13). Then substituting the expectation term in (B.2.11) by (B.2.12) provides the first assertion.

Next we prove the second assertion. Similarly to (B.2.11), we have

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathbb{S}^1} \mathbb{E}\mathbb{1}_{\mathcal{C}_j}(\boldsymbol{g})\langle \boldsymbol{g}, \boldsymbol{w}\rangle_+ \\
&= \sup_{\boldsymbol{w}\in\mathbb{S}^1} P\left\{\boldsymbol{g}\in\mathcal{C}\right\} \mathbb{E}\left[\|\boldsymbol{g}\|_2 \cdot \left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle_+ \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right] \\
&\overset{(a)}{=} \sup_{\boldsymbol{w}\in\mathbb{S}^1} P\left\{\boldsymbol{g}\in\mathcal{C}\right\} \mathbb{E}\left[\|\boldsymbol{g}\|_2\right] \mathbb{E}\left[\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle_+ \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right] \\
&\overset{(b)}{=} \frac{\sqrt{2}\Gamma(3/2)}{\Gamma(2)} \sup_{\boldsymbol{w}\in\mathbb{S}^1} \frac{\theta_{\mathcal{C}}}{2\pi} \mathbb{E}\left[\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle_+ \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right]\,,
\end{aligned}$$

where (a) holds since $\|\boldsymbol{g}\|_2$ and $\boldsymbol{g}/\|\boldsymbol{g}\|_2$ are independent, (b) follows from $\mathbb{E}\|\boldsymbol{g}\|_2 = \sqrt{2}\Gamma(3/2)/\Gamma(2)$, and

$$\mathbb{P}\{\boldsymbol{g} \in \mathcal{C}\} = \mathbb{P}\left\{\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right\} = \frac{\theta_{\mathcal{C}}}{2\pi}\,.$$

If suffices to show that

$$\max_{\boldsymbol{w}\in\mathbb{S}^1} \mathbb{E}\left[\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle_+ \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right] = \frac{2}{\theta_{\mathcal{C}}}\sin\left(\frac{\theta_{\mathcal{C}}}{2}\right)\,. \tag{B.2.24}$$

Since $\boldsymbol{g}/\|\boldsymbol{g}\|_2$ is uniformly distributed on the unit sphere $\mathbb{S}^1$ and $u_+ = (u + |u|)/2$ for all $u \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle_+ \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right] \\
&= \int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} \frac{\cos\theta + |\cos\theta|}{2\theta_{\mathcal{C}}} d\theta \\
&= \frac{1}{2}\left(\frac{1}{\theta_{\mathcal{C}}}\int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} |\cos\theta| d\theta + \frac{1}{\theta_{\mathcal{C}}}\int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} \cos\theta d\theta\right) \\
&= \frac{1}{2}\Bigg(\mathbb{E}\left[\left|\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle\right| \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right] \\
&\qquad\qquad + \mathbb{E}\left[\left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \boldsymbol{w}\right\rangle \,\middle|\, \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \in \mathcal{C}\right]\Bigg).
\end{aligned} \tag{B.2.25}$$

As shown above, the first term in (B.2.25) is maximized at $\phi(\boldsymbol{w}) = 0$ and the maximum is given in (B.2.13). Furthermore, the second term in (B.2.25) is rewritten as

$$\begin{aligned}
\int_{\phi(\boldsymbol{w})-\theta_{\mathcal{C}}/2}^{\phi(\boldsymbol{w})+\theta_{\mathcal{C}}/2} \cos\theta d\theta &= \sin\left(\phi(\boldsymbol{w}) + \frac{\theta_{\mathcal{C}}}{2}\right) - \sin\left(\phi(\boldsymbol{w}) - \frac{\theta_{\mathcal{C}}}{2}\right) \\
&= 2\cos\phi(\boldsymbol{w})\sin\left(\frac{\theta_{\mathcal{C}}}{2}\right)\,.
\end{aligned} \tag{B.2.26}$$

Since $\sin(\theta_{\mathcal{C}}/2) \geq 0$, the expression in (B.2.26) is a decreasing function of $\phi(\boldsymbol{w}) \in [0, \pi]$. Hence, the maximum is attained at $\phi(\boldsymbol{w}) = 0$ as

$$\max_{\boldsymbol{w}\in\mathbb{S}^1} 2\cos\phi(\boldsymbol{w})\sin\left(\frac{\theta_{\mathcal{C}}}{2}\right) = 2\sin\left(\frac{\theta_{\mathcal{C}}}{2}\right)\,. \tag{B.2.27}$$

Since the two terms in (B.2.25) are maximized simultaneously, by plugging in the above results to (B.2.24), the second assertion is obtained.

B.2.3 Proof of Lemma 2

By construction, we have

$$\frac{1}{n}\sum_{i=1}^{N} z_i \boldsymbol{x}_i \sim \text{Normal}\left(\boldsymbol{0}, \frac{\|\boldsymbol{w}\|_2^2}{n^2}\boldsymbol{I}_d\right).$$

Then, the concentration of the Euclidean norm of a standard Gaussian vector guarantees, with probability at least $1-\delta/2$, that

$$\left\|\frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{x}_i\right\|_2 \lesssim \frac{\|\boldsymbol{w}\|_2}{n}(\sqrt{d}+\sqrt{\log(1/\delta)}) \tag{B.2.28}$$

for some absolute constant C. This implies the first bound in (2.1.12).

Next, we want to obtain an upper bound on the second term in (2.1.12). By the variational characterization of the spectral norm, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n} z_i\left(\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}} - \boldsymbol{I}_d\right)\right\| \leq \sup_{\boldsymbol{u}\in\mathbb{B}_2^d}\left|\frac{1}{n}\sum_{i=1}^{n} z_i\left((\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{u})^2-1\right)\right|. \tag{B.2.29}$$

For brevity, we introduce a shorthand notation to denote the following random process

$$Y_{\boldsymbol{u}} := \sum_{i=1}^{n} z_i\left((\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{u})^2-1\right),$$

indexed by $\boldsymbol{u}\in\mathbb{B}_2^d$. Then, for $\boldsymbol{u},\boldsymbol{u}'\in\mathbb{B}_2^d$, we have

$$Y_{\boldsymbol{u}} - Y_{\boldsymbol{u}} = \sum_{i=1}^{n} z_i\langle\boldsymbol{x}_i, \boldsymbol{u}-\boldsymbol{u}'\rangle\langle\boldsymbol{x}_i, \boldsymbol{u}+\boldsymbol{u}'\rangle.$$

Therefore, we bound the subexponential norm of each summand as

$$\begin{aligned}&\|z_i\langle\boldsymbol{x}_i, \boldsymbol{u}-\boldsymbol{u}'\rangle\langle\boldsymbol{x}_i, \boldsymbol{u}+\boldsymbol{u}'\rangle\|_{\psi_1}\\ &\leq z_i\|\langle\boldsymbol{x}_i, \boldsymbol{u}-\boldsymbol{u}'\rangle\|_{\psi_2}\cdot\|\boldsymbol{x}_i, \boldsymbol{u}+\boldsymbol{u}'\|_{\psi_2} \lesssim z_i\|\boldsymbol{u}-\boldsymbol{u}'\|_2.\end{aligned}$$

Applying the Bernstein inequality (e.g. see [93, Theorem 2.8.1]) then yields

$$\begin{aligned}&\mathbb{P}\left(|Y_{\boldsymbol{u}} - Y_{\boldsymbol{u}'}| \geq c\left(\sqrt{t}\|\boldsymbol{w}\|_2\|\boldsymbol{u}-\boldsymbol{u}'\|_2 + t\|\boldsymbol{w}\|_\infty\|\boldsymbol{u}-\boldsymbol{u}'\|_2\right)\right)\\ &\qquad\leq 2\exp(-t),\end{aligned} \tag{B.2.30}$$

for any $t \geq 0$ and an absolute constant c. Then, the process $Y_{\boldsymbol{u}}$ has mixed tail increments (i.e, see [29, Equation 12]) with respect to the metrics (d_1, d_2) where $d_1(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{w}\|_\infty \|\boldsymbol{a} - \boldsymbol{b}\|_2$ and $d_2(a, b) = \|\boldsymbol{w}\|_2 \|\boldsymbol{a} - \boldsymbol{b}\|_2$ for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{B}_2^d$. Hence, applying [29, Corollary 5.2] with the bound on γ-functional (i.e, see [29, Equation 4]) provides

$$\begin{aligned}
&\sup_{\boldsymbol{u} \in \mathbb{B}_2^d} |Y_{\boldsymbol{u}}| \\
&\lesssim \|\boldsymbol{w}\|_2 \left(\int_0^\infty \sqrt{\log N\left(\mathbb{B}_2^d, \|\cdot\|_2, \eta\right)} d\eta + \sqrt{\log(1/\delta)} \right) \\
&\qquad + \|\boldsymbol{w}\|_\infty \left(\int_0^\infty \log N\left(\mathbb{B}_2^d, \|\cdot\|_2, \eta\right) d\eta + \log(1/\delta) \right) \\
&\overset{\text{(b)}}{\leq} \|\boldsymbol{w}\|_2 (\sqrt{d} + \sqrt{\log(1/\delta)}) + \|\boldsymbol{w}\|_\infty (p + \log(1/\delta)),
\end{aligned}$$

holds with probability at least $1 - \delta/2$ where (b) holds due to an upper bound on the covering number $N(\mathbb{B}_2^d, \|\cdot\|_2, \eta) \leq (3/\eta)^d$ (e.g. see [93, Example 8.1.11]). This implies the second bound in (2.1.12).

B.2.4 Proof of Lemma 3

For any $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\|_{1,2} = 1$, we have

$$V_{\boldsymbol{h}} \geq \min_{\|\boldsymbol{h}\|_{1,2}=1} \mathbb{E} V_{\boldsymbol{h}} - \sup_{\boldsymbol{h} \in B_{1,2}} |V_{\boldsymbol{h}} - \mathbb{E} V_{\boldsymbol{h}}| . \tag{B.2.31}$$

In what follows, we derive lower estimates of the summands in the right-hand side of (B.2.31).

First, we derive a lower bound on $\min_{\|\boldsymbol{h}\|_{1,2}=1} \mathbb{E} V_{\boldsymbol{h}}$. Since $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are i.i.d. Normal$(\boldsymbol{0}, \boldsymbol{I}_d)$, we have

$$\begin{aligned}
\mathbb{E} V_{\boldsymbol{h}} &= \mathbb{E} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i) |\langle \boldsymbol{x}_i, \boldsymbol{h}_j \rangle| = \mathbb{E} \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{g}) |\langle \boldsymbol{g}, \boldsymbol{h}_j \rangle| \\
&= \sum_{j=1}^k \|\boldsymbol{h}_j\|_2 \mathbb{E}\, \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{g}) \left| \left\langle \boldsymbol{g}, \frac{\boldsymbol{h}_j}{\|\boldsymbol{h}_j\|_2} \right\rangle \right| ,
\end{aligned}$$

where $\boldsymbol{h} = [\boldsymbol{h}_1;\ \ldots;\ \boldsymbol{h}_k]$. Then $\mathbb{E}V_{\boldsymbol{h}}$ is lower-bounded by

$$\mathbb{E}V_{\boldsymbol{h}} \geq \|\boldsymbol{h}\|_{1,2} \inf_{j\in[k],\boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\,\mathbb{1}_{\mathcal{C}_j}(\boldsymbol{g})\,|\langle \boldsymbol{g}, \boldsymbol{w}\rangle| \ .$$

Next, we show that $(V_{\boldsymbol{h}} - \mathbb{E}V_{\boldsymbol{h}})_{\boldsymbol{h}\in B_{1,2}}$ is concentrated around 0 with high probability by using the following lemma.

Lemma 7 *Suppose that $\mathcal{A}_1, \ldots, \mathcal{A}_k$ be disjoint subsets in $\mathbb{R}^d$. Let $(U_{\boldsymbol{h}})_{\boldsymbol{h}\in B_{1,2}}$ be a random process defined by*

$$U_{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \mathbb{1}_{\mathcal{A}_j}(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, \boldsymbol{h}_j\rangle_+ \,, \tag{B.2.32}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are i.i.d. $\mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_d)$*. Then, for any $\delta \in (0,1)$, there exists an absolute constant $c > 0$ such that*

$$\sup_{\boldsymbol{h}\in B_{1,2}} |U_{\boldsymbol{h}} - \mathbb{E}U_{\boldsymbol{h}}| \leq c\left(\frac{d\log^3 d \log^5 k + \log(\delta^{-1})\log k}{n}\right)^{1/2} \tag{B.2.33}$$

holds with probability at least $1-\delta$.

Proof 14 *We first show that $U_{\boldsymbol{h}}$ has sub-Gaussian increments with respect to the $\ell_\infty^k(\ell_2^d)$-norm, i.e.*

$$\|U_{\boldsymbol{h}} - U_{\boldsymbol{h}'}\|_{\psi_2} \lesssim \frac{\sqrt{\log k}}{\sqrt{n}} \left\|(\boldsymbol{h}_j)_{j=1}^k - \left(\boldsymbol{h}'_j\right)_{j=1}^k\right\|_{\ell_\infty^k(\ell_2^d)} . \tag{B.2.34}$$

Since $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are disjoint, it follows that

$$\begin{aligned} |U_{\boldsymbol{h}} - U_{\boldsymbol{h}'}| &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \mathbb{1}_{\mathcal{A}_j}(\boldsymbol{x}_i)\left|\langle \boldsymbol{x}_i, \boldsymbol{h}_j - \boldsymbol{h}'_j\rangle\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \max_{1\leq j\leq k}\left|\langle \boldsymbol{x}_i, \boldsymbol{h}_j - \boldsymbol{h}'_j\rangle\right| \end{aligned} \tag{B.2.35}$$

holds almost surely, where the last step follows from Hölder's inequality. We proceed with the following lemma.

Lemma 8 ([88, Lemma 2.2.2]) *Let* $\boldsymbol{g} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$ *and* $\boldsymbol{a}_1, \dots, \boldsymbol{a}_k \in \mathbb{R}^d$. *Then*

$$\left\| \max_{j\in[k]} |\langle \boldsymbol{g}, \boldsymbol{a}_j \rangle| \right\|_{\psi_2} \lesssim \sqrt{\log k} \max_{j\in[k]} \|\boldsymbol{a}_j\|_2 .$$

It follows from (B.2.35) *and Lemma 8 that*

$$\begin{aligned}
\|U_{\boldsymbol{h}} - U_{\boldsymbol{h}'}\|_{\psi_2} &\le \left\| \frac{1}{n} \sum_{i=1}^n \max_{j\in[k]} \left| \langle \boldsymbol{x}_i, \boldsymbol{h}_j - \boldsymbol{h}'_j \rangle \right| \right\|_{\psi_2} \\
&\lesssim \frac{1}{n} \sqrt{\sum_{i=1}^n \left\| \max_{j\in[k]} \left| \langle \boldsymbol{x}_i, \boldsymbol{h}_j - \boldsymbol{h}'_j \rangle \right| \right\|_{\psi_2}^2} \\
&\lesssim \frac{\sqrt{\log k}}{\sqrt{n}} \max_{j\in[k]} \left\| \boldsymbol{h}_j - \boldsymbol{h}'_j \right\|_2 \\
&= \frac{\sqrt{\log k}}{\sqrt{n}} \left\| (\boldsymbol{h}_j)_{j=1}^k - \left(\boldsymbol{h}'_j\right)_{j=1}^k \right\|_{\ell_\infty^k(\ell_2^d)},
\end{aligned}$$

where the second inequality follows from [93, Proposition 2.6.1].

Since U_z *has a sub-Gaussian increment as in* (B.2.34)*, by [93, Lemma 2.6.8], which says that centering does not harm the sub-gaussianity, we also have*

$$\begin{aligned}
&\|(U_{\boldsymbol{h}} - \mathbb{E}U_{\boldsymbol{h}}) - (U_{\boldsymbol{h}'} - \mathbb{E}U_{\boldsymbol{h}'})\|_{\psi_2} \\
&\qquad \lesssim \frac{\sqrt{\log k}}{\sqrt{n}} \left\| (\boldsymbol{h}_j)_{j=1}^k - \left(\boldsymbol{h}'_j\right)_{j=1}^k \right\|_{\ell_\infty^k(\ell_2^d)} .
\end{aligned} \tag{B.2.36}$$

Therefore Dudley's inequality [32] applies to provide a tail bound on the left-hand side of (B.2.33)*. Specifically it follows from a version of Dudley's inequality [93, Theorem 8.1.6] that*

$$\begin{aligned}
&\sup_{\boldsymbol{h}\in B_{1,2}} |U_{\boldsymbol{h}} - \mathbb{E}U_{\boldsymbol{h}}| \lesssim \\
&\frac{\sqrt{\log k}}{\sqrt{n}} \left(\int_0^\infty \sqrt{\log N(B_{1,2}, \|\cdot\|_{\ell_\infty^k(\ell_2^d)}, \eta)} d\eta + u \operatorname{diam}(B_{1,2}) \right)
\end{aligned} \tag{B.2.37}$$

holds with probability at least $1 - 2\exp(-u^2)$. *Note that the diameter term in* (B.2.37) *is trivially upper-bounded by*

$$diam(B_{1,2}) = \sup_{\boldsymbol{h},\boldsymbol{h}' \in B_{1,2}} \|\boldsymbol{h} - \boldsymbol{h}'\|_{\ell_\infty^k(\ell_2^d)} \leq 2\,.$$

Moreover, since $B_{1,2} \subseteq \sqrt{d}B_1$, *where* B_1 *denotes the unit ball in* ℓ_1, *we have*

$$\begin{aligned}
&\int_0^\infty \sqrt{\log N(B_{1,2}, \|\cdot\|_{\ell_\infty^k(\ell_2^d)}, \eta)}d\eta \\
&\qquad \leq \int_0^\infty \sqrt{\log N(\sqrt{d}B_1, \|\cdot\|_{\ell_\infty^k(\ell_2^d)}, \eta)}d\eta \\
&\qquad \lesssim \sqrt{d}\log^{3/2} d \log^2 k\,,
\end{aligned}$$

where the second inequality follows from Maurey's empirical method [19] (also see [57, Lemma 3.4]). By plugging in these estimates to (B.2.37), *we obtain that*

$$\sup_{\boldsymbol{h}\in B_{1,2}} |U_{\boldsymbol{h}} - \mathbb{E}U_{\boldsymbol{h}}| \lesssim \left(\frac{d\log^3 d \log^5 k + \log(\delta^{-1})\log k}{n}\right)^{1/2}$$

holds with probability at least $1 - \delta$.

Note that $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are disjoint except on a boundary, which corresponds to a set of measure zero. Since the standard multivariate normal distribution is absolutely continuous relative to the Lebesgue measure, these null sets can be ignored in getting a tail bound on the infimum of the random process $(V_{\boldsymbol{h}})_{\boldsymbol{h}\in B_{1,2}}$. Moreover, $V_{\boldsymbol{h}}$ is written as $V_{\boldsymbol{h}} = V_{\boldsymbol{h}}^+ + V_{\boldsymbol{h}}^-$, where

$$V_{\boldsymbol{h}}^+ := \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, \boldsymbol{h}_j\rangle_+$$

and

$$V_{\boldsymbol{h}}^- := \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j}(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, -\boldsymbol{h}_j\rangle_+\,.$$

Since $(V_{\boldsymbol{h}}^+)_{\boldsymbol{h}\in B_{1,2}}$ and $(V_{\boldsymbol{h}}^-)_{\boldsymbol{h}\in B_{1,2}}$ are in the form of (B.2.32), by Lemma 7, we obtain that

$$\begin{aligned}\sup_{\boldsymbol{h}\in B_{1,2}} |V_{\boldsymbol{h}} - \mathbb{E}V_{\boldsymbol{h}}| &\le \sup_{\boldsymbol{h}\in B_{1,2}} \left|V_{\boldsymbol{h}}^+ - \mathbb{E}V_{\boldsymbol{h}}^+\right| + \sup_{\boldsymbol{h}\in B_{1,2}} \left|V_{\boldsymbol{h}}^- - \mathbb{E}V_{\boldsymbol{h}}^-\right| \\ &\lesssim \left(\frac{d\log^3 d\log^5 k + \log(\delta^{-1})\log k}{n}\right)^{1/2} \end{aligned} \tag{B.2.38}$$

holds with probability at least $1-\delta/2$.

Finally, the assertion is obtained by plugging in the above estimates to (B.2.31).

B.2.5 Proof of Lemma 4

Note that $Q_{\boldsymbol{h}}$ is decomposed into

$$\begin{aligned}Q_{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\widetilde{\mathcal{C}}_j\setminus\mathcal{C}_j}(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, \boldsymbol{h}_j\rangle \\ + \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j\setminus\widetilde{\mathcal{C}}_j}(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, -\boldsymbol{h}_j\rangle\,. \end{aligned} \tag{B.2.39}$$

Then the summands in the right-hand side of (B.2.39) are respectively upper-bounded by

$$Q'_{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\widetilde{\mathcal{C}}_j\setminus\mathcal{C}_j}(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, \boldsymbol{h}_j\rangle_+$$

and

$$Q''_{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j\setminus\widetilde{\mathcal{C}}_j}(\boldsymbol{x}_i)\langle \boldsymbol{x}_i, -\boldsymbol{h}_j\rangle_+\,.$$

We upper-bound $\sup_{\boldsymbol{h}\in B_{1,2}} Q'_{\boldsymbol{h}}$ and $\sup_{\boldsymbol{h}\in B_{1,2}} Q''_{\boldsymbol{h}}$ to get an upper bound on $\sup_{\boldsymbol{h}\in B_{1,2}} Q_{\boldsymbol{h}}$ through (B.2.39) by the triangle inequality. Specifically, we show that there exists an absolute constant $c>0$ such that

$$\begin{aligned}\sup_{\|\boldsymbol{h}\|_{1,2}=1} Q'_{\boldsymbol{h}} \le \sup_{j\in[k],\boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\mathbb{1}_{\widetilde{\mathcal{C}}_j\setminus\mathcal{C}_j}(\boldsymbol{g})\langle \boldsymbol{g}, \boldsymbol{w}\rangle_+ \\ + c\left(\frac{d\log^3 d\log^5 k + \log(\delta^{-1})\log k}{n}\right)^{1/2} \end{aligned} \tag{B.2.40}$$

and

$$\sup_{\|\boldsymbol{h}\|_{1,2}=1} Q''_{\boldsymbol{h}} \leq \sup_{j\in[k],\boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\mathbb{1}_{\mathcal{C}_j\setminus\widetilde{\mathcal{C}}_j}(\boldsymbol{g})\langle \boldsymbol{g},\boldsymbol{w}\rangle_+ + c\left(\frac{d\log^3 d\log^5 k+\log(\delta^{-1})\log k}{n}\right)^{1/2}$$

hold simultaneously with probability at least $1-\delta/2$.

Due to the symmetry, it suffices to show that (B.2.40) holds with probability $1-\delta/4$. By the triangle inequality, it follows that

$$\sup_{\|\boldsymbol{h}\|_{1,2}=1} Q'_{\boldsymbol{h}} \leq \sup_{\|\boldsymbol{h}\|_{1,2}=1} \mathbb{E}Q'_{\boldsymbol{h}} + \sup_{\boldsymbol{h}\in B_{1,2}} \left|Q'_{\boldsymbol{h}} - \mathbb{E}Q'_{\boldsymbol{h}}\right|.$$

Then, similar to Lemma 3, we derive (B.2.40) through the concentration of the maximum deviation, that is, $\sup_{\boldsymbol{h}\in B_{1,2}} |Q'_{\boldsymbol{h}} - \mathbb{E}Q'_{\boldsymbol{h}}|$, and an upper bound on $\sup_{\boldsymbol{h}\in B_{1,2}} \mathbb{E}Q'_{\boldsymbol{h}}$. The supremum of the expectation is upper-bounded as

$$\begin{aligned}\mathbb{E}Q'_{\boldsymbol{h}} &= \mathbb{E}\sum_{j=1}^{k} \mathbb{1}_{\widetilde{\mathcal{C}}_j\setminus\mathcal{C}_j}(\boldsymbol{g})\langle \boldsymbol{g},\boldsymbol{h}_j\rangle_+ \\ &\leq \max_{j\in[k],\boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}\mathbb{1}_{\widetilde{\mathcal{C}}_j\setminus\mathcal{C}_j}(\boldsymbol{g})\langle \boldsymbol{g},\boldsymbol{w}\rangle_+ \sum_{j=1}^{k}\|\boldsymbol{h}_j\|_2.\end{aligned}$$

Moreover, since $\widetilde{\mathcal{C}}_1,\ldots,\widetilde{\mathcal{C}}_k$ are disjoint (except on a set of measure zero), by Lemma 7, we obtain that

$$\sup_{\boldsymbol{h}\in B_{1,2}} |Q'_{\boldsymbol{h}} - \mathbb{E}Q'_{\boldsymbol{h}}| \lesssim \left(\frac{d\log^3 d\log^5 k+\log(\delta^{-1})\log k}{n}\right)^{1/2} \tag{B.2.41}$$

holds with probability at least $1-\delta/4$. This provides the assertion in (B.2.40).

Appendix C: Proofs for Chapter 3

C.1 Tools

This section collects a set of standard results on concentration inequalities, which will be used in the proofs of Theorem 8. The following lemma provides the concentration of extreme singular values of sub-Gaussian matrices.

Lemma 17 ([93, Theorem 4.6.1]) *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be independent isotropic η-sub-Gaussian random vectors in $\mathbb{R}^d$. Then there exists an absolute constant $C > 0$ such that*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \boldsymbol{I}_p\right\| > \eta^2\max(\epsilon,\epsilon^2)\right) \leq \delta \quad \textit{where} \quad \epsilon = \sqrt{\frac{C(d+\log(2/\delta))}{n}}.$$

Remark 18 *It has been shown that Lemma 17 continues to hold when $\boldsymbol{x}_i$ is substituted by $\boldsymbol{\xi} = [\boldsymbol{x}_i;\ 1]$ [37]. Indeed, multiplying a random sign to the last coordinate of $\boldsymbol{\xi}_i$ does not modify the outer product $\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top$ whereas $\boldsymbol{\xi}_i$ remains a sub-Gaussian vector.*

Furthermore, we also use the results from the standard Vapnik–Chervonenkis (VC) theory stated in the following lemmas.

Lemma 19 ([89, Theorem 2]) *Let $\mathcal{V}$ be a collection of subsets of a set $\mathcal{X}$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ be n independent copies of a random variable $\boldsymbol{x} \in \mathcal{X}$. Then it holds for all*

$\epsilon > 0$ and $n \geq 2/\epsilon^2$ that

$$\mathbb{P}\left(\sup_{V\in\mathcal{V}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in V\}} - \mathbb{P}(\boldsymbol{x}\in V)\right| \geq \epsilon\right) \leq 4\Pi_{\mathcal{V}}(2n)\exp(-n\epsilon^2/16),$$

where $\Pi_{\mathcal{V}}(n)$ denotes the growth function defined by

$$\Pi_{\mathcal{V}}(n) := \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\in\mathcal{X}} \left|\left\{\left(\mathbb{1}_{\{\boldsymbol{x}_1\in V\}},\ldots,\mathbb{1}_{\{\boldsymbol{x}_n\in V\}}\right) : V\in\mathcal{V}\right\}\right|.$$

Lemma 20 ([68, Corollary 3.18]) *Let $\mathcal{V}$ be a collection of subsets having VC dimension d. Then, for all $n \geq d$, the growth function of $\mathcal{V}$ is upper-bounded by*

$$\Pi_{\mathcal{V}}(n) \leq \left(\frac{en}{d}\right)^d.$$

The VC dimension of the k-fold intersection has been known in the literature (e.g. see [11]). We will use the following lemma for the result for the intersection of size two. Since it was given as an exercise in [68], we provide a proof for the sake of completeness.

Lemma 21 ([68, Equation (3.53)]) *Let $\mathcal{V}$ and $\mathcal{W}$ be collections of subsets of a common set. Then their intersection given by $\mathcal{V}\cap\mathcal{W} := \{V\cap W : V\in\mathcal{V}, W\in\mathcal{W}\}$ satisfies that*

$$\Pi_{\mathcal{V}\cap\mathcal{W}}(n) \leq \Pi_{\mathcal{V}}(n)\Pi_{\mathcal{W}}(n), \quad \forall n\in\mathbb{N}.$$

Proof 15 *For any $V\cap W\in\mathcal{V}\cap\mathcal{W}$, we have*

$$\left(\mathbb{1}_{\{\boldsymbol{x}_1\in V\cap W\}},\ldots,\mathbb{1}_{\{\boldsymbol{x}_n\in V\cap W\}}\right) = \left(\mathbb{1}_{\{\boldsymbol{x}_1\in V\}},\ldots,\mathbb{1}_{\{\boldsymbol{x}_n\in V\}}\right)\odot\left(\mathbb{1}_{\{\boldsymbol{x}_1\in W\}},\ldots,\mathbb{1}_{\{\boldsymbol{x}_n\in W\}}\right),$$

where $\odot$ denotes the pointwise product. Therefore, the claim follows from the definition of the growth function.

Lemma 22 *Let $\mathcal{P}_k$ be the collection of all polytopes constructed by the intersection of k half spaces in $\mathbb{R}^d$. Then the growth function of $\mathcal{P}_k$ satisfies*

$$\Pi_{\mathcal{P}_k}(n) \leq \left(\frac{en}{d+1}\right)^{k(d+1)}. \tag{C.1.1}$$

Proof 16 *Let $\mathcal{H}_j$ be the collection of all half spaces in $\mathbb{R}^d$ for $j \in [k]$. Then, by the construction of $\mathcal{P}_k$, we have $\mathcal{P}_k = \cap_{j=1}^k \mathcal{H}_j$. Therefore, by inductive application of Theorem 21, the growth function of $\mathcal{P}_k$ satisfies*

$$\Pi_{\mathcal{P}_k}(n) \leq \prod_{j=1}^{k} \Pi_{\mathcal{H}_j}(n). \tag{C.1.2}$$

Furthermore, since the VC dimensions of half spaces in $\mathbb{R}^d$ is $d+1$ (e.g. see [68, Section 3]), Theorem 20 implies

$$\Pi_{\mathcal{H}_j}(n) \leq \left(\frac{en}{d+1}\right)^{d+1}, \quad \forall j \in [k]. \tag{C.1.3}$$

The assertion is obtained by plugging in (C.1.3) *into* (C.1.2).

Finally, the following corollary is a direct consequence of Lemmas 19, 20, and 21.

Corollary 23 *Let $\delta \in (0,1)$ and $\mathcal{P}_k$ be the collection of all polytopes constructed by the intersection of k half-spaces in $\mathbb{R}^d$. Suppose that $\{\boldsymbol{x}_i\}_{i=1}^n$ are independent copies of a random vector $\boldsymbol{x} \in \mathbb{R}^d$. Then it holds with probability at least $1-\delta$ that*

$$\sup_{Z \in \mathcal{P}_k} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in Z\}} - \mathbb{P}(\boldsymbol{x} \in Z) \right| \leq 4\sqrt{\frac{\log(4/\delta) + 2k(d+1)\log(2en/(d+1))}{n}}. \tag{C.1.4}$$

C.2 Supporting lemmas

In this section, we list lemmas to prove Theorem 8. These lemmas are borrowed from [85] and [37]. We improve on a subset of these results derived with a streamlined proof.

C.2.1 Worst-case extreme eigenvalues of partial sum of outer products of covariates

A partial sum of the outer products of covariates, $\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top$ appears frequently in the proof. The summation indices in $\mathcal{I}$ often depend on covariates. The following lemma by Tan and Vershynin [85] provides a tail bound on the worst-case largest eigenvalue of $\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top$ when the cardinality of $\mathcal{I}$ is bounded from above.

Lemma 24 ([85, Theorem 5.7]) *Let $\delta \in (0, 1/e)$, $\alpha \in (0, 1)$, and $\boldsymbol{\xi}_i = [\boldsymbol{x}_i, 1] \in \mathbb{R}^{d+1}$ for $i \in [n]$. Suppose that Assumption 1 holds. Then it holds with probability at least $1 - \delta$ that*

$$\sup_{\mathcal{I}:|\mathcal{I}|\le\alpha n} \lambda_1\left(\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\mathsf{T}\right) \le C_4(\eta^2 \vee 1)\sqrt{\alpha}n$$

for some absolute constant $C_4 > 0$, provided

$$n \ge \left(d \vee \frac{\log(1/\delta)}{\alpha}\right). \tag{C.2.1}$$

Remark 25 *In the original result, Tan and Vershynin assumed that $\{\boldsymbol{\xi}_i\}_{i=1}^n$ are isotropic η-sub-Gaussian random vectors [85, Theorem 5.7]. Later, Ghosh et al. [37] showed that the result also applies to the setting in Lemma 24 through the following argument. The outer product $\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top$ remains the same as one multiplies a random sign to the last entry of $\boldsymbol{\xi}_i$ which makes the random vector $\widetilde{\eta}$-sub-Gaussian with $\widetilde{\eta} = \max(\eta, 1)$.*

Moreover, Ghosh et al. also derived analogous lower tail bound on the smallest eigenvalue when the index set $\mathcal{I}$ exceeds a threshold [37, Lemma 7]. Their proof strategy adopted an epsilon-net approximation and a union bound argument. Our lemma below, derived by using the small-ball method [60], provides a streamlined proof and a sharper bound.

Lemma 26 *Let $\alpha, \delta \in (0,1)$ and $\boldsymbol{\xi}_i = [\boldsymbol{x}_i, 1] \in \mathbb{R}^{d+1}$ for $i \in [n]$. Suppose that Assumption 2 holds. Then there exists an absolute constant $C > 0$ such that if*

$$n \geq C\alpha^{-2}(d\log(n/d) \vee \log(1/\delta)) \tag{C.2.2}$$

then it holds with probability at least $1-\delta$ that

$$\inf_{\mathcal{I}\subset[n]:|\mathcal{I}|\geq\alpha n} \lambda_{d+1}\left(\sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right) \geq \frac{2n}{\gamma}\left(\frac{\alpha}{4}\right)^{1+\zeta^{-1}}. \tag{C.2.3}$$

We compare Lemma 26 to the previous result by Ghosh et al. [37, Lemma 7] when the parameter γ is treated as a fixed constant. They demonstrated that the worst-case minimum eigenvalue in the left-hand side of (C.2.3) satisfies $\Omega(n\alpha^{1+2\zeta^{-1}})$ if $n \geq \alpha^{-1}\max(4p, \zeta^{-1}(d+1))$. On one hand, their requirement in the sample complexity is less stringent than that in (C.2.2). On the other hand, the lower bound in (C.2.3) is tighter than theirs by a factor of $\alpha^{\zeta^{-1}}$. When these two results are applied to derive Theorem 8 with α substituted by $\pi_{\min}$, the resulting sample complexity $\widetilde{O}(\pi_{\min}^{-4(1+\zeta^{-1})}d)$ by Lemma 26 is smaller than $\widetilde{O}(\pi_{\min}^{-4(1+2\zeta^{-1})}d)$ by [37, Lemma 7]. The gain due to Lemma 26 is $\pi_{\min}^{-4\zeta^{-1}}$, which is no less than $k^{4\zeta^{-1}}$. For example, if the covariates are Gaussian $\zeta = 1/2$, then the gain is k^8.

Proof 17 *Let $T > 0$ be an arbitrarily fixed threshold. If*

$$N(\boldsymbol{v}) := \sum_{i=1}^{n} \mathbb{1}_{\{\langle\boldsymbol{\xi}_i,\boldsymbol{v}\rangle^2 > T\}} > n - \frac{\alpha n}{2} \tag{C.2.4}$$

then it follows that

$$\frac{1}{n}\sum_{i\in\mathcal{I}}\langle\boldsymbol{\xi}_i,\boldsymbol{v}\rangle^2 \geq \frac{\alpha T}{2}, \quad \forall\mathcal{I}\subset[n]:|\mathcal{I}|\geq\alpha n.$$

Therefore, it suffices to show that (C.2.4) *holds for all $\boldsymbol{v} \in \mathbb{S}^d$ with probability $1-\delta$. Let $\mathcal{H}$ denote the collection of half-spaces in $\mathbb{R}^d$ given by $\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}^\intercal\boldsymbol{u} > \sqrt{T} - w\}$*

for all $\boldsymbol{v} = [\boldsymbol{u};\ w] \in \mathbb{S}^d$. *Since the VC dimension of all half-spaces in* $\mathbb{R}^d$ *is at most* $d+1$, *by Lemmas 19 and 20, it holds with probability at least* $1-\delta/2$ *that*

$$\frac{1}{n}N(\boldsymbol{v}) \geq \frac{1}{n}\mathbb{E}N(\boldsymbol{v}) - C'\sqrt{\frac{d\log(n/d)+\log(1/\delta)}{n}}, \quad \forall \boldsymbol{v} \in \mathbb{S}^d, \tag{C.2.5}$$

where $C' > 0$ *is an absolute constant.*

Moreover, it follows from Assumption 2 that

$$\frac{1}{n}\mathbb{E}N(\boldsymbol{v}) = \mathbb{P}\left(|\langle \boldsymbol{x}, \boldsymbol{u}\rangle + w|^2 > T\right) \geq 1 - (T\gamma)^{\zeta}. \tag{C.2.6}$$

By plugging in (C.2.6) *into* (C.2.5), *we obtain that*

$$\frac{1}{n}N(\boldsymbol{v}) \geq 1 - (T\gamma)^{\zeta} - C'\sqrt{\frac{d\log(n/d)+\log(1/\delta)}{n}}, \quad \forall \boldsymbol{w} \in \mathbb{S}^d.$$

Then (C.2.4) *is satisfied for all* $\boldsymbol{v} \in \mathbb{S}^d$ *when* $T = \frac{1}{\gamma}\left(\frac{\alpha}{4}\right)^{\zeta^{-1}}$ *and* $C = (4C')^2$. *This completes the proof.*

C.2.2 Local estimates

In this section, we present local tail bounds which arise in the proof of the main result. The following lemma, obtained as a direct consequence of the triangle inequality and the definition of κ in (3.1.10), provides a basic inequality that will be used frequently throughout this section.

Lemma 27 *Suppose that* $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, *where* $\mathcal{N}(\boldsymbol{\beta}^\star)$ *is defined as in* (3.1.11). *Then we have*

$$\|(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}) - (\boldsymbol{\beta}^\star_j - \boldsymbol{\beta}^\star_{j'})\|_2 \leq 2\rho\|(\boldsymbol{\beta}^\star_j - \boldsymbol{\beta}^\star_{j'})_{1:d}\|_2, \quad \forall j \neq j' \in [k].$$

Proof 18 *Since* $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, *by the triangle inequality, we have*

$$\|(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}) - (\boldsymbol{\beta}^\star_j - \boldsymbol{\beta}^\star_{j'})\|_2 \leq \|\boldsymbol{\beta}_j - \boldsymbol{\beta}^\star_j\|_2 + \|\boldsymbol{\beta}_{j'} - \boldsymbol{\beta}^\star_{j'}\|_2 \leq 2\kappa\rho, \quad \forall j, j' \in [k].$$

Furthermore, it follows from the definition of κ *in* (3.1.10) *that*

$$\kappa \le \|(\boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star)_{1:d}\|_2, \quad \forall j \ne j' \in [k].$$

Then the assertion follows.

We also use the following lemma by Ghosh et al. [37], which is a consequence of Assumptions 1 and 2 respectively for the sub-Gaussianity and anti-concentration.

Lemma 28 ([37, Lemma 17]) *Suppose that* $\boldsymbol{x} \in \mathbb{R}^d$ *satisfies Assumptions 1 and 2. If*

$$\|\boldsymbol{v} - \boldsymbol{v}^\star\|_2 \le \frac{1}{2}\|(\boldsymbol{v}^\star)_{1:d}\|_2,$$

then

$$\mathbb{P}\left(\langle[\boldsymbol{x};1], \boldsymbol{v}^\star\rangle^2 \le \langle[\boldsymbol{x};1], \boldsymbol{v} - \boldsymbol{v}^\star\rangle^2\right) \lesssim \left(\left(\frac{\|\boldsymbol{v} - \boldsymbol{v}^\star\|_2}{\|(\boldsymbol{v}^\star)_{1:d}\|_2}\right)^2 \cdot \log\left(\frac{2\|(\boldsymbol{v}^\star)_{1:d}\|_2}{\|\boldsymbol{v} - \boldsymbol{v}^\star\|_2}\right)\right)^\zeta.$$

Intuitively, when the parameter vector $\boldsymbol{\beta}$ belongs to a small neighborhood of the ground-truth, the partition sets $(\mathcal{C}_j)_{j=1}^k$ by $\boldsymbol{\beta}$ and $\left(\mathcal{C}_j^\star\right)_{j=1}^k$ by the ground-truth $\boldsymbol{\beta}^\star$ will be similar. The next lemmas quantify the empirical measure on the event of $\boldsymbol{x} \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star$ for distinct indices j and j', and quadratic forms given as a partial summation indexed by the indicator functions on this event.

Lemma 29 *Let* $(\mathcal{C}_j)_{j=1}^k$ *and* $\left(\mathcal{C}_j^\star\right)_{j=1}^k$ *be defined as in* (1.2.4) *and* (3.1.9) *respectively by* $\boldsymbol{\beta}$ *and* $\boldsymbol{\beta}^\star$*. Furthermore, let* $\pi_{\min}$ *be defined as in* (3.1.8) *by* $\boldsymbol{\beta}^\star$*. Suppose that* $\boldsymbol{x} \in \mathbb{R}^d$ *and* $\{\boldsymbol{x}_i\}_{i=1}^n$ *satisfy Assumptions 1 and 2, and that the parameter* ρ *of* $\mathcal{N}(\boldsymbol{\beta}^\star)$ *in* (3.1.11) *satisfies* (3.1.12) *for some numerical constant* $R > 0$*. Then there exists an absolute constant* C *such that if*

$$n \ge C\pi_{\min}^{-2} \cdot (kd\log(n/d) \vee \log(1/\delta)) \tag{C.2.7}$$

then with probability at least $1-\delta$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_j^\star\}}\geq\frac{\pi_{\min}}{4} \tag{C.2.8}$$

holds for all $j\in[k]$, $\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star)$, *and* $\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}$.

Proof 19 *Note that the left-hand side of* (C.2.8) *is an empirical measure on the event* $\boldsymbol{x}\in\mathcal{C}_j\cap\mathcal{C}_j^\star$. *We first derive a lower bound on its expectation, which is written as*

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{x}\in\mathcal{C}_j,\boldsymbol{x}\in\mathcal{C}_j^\star\right)&=\mathbb{P}\left(\boldsymbol{x}\in\mathcal{C}_j|\boldsymbol{x}\in\mathcal{C}_j^\star\right)\cdot\mathbb{P}\left(\boldsymbol{x}\in\mathcal{C}_j^\star\right)\\&=\left(1-\mathbb{P}\left(\boldsymbol{x}\notin\mathcal{C}_j|\boldsymbol{x}\in\mathcal{C}_j^\star\right)\right)\cdot\mathbb{P}\left(\boldsymbol{x}\in\mathcal{C}_j^\star\right).\end{aligned} \tag{C.2.9}$$

Then, by the construction of $(\mathcal{C}_j)_{j=1}^k$ *in* (1.2.4), *we have*

$$\begin{aligned}&\mathbb{P}\left(\boldsymbol{x}\notin\mathcal{C}_j|\boldsymbol{x}\in\mathcal{C}_j^\star\right)\\&=\frac{\mathbb{P}(\boldsymbol{x}\notin\mathcal{C}_j,\boldsymbol{x}\in\mathcal{C}_j^\star)}{\mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j^\star)}\\&\leq\frac{1}{\mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j^\star)}\sum_{j'\neq j}\mathbb{P}\left(\langle[\boldsymbol{x};\ 1],\boldsymbol{\beta}_{j'}\rangle\geq\langle[\boldsymbol{x};\ 1],\boldsymbol{\beta}_j\rangle,\langle[\boldsymbol{x};\ 1],\boldsymbol{\beta}_j^\star\rangle\geq\langle[\boldsymbol{x};\ 1],\boldsymbol{\beta}_{j'}^\star\rangle\right)\\&\leq\frac{1}{\mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j^\star)}\sum_{j'\neq j}\mathbb{P}\left(\langle[\boldsymbol{x};\ 1],\boldsymbol{v}_{j,j'}\rangle\langle[\boldsymbol{x};\ 1],\boldsymbol{v}_{j,j'}^\star\rangle\leq 0\right)\\&\leq\frac{1}{\mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j^\star)}\sum_{j'\neq j}\mathbb{P}\left(\langle[\boldsymbol{x};1],\boldsymbol{v}_{j,j'}^\star\rangle^2\leq\langle[\boldsymbol{x};1],\boldsymbol{v}_{j,j'}-\boldsymbol{v}_{j,j'}^\star\rangle^2\right),\end{aligned}$$

where the second inequality holds since $\boldsymbol{v}_{j,j'}=\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j'}$ *and* $\boldsymbol{v}_{j,j'}^\star=\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star$, *and the last inequality follows from the fact that* $ab\leq 0$ *implies* $|b|\leq|a-b|$ *for* $a,b\in\mathbb{R}$. *Recall that* $\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star)$ *implies* $\|\boldsymbol{v}_{j,j'}-\boldsymbol{v}_{j,j'}^\star\|_2\leq 2\rho\|(\boldsymbol{v}_{j,j'}^\star)_{1:d}\|_2$ *due to Lemma 27. Furthermore, one can choose the numerical constant* $R>0$ *in* (3.1.12) *sufficiently*

small (but independent of k and p) so that $2\rho \le 0.1$. Then it follows that

$$\begin{aligned}
\mathbb{P}(\boldsymbol{x} \notin \mathcal{C}_{j'} \mid \boldsymbol{x} \in \mathcal{C}_{j'}^\star) &\overset{\text{(i)}}{\lesssim} \frac{k}{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \left(\frac{\|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2^2}{\|(\boldsymbol{v}_{j,j'}^\star)_{1:d}\|_2^2} \log\left(\frac{2\|(\boldsymbol{v}_{j,j'}^\star)_{1:d}\|_2}{\|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2} \right) \right)^\zeta \\
&\overset{\text{(ii)}}{\le} \frac{k}{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \left((2\rho)^2 \log\left(\frac{1}{\rho}\right) \right)^\zeta \\
&\overset{\text{(iii)}}{\le} \frac{k}{\pi_{\min}} \left(\frac{R^2 \pi_{\min}^{2\zeta^{-1}(1+\zeta^{-1})}}{k^{2\zeta^{-1}}} \right)^\zeta \\
&\le \frac{R^{2\zeta} \pi_{\min}^{1+2\zeta^{-1}}}{k}, \qquad \text{(C.2.10)}
\end{aligned}$$

where (i) follows from Lemma 28; (ii) holds since $a \log^{1/2}(2/a)$ is monotone increasing for $a \in (0,1]$; (iii) follows from the fact that $a \le \frac{b}{2}\log^{-1/2}(1/b)$ implies $a\log^{1/2}(2/a) \le b$ for $b \in (0, 0.1]$. Since $\pi_{\min} \le \frac{1}{k}$, once again $R > 0$ can be made sufficiently small so that the right-hand side of (C.2.10) *is at most $\frac{1}{2}$. Then plugging in this upper bound by* (C.2.10) *into* (C.2.9) *yields*

$$\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_{j'} \cap \mathcal{C}_{j'}^\star) \ge \frac{1}{2} \cdot \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_{j'}^\star). \qquad \text{(C.2.11)}$$

It remains to show the concentration of the left-hand side of (C.2.8) *around the expectation. Recall that $\mathcal{C}_j$ and $\mathcal{C}_j^\star$ are constructed as the intersection of at most k half-spaces. Then $\mathcal{C}_j \cap \mathcal{C}_j^\star$ belongs to the set $\mathcal{P}_{2k}$ defined in Lemma 22 and, hence, we have*

$$\sup_{\substack{j\in[k], \boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star) \\ \boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} - \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j \cap \mathcal{C}_j^\star) \right| \le \sup_{\mathcal{Z}\in\mathcal{P}_{2k}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{Z}\}} - \mathbb{P}(\boldsymbol{x} \in \mathcal{Z}) \right|.$$

Therefore, it follows from Corollary 23 that with probability at least $1-\delta$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} \ge \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j \cap \mathcal{C}_j^\star) - 4\sqrt{\frac{\log(4/\delta) + 2k(d+1)\log(2en/(d+1))}{n}} \qquad \text{(C.2.12)}$$

holds for all $j \in [k]$, $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, *and* $\boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}$. *The first summand in the right-hand side of* (C.2.12) *is bounded from below as in* (C.2.11). *Then choosing* C *in* (C.2.7) *large enough makes the second summand less than half of the lower bound in* (C.2.11). *This completes the proof.*

Next, the following lemma provides a slightly improved upper bound compared to the analogous previous result [37, Lemma 6]. Moreover, Lemma 30 is derived by using the VC theory and provides a streamlined and shorter proof compared to previous work [37].

Lemma 30 *Suppose that Assumptions 1 and 2 hold, and that* ρ *satisfies* (3.1.12) *for some numerical constant* $R > 0$. *Let* $\delta \in (0, 1/e)$. *There exists an absolute constant* C *such that if*

$$n \geq Ck^4 \pi_{\min}^{-4(1+\zeta^{-1})} (\log(k/\delta) \vee d \log(n/d)) \tag{C.2.13}$$

then with probability at least $1 - \delta$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \langle [\boldsymbol{x}_i;\ 1], \boldsymbol{v}_{j,j'}^\star \rangle^2 \leq \frac{2}{5\gamma k} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2^2 \tag{C.2.14}$$

holds for all $j \in [k]$, $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, *and* $\boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}$ *where* $\boldsymbol{v}_{j,j'} = \boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}$ *and* $\boldsymbol{v}_{j,j'}^\star = \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star$.

The previous result [37, Lemma 6] showed that with probability at least $1-\delta$ the left-hand side of (C.2.14) is bounded from above by $\widetilde{O}((\pi_{\min}^{1+\zeta^{-1}}/k) \log^{\zeta/2+1}(k/(\pi_{\min}{}^{1+\zeta^{-1}})))$ if $n \geq O(\max(p, \log(1/\delta)))$. In contrast, Lemma 30 provides a smaller upper bound by a logarithmic factor at the cost of increased sample complexity. However, the condition in (C.2.13) is implied by another sufficient condition from another step of the analysis; hence, it does not affect the main result in Theorem 8.

Proof 20 *By the definition of* $(\mathcal{C}_j)_{j=1}^k$ *in* (1.2.4)*, it holds for any* $j \neq j'$ *that*

$$\begin{aligned} \boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star &\iff \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle \geq \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_{j'} \rangle,\ \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_{j'}^\star \rangle \geq \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star \rangle \\ &\iff \langle \boldsymbol{\xi}_i, \boldsymbol{v}_{j,j'} \rangle \geq 0,\ \langle \boldsymbol{\xi}_i, \boldsymbol{v}_{j,j'}^\star \rangle \leq 0 \\ &\implies \langle \boldsymbol{\xi}_i, \boldsymbol{v}_{j,j'} \rangle \langle \boldsymbol{\xi}_i, \boldsymbol{v}_{j,j'}^\star \rangle \leq 0. \end{aligned} \tag{C.2.15}$$

Furthermore, by Lemma 27, every $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$ *satisfies* $\|\boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star\|_2 \leq 2\rho \|(\boldsymbol{v}_{j,j'}^\star)_{1:d}\|_2$*. Therefore, it suffices to show that with probability at least* $1 - \delta$

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v} \rangle \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle \leq 0\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle^2 \leq \frac{2}{5\gamma k} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \|\boldsymbol{v} - \boldsymbol{v}^\star\|_2^2 \tag{C.2.16}$$

holds for all $(\boldsymbol{v}, \boldsymbol{v}^\star) \in \mathcal{M}$*, where*

$$\mathcal{M} := \{(\boldsymbol{v}, \boldsymbol{v}^\star) \in \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} : \|\boldsymbol{v} - \boldsymbol{v}^\star\| \leq 2\rho \|(\boldsymbol{v})_{1:d}\|_2\}.$$

Since $ab \leq 0$ *implies* $|b| \leq |a - b|$ *for* $a, b \in \mathbb{R}$*, each summand in the left-hand side of* (C.2.16) *is upper-bounded by*

$$\begin{aligned} \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v} \rangle \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle \leq 0\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle^2 &\leq \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle^2 \leq \langle \boldsymbol{\xi}_i, \boldsymbol{v} - \boldsymbol{v}^\star \rangle^2\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle^2 \\ &\leq \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle^2 \leq \langle \boldsymbol{\xi}_i, \boldsymbol{v} - \boldsymbol{v}^\star \rangle^2\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v} - \boldsymbol{v}^\star \rangle^2. \end{aligned}$$

Before we proceed to the next step, for brevity, we introduce a shorthand notation given by

$$\mathcal{S}_{\boldsymbol{v}, \boldsymbol{v}^\star} := \{\boldsymbol{\xi} \in \mathbb{R}^{d+1} : \langle \boldsymbol{\xi}, \boldsymbol{v} - \boldsymbol{v}^\star \rangle^2 \geq \langle \boldsymbol{\xi}, \boldsymbol{v}^\star \rangle^2\}. \tag{C.2.17}$$

Then the left-hand side of (C.2.16) *is bounded from above as*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\langle \boldsymbol{\xi}_i, \boldsymbol{v} \rangle \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle \leq 0\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star \rangle^2 \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{\xi}_i \in \mathcal{S}_{\boldsymbol{v}, \boldsymbol{v}^\star}\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v} - \boldsymbol{v}^\star \rangle^2.$$

Next, we derive a tail bound on the empirical measure $\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{\xi}_i \in \mathcal{S}_{\boldsymbol{v}, \boldsymbol{v}^\star}\}}$ *on the event for* $\boldsymbol{\xi} \in \mathcal{S}_{\boldsymbol{v}, \boldsymbol{v}^\star}$*. Let* $\mathcal{P}_2$ *denote the collection of all polytopes given by the intersections*

of two half-spaces. Then $\mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}$ belongs to $\mathcal{P}_2 \cup \mathcal{P}_2$. It follows from Lemma 22 and [24, Theorem A] that

$$\Pi_{\mathcal{P}_2 \cup \mathcal{P}_2}(n) \leq \left(\frac{en}{C'(d+1)}\right)^{C'(d+1)} \tag{C.2.18}$$

for some absolute constant C'. Therefore, by Lemma 19 and (C.2.18)*, we obtain that*

$$\sup_{(\boldsymbol{v},\boldsymbol{v}^\star)\in\mathcal{M}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{\xi}_i \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}\}} - \mathbb{P}(\boldsymbol{\xi} \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}) \right| \lesssim \sqrt{\frac{\log(1/\delta) + d\log(n/d)}{n}} \tag{C.2.19}$$

holds with probability at least $1 - \frac{\delta}{2}$.

Similar to (C.2.10)*, we obtain an upper bound on the probability by using Lemma 28 as follows:*

$$\begin{aligned} \sup_{(\boldsymbol{v},\boldsymbol{v}^\star)\in\mathcal{M}} \mathbb{P}(\boldsymbol{\xi} \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}) &\leq C_1 \left((2\rho)^2 \log\left(\frac{1}{\rho}\right)\right)^{\zeta} \\ &\leq C_1 \left(\frac{R^2 \pi_{\min}^{2\zeta^{-1}(1+\zeta^{-1})}}{k^{2\zeta^{-1}}}\right)^{\zeta} \\ &\leq \underbrace{\frac{C_1 R^{2\zeta} \pi_{\min}^{2+2\zeta^{-1}}}{k^2}}_{\alpha} \end{aligned} \tag{C.2.20}$$

where $C_1 > 0$ is an absolute constant. By choosing the numerical constant $C > 0$ in (C.2.13) *sufficiently large, we obtain from* (C.2.19) *and* (C.2.20) *that*

$$\mathbb{P}\left(\sup_{(\boldsymbol{v},\boldsymbol{v}^\star)\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{\xi}_i \in \mathcal{S}_{\boldsymbol{v},\boldsymbol{v}^\star}\}} > \frac{\alpha}{2}\right) \leq \frac{\delta}{2}. \tag{C.2.21}$$

Furthermore, one can choose the numerical constant $R > 0$ small enough so that $\alpha \in (0,1)$. Then, since (C.2.13) *and* (3.1.12) *imply* (C.2.1)*, by Lemma 24, it holds with probability at least $1 - \delta/2$ that*

$$\sup_{\mathcal{I}:|\mathcal{I}|\leq\frac{\alpha n}{2}} \left\| \sum_{i\in\mathcal{I}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \right\| \lesssim (\eta^2 \vee 1)\sqrt{\alpha} n. \tag{C.2.22}$$

Finally, by combining the results in (C.2.21) *and* (C.2.22)*, we obtain that with probability at least* $1-\delta$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\langle\boldsymbol{\xi}_i,\boldsymbol{v}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star\rangle\le 0\}}\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star\rangle^2 &\le \sup_{\mathcal{I}:|\mathcal{I}|\le\frac{\alpha n}{2}}\frac{1}{n}\sum_{i\in\mathcal{I}}\langle\boldsymbol{\xi}_i,\boldsymbol{v}-\boldsymbol{v}^\star\rangle^2\\ &\le \sup_{\mathcal{I}:|\mathcal{I}|\le\frac{\alpha n}{2}}\left\|\frac{1}{n}\sum_{i\in\mathcal{I}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|\cdot\|\boldsymbol{v}-\boldsymbol{v}^\star\|_2^2\\ &\le C_2(\eta^2\vee 1)R^\zeta\left(\frac{\pi_{\min}^{(1+\zeta^{-1})}}{k}\right)\cdot\|\boldsymbol{v}-\boldsymbol{v}^\star\|_2^2\end{aligned}$$

holds for all $(\boldsymbol{v},\boldsymbol{v}^\star)\in\mathcal{M}$*, where* C_2 *is an absolute constant. By choosing* $R>0$ *sufficiently small so that*

$$C_2(\eta^2\vee 1)R^\zeta\le\frac{2}{5\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}},$$

we obtain the assertion in (C.2.16)*.*

C.3 Proof of Theorem 8

The loss function $\ell(\boldsymbol{\beta})$ is decomposed as

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \frac{1}{2n}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle-\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle-z_i\right)^2\\ &= \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle-\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle\right)^2}_{\ell^{\mathrm{clean}}(\boldsymbol{\beta})}\\ &\quad-\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}z_i\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle-\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle\right)-\frac{1}{2n}\sum_{i=1}^{n}z_i^2\right)}_{\ell^{\mathrm{noise}}(\boldsymbol{\beta})}.\end{aligned}$$

Then the partial gradient of $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_l$ is written as

$$\begin{aligned}\nabla_{\boldsymbol{\beta}_l}\ell(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle-\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle-z_i\right)\boldsymbol{\xi}_i\\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle-\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle\right)\boldsymbol{\xi}_i}_{\nabla_{\boldsymbol{\beta}_l}\ell^{\mathrm{clean}}(\boldsymbol{\beta})}-\underbrace{\frac{1}{n}\sum_{i=1}^{n}z_i\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\boldsymbol{\xi}_i}_{\nabla_{\boldsymbol{\beta}_l}\ell^{\mathrm{noise}}(\boldsymbol{\beta})}}\end{aligned}\tag{C.3.1}$$

where $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are determined by $\boldsymbol{\beta}$ as in (1.2.4).

In the remainder of the proof, we will use the following shorthand notation to denote the pairwise difference of parameter vectors and the probability measure on the largest partition by the ground-truth model:

$$\boldsymbol{v}_{j,j'} := \boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}, \quad \boldsymbol{v}^\star_{j,j'} := \boldsymbol{\beta}^\star_j - \boldsymbol{\beta}^\star_{j'}, \quad \text{and} \quad \pi_{\max} := \max_{j \in [k]} \mathbb{P}\left(\boldsymbol{x} \in \mathcal{C}^\star_j\right).$$

Below we show that the following lemmas hold under the condition in (3.1.14). The proof is provided in Appendix C.3.1.

Lemma 31 *Under the hypothesis of Theorem 8, if* (3.1.14) *is satisfied, then with probability at least* $1 - \delta$ *the following inequalities hold for all* $j \in [k]$, $\boldsymbol{\beta}^\star \in \mathbb{R}^{k(d+1)}$, *and* $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$*:*

$$\langle \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}^t_j - \boldsymbol{\beta}^\star_j \rangle \geq \frac{2}{\gamma} \left(\frac{\pi_{\min}}{16}\right)^{1+\zeta^{-1}} \left(\|\boldsymbol{\beta}^t_j - \boldsymbol{\beta}^\star_j\|_2^2 - \frac{1}{10k} \sum_{j' : j' \neq j} \left\|\boldsymbol{v}^t_{j,j'} - \boldsymbol{v}^\star_{j,j'}\right\|_2^2 \right), \tag{C.3.2}$$

$$\|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 \lesssim \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right) \left\|\boldsymbol{\beta}^t_j - \boldsymbol{\beta}^\star_j\right\|_2^2 + \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{k^2} \sum_{j' : j' \neq j} \left\|\boldsymbol{v}^t_{j,j'} - \boldsymbol{v}^\star_{j,j'}\right\|_2^2, \tag{C.3.3}$$

and

$$\left\|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2 \lesssim \frac{\sigma \sqrt{kd \log(n/d) + \log(1/\delta)}}{\sqrt{n}}. \tag{C.3.4}$$

The remainder of the proof shows that the assertion of the theorem is obtained from (C.3.2), (C.3.3) and (C.3.4) via the following three steps.

Step 1: We prove by induction that all iterates remain within the neighborhood $\mathcal{N}(\boldsymbol{\beta}^\star)$. Suppose that $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$ holds for a fixed $t \in \mathbb{N}$. By the triangle inequality,

for any $j \in [k]$, the next iterate $\boldsymbol{\beta}^{t+1}$ satisfies

$$\begin{aligned}\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^\star\|_2 &= \|\boldsymbol{\beta}_j^t - \mu\nabla_{\boldsymbol{\beta}_j}\ell(\boldsymbol{\beta}^t) - \boldsymbol{\beta}_j^\star\|_2 \\ &\le \underbrace{\|\boldsymbol{\beta}_j^t - \mu\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t) - \boldsymbol{\beta}_j^\star\|_2}_{A_{\text{clean}}} + \underbrace{\mu\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta}^t)\|_2}_{A_{\text{noise}}}. \end{aligned} \tag{C.3.5}$$

Then it remains to show

$$\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^\star\|_2 \le A_{\text{clean}} + A_{\text{noise}} \le \kappa\rho, \quad \forall j \in [k]. \tag{C.3.6}$$

Note that the first summand in the right-hand side of (C.3.5) satisfies

$$A_{\text{clean}}^2 = \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - 2\mu\langle\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\rangle + \mu^2\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2.$$

Therefore, it follows from (C.3.2) and (C.3.3) that

$$\begin{aligned} A_{\text{clean}}^2 &\le \left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2^2 - \frac{4\mu}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\pi_{\min}^{1+\zeta^{-1}}\left(\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - \frac{1}{10k}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2\right) \\ &\quad + \mu^2 C_1\left(\left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right)\left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2^2 + \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{k^2}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2\right) \\ &= \left(1 - \frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\mu\pi_{\min}^{1+\zeta^{-1}} + C_1\mu^2\left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right)\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\ &\quad + \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\mu\pi_{\min}^{1+\zeta^{-1}}}{5k} + \frac{C_1\mu^2\pi_{\min}^{2(1+\zeta^{-1})}}{k^2}\right)\sum_{j'^*:j'\neq j}\left\|\boldsymbol{v}_{j,j}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2. \end{aligned} \tag{C.3.7}$$

We set the step size μ to be

$$\mu = \frac{\omega\pi_{\min}^{1+\zeta^{-1}}}{\tau} \tag{C.3.8}$$

where ω is a constant that will be specified later and τ is given by

$$\tau := \pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}. \tag{C.3.9}$$

Putting the choices of μ and τ respectively by (C.3.8) and (C.3.9) into (C.3.7) yields

$$\begin{aligned}
A_{\text{clean}}^2 &\leq \left(1 - \frac{\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} + \frac{C_1\omega^2\pi_{\min}^{2(1+\zeta^{-1})}\left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right)}{\tau^2}\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{5\tau k} + \frac{C_1\omega^2\pi_{\min}^{4(1+\zeta^{-1})}}{\tau^2 k^2}\right)\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2 \\
&\leq \left(1 - \frac{\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} + \frac{C_1\omega^2\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&\quad + \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\omega\pi_{\min}^{2(1+\zeta^{-1})}}{5\tau} + \frac{C_1 w^2\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\right)\max_{1\leq j\neq j'\leq k}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2.
\end{aligned} \tag{C.3.10}$$

Next, since $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$, by the definition of $\mathcal{N}(\boldsymbol{\beta}^\star)$ in (3.1.11), we have

$$\max_{j\in[k]}\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2 \leq \kappa\rho. \tag{C.3.11}$$

Furthermore, by Lemma 27, we also have

$$\max_{1\leq j\neq j'\leq k}\left\|\boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2 \leq 2\kappa\rho. \tag{C.3.12}$$

Then plugging in (C.3.11) and (C.3.12) into (C.3.10) yields

$$\begin{aligned}
(\kappa\rho)^{-2}A_{\text{clean}}^2 &\leq 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})}\omega}{\tau}\left(\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\left(2 - \frac{4}{5}\right) + C_1\omega\left(1+4\right)\right) \\
&\leq 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\cdot\omega\left(\frac{\frac{12}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}}{5} + 5\omega C_1\right) \\
&\leq 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\cdot\omega\underbrace{\left(\frac{\frac{12}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}}{5}\right)}_{c_0},
\end{aligned} \tag{C.3.13}$$

which is rewritten as

$$A_{\text{clean}}^2 \leq (\kappa\rho)^2\left(1 - \frac{c_0\omega\pi_{\min}^{2(1+\zeta^{-1})}}{\tau}\right). \tag{C.3.14}$$

For fixed γ and ζ, c_0 is a positive numerical constant. Due to the choice of τ by (C.3.9), we have

$$\frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} = \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}} < 1,$$

Furthermore, one can choose $\omega > 0$ sufficiently small so that $\omega c_0 < 1$. Then the upper bound in the right-hand side of (C.3.14) is valid as a positive number.

If A_{noise} is upper-bounded as

$$A_{\text{noise}} \leq \kappa\rho \frac{c_0 \omega \pi_{\min}^{2(1+\zeta^{-1})}}{2\tau}, \tag{C.3.15}$$

then, by the elementary inequality $1 - \sqrt{1-\alpha} \geq \alpha/2$ that holds for any $\alpha \in (0,1)$, we have

$$A_{\text{noise}} \leq \kappa\rho \left(1 - \sqrt{1 - \frac{c_0 \omega \pi_{\min}^{2(1+\zeta^{-1})}}{\tau}}\right). \tag{C.3.16}$$

Then (C.3.14) and (C.3.16) yield (C.3.6). Therefore, it suffices to show that (C.3.15) holds.

Due to the inequality in (C.3.4), we have

$$\left\|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2 \lesssim \frac{\sigma\sqrt{kd\log(n/d) + \log(1/\delta)}}{\sqrt{n}}, \quad \forall j \in [k].$$

By the choice of μ in (C.3.8), we obtain an upper bound on A_{noise} given by

$$A_{\text{noise}} = \mu \left\|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2 \lesssim \frac{\omega \pi_{\min}^{1+\zeta^{-1}}}{\tau} \cdot \frac{\sigma\sqrt{kd\log(n/d) + \log(1/\delta)}}{\sqrt{n}}. \tag{C.3.17}$$

The condition in (3.1.14) implies

$$n \geq C \cdot \frac{\sigma^2 \pi_{\min}^{-2(1+\zeta^{-1})} \left(kd\log(n/d) + \log(1/\delta)\right)}{\kappa^2\rho^2}. \tag{C.3.18}$$

One can choose the absolute constant $C > 0$ in (3.1.14) and (C.3.18) as large enough so that (C.3.18) and (C.3.17) imply (C.3.15). This completes the induction argument

in Step 1.

Step 2: Next we show that all iterates also satisfy

$$\left\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^{\star}\right\|_2 \leq \sqrt{1-\nu}\left\|\boldsymbol{\beta}^{t}-\boldsymbol{\beta}^{\star}\right\|_2 + C'\mu\sigma\sqrt{\frac{k\left(kd\log(n/d)+\log(1/\delta)\right)}{n}}. \quad \text{(C.3.19)}$$

We use the fact that $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$, which has been shown in Step 1. By the update rule of gradient descent and the triangle inequality, the left-hand side of (C.3.19) satisfies

$$\begin{aligned}
\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^{\star}\|_2 &= \|\boldsymbol{\beta}^{t}-\mu\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^t)-\boldsymbol{\beta}^{\star}\|_2 \\
&\leq \|\boldsymbol{\beta}^{t}-\mu\nabla_{\boldsymbol{\beta}}\ell^{\text{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}^{\star}\|_2 + \mu\|\nabla_{\boldsymbol{\beta}}\ell^{\text{noise}}(\boldsymbol{\beta}^t)\|_2 \\
&= \underbrace{\sqrt{\sum_{j=1}^{k}\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star-\mu\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2}}_{B_{\text{clean}}} + \underbrace{\sqrt{\mu^2\sum_{j=1}^{k}\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta}^t)\|_2^2}}_{B_{\text{noise}}}.
\end{aligned} \quad \text{(C.3.20)}$$

Below we derive an upper bound on each of the summands on the right-hand side of (C.3.20). First we show that

$$B_{\text{clean}}^2 \leq (1-\nu)\sum_{j=1}^{k}\left\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\|_2^2. \quad \text{(C.3.21)}$$

Since $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$, the inequality in (C.3.21) holds if there exist constants $\mu, \lambda \in (0,1)$ such that

$$\sum_{j=1}^{k}\langle\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\rangle \geq \frac{\mu}{2}\sum_{j=1}^{k}\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{k}\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2, \quad \forall\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star). \quad \text{(C.3.22)}$$

Indeed, the condition in (C.3.22) and $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$ imply

$$\begin{aligned}
B_{\text{clean}}^2 &= \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \mu \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t) - \boldsymbol{\beta}_j^\star\|_2^2 \\
&= \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 + \sum_{j=1}^k \mu^2 \|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 - 2\mu \sum_{j=1}^k \langle \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star, \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\rangle \\
&\le \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - \mu\lambda \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
&= (1 - \mu\lambda) \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2. \qquad \text{(C.3.23)}
\end{aligned}$$

Next we show that (C.3.22) holds. Due to (C.3.2) and the elementary inequality $\|\boldsymbol{a} + \boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$, it holds for all $j \in [k]$ that

$$\begin{aligned}
&\langle \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star \rangle \\
&\ge \frac{2}{\gamma} \left(\frac{1}{16}\right)^{1+\zeta^{-1}} \pi_{\min}^{1+\zeta^{-1}} \left(\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - \frac{1}{5k} \sum_{j':j' \ne j} \left(\left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2^2 + \left\|\boldsymbol{\beta}_{j'}^t - \boldsymbol{\beta}_{j'}^\star\right\|_2^2 \right) \right). \qquad \text{(C.3.24)}
\end{aligned}$$

By taking the summation of (C.3.24) over $j \in [k]$, we obtain

$$\sum_{j=1}^k \langle \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star \rangle \ge \frac{\frac{6}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \pi_{\min}^{1+\zeta^{-1}}}{5} \sum_{j=1}^k \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2. \qquad \text{(C.3.25)}$$

Furthermore, by using (C.3.3) and the elementary inequality $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$ again, we obtain

$$\begin{aligned}
\|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 \le C_1 \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})}\right) \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\
+ \frac{2C_1 \pi_{\min}^{2(1+\zeta^{-1})}}{k^2} \sum_{j':j' \ne j} \left(\left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2^2 + \left\|\boldsymbol{\beta}_{j'}^t - \boldsymbol{\beta}_{j'}^\star\right\|_2^2 \right). \qquad \text{(C.3.26)}
\end{aligned}$$

Summing the equation in (C.3.26) over $j \in [k]$ yields

$$\sum_{j=1}^{k} \|\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}^t)\|_2^2 \leq C_1 \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} + \frac{4(k-1)\pi_{\min}^{2(1+\zeta^{-1})}}{k^2} \right) \sum_{j=1}^{k} \left\| \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star \right\|_2^2$$
$$\leq C_1 \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} + 4\pi_{\min}^{2(1+\zeta^{-1})} \right) \sum_{j=1}^{k} \left\| \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star \right\|_2^2 . \tag{C.3.27}$$

By combining (C.3.25) and (C.3.27) with μ as in (C.3.8), we obtain a sufficient condition for (C.3.22) given by

$$\frac{\frac{6}{\gamma} \left(\frac{1}{16}\right)^{1+\zeta^{-1}} \pi_{\min}^{1+\zeta^{-1}}}{5} \geq \frac{\omega \pi_{\min}^{1+\zeta^{-1}} C_1 \left(\pi_{\max} + 5\pi_{\min}^{2(1+\zeta^{-1})} \right)}{2 \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} \right)} + \frac{\lambda}{2}. \tag{C.3.28}$$

By choosing $\omega > 0$ small enough, (C.3.28) is satisfied when λ is chosen as

$$\lambda = \min(c_2 \pi_{\min}^{1+\zeta^{-1}}, 1) \tag{C.3.29}$$

for an absolute constant $c_2 > 0$. Hence, we have shown that the condition in (C.3.22) holds with μ and λ specified by (C.3.8) and (C.3.29).

Next we consider the second summand on the right-hand side of (C.3.20). The inequality in (C.3.4) implies

$$B_{\text{noise}}^2 = \mu^2 \sum_{j=1}^{k} \left\| \nabla_{\boldsymbol{\beta}_j} \ell^{\text{noise}}(\boldsymbol{\beta}^t) \right\|_2^2 \lesssim \frac{\mu^2 \sigma^2 k (kd \log(n/d) + \log(1/\delta))}{n}. \tag{C.3.30}$$

Finally, plugging in (C.3.23) and (C.3.30) into (C.3.20) provides the assertion (C.3.19). This completes the proof of Step 2.

Step 3: We finish the proof of Theorem 8 by applying the results in Step 1 and Step 2. Plugging in the expression of $\nu = \mu\lambda$ with μ and λ as in (C.3.8) and (C.3.29) provides

$$\begin{aligned}
\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|_2 &\leq (1-\mu\lambda)^{t/2} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2 + C_2 \cdot \frac{\mu\sigma}{1-\sqrt{1-\mu\lambda}} \cdot \sqrt{\frac{k\left(kd\log(n/d) + \log(1/\delta)\right)}{n}} \\
&\overset{\text{(a)}}{\leq} (1-\mu\lambda)^{t/2} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2 + C_2 \cdot \frac{2\sigma}{\lambda} \cdot \sqrt{\frac{k\left(kd\log(n/d) + \log(1/\delta)\right)}{n}} \\
&\overset{\text{(b)}}{\leq} (1-\mu\lambda)^{t/2} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2 + C_3 \cdot \frac{\sigma}{\pi_{\max}} \cdot \sqrt{\frac{k\left(kd\log(n/d) + \log(1/\delta)\right)}{n}} \\
&\overset{\text{(c)}}{\leq} (1-\mu\lambda)^{t/2} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2 + C_3 \cdot \sigma k \sqrt{\frac{k\left(kd\log(n/d) + \log(1/\delta)\right)}{n}},
\end{aligned}$$

where (a) follows from the elementary inequality $\sqrt{1-t} < 1 - t/2$ for any $t \in (0,1)$; (b) holds by the choice of τ in (C.3.9); (c) holds since $\pi_{\max}^{-1} \leq k$.

C.3.1 Proof of Lemma 31

We show that each of (C.3.2), (C.3.3), and (C.3.4) holds with probability at least $1 - \delta/3$. We also note that for simplicity, we proceed on the proofs using $\boldsymbol{\beta}$ and $\boldsymbol{v}_{j,j'}$. Therefore, the assertions in (C.3.2), (C.3.3), and (C.3.4) can be completed by substituting $\boldsymbol{\beta}$ and $\boldsymbol{v}_{j,j'}$ with $\boldsymbol{\beta}^t$ and $\boldsymbol{v}_{j,j'}^t$ respectively.

Proof of (C.3.2)**:** We show that (C.3.2) holds with high probability under the following condition

$$n \geq C_1 \left(\log(k/\delta) \vee d\log(n/d)\right) k^4 \pi_{\min}^{-4(1+\zeta^{-1})}, \tag{C.3.31}$$

which is implied by the assumption in (3.1.14). We proceed with the proof under the following three events, each of which holds with probability at least $1 - \delta/9$. First, since (C.3.31) implies (C.2.13), by Lemma 30, it holds with probability at least $1 - \delta/9$

that

$$\frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_{j'}\}}\langle\boldsymbol{\xi}_i,\boldsymbol{v}^\star_{j,j'}\rangle^2$$
$$\leq\frac{2}{5\gamma k}\left(\frac{\pi_{\min}}{16}\right)^{1+\zeta^{-1}}\sum_{j':j'\neq j}\|\boldsymbol{v}_{j,j'}-\boldsymbol{v}^\star_{j,j'}\|_2^2,\quad\forall j\in[k],\ \forall\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star),\ \forall\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}. \tag{C.3.32}$$

Moreover, since (C.3.31) also implies (C.2.7), by Lemma 29, it holds with probability at least $1-\delta/3$ that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_j\}}\geq\frac{\pi_{\min}}{4},\quad\forall j\in[k],\ \forall\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star),\ \forall\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}. \tag{C.3.33}$$

Lastly, since (C.3.31) is a sufficient condition to invoke Lemma 26 with $\alpha=\pi_{\min}/4$, it holds with probability at least $1-\delta/9$ that

$$\inf_{\mathcal{I}\subset[n]:|\mathcal{I}|\geq\frac{\pi_{\min}n}{4}}\lambda_{d+1}\left(\frac{1}{n}\sum_{i\in\mathcal{I}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right)\geq\frac{2}{\gamma}\left(\frac{\pi_{\min}}{16}\right)^{1+\zeta^{-1}}. \tag{C.3.34}$$

Therefore, we have shown that (C.3.32), (C.3.33), and (C.3.34) hold with probability at least $1-\delta/3$. The remainder of the proof is conditioned on the event that $\{\boldsymbol{\xi}_i\}_{i=1}^n$ satisfy (C.3.32), (C.3.33), and (C.3.34).

Let $\boldsymbol{\beta}^\star\in\mathbb{R}^{d+1}$, $\boldsymbol{\beta}\in\mathcal{N}(\boldsymbol{\beta}^\star)$, and $j\in[k]$ be arbitrarily fixed. For brevity, we will use the shorthand notation $\boldsymbol{h}_j:=\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_j$. Then the left-hand side of (C.3.2) is rewritten as

$$\begin{aligned}\langle\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}),\boldsymbol{h}_j\rangle&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\left(\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle-\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}^\star_j\rangle\right)\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle\\&=\frac{1}{n}\sum_{j'=1}^{k}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_{j'}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle\\&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle^2+\frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}^\star_{j'}\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}^\star_{j'}\rangle\langle\boldsymbol{\xi}_i,\boldsymbol{h}_j\rangle.\end{aligned}$$

By the inequality of arithmetic and geometric means, we have

$$\begin{aligned}\langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}^\star_{j'} \rangle \langle \boldsymbol{\xi}_i, \boldsymbol{h}_j \rangle &= \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}^\star_j + \boldsymbol{\beta}^\star_j - \boldsymbol{\beta}^\star_{j'} \rangle \langle \boldsymbol{\xi}_i, \boldsymbol{h}_j \rangle \\ &= \langle \boldsymbol{\xi}_i, \boldsymbol{h}_j + \boldsymbol{v}^\star_{j,j'} \rangle \langle \boldsymbol{\xi}_i, \boldsymbol{h}_j \rangle \\ &\geq \frac{\langle \boldsymbol{\xi}_i, \boldsymbol{h}_j \rangle^2}{2} - \frac{\langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star_{j,j'} \rangle^2}{2} \geq -\frac{\langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star_{j,j'} \rangle^2}{2}.\end{aligned}$$

Therefore, we obtain

$$\langle \nabla_{\boldsymbol{\beta}_j} \ell^{\mathrm{clean}}(\boldsymbol{\beta}), \boldsymbol{h}_j \rangle \geq \underbrace{\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}^\star_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{h}_j \rangle^2}_{(*)} - \underbrace{\frac{1}{2n} \sum_{j':j' \neq j} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}^\star_{j'}\}} \langle \boldsymbol{\xi}_i, \boldsymbol{v}^\star_{j,j'} \rangle^2}_{(**)}. \tag{C.3.35}$$

By (C.3.33) and (C.3.34), the first summand in the right-hand side of (C.3.35) is bounded from below as

$$(*) \geq \frac{2}{\gamma} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \|\boldsymbol{h}_j\|_2^2. \tag{C.3.36}$$

Moreover, due to (C.3.32), $(**)$ is bounded from above as

$$(**) \leq \frac{1}{5\gamma k} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \sum_{j':j' \neq j} \|\boldsymbol{v}_{j,j'} - \boldsymbol{v}^\star_{j,j'}\|_2^2. \tag{C.3.37}$$

Then, plugging in (C.3.36) and (C.3.37) into (C.3.35) provides

$$\begin{aligned}&\langle \nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}), \boldsymbol{h}_j \rangle \\ &\geq \frac{2}{\gamma} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \|\boldsymbol{h}_j\|_2^2 - \frac{1}{5\gamma} \left(\frac{1}{16} \right)^{1+\zeta^{-1}} \left(\frac{\pi_{\min}^{1+\zeta^{-1}}}{k} \right) \sum_{j':j' \neq j} \|\boldsymbol{v}_{j,j'} - \boldsymbol{v}^\star_{j,j'}\|_2^2 \\ &= \frac{2}{\gamma} \left(\frac{\pi_{\min}}{16} \right)^{1+\zeta^{-1}} \left(\|\boldsymbol{h}_j\|_2^2 - \frac{1}{10k} \sum_{j':j' \neq j} \left\|\boldsymbol{v}_{j,j'} - \boldsymbol{v}^\star_{j,j'}\right\|_2^2 \right).\end{aligned}$$

This completes the proof.

Proof of (C.3.3)**:** The proof is based on the condition

$$n \geq C_2 \left(\log(k/\delta) \vee d \log(n/d) \right) k^4 \pi_{\min}^{-4(1+\zeta^{-1})}, \tag{C.3.38}$$

which is implied by (3.1.14). We will proceed under the following four events, each of which holds with probability at least $1 - \delta/12$. First, since (C.3.38) implies (C.2.13), by Lemma 30, (C.3.32) holds with probability at least $1 - \delta/12$. Next, since $\left(\mathcal{C}_j^\star\right)_{j=1}^k$ are included in the set of intersection of k half-spaces in $\mathbb{R}^d$, by Corollary 23 and (C.3.38), it holds with probability at least $1 - \delta/12$ that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^\star\}} \le 2\mathbb{P}\left(\boldsymbol{x} \in \mathcal{C}_j^\star\right), \quad \forall j \in [k]. \tag{C.3.39}$$

We also consider the event given by

$$\sum_{i=1}^n \mathbb{1}_{\left\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\right\}} \le 2nc\left(\frac{\pi_{\min}^{2(1+\zeta^{-1})}}{k^2}\right), \quad \forall j \neq j',\ \forall \boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star) \tag{C.3.40}$$

for some numerical constant $c \in (0,1)$. Note that (C.3.38) is a sufficient condition to invoke Lemma 30 with probability at least $1 - \delta/12$. Therefore, all intermediate steps in the proof of Lemma 30 hold. In particular, due to the inclusion argument in (C.2.15), $\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star$ implies $\boldsymbol{\xi}_i = [\boldsymbol{x}_i; 1] \in \mathcal{S}_{\boldsymbol{v}_{j,j'}, \boldsymbol{v}_{j,j'}^\star}$ for any $j \neq j'$, where $\mathcal{S}_{\boldsymbol{v}_{j,j'}, \boldsymbol{v}_{j,j'}^\star}$ is defined in (C.2.17). Then, (C.2.21) with α as in (C.2.20) implies (C.3.40). The last event is defined by

$$\max_{\substack{\mathcal{I} \subset [n] \\ |\mathcal{I}| \le 2\alpha n}} \lambda_{\max}\left(\frac{1}{n}\sum_{i \in \mathcal{I}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T}\right) \le C_4(\eta^2 \vee 1)\sqrt{\alpha}, \quad \forall \alpha \in \left\{\frac{c\pi_{\min}^{2(1+\zeta^{-1})}}{k^2}\right\} \cup \left\{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)\right\}_{j=1}^k. \tag{C.3.41}$$

By (C.3.38), Lemma 24, and the union bound over $j \in [k]$, (C.3.41) holds with probability at least $1 - \delta/12$. Thus far we have shown that (C.3.32), (C.3.39), (C.3.40), and (C.3.41) hold with probability at least $1 - \delta/3$. We proceed conditioned on the event that $\{\boldsymbol{\xi}_i\}_{i=1}^n$ satisfy these conditions.

Let $\boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}$, $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, and $j \in [k]$ be arbitrarily fixed. Then the partial gradient of $\ell^{\text{clean}}(\boldsymbol{\beta})$ with respect to the jth block $\boldsymbol{\beta}_j \in \mathbb{R}^{d+1}$ of $\boldsymbol{\beta} \in \mathbb{R}^{k(d+1)}$ is written

as

$$\begin{aligned}\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\left(\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle\right)\boldsymbol{\xi}_i \\ &= \frac{1}{n}\sum_{j'\in[k]}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\left(\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j\rangle - \langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_{j'}^\star\rangle\right)\boldsymbol{\xi}_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\rangle\boldsymbol{\xi}_i + \frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j'}^\star\rangle\boldsymbol{\xi}_i.\end{aligned} \tag{C.3.42}$$

By using the identity $\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j'}^\star\rangle = \langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star+\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star\rangle$, (C.3.42) is rewritten as

$$\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\rangle\boldsymbol{\xi}_i + \frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star\rangle\boldsymbol{\xi}_i. \tag{C.3.43}$$

Then it follows from (C.3.43) that

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta})\right\|_2^2 \\ &\overset{\text{(i)}}{\leq} 2\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\rangle\boldsymbol{\xi}_i\right\|_2^2 + 2\left\|\frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star\rangle\boldsymbol{\xi}_i\right\|_2^2 \\ &\overset{\text{(ii)}}{\leq} 2\cdot\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\rangle^2 \\ &\quad + 2\cdot\sum_{j':j'\neq j}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star\rangle^2 \\ &\leq 2\cdot\underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|^2}_{\text{(a)}}\cdot\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\|_2^2 \\ &\quad + 2\cdot\underbrace{\max_{j':j'\neq j}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\|}_{\text{(b)}}\cdot\underbrace{\frac{1}{n}\sum_{j':j'\neq j}\sum_{i=1}^{n}\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\cap\mathcal{C}_{j'}^\star\}}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star-\boldsymbol{\beta}_{j'}^\star\rangle^2}_{\text{(c)}},\end{aligned} \tag{C.3.44}$$

where (i) holds since $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2\leq 2\|\boldsymbol{a}\|_2^2+2\|\boldsymbol{b}\|_2^2$ and (ii) holds since $\mathcal{C}_j\cap\mathcal{C}_l^\star$ and $\mathcal{C}_j\cap\mathcal{C}_{l'}^\star$ are disjoint for any $l\neq l'\in[k]$. An upper bound on (b) is provided by (C.3.32). It remains to derive upper bounds on (a) and (c).

First, we derive an upper bound on (a). By the triangle inequality, we have

$$\sqrt{(\mathrm{a})} \leq \sum_{j'=1}^{k} \left\| \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\|. \tag{C.3.45}$$

For the summand indexed by $j' = j$, due to the set inclusion $\mathcal{C}_j \cap \mathcal{C}_j^\star \subset \mathcal{C}_j^\star$, we obtain that

$$\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \preceq \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T}.$$

Therefore, by (C.3.39) and (C.3.41), we have

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| &\leq \max_{\mathcal{I}: |\mathcal{I}| \leq 2n\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \left\| \frac{1}{n} \sum_{i \in \mathcal{I}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| \\ &\lesssim (\eta^2 \vee 1) \sqrt{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \\ &\leq (\eta^2 \vee 1) \sqrt{\pi_{\max}}, \end{aligned} \tag{C.3.46}$$

where the last inequality holds by the definition of $\pi_{\max}$. Similarly, by (C.3.40) and (C.3.41), we have

$$\left\| \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| \lesssim (\eta^2 \vee 1) \sqrt{c} \left(\frac{\pi_{\min}^{1+\zeta^{-1}}}{k} \right), \quad \forall j' \neq j. \tag{C.3.47}$$

Then by plugging in (C.3.46) and (C.3.47) to (C.3.45), we obtain

$$(\mathrm{a}) \lesssim \left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} \right) \left\| \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \right\|_2^2$$

for an absolute constant C_1. Finally, since an upper bound on (b) is given by (C.3.47), plugging in the obtained upper bounds to (C.3.44) provides the assertion.

Proof of (C.3.4)**:** By the variational characterization of the Euclidean norm and the triangle inequality, we have

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta})\right\|_2 &= \sup_{[\boldsymbol{u};\ w]\in B_2^{d+1}}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}(\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle+w)\right| \\ &\le \underbrace{\sup_{\boldsymbol{u}\in B_2^p}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle\right|}_{\text{(A)}} + \underbrace{\sup_{|w|\le 1}\left|\frac{1}{n}\sum_{i=1}^n z_i\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}w\right|}_{\text{(B)}}, \quad \text{(C.3.48)}\end{aligned}$$

where B_2^d denotes the unit ball in ℓ_2^d. Note that (A) and (B) depend on $\boldsymbol{\beta}$ only through $\mathcal{C}_j$, which are determined by $\boldsymbol{\beta}$ according to (1.2.4). For any $\boldsymbol{\beta}$ and any $j\in[k]$, the corresponding $\mathcal{C}_j$ is given as the intersection of up to k affine spaces. Therefore, it suffices to maximize $\left\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta})\right\|_2$ over $\mathcal{C}_j\in\mathcal{P}_{k-1}$ for a fixed j, where $\mathcal{P}_{k-1}$ is defined in the statement of Theorem 22.

We proceed under the event that the following inequalities hold:

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\right\| \le 1+\epsilon \quad \text{(C.3.49)}$$

and

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}} - \mathbb{P}(\boldsymbol{x}\in\mathcal{C}_j)\right| \le \epsilon, \quad \forall\mathcal{C}_j\in\mathcal{P}_{k-1} \quad \text{(C.3.50)}$$

for some constant ϵ, which we specify later. The remainder of the proof is given conditioned on $(\boldsymbol{x}_i)_{i=1}^n$ satisfying (C.3.49) and (C.3.50).

First, we derive an upper bound on (A) in (C.3.48). Note that (A) corresponds to the supremum of the random process

$$Z_{\boldsymbol{u}} := \frac{1}{n}\sum_{i=1}^n z_i\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_j\}}\langle\boldsymbol{x}_i,\boldsymbol{u}\rangle$$

over $\boldsymbol{u} \in B_2^p$. The sub-Gaussian increment satisfies

$$\begin{aligned}
\|Z_{\boldsymbol{u}} - Z_{\boldsymbol{u}'}\|_{\psi_2} &\lesssim \frac{\sigma}{\sqrt{n}} \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{x}_i, \boldsymbol{u} - \boldsymbol{u}' \rangle^2} \\
&\leq \frac{\sigma}{\sqrt{n}} \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right\|^{1/2} \cdot \|\boldsymbol{u} - \boldsymbol{u}'\|_2 \\
&\leq \frac{\sigma}{\sqrt{n}} \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top \right\|^{1/2} \cdot \|\boldsymbol{u} - \boldsymbol{u}'\|_2 \\
&\leq \frac{\sigma\sqrt{1+\epsilon}}{\sqrt{n}} \cdot \|\boldsymbol{u} - \boldsymbol{u}'\|_2,
\end{aligned}$$

where the third step follows from the inequality

$$\left\| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right\| \leq \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top \right\|,$$

which holds deterministically, and the last step follows from (C.3.49). Then, by applying a version of Dudley's inequality [93, Theorem 8.1.6], we obtain that

$$\mathbb{P}\left(\sup_{\boldsymbol{u} \in B_2^p} |Z_{\boldsymbol{u}}| > \frac{C_1 \sigma \sqrt{1+\epsilon}}{\sqrt{n}} \left(\int_0^\infty \sqrt{\log N(B_2^p, \|\cdot\|_2, \eta)} d\eta + \sqrt{\log(1/\delta)} \right) \right) \leq \delta.$$

By the elementary upper bound on the covering number $N(B_2^p, \|\cdot\|_2, \eta) \leq (3/\eta)^p$ (e.g. see [93, Example 8.1.11]) and the definition of (A) in (C.3.48), we have

$$\text{(A)} \lesssim \sqrt{\frac{\sigma^2(1+\epsilon)(d + \log(1/\delta))}{n}}, \tag{C.3.51}$$

holds with probability $1 - \delta/3$. Then we apply the union bound over $\mathcal{C}_j \in \mathcal{P}_{k-1}$. It follows from (C.1.1) that

$$\sup_{\mathcal{C}_j \in \mathcal{P}_{k-1}} \text{(A)} \lesssim \sqrt{\frac{\sigma^2(1+\epsilon)(\log(1/\delta) + kd\log(n/d))}{n}}$$

holds with probability $1 - \delta/9$.

Next we derive an upper bound on (B) in (C.3.48). Note that (B) is rewritten as the absolute value of

$$\varrho = \frac{1}{n}\sum_{i=1}^{n} z_i \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}}.$$

Conditioned on $(\boldsymbol{x}_i)_{i=1}^n$ satisfying (C.3.50), ϱ is a sub-Gaussian random variable that satisfies $\mathbb{E}\varrho = 0$ and

$$\mathbb{E}\varrho^2 = \frac{\sigma^2}{n} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \right) \leq \frac{\sigma^2(\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j) + \epsilon)}{n}.$$

The standard sub-Gaussian tail bound implies

$$\mathbb{P}\left(|\varrho| > \sqrt{\frac{C_2\sigma^2(\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j) + \epsilon)\log(1/\delta)}{n}} \right) \leq \delta.$$

By taking the union bound over $\mathcal{C}_j \in \mathcal{P}_{k-1}$ and utilizing the inequality in (C.1.1), we obtain that

$$\begin{aligned} \sup_{\mathcal{C}_j \in \mathcal{P}_{k-1}} (\mathrm{B}) &\lesssim \sqrt{\frac{\sigma^2(\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j) + \epsilon)\left(kd\log(n/d) + \log(1/\delta)\right)}{n}} \\ &\leq \sqrt{\frac{\sigma^2(1+\epsilon)\left(kd\log(n/d) + \log(1/\delta)\right)}{n}} \end{aligned} \tag{C.3.52}$$

holds with probability $1 - \delta/9$.

Finally it remains to show that (C.3.49) and (C.3.50) hold with probability $1 - \delta/3$ for ϵ satisfying

$$\epsilon \lesssim \sqrt{\frac{kp(\log(n/d) + \log(1/\delta))}{n}}.$$

This is obtained as a direct consequence of Lemmas 17 and 19. One can choose the absolute constant C in (3.1.14) large enough so that $\epsilon < 1$. Then the parameter ϵ in (C.3.51) and (C.3.52) will be dropped. This completes the proof.

C.4 Proof of Theorem 9

The proof will be similar to that for Theorem 8. We will focus on the distinction due to the modification of the algorithm with random sampling. The partial subgradient in the update for the mini-batch stochastic gradient descent algorithm is given by

$$\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_l}\ell_i(\boldsymbol{\beta}^t)=\frac{1}{m}\sum_{i\in I_t}\underbrace{\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\left(\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^t\rangle-\max_{j\in[k]}\langle\boldsymbol{\xi}_i,\boldsymbol{\beta}_j^\star\rangle\right)\boldsymbol{\xi}_i}_{\nabla_{\boldsymbol{\beta}_l}\ell_i^{\mathrm{clean}}(\boldsymbol{\beta}^t)}-\frac{1}{m}\sum_{i\in I_t}\underbrace{z_i\mathbb{1}_{\{\boldsymbol{x}_i\in\mathcal{C}_l\}}\boldsymbol{\xi}_i}_{\nabla_{\boldsymbol{\beta}_l}\ell_i^{\mathrm{noise}}(\boldsymbol{\beta}^t)},$$

where $\mathcal{C}_1,\ldots,\mathcal{C}_k$ are determined by $\boldsymbol{\beta}^t$ as in (1.2.4).

As shown in Appendix C.3, (3.1.14) invokes Lemma 31 and hence (C.3.2) holds with probability $1-\delta/3$. Next, we show that under the condition (3.1.14), the statements of the following lemma hold with probability $1-2\delta/3$. The proof is provided in Appendix C.4.1.

Lemma 32 *Suppose that the hypothesis of Theorem 9 holds. If* (3.1.14) *is satisfied, then the following statement holds with probability at least* $1-2\delta/3$*: For all* $j\in[k]$*,* $\boldsymbol{\beta}^\star\in\mathbb{R}^{k(d+1)}$*, and* $\boldsymbol{\beta}^t\in\mathcal{N}(\boldsymbol{\beta}^\star)$*, we have*

$$\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\mathrm{clean}}(\boldsymbol{\beta}^t)\right\|_2^2\lesssim$$
$$\left(1\vee\frac{d+\log(n/\delta)}{m}\right)\left(\left(\sqrt{\pi_{\max}}+\pi_{\min}^{1+\zeta^{-1}}\right)\left\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\|_2^2+\frac{\pi_{\min}^{1+\zeta^{-1}}}{k}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}^t-\boldsymbol{v}_{j,j'}^\star\right\|_2^2\right),\tag{C.4.1}$$

and

$$\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\mathrm{noise}}(\boldsymbol{\beta}^t)\right\|_2^2\lesssim\sigma^2\left(\frac{d+\log(n/\delta)}{m}\vee\frac{kd\log(n/d)+\log(1/\delta)}{n}\right).\tag{C.4.2}$$

Then we show that the assertion of the theorem follows from (C.3.2), (C.4.1), and (C.4.2) via the following three steps.

Step 1: We show that every iterate remains within the neighborhood $\mathcal{N}(\boldsymbol{\beta}^\star)$ by the induction argument. Therefore, we illustrate that if we suppose $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$ holds for a fixed $t \in \mathbb{N}$, we show $\boldsymbol{\beta}^{t+1} \in \mathcal{N}(\boldsymbol{\beta}^\star)$ in expectation. By the update rule of SGD with batch size m, the triangle inequality gives

$$\mathbb{E}_{I_t}\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^\star\|_2 \leq \underbrace{\mathbb{E}_{I_t}\left\|\boldsymbol{\beta}_j^t - \mu\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t) - \boldsymbol{\beta}_j^\star\right\|_2}_{A_{\text{clean}}} + \underbrace{\mu\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2}_{A_{\text{noise}}}. \tag{C.4.3}$$

We will show that

$$\mathbb{E}_{I_t}\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^\star\|_2 \leq A_{\text{clean}} + A_{\text{noise}} \leq \kappa\rho, \quad \forall j \in [k]. \tag{C.4.4}$$

By applying Jensen's inequality, we can obtain an upper-bound A_{clean} in (C.4.3):

$$\begin{aligned} A_{\text{clean}}^2 &\leq \mathbb{E}_{I_t}\left\|\boldsymbol{\beta}_j^t - \mu\cdot\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t) - \boldsymbol{\beta}_j^\star\right\|_2^2 \\ &= \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 - 2\mu\mathbb{E}_{I_t}\left\langle\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\rangle + \mu^2\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i(\boldsymbol{\beta}^t)\right\|_2^2. \end{aligned} \tag{C.4.5}$$

Due to the expectation, the second term in (C.4.5) simplifies to

$$\mathbb{E}_{I_t}\left\langle\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\rangle = \langle\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\rangle, \tag{C.4.6}$$

where $\nabla_{\boldsymbol{\beta}_j}\ell^{\text{clean}}(\boldsymbol{\beta}^t)$ is defined in (C.3.1). Then, (C.3.2) gives a lower bound on (C.4.6). Furthermore, an upper bound on the third term in (C.4.5) is given by (C.4.1). Putting the bounds (C.3.2) and (C.4.1) in (C.4.5) provides

$$\begin{aligned} &A_{\text{clean}}^2 \leq \\ &\left(1 - \frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}\mu\pi_{\min}^{1+\zeta^{-1}} + C_1\mu^2\left(1 \vee \frac{d+\log(n/\delta)}{m}\right)\left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}\right)\right)\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\ &+ \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}^{1+\zeta^{-1}}\right)\mu\pi_{\min}^{1+\zeta^{-1}}}{5k} + C_1\left(1 \vee \frac{d+\log(n/\delta)}{m}\right)\frac{\mu^2\pi_{\min}^{1+\zeta^{-1}}}{k}\right)\sum_{j'^*:j'\neq j}\left\|\boldsymbol{v}_{j,j}^t - \boldsymbol{v}_{j,j'}^\star\right\|_2^2. \end{aligned} \tag{C.4.7}$$

Let us choose the step size μ following

$$\mu = \frac{\omega \pi_{\min}^{1+\zeta^{-1}}}{\tau} \cdot \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \tag{C.4.8}$$

for a numerical constant ω, which we specify later, and τ defined as

$$\tau := \sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}. \tag{C.4.9}$$

Taking μ by (C.4.8) and τ by (C.4.9) in (C.4.7) yields

$$\begin{aligned}
&A_{\text{clean}}^2 \\
&\leq \left(1 - \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \cdot \right. \\
&\qquad \left. \left(\frac{\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \omega \pi_{\min}^{2(1+\zeta^{-1})}}{\tau} - \frac{C_1 \omega^2 \pi_{\min}^{2(1+\zeta^{-1})} \left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}\right)}{\tau^2} \right) \right) \left\| \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star \right\|_2^2 \\
&+ \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \cdot \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \omega \pi_{\min}^{2(1+\zeta^{-1})}}{5\tau k} + \frac{C_1 \omega^2 \pi_{\min}^{3(1+\zeta^{-1})}}{\tau^2 k} \right) \sum_{j' : j' \neq j} \left\| \boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star \right\|_2^2 \\
&\leq \left(1 - \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \cdot \left(\frac{\frac{4}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \omega \pi_{\min}^{2(1+\zeta^{-1})}}{\tau} - \frac{C_1 \omega^2 \pi_{\min}^{2(1+\zeta^{-1})}}{\tau} \right) \right) \left\| \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star \right\|_2^2 \\
&+ \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \cdot \left(\frac{\frac{2}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}} \omega \pi_{\min}^{2(1+\zeta^{-1})}}{5\tau} + \frac{C_1 \omega^2 \pi_{\min}^{2(1+\zeta^{-1})}}{\tau} \right) \max_{j \neq j'} \left\| \boldsymbol{v}_{j,j'}^t - \boldsymbol{v}_{j,j'}^\star \right\|_2^2 .
\end{aligned} \tag{C.4.10}$$

Due to $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$ defined in (3.1.11), we have (C.3.11) and (C.3.12) by Theorem 27.

Inserting (C.3.11) and (C.3.12) into (C.4.10) gives

$$\begin{aligned}
(\kappa\rho)^{-2} A_{\text{clean}}^2 &\leq 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})} \omega}{\tau} \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \left(\frac{4}{\gamma} \left(\frac{1}{16}\right)^{1+\zeta^{-1}} \left(1 - \frac{2}{5}\right) + C_1 \omega \left(1 + 4\right) \right) \\
&= 1 - \frac{\pi_{\min}^{2(1+\zeta^{-1})} \omega}{\tau} \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \left(\frac{\frac{12}{\gamma}\left(\frac{1}{16}\right)^{1+\zeta^{-1}}}{5} + 5\omega C_1 \right) \\
&\leq 1 - \frac{c_0 \omega \pi_{\min}^{2(1+\zeta^{-1})}}{\tau} \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right),
\end{aligned} \tag{C.4.11}$$

where c_0 is the numerical constant defined in (C.3.13). We represent (C.4.11) as

$$A_{\text{clean}}^2 \leq (\kappa\rho)^2 \left(1 - \frac{c_0 \omega \pi_{\min}^{2(1+\zeta^{-1})}}{\tau} \cdot \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right)\right). \tag{C.4.12}$$

We note that by (C.3.13), c_0 is a positive absolute constant given γ and ζ. On the other hand, the choice of τ in (C.4.9) provides a bound

$$\frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\tau} = \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}} < 1.$$

Since $(1 \wedge m/(d + \log(n/\delta)) < 1$, one can set $\omega > 0$ such that $\omega c_0 < 1$, which makes the upper bound in the right-hand side of (C.4.12) a positive scalar belonging in $(0, 1)$.

By following the arguments in (C.3.15) and (C.3.16), if

$$A_{\text{noise}} \leq \kappa\rho \left(\frac{c_0 \omega \pi_{\min}^{2(1+\zeta^{-1})}}{2\tau}\right) \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \tag{C.4.13}$$

holds, we have

$$A_{\text{noise}} \leq \kappa\rho \left(1 - \sqrt{1 - \frac{c_0 \omega \pi_{\min}^{2(1+\zeta^{-1})}}{\tau} \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right)}\right). \tag{C.4.14}$$

Since the upper bounds (C.4.12) and (C.4.14) satisfies (C.4.4) it suffices to show (C.4.13).

By (C.4.2), we have

$$\sqrt{\mathbb{E}_{I_t} \left\| \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{noise}}(\boldsymbol{\beta}^t) \right\|_2^2} \lesssim \sigma \sqrt{\left(\frac{d + \log(n/\delta)}{m} \vee \frac{kd \log(n/d) + \log(1/\delta)}{n}\right)}$$

for all $j \in [k]$. After applying Jensen's inequality, we consider the choice of μ given in (C.4.8). Then, we have

$$A_{\text{noise}} = \mu \mathbb{E}_{I_t} \left\| \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{noise}}(\boldsymbol{\beta}^t) \right\|_2 \leq \mu \sqrt{\mathbb{E}_{I_t} \left\| \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{noise}}(\boldsymbol{\beta}^t) \right\|_2^2} \lesssim$$
$$\frac{\sigma \omega \pi_{\min}^{1+\zeta^{-1}}}{\tau} \left(1 \wedge \frac{m}{d + \log(n/\delta)}\right) \sqrt{\left(\frac{d + \log(n/\delta)}{m} \vee \frac{kd \log(n/d) + \log(1/\delta)}{n}\right)}. \tag{C.4.15}$$

Since (3.1.14) implies (C.3.18), we can choose a sufficiently large absolute constant $C > 0$ in (C.3.18) such that (C.3.18) and (C.4.15) result in (C.4.13). We complete the proof of induction argument in Step 1.

Step 2: In this step, we show that every iterate obeys

$$\begin{aligned}&\mathbb{E}_{I_t}\left\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^\star\right\|_2 \leq \\ &\sqrt{1-\nu}\left\|\boldsymbol{\beta}^{t}-\boldsymbol{\beta}^\star\right\|_2 + C'\mu\sigma\sqrt{k}\cdot\left(\sqrt{\frac{d+\log(n/\delta)}{m}}\vee\sqrt{\frac{kd\log(n/d)+\log(1/\delta)}{n}}\right).\end{aligned} \tag{C.4.16}$$

In Step 1, we showed $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$. By following the argument (C.4.3), we have

$$\begin{aligned}\mathbb{E}_{I_t}\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^\star\|_2 &\leq \mathbb{E}_{I_t}\left\|\boldsymbol{\beta}^t-\mu\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}^\star\right\|_2 + \mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I}\nabla_{\boldsymbol{\beta}}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2 \\ &\leq \underbrace{\sqrt{\mathbb{E}_{I_t}\left\|\boldsymbol{\beta}^t-\mu\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)-\boldsymbol{\beta}^\star\right\|_2^2}}_{B_{\text{clean}}} + \underbrace{\sqrt{\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I}\nabla_{\boldsymbol{\beta}}\ell_i^{\text{noise}}(\boldsymbol{\beta}^t)\right\|_2^2}}_{B_{\text{noise}}},\end{aligned} \tag{C.4.17}$$

where the last inequality holds by the Jensen's inequality. We first show an upper bound on B_{clean} in (C.4.17):

$$B_{\text{clean}}^2 \leq (1-\nu)\sum_{j=1}^k\left\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\|_2^2. \tag{C.4.18}$$

By following the argument in (C.3.23), (C.4.18) holds if there exist constants $\mu, \lambda \in (0,1)$ such that for all $\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star)$,

$$\begin{aligned}&\sum_{j=1}^k\mathbb{E}_{I_t}\left\langle\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\right\rangle \\ &\qquad\geq \frac{\mu}{2}\sum_{j=1}^k\mathbb{E}_{I_t}\left\|\frac{1}{m}\sum_{i\in I_t}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{clean}}(\boldsymbol{\beta}^t)\right\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^k\|\boldsymbol{\beta}_j^t-\boldsymbol{\beta}_j^\star\|_2^2.\end{aligned} \tag{C.4.19}$$

Hence, we show (C.4.19).First, since (C.3.2) holds, (C.3.25) holds. Also, the left-hand side in (C.4.19) can be computed as (C.4.6). Thus, by (C.4.6) and (C.3.25), we obtain

a lower bound on the left-hand side of (C.4.19):

$$\sum_{j=1}^{k} \mathbb{E}_{I_t} \left\langle \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}^t), \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star \right\rangle \geq \frac{\frac{6}{\gamma} \left(\frac{1}{16}\right)^{1+\zeta^{-1}} \pi_{\min}^{1+\zeta^{-1}}}{5} \sum_{j=1}^{k} \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2. \tag{C.4.20}$$

Furthermore, to obtain an upper bound on first term in the right-hand side of (C.4.19), applying (C.4.1) with the elementary inequality $\|\boldsymbol{a} + \boldsymbol{b}\|_2^2 \leq 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$ provides

$$\begin{aligned} \mathbb{E}_{I_t} \left\| \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}^t) \right\|_2^2 \leq C_1 \left(1 \vee \frac{d + \log(n/\delta)}{m}\right) \Bigg(& \left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}\right) \|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 \\ & + \frac{2\pi_{\min}^{1+\zeta^{-1}}}{k} \sum_{j':j' \neq j} \left(\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\|_2^2 + \|\boldsymbol{\beta}_{j'}^t - \boldsymbol{\beta}_{j'}^\star\|_2^2\right) \Bigg). \end{aligned} \tag{C.4.21}$$

Taking summation on (C.4.21) over $j \in [k]$ yields

$$\begin{aligned} & \sum_{j=1}^{k} \mathbb{E}_{I_t} \left\| \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}^t) \right\|_2^2 \\ & \leq C_1 \left(1 \vee \frac{d + \log(n/\delta)}{m}\right) \left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}} + 4\pi_{\min}^{1+\zeta^{-1}}\right) \sum_{j=1}^{k} \left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2^2. \end{aligned} \tag{C.4.22}$$

Putting the bounds (C.4.20) and (C.4.22) in (C.4.19) with μ chosen in (C.4.8), we have a sufficient condition for (C.4.19):

$$\frac{\frac{6}{\gamma} \left(\frac{1}{16}\right)^{1+\zeta^{-1}} \pi_{\min}^{1+\zeta^{-1}}}{5} \geq \frac{\omega \pi_{\min}^{1+\zeta^{-1}} C_1 \left(\sqrt{\pi_{\max}} + 5\pi_{\min}^{1+\zeta^{-1}}\right)}{2\left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}}\right)} + \frac{\lambda}{2}. \tag{C.4.23}$$

(C.4.23) is satisfied when we choose $\omega > 0$ small enough and λ as in (C.3.29). Hence, we have shown (C.4.18) with $\nu = \mu\lambda$ where μ and λ are chosen by (C.4.8) and (C.3.29).

Next, we bound B_{noise} in (C.4.17). By (C.4.2), we obtain an upper bound on B_{noise}:

$$\begin{aligned} B_{\text{noise}}^2 &= \mu^2 \sum_{j=1}^{k} \mathbb{E}_{I_t} \left\| \frac{1}{m} \sum_{i \in I_t} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{noise}}(\boldsymbol{\beta}^t) \right\|_2^2 \\ &\lesssim k\mu^2\sigma^2 \left(\frac{d + \log(n/\delta)}{m} \vee \frac{kd\log(n/d) + \log(1/\delta)}{n} \right). \end{aligned} \tag{C.4.24}$$

Finally, putting (C.4.18) and (C.4.24) in (C.4.17) gives (C.4.16). We complete the proof of Step 2.

Step 3: We finish the proof of Theorem 9 using the results demonstrated in Step 1 and Step 2. By substituting the expression $\nu = \mu\lambda$, where we choose μ and λ according to (C.4.8) and (C.3.29) respectively, into (C.4.16), we obtain

$$\begin{aligned} &\mathbb{E}_{I_t} \|\boldsymbol{\beta}^t - \boldsymbol{\beta}^\star\|_2 \\ &(1-\mu\lambda)^{t/2} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2 + C_2 \cdot \frac{\mu\sigma}{1-\sqrt{1-\mu\lambda}} \cdot \sqrt{k \cdot \left(\frac{d + \log(n/\delta)}{m} \vee \frac{kd\log(n/d) + \log(1/\delta)}{n} \right)} \\ &\overset{\text{(a)}}{\leq} (1-\mu\lambda)^{t/2} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2 + C_2 \cdot \frac{2\sigma}{\lambda} \cdot \sqrt{k \cdot \left(\frac{d + \log(n/\delta)}{m} \vee \frac{kd\log(n/d) + \log(1/\delta)}{n} \right)} \\ &\overset{\text{(b)}}{\leq} (1-\mu\lambda)^{t/2} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2 + C_3 \cdot \frac{\sigma}{\pi_{\max}} \cdot \sqrt{k \cdot \left(\frac{d + \log(n/\delta)}{m} \vee \frac{kd\log(n/d) + \log(1/\delta)}{n} \right)} \\ &\overset{\text{(c)}}{\leq} (1-\mu\lambda)^{t/2} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^\star\|_2 + C_3 \cdot \sigma k \cdot \sqrt{k \cdot \left(\frac{d + \log(n/\delta)}{m} \vee \frac{kd\log(n/d) + \log(1/\delta)}{n} \right)}, \end{aligned}$$

where i) (a) follows from the inequality $\sqrt{1-t} < -t/2 + 1$ for any $t \in (0,1)$; ii) (b) holds by the choice of τ in (C.4.9); iii) (c) is a result of $\pi_{\max}^{-1} \leq k$.

C.4.1 Proof of Lemma 32

We will show that both (C.4.1) and (C.4.2) hold with probability at least $1 - \delta/3$. Furthermore, for simplicity, we proceed on the proofs using $\boldsymbol{\beta}$ and $\boldsymbol{v}_{j,j'}$ instead of

using $\boldsymbol{\beta}^t$ and $\boldsymbol{v}_{j,j'}^t$ in the statements of Theorem 32. Thus, we complete the assertions in (C.4.1) and (C.4.2) by substituting $\boldsymbol{\beta}$ and $\boldsymbol{v}_{j,j'}$ with $\boldsymbol{\beta}^t$ and $\boldsymbol{v}_{j,j'}^t$ respectively.

Proof of (C.4.1)**:** We show that with high probability, (C.4.1) holds if

$$n \geq C_1 \left(\log(k/\delta) \vee d \log(n/d)\right) k^4 \pi_{\min}^{-4(1+\zeta^{-1})}, \tag{C.4.25}$$

Note that (3.1.14) is a sufficient condition for (C.4.25). We proceed with the proof under the following six events, each of which holds with probability at least $1 - \delta/18$. First, by the proof of (C.3.3) in Appendix C.3.1, (C.4.25) is a sufficient condition to invoke (C.3.3) with probability at least $1 - \delta/18$. Next, by following the argument for (C.3.39), (C.4.25) is a sufficient condition to invoke (C.3.39) with probability at least $1 - \delta/18$. Furthermore, (C.4.25) implies (C.2.13) and is a sufficient condition to invoke Lemma 30 and Lemma 24 with probability at least $1 - \delta/18$ respectively. Hence, by following the arguments for (C.3.40), (C.3.41), and (C.3.32), (C.3.40), (C.3.41), and (C.3.32) hold with probability at least $1 - \delta/18$ respectively. The last event is defined as

$$\max_{i \in [n]} \|\boldsymbol{\xi}_i \boldsymbol{\xi}_i^{\mathsf{T}}\| \lesssim d + \log(n/\delta). \tag{C.4.26}$$

By Lemma 17 and the union bound over $i \in [n]$, (C.4.26) holds with probability at least $1 - \delta/18$.

Since we showed that (C.3.3), (C.3.39), (C.3.40), (C.3.41), (C.3.32), and (C.4.26) hold with probability at least $1 - \delta/3$, we will move forward with the remainder of the proof by assuming those conditions are satisfied.

Let $\boldsymbol{\beta}^\star \in \mathbb{R}^{d+1}$, $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}^\star)$, and $j \in [k]$ be arbitrarily fixed. By the argument in [63, Equation 7], we decompose

$$\mathbb{E}_I \left\| \frac{1}{m} \sum_{i \in I} \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}) \right\|_2^2 = \underbrace{\frac{1}{m} \mathbb{E}_{i_1} \left\| \nabla_{\boldsymbol{\beta}_j} \ell_{i_1}^{\text{clean}}(\boldsymbol{\beta}) \right\|_2^2}_{\text{(A)}} + \underbrace{\frac{m-1}{m} \| \nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta}) \|_2^2}_{\text{(B)}}, \quad \text{(C.4.27)}$$

where we define $I := \{i_1, \ldots, i_m\} \subset [n]$ and $\nabla_{\boldsymbol{\beta}_j} \ell^{\text{clean}}(\boldsymbol{\beta})$ in (C.3.1).

Note that (C.3.3) gives an upper bound on (B):

$$\text{(B)} \lesssim \frac{m-1}{m} \left(\left(\pi_{\max} + \pi_{\min}^{2(1+\zeta^{-1})} \right) \left\| \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \right\|_2^2 + \frac{\pi_{\min}^{2(1+\zeta^{-1})}}{k^2} \sum_{j' : j' \neq j} \left\| \boldsymbol{v}_{j,j'} - \boldsymbol{v}_{j,j'}^\star \right\|_2^2 \right). \quad \text{(C.4.28)}$$

It remains to show the bound on (A). By following arguments (C.3.43), we decompose $\nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta})$ following

$$\nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}) = \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle \boldsymbol{\xi}_i + \sum_{j' : j' \neq j} \mathbb{1}_{\left\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\right\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle \boldsymbol{\xi}_i, \quad \forall i \in [n]. \quad \text{(C.4.29)}$$

Then it follows from (C.4.29) that for any $i \in [n]$,

$$\begin{aligned}
&\left\| \nabla_{\boldsymbol{\beta}_j} \ell_i^{\text{clean}}(\boldsymbol{\beta}) \right\|_2^2 \\
&\overset{\text{(i)}}{\leq} 2 \left\| \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle \boldsymbol{\xi}_i \right\|_2^2 + 2 \left\| \sum_{j' : j' \neq j} \mathbb{1}_{\left\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\right\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle \boldsymbol{\xi}_i \right\|_2^2 \\
&\overset{\text{(ii)}}{=} 2 \cdot \left\| \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \right\| \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle^2 + 2 \cdot \left\| \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \right\| \cdot \sum_{j' : j' \neq j} \mathbb{1}_{\left\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\right\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle^2 \\
&\overset{\text{(iii)}}{\lesssim} (d + \log(n/\delta)) \cdot \left(\mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle^2 + \sum_{j' : j' \neq j} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle^2 \right),
\end{aligned} \quad \text{(C.4.30)}$$

where (i) holds due to $\|\boldsymbol{a} + \boldsymbol{b}\|_2^2 \leq 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$; (ii) holds since $\mathcal{C}_j \cap \mathcal{C}_l^\star$ and $\mathcal{C}_j \cap \mathcal{C}_{l'}^\star$ are disjoint for any $l \neq l' \in [k]$; and (iii) holds by (C.4.26).

Applying the expectation on (C.4.30) yields

$$\mathbb{E}_{i_1} \left\|\nabla_{\boldsymbol{\beta}_j} \ell_{i_1}(\boldsymbol{\beta})\right\|_2^2 \lesssim$$

$$(d+\log(n/\delta)) \cdot \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \rangle^2}_{\text{(a)}} + \underbrace{\frac{1}{n} \sum_{j':j' \neq j} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star \rangle^2}_{\text{(b)}} \right). \tag{C.4.31}$$

An upper bound on (b) is provided by (C.3.32). It remains to derive an upper bound on (a).

The triangle inequality provides

$$\text{(a)} \leq \sum_{j'=1}^{k} \left\| \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| \cdot \left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\right\|_2^2 \tag{C.4.32}$$

For the summand indexed by $j' = j$, the set inclusion, $\mathcal{C}_j \cap \mathcal{C}_j^\star \subseteq \mathcal{C}_j^\star$ yields

$$\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \preceq \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T}.$$

Therefore, by (C.3.39) and (C.3.41), we have

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| &\leq \max_{\mathcal{I}:|\mathcal{I}| \leq 2n\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \left\| \frac{1}{n} \sum_{i \in \mathcal{I}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| \\ &\lesssim (\eta^2 \vee 1) \sqrt{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)} \\ &\leq (\eta^2 \vee 1) \sqrt{\pi_{\max}}, \end{aligned} \tag{C.4.33}$$

where the last inequality holds by the definition of $\pi_{\max}$. Similarly, by (C.3.40) and (C.3.41), we have

$$\left\| \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j \cap \mathcal{C}_{j'}^\star\}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\mathsf{T} \right\| \lesssim (\eta^2 \vee 1) \sqrt{c} \left(\frac{\pi_{\min}^{1+\zeta^{-1}}}{k} \right), \quad \forall j' \neq j. \tag{C.4.34}$$

Then by plugging in (C.4.33) and (C.4.34) into (C.4.32), we obtain

$$\text{(a)} \lesssim \left(\sqrt{\pi_{\max}} + \pi_{\min}^{1+\zeta^{-1}} \right) \left\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star\right\|_2^2.$$

Finally, applying obtained upper bounds on (a) and (b) in (C.4.31) gives

$$\text{(A)} \lesssim \frac{(d+\log(n/\delta))}{m}\left(\left(\sqrt{\pi_{\max}}+\pi_{\min}^{(1+\zeta^{-1})}\right)\left\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^\star\right\|_2^2+\frac{\pi_{\min}^{(1+\zeta^{-1})}}{k}\sum_{j':j'\neq j}\left\|\boldsymbol{v}_{j,j'}-\boldsymbol{v}_{j,j'}^\star\right\|_2^2\right). \tag{C.4.35}$$

Putting (C.4.28) and (C.4.35) in (C.4.27) completes the proof.

Proof of (C.4.2)**:** We proceed with the proof under the following three events, each of which holds with probability at least $1-\delta/9$. First, (3.1.14) invokes (C.3.4) with probability at least $1-\delta/9$. Next, by following the same argument in the proof of (C.4.1), (C.4.26) holds with probability at least $1-\delta/9$. The last event is the following:

$$\frac{1}{n}\sum_{i=1}^n z_i^2 \leq \sigma^2\left(1+\sqrt{\frac{C\log(1/\delta)}{n}}\right). \tag{C.4.36}$$

Since $\{z_i\}_{i=1}^n$ are i.i.d σ-sub-Gaussian random variables, the Bernstein's inequality yields that (C.4.36) holds with probability at least $1-\delta/9$.

We have shown that (C.3.4), (C.4.26), and (C.4.36) hold with probability at least $1-\delta/3$. For the remainder of the proof, we assume that those conditions are satisfied.

Then, by the argument in [63, Equation 7], we decompose

$$\mathbb{E}_I\left\|\frac{1}{m}\sum_{i\in I}\nabla_{\boldsymbol{\beta}_j}\ell_i^{\text{noise}}(\boldsymbol{\beta})\right\|_2^2=\underbrace{\frac{1}{m}\mathbb{E}_{i_1}\left\|\nabla_{\boldsymbol{\beta}_j}\ell_{i_1}^{\text{noise}}(\boldsymbol{\beta})\right\|_2^2}_{\text{(A)}}+\underbrace{\frac{m-1}{m}\|\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta})\|_2^2}_{\text{(B)}}, \tag{C.4.37}$$

where we define $I:=\{i_1,\ldots,i_m\}\subset[n]$ and $\nabla_{\boldsymbol{\beta}_j}\ell^{\text{noise}}(\boldsymbol{\beta})$ in (C.3.1).

(C.3.4) gives an upper bound on (B):

$$\text{(B)} \lesssim \frac{\sigma^2 kd\log(n/d)+\log(k/\delta)}{n}. \tag{C.4.38}$$

The remaining step is to obtain a bound on (A). Since we have

$$\left\|\nabla_{\boldsymbol{\beta}_j}\ell_{i_1}^{\text{noise}}(\boldsymbol{\beta})\right\|_2^2 \leq \|z_{i_1}\boldsymbol{\xi}_{i_1}\|_2^2 \leq \|\boldsymbol{\xi}_{i_1}\boldsymbol{\xi}_{i_1}^\top\|z_{i_1}^2 \lesssim d+\log(n/\delta)z_{i_1}^2,$$

where the last inequality holds by (C.4.26), applying the expectation and (C.4.36) gives an upper bound on (A):

$$\begin{aligned}(\mathrm{A}) \lesssim \frac{1}{n}\sum_{i=1}^{n} z_i^2 \left(\frac{d+\log(n/\delta)}{m}\right) &\lesssim \sigma^2\left(1 \vee \left(\frac{\log(1/\delta)}{n}\right)^{1/2}\right)\left(\frac{d+\log(n/\delta)}{m}\right) \\ &\leq \sigma^2\left(\frac{d+\log(n/\delta)}{m}\right),\end{aligned} \tag{C.4.39}$$

where the last inequality hold by (3.1.14). Putting the results (C.4.38) and (C.4.39) into (C.4.37) reduces to (C.4.2).

C.5 Discussion on the proofs of [38, Theorem 1] and [36, Theorem 1]

In the proof of [38, Theorem 1], they claimed that $n \gtrsim \delta^{-2}$ implies [38, Equation (45)]. They showed that [38, Equation (45)] follows from [38, Lemmas 10 and 11]. Their [38, Lemma 10] presents the concentration of the supremum of an empirical measure via the VC dimension and [38, Lemma 11] computes an upper bound on the VC dimension of the feasible set of the maximization. According to their proof argument, the number of observations n should be proportional to the VC dimension $d\log(n/d)$ to obtain the concentration in [38, Equation (45)]. Their sufficient condition $n \gtrsim \delta^{-2}$ for [38, Equation (45)] missed the dependence on the VC dimension. We suspect that this is a typo. While it does not ruin their main result, the sample complexity in [38, Theorem 1] might need to be corrected accordingly. Specifically, between [38, Equation (32) and (33)], the parameter δ in [38, Lemma 6] was set to $\delta = Ck^{-2}\pi_{\min}^6$ to upper-bound the second summand in the right-hand side of [38, Equation (32)]. Therefore, the corrected sample complexity of [38, Lemma 6] increases to

$\widetilde{O}(k^4 d\pi_{\min}^{-12})$ so that it dominates the sample complexity for part (b) in [38, Proposition 1] ($n \gtrsim kd\pi_{\min}^{-3}$). Consequently, the sample complexity in [38, Theorem 1] will increase by a factor $k^3\pi_{\min}^{-9}$.

Next, we report another mistake in their analysis under the generalized covariate model [36, Theorem 1]. They mistakenly omitted the dependence of σ in the sample complexity. A careful examination of their proof on page 48 in [37] will reveal that they use the same technique as in their other analysis in the Gaussian covariates case [38]. Therefore, we expect that their sample complexity should depend on the noise variance σ^2 to ensure that the next iterate belongs to the local neighborhood of the ground truth (refer to the proof of their Theorem 1 on page 1865 in [38]).

Bibliography

[1] Sydney N Afriat. The construction of utility functions from expenditure data. *International economic review*, 8(1):67–77, 1967.

[2] Sohail Bahmani. Estimation from nonlinear observations via convex programming with application to bilinear regression. *Electronic Journal of Statistics*, 13(1):1978–2011, 2019.

[3] Sohail Bahmani and Justin Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. In *Artificial Intelligence and Statistics*, pages 252–260, Fort Lauderdale, FL, USA, 2017. PMLR.

[4] Sohail Bahmani and Justin Romberg. Solving equations of random convex functions via anchored regression. *Foundations of Computational Mathematics*, 19(4):813–841, 2019.

[5] Gábor Balázs. Adaptively partitioning max-affine estimators for convex regression. In *International Conference on Artificial Intelligence and Statistics*, pages 860–874. PMLR, 2022.

[6] Gábor Balázs, András György, and Csaba Szepesvári. Near-optimal max-affine estimators for convex regression. In *Artificial Intelligence and Statistics*, pages 56–64. PMLR, 2015.

[7] Gábor Balázs. *Convex Regression: Theory, Practice, and Applications*. PhD thesis, Dept. Comput. Sci., University of Alberta, Edmonton, AB, Canada, 2016.

[8] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4331–4339, 2019.

[9] M Bertero and Michele Piana. Inverse problems in biomedical imaging: modeling and methods of solution. *Complex systems in biomedicine*, pages 1–33, 2006.

[10] Peter Bloomfield and William L Steiger. *Least absolute deviations: theory, applications, and algorithms*. Springer, 1983.

[11] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[12] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[13] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[14] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[15] Oliver Bunk, Ana Diaz, Franz Pfeiffer, Christian David, Bernd Schmitt, Dillip K Satapathy, and J Friso Van Der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A: Foundations of Crystallography*, 63(4):306–314, 2007.

[16] James V Burke and Michael C Ferris. A Gauss–Newton method for convex composite optimization. *Mathematical Programming*, 71(2):179–194, 1995.

[17] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[18] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[19] Bernd Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. In *Annales de l'institut Fourier*, volume 35, pages 79–118, 1985.

[20] Anwei Chai, Miguel Moscoso, and George Papanicolaou. Array imaging using intensity-only measurements. *Inverse Problems*, 27(1):015005, 2010.

[21] F.H. Clarke. *Optimization and Nonsmooth Analysis*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990.

[22] Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

[23] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

[24] Monika Csikos, Andrey Kupavskii, and Nabil H Mustafa. Optimal bounds on the vc-dimension. *arXiv preprint arXiv:1807.07924*, 2018.

[25] Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pages 1155–1164. PMLR, 2018.

[26] Amit Daniely, Sivan Sabato, and Shai Shwartz. Multiclass learning approaches: A theoretical comparison with implications. *Advances in Neural Information Processing Systems*, 25, 2012.

[27] John E Dennis Jr and Roy E Welsch. Techniques for nonlinear least squares and robust regression. *Communications in Statistics-simulation and Computation*, 7(4):345–359, 1978.

[28] Ilias Diakonikolas, Jong Ho Park, and Christos Tzamos. Relu regression with massart noise. *Advances in Neural Information Processing Systems*, 34, 2021.

[29] Sjoerd Dirksen. Tail bounds via generic chaining. 2015.

[30] Jonathan Dong, Lorenzo Valzania, Antoine Maillard, Thanh-an Pham, Sylvain Gigan, and Michael Unser. Phase retrieval: From computational imaging to machine learning: A tutorial. *IEEE Signal Processing Magazine*, 40(1):45–57, 2023.

[31] John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

[32] Richard M Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[33] Andreas Engel. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[34] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[35] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237, 1972.

[36] Avishek Ghosh, Ashwin Pananjady, Aditya Guntuboyina, and Kannan Ramchandran. Max-affine regression with universal parameter estimation for small-ball designs. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2706–2710. IEEE, 2020.

[37] Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. Max-affine regression: Provable, tractable, and near-optimal statistical estimation. *arXiv preprint arXiv:1906.09255*, 2019.

[38] Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. Max-affine regression: Parameter estimation for gaussian designs. *IEEE Transactions on Information Theory*, 68(3):1851–1885, 2021.

[39] Tom Goldstein and Christoph Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, 2018.

[40] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[42] Adityanand Guntuboyina and Bodhisattva Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, 2012.

[43] LLC Gurobi Optimization. Gurobi optimizer reference manual, 2021.

[44] LLC Gurobi Optimization. Gurobi optimizer reference manual, 2021.

[45] Qiyang Han and Jon A Wellner. Multivariate convex regression: global risk bounds and adaptation. *arXiv preprint arXiv:1601.06844*, 2016.

[46] Paul Hand and Vladislav Voroninski. Corruption robust phase retrieval via linear programming. *arXiv preprint arXiv:1612.03547*, 2016.

[47] Lauren A Hannah and David B Dunson. Multivariate convex regression with adaptive partitioning. *The Journal of Machine Learning Research*, 14(1):3261–3294, 2013.

[48] David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

[49] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[50] JB Hiriart-Urruty. New concepts in nondifferentiable programming. *Bull. Soc. Math. France*, 60:57–85, 1979.

[51] Kenneth Holmström, Anders O Göran, and Marcus M Edvall. User's guide for tomlab/cplex v12. 1. *Tomlab Optim. Retrieved*, 1:2017, 2009.

[52] Tianyang Hu, Zuofeng Shang, and Guang Cheng. Sharp rate of convergence for deep neural network classifiers under the teacher-student setting. *arXiv preprint arXiv:2001.06892*, 2020.

[53] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[54] Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

[55] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

[56] Elizabeth John and E Alper Yıldırım. Implementation of warm-start strategies in interior-point methods for linear programming in fixed dimension. *Computational Optimization and Applications*, 41(2):151–183, 2008.

[57] Marius Junge and Kiryung Lee. Generalized notions of sparsity and restricted isometry property. Part I: A unified framework. *Information and Inference: A Journal of the IMA*, 9(1):157–193, 2020.

[58] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, and Rachel Cummings. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[59] Seyed Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and A Salman Avestimehr. Fitting relus via sgd and quantized sgd. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2469–2473. IEEE, 2019.

[60] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[61] Gil Kur, Yuval Dagan, and Alexander Rakhlin. Optimality of maximum likelihood for log-concave density estimation and bounded convex regression. *arXiv preprint arXiv:1903.05315*, 2019.

[62] Eunji Lim and Peter W Glynn. Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208, 2012.

[63] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. *arXiv preprint arXiv:1712.06559*, 2017.

[64] CWH Mace and ACC Coolen. Statistical mechanical analysis of the dynamics of learning in perceptrons. *Statistics and Computing*, 8:55–88, 1998.

[65] Alessandro Magnani and Stephen P Boyd. Convex piecewise-linear fitting. *Optimization and Engineering*, 10(1):1–17, 2009.

[66] Fotios D Mandanas and Constantine L Kotropoulos. Robust multidimensional scaling using a maximum correntropy criterion. *IEEE Transactions on Signal Processing*, 65(4):919–932, 2016.

[67] Ali Mohamad-Djafari. *Inverse problems in vision and 3D tomography*. John Wiley & Sons, 2013.

[68] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[69] Jamie Morgenstern and Tim Roughgarden. Learning simple auctions. In *Conference on Learning Theory*, pages 1298–1318. PMLR, 2016.

[70] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

[71] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[72] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems*, 26, 2013.

[73] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017.

[74] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.

[75] Yaniv Plan and Roman Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461, 2014.

[76] F. L. Ramsey and D. W. Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury, 2nd edition, 2002. Data sets: `https://cran.r-project.org/web/packages/Sleuth2/`.

[77] Aviad Rubinstein and S Matthew Weinberg. Simple mechanisms for a subadditive buyer and applications to revenue monotonicity. *ACM Transactions on Economics and Computation (TEAC)*, 6(3-4):1–25, 2018.

[78] Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

[79] Ali Siahkamari, Durmus Alp Emre Acar, Christopher Liao, Kelly L Geyer, Venkatesh Saligrama, and Brian Kulis. Faster algorithms for learning convex functions. In *International Conference on Machine Learning*, pages 20176–20194. PMLR, 2022.

[80] Ali Siahkamari, Venkatesh Saligrama, David Castanon, and Brian Kulis. Learning Bregman divergences. *arXiv preprint arXiv:1905.11545*, 2019.

[81] Ali Siahkamari, Xide Xia, Venkatesh Saligrama, David Castañón, and Brian Kulis. Learning to approximate a bregman divergence. *Advances in Neural Information Processing Systems*, 33:3603–3612, 2020.

[82] Mahdi Soltanolkotabi. Learning relus via gradient descent. *Advances in neural information processing systems*, 30, 2017.

[83] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[84] Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *arXiv preprint arXiv:1910.12837*, 2019.

[85] Yan Shuo Tan and Roman Vershynin. Phase retrieval via randomized kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, 8(1):97–123, 2019.

[86] Alejandro Toriello and Juan Pablo Vielma. Fitting piecewise linear continuous functions. *European Journal of Operational Research*, 219(1):86–95, 2012.

[87] Jan van den Brand. A deterministic linear program solver in current matrix multiplication time. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 259–278. SIAM, 2020.

[88] Aad W van der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer, 1996.

[89] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[90] Gal Vardi, Gilad Yehudai, and Ohad Shamir. Learning a single neuron with bias using gradient descent. *Advances in Neural Information Processing Systems*, 34, 2021.

[91] Hal R Varian. The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, pages 945–973, 1982.

[92] Hal R Varian. The nonparametric approach to production analysis. *Econometrica: Journal of the Econometric Society*, pages 579–597, 1984.

[93] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[94] Irene Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *IEEE Transactions on Information Theory*, 64(5):3301–3312, 2018.

[95] Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[96] Adriaan Walther. The question of phase retrieval in optics. *Optica Acta: International Journal of Optics*, 10(1):41–49, 1963.

[97] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017.

[98] Gang Wang, Georgios B Giannakis, Yousef Saad, and Jie Chen. Phase retrieval via reweighted amplitude flow. *IEEE Transactions on Signal Processing*, 66(11):2818–2833, 2018.

[99] Sinong Wang and Ness Shroff. A new alternating direction method for linear programming. *Advances in Neural Information Processing Systems*, 30, 2017.

[100] Zhi-Yong Wang, Hing Cheung So, and Abdelhak M Zoubir. Robust low-rank matrix recovery via hybrid ordinary-welsch function. *IEEE Transactions on Signal Processing*, 2023.

[101] Daniel S Weller, Ayelet Pnueli, Gilad Divon, Ori Radzyner, Yonina C Eldar, and Jeffrey A Fessler. Undersampled phase retrieval with outliers. *IEEE Transactions on Computational Imaging*, 1(4):247–258, 2015.

[102] Yinyu Ye and Edison Tse. An extension of karmarkar's projective algorithm for convex quadratic programming. *Mathematical programming*, 44:157–179, 1989.

[103] Gilad Yehudai and Shamir Ohad. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pages 3756–3786. PMLR, 2020.

[104] Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Median-truncated nonconvex approach for phase retrieval with outliers. *IEEE Transactions on Information Theory*, 64(11):7287–7310, 2018.

[105] Huishuai Zhang, Yi Zhou, Yingbin Liang, and Yuejie Chi. A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 18, 2017.

[106] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pages 1524–1534. PMLR, 2019.

[107] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International conference on machine learning*, pages 4140–4149. PMLR, 2017.