Beyond Sub-Gaussian and Independent Data in High Dimensional Regression

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Vidyashankar Sivakumar

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Prof. Arindam Banerjee, Advisor

October, 2020

Acknowledgements

My sincere and deepest gratitude goes first and foremost to my advisor Prof. Arindam Banerjee for his mentorship, guidance, support, encouragement and patience throughout my journey as a graduate student. One of the great joys and pleasures of graduate school life for me was the time I spent with him learning, discussing and solving statistical machine learning problems. His dedication to work, infectious enthusiasm, optimism and positive disposition in all situations have been a source of inspiration for me. Apart from his wide and deep knowledge of all things machine learning, I have learnt and benefitted immensely from his clarity of thought, attention to detail and precise communication. I was extremely fortunate to have him as my advisor.

I am grateful to my thesis committee members - Prof. Vipin Kumar, Prof. Jarvis Haupt and Prof. Steven Wu - for their guidance, support, feedback and encouragement. I am also thankful to my terrific collaborators - Pradeep Ravikumar, Steven Wu, Andre Goncalves, Soumyadeep Chatterjee, Puja Das, Farideh Fazayeli, Sheng Chen, Nicholas Johnson, Qilong Gu, Sijie He and Xinyan Li - for the personally enjoyable times learning, discussing and working on research problems.

A big thank you to all my teachers at the University of Minnesota. The wonderful courses on data mining, machine learning, matrix theory, optimization and statistics were the foundational building blocks of my research work. I would also like to express my thanks and appreciation for the staff in the computer science department, graduate school and International Student and Scholar Services for their patience and wonderful service welcoming new graduate students, advising on degree progress, course registrations, visa issues, conference travel etc., and in general making lives for all students much easier.

I would like to thank my mentors at IBM Research - Dr. Jayant Kalagnanam, Dr. Kyongmin Yeo and Dr. Nam Nguyen - for the opportunity, guidance and patience during my two internship stints where I learnt the ropes for working with practical real world datasets.

I also want to thank the people who have positively impacted various stages of my life leading up to graduate studies: Prof. Sutanu Chakraborti for his mentorship, guidance and the opportunity to work on a text analytics project with him at IIT Madras; Prof. Ashish Tendulkar and Prof. Ravindran Balaraman for the wonderful courses on Natural Language Processing and Machine Learning which helped me prepare for my graduate school journey; Gadgil sir for his guidance and teaching of high school mathematics and all my school teachers for their immense contributions to my overall development.

Life in graduate school was enriched by the friendship of my labmates: Amir, Andre, Ankush, Chintan, Farideh, Guruprasad, Hardik, Huahua, Igor, Jared, Karthik, Konstantina, Kshitij, Nicholas, Puja, Qilong, Rahul, Robert, Saurabh, Sheng, Sijie, Soumyadeep, Xi, Xiaoli, Xiaowei, Xinyan, Yingxue and Zhihao. Thank you all for your friendship and support. I will have fond memories of our interactions during the group meetings and many informal meetups in the corridors of the 6th floor in Keller Hall.

Thank you to my friends - Aritra, Abhilash, Anand, David, Harishankar, Rohan, Rupak, Sahil, Santosh, Sathyanarayan, Shivram and Sriram - for sharing in the joys and being supportive during stressful times of my graduate student life. I would also like to take this opportunity to thank Michael and Sara Teachout, my hosts when I first arrived in Minneapolis, and the entire ANCF organization for their love, help and wonderful service welcoming, hosting and guiding new graduate students from foreign countries until they find their feet.

Finally, a big thank you to my family for their unconditional love and support. I am who I am because of my family. Thank you to my mother for being my first teacher, instilling in me the joy of solving mathematical problems and building a strong foundation in my education; my father for inculcating in me the values of hard work, discipline, education and for his advice to pursue graduate studies; my brother for being a role model, supportive and a huge positive influence in my life.

Dedication

To God and my family

Abstract

The past three decades has seen major developments in high-dimensional regression models leading to their successful use in applications from multiple domains including climate science, finance, recommendation systems, computational biology, signal processing to name a few. The underlying assumption in high-dimensional regression models is that the phenomenon under study can be explained by a simple model with few variables. In high-dimensional parametric regression models with parameters existing in high-dimensional space, the simplicity assumption is encoded by a sparsity constraint to be satisfied by the parameter vector. Statistical analysis of high-dimensional regression models delves into the study of the properties of the models, including how faithfully the models recover the assumed true sparse parameter and model sensitivity to different data assumptions.

While major progress has been made over the past several years, non-asymptotic statistical analysis of high-dimensional regression models still makes standard data assumptions of (sub)-Gaussianity and independence which do not hold in some practical applications. For example, datasets in climate and finance are known to have variables with heavier tails than Gaussian or bandit algorithms have data that is sequentially chosen thus violating the independence assumption. The topic of this thesis is the non-asymptotic statistical analysis and study of high-dimensional regression estimators under non-standard data assumptions, including analysis of traditional estimators like regularized least squares as also design of new algorithms to improve estimation performance. Our technical results highlight geometric properties of high-dimensional models and hence all results are expressed in terms of geometric quantities associated with the sparsity structure assumed for the parameter. Much of the analysis borrows tools and techniques from random matrix analysis, probability tools like generic chaining and, in general, probability results for behavior of random variables, vectors in high-dimensional space. We analyze four problems:

- **Regularized least squares with sub-exponential data:** Data in multiple domains like finance, climate science are known to be sub-exponential, which have probability distributions with tails heavier than Gaussians but dominated by a suitably scaled centered exponential distribution. We study non-asymptotic estimation performance of the regularized least squares estimator with i.i.d. sub-exponential data showing that the estimation performance is slightly worse compared to the i.i.d. sub-Gaussian setting.

- **High-dimensional quantile regression:** We study the quantile regression problem in high dimensions which models the conditional quantile of a response given covariates. While least squares regression is ideal to model the conditional mean of a response variable which is symmetric (sub)-Gaussian, there are multiple applications where it is imperative/of interest to model conditional quantiles of the response given covariates to completely understand the behavior of the conditional response, e.g., in meteorology there is more interest in modeling the extremes of climate variables like precipitation rather than the mean. We show that sample complexity for parameter recovery for quantile regression is the same as the least squares loss and the estimation is robust to heavy tailed/outliers in response conforming with traditional wisdom.

- **Unified analysis of robust high-dimensional semi-parametric single index models:** High-dimensional semiparametric single index models are a tradeoff between linear parametric models, which is too restrictive in many applications, and traditional nonparametric regression which cannot be used in high-dimensions due to the curse of dimensionality. We unify the analysis of multiple existing estimators for high-dimensional single-index models, highlight their strengths and weakness and also propose new estimators which simultaneously overcome highlighted negatives of the previously proposed estimators. We also study the non-asymptotic estimation performance of high-dimensional quantile single index models.

- **Smoothed analysis of high-dimensional structured stochastic linear bandits:** We analyze the performance of the greedy algorithm under a 'smoothed' framework for the stochastic linear bandits problem assuming the parameter has sparsity inducing structure. The smoothed setting has adverserial contexts perturbed by independent Gaussian noise. Due to the sequential nature of the algorithm, novel analysis and results are required since the rows of the design matrix are not independent and can have arbitrary correlations. We show there is implicit exploration in the smoothed setting where a simple greedy algorithm works precluding the exploration for exploration strategies for high-dimensional linear bandit problems!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Advances in data collection and storage technologies has led to an explosion in the volume and variety of data collected in various fields of science, technology, business and industry. There is a crucial requirement to leverage the enormous datasets to further scientific analysis, discovery and improve efficiency and effectiveness of business and industrial operations. For example, bioinformatics scientists can use data from cancer patients to identify and further analyze human genes responsible for cancer [32]. Climate scientists are interested, for example, on the relationship between precipitation in North America with other meteorological and climate variables like sea surface temperature and pressure [36]. Online movie recommendation companies endeavour to improve on their recommendations by learning users personal preferences based on their movie ratings. An underlying belief in the analysis of many high-dimensional datasets having thousands to millions of variables is sparsity. The assumption of sparsity implies simpler or parsimonious models where only a few variables can explain the phenomemon under study. For example, we expect only a few of the 40,000 genes to be involved in the process that leads to development of cancer. Similarly, it is logical to assume that precipitation in North America can be predicted using a few of the thousands of climate variables or a users likes and dislikes can be learnt from ratings to a few out of thousands of movies. High-dimensional sparse statistical models seek to exploit sparsity for knowledge inference from datasets having large number of variables.

1.1 High-Dimensional Regression

1.1.1 Linear Regression in High Dimensions

High-dimensional parametric linear regression models and variants are the standard workhorse models which have been successfully applied in many domains like computational biology

[9, 32, 117, 142], climate science [34, 36, 37, 45], ecology [47, 58], astronomy [10, 120], medical imaging [46, 87, 95, 61] to name a few. We will hence use the parametric linear regression models to explain the main ideas behind high-dimensional models.

The parametric linear regression problem can be posed as follows. We are given n samples $\{x_i, y_i\}_{i=1}^n$, where each $x_i = \{x_{i1}, \ldots, x_{ip}\}$ is a p-dimensional vector of features (predictors) and $y_i \in \mathbb{R}$ is the response variable (or predictand). Each sample is assumed to satisfy the following linear relationship,

$$y_i = \sum_{j=1}^{p} x_{ij}\theta_j^* + \omega_i = \langle x_i, \theta^* \rangle + \omega_i \ , \tag{1.1}$$

where θ^* is the unknown parameter and ω_i is some unknown noise observed in the response variable. An estimate for the regression parameter can be obtained using the standard least squares estimator:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmin}} \frac{1}{2n} \|y - X\theta\|_2^2 \\ &= \frac{1}{n}(X^T X)^{-1} X^T y \ , \end{aligned} \tag{1.2}$$

where $X \in \mathbb{R}^{n \times p}$ is the design matrix obtained by stacking the n feature vectors as rows and $y \in R^n$ is the corresponding response vector. While the least squares estimator performs reasonably well in many applications in the high-sample, low-dimension regime when $n >> p$, it suffers from a few drawbacks in the high-dimension regime when p is large. First without the assumption of sparsity, the least squares estimate has low bias and high variance leading to poor prediction accuracy. Infact when $p > n$, as commonly encountered in fields like cancer genomics constrained by scarce cancer patient data, the least squares estimate is not even unique. There are an infinite number of solutions where the least squares loss function evaluates to zero. Second the least squares estimate is generally not sparse and hence suffers from poor interpretability. Often in practical applications, the goal is to identify the subset of variables which have the highest influence on the response variable. Practically it is observed that increasing bias by assuming sparse parameter vector reduces variability hence improving prediction accuracy as also interpretability.

Computationally, the least squares regression with a hard sparsity constraint on the parameter takes the following form:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2n} \|y - X\theta\|_2^2 \quad s.t. \quad \|\theta\|_0 \leq s \ , \tag{1.3}$$

where the assumption is that less than s out of p variables are non-zero. Formulation (1.3) is a non-convex NP-hard combinatorial problem. While earlier literature focused on developing heuristics or use domain knowledge [15, 66, 52, 50] to solve the problem, the non-convexity precluded its widespread adoption.

1.1.2 A Convex Formulation

The constraint $\|\theta\|_0 \leq s$ makes the formulation (1.3) non-convex. In the late 1980's and early 1990's [116, 42, 123], it was proposed to replace the ℓ_0 norm constraint with the ℓ_1 norm constraint:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2n} \|y - X\theta\|_2^2 \quad s.t. \quad \|\theta\|_1 \leq \lambda \,. \tag{1.4}$$

It equivalent to the more widely popular regularized formulation which is called the Lasso:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \,. \tag{1.5}$$

Why use the ℓ_q norm with $q = 1$? It turns out that using $0 < q < 1$ gives a sparse solution but the non-convex formulation makes the problem computationally difficult. On the other hand with $q > 1$, the problem although convex, gives non-sparse solutions. Using $q = 1$ achieves the dual objectives of obtaining sparse solutions with a computationally feasible formulation. In the regularized formulation, the ℓ_1-norm penalty encourages sparse solutions which minimize the squared loss with λ controlling the strength of the penalty. $\lambda = 0$ gives a non-sparse least squares estimate with increasing λ leading to sparser solutions. The Lasso formlulation was the harbinger of huge theoretical advances and widespeard adoption of high-dimensional statistical models in many application domains over the next couple of decades.

1.1.3 Further Advances

Considerable progress has been made over the past two decades on various aspects of high-dimensional regression models. Motivated by practical applications multiple papers focus on replacing the ℓ_1-norm constraint with other constraints which better capture structural assumptions on the parameter. For example, the group sparse regularizers [143, 71] ensure that groups of variables are simultaneously activated or assigned zero values. Examples of other structural prior constraints include the k-support norm [7], OWL norm [19], matrix norms like nuclear norm [28, 114] etc. In parallel, optimization methods like proximal gradient methods [94, 74, 8], forward-backward splitting algorithms [96, 82], alternating direction method of multipliers (ADMM) [22, 132] were proposed for efficient and fast computations. High-dimensional regression techniques were extended for time-series analysis [133, 101], estimation in latent structure models [107] and semi-parametric models [111, 40]. Multiple works focus on particular applications in varied domains like compressed sensing [110, 21], graphical model estimation [99, 113, 140, 141], computational biology [9, 32, 117, 142], climate science [34, 36, 37, 45], ecology [47, 58], astronomy [10, 120], medical imaging [46, 87, 95, 61] to name a few.

Statistical analysis of high-dimensional models, which is the focus of this work, studies properties of the estimator and how faithfully the assumed true parameter is recovered making suitable assumptions on the data. Initial work in high-dimensional statistics established bounds on the ℓ_2-norm estimation error for the Lasso estimator in (1.5). Assuming a linear model, i.e., $y = X\theta^* + \omega$, data in the design matrix, noise to be independent and identically distributed (i.i.d.) Gaussian, the true parameter is s-sparse, i.e., has only s non-zero coefficients the following result was shown to hold with high probability assuming a suitable choice for the regularization parameter λ [105, 11, 17]:

$$\|\hat{\theta} - \theta^*\|_2 \leq O\left(\sqrt{\frac{s \log p}{n}}\right) , \tag{1.6}$$

Contrast this with the upper bound on estimation error of $O\left(\sqrt{\frac{p}{n}}\right)$ for the least squares without the ℓ_1 constraint, even when θ^* is sparse, the advantages of the ℓ_1 constraint is immediately evident especially when p is large. Over the past few years the statistical analysis has been unified and extended to obtain estimation error bounds with different loss functions like generalized linear models [105, 11], different prior structure constraints like group-sparse norms, k-support norm [143, 7] and different assumptions on data like semi-parametric models [111, 40]. But an underlying assumption in most early and recent work is that data is i.i.d. sub-Gaussian. Informally speaking, sub-Gaussian distributions are probability distributions whose tails are dominated by a suitably scaled Gaussian distribution.

The starting point of this work is the observation that the data assumptions made in prior work does not hold in many practical applications. For example, in climate science there are multiple variables like maximum daily temperature where the fitted empirical distribution is observed to have heavier tails than Gaussians [34] or the response variable is bimodal [80]. In online learning and bandit problems [26], the rows of the design matrix are adaptively chosen based on previously observed rows violating the independence aasumption. While high-dimensional techniques have been empirically applied in applications which do not satisfy traditional data assumptions, theoretical advances are required to understand the behavior of the model to make them more efficient as also to draw correct inferences from their results.

1.2 Contributions and Roadmap of Thesis

The focus of the thesis is development and statistical analysis of high-dimensional regression models under more relaxed assumptions on the data. We recap assumptions made in prior literature before highlighting our contributions.

Prior literature predominantly assumes the linear model:

$$y = X\theta^* + \omega \;, \tag{1.7}$$

where $X \in \mathbb{R}^{n \times p}$ is a sub-Gaussian design matrix, $y \in \mathbb{R}^n$ is the observed response variable, $\omega \in \mathbb{R}^n$ is the unknown noise vector assumed to be Gaussian and $\theta^* \in \mathbb{R}^p$ is the unknown parameter vector. The widely analyzed estimator is the regularized least squares estimator:

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n R(\theta) \;, \tag{1.8}$$

where $R(\theta)$ encodes prior structural constraint assumptions like ℓ_1, group-sparse or k-support norm. A related estimator is the following constrained estimator [33, 109]:

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \frac{1}{2n} \|y - x\theta\|_2^2 \quad s.t. \quad R(\theta) \leq R(\theta^*) \;. \tag{1.9}$$

We make the following contributions:

- **Least squares regression with sub-exponential data:** In Chapter 3, we obtain bounds on $\|\hat{\theta} - \theta^*\|_2$ for the regularized least squares formulation when the rows of the design matrix and elements of the noise are drawn i.i.d. from a sub-exponential distribution. Informally speaking, sub-exponential distributions are probability distributions whose tails are dominated by a suitably scaled exponential distribution. The results cover all log-concave and extreme-value distributions which are part of the sub-exponential family of distributions. The results are useful for applications in domains like finance, climate science, ecology, social network analysis etc. which frequently have sub-exponential data.

- **High-dimensional linear quantile regression:** In Chapter 4, we analyze and obtain estimation error bounds for the regularized quantile regression estimator. Quantile regression models the conditional median and quantiles of a response variable given predictor variables. Least squares estimator assumes $E[y_i|x_i] = \langle x_i, \theta^* \rangle$, where $E[\cdot]$ denotes the expectation of a random variable, and is ideal when the noise is homoscedastic (homoscedasticity implies the variance of the noise does not change with covariates x_i), symmetric (sub)-Gaussian and has no outliers. On the other hand linear quantile regression assumes the following linear relationship between the quantiles of the response variable and the covariates: $F^{-1}_{y_i|x_i}(\tau|x_i) = \langle x_i, \theta^*_\tau \rangle$, $\tau \in (0,1)$, where $F^{-1}_{y_i|x_i}$ is the inverse of the conditional distribution functions of the response given the covariates. Assuming different parameters for each quantile, we can model the complete conditional distribution of the response given the covariates. We characterize the sample complexity for consistent recovery and give non-asymptotic bounds on the estimation error and, conforming to conventional wisdom, show quantile regression is robust to noise with heavy-tails and/or outliers. Unlike previous treatments of this topic, our analysis and results are geometric in nature and holds for the class of atomic norms.

- **Robust high-dimensional single index models:** In Chapter 5, we consider the problem of parameter estimation in high-dimensional (conditional) mean and quantile single index models (SIMs) when the data is assumed to be generated by a semi-parametric model $y_i = f(\langle x_i, \theta^* \rangle) + \omega_i$ for some unknown transfer function $f(\cdot)$ and under general structural assumptions on the parameter. High-dimensional semiparametric single index models are a tradeoff between linear parametric models, which is too restrictive in many applications, and traditional non-parametric regression which cannot be used in high-dimensions due to the curse of dimensionality. Although multiple estimators for mean SIMs have been proposed in the past, limited technical comparisons among the estimators exist. In this chapter, we present a unified analysis of the different estimators for mean SIMs and their performance under different assumptions on the properties of the data and the nonlinear function. We also propose a new robust estimator for mean SIMs using ideas from sliced inverse regression, a previously proposed algorithm for parameter estimation in mean SIM. Unlike past estimators, the proposed estimator is robust to heavy-tailed noise and outliers in the response and can efficiently estimate the parameter even when the function is non-monotonic. For quantile SIMs, we show that the linear quantile regression estimator gives a consistent estimate of the true parameter and obtain non-asymptotic estimation error bounds. The theoretical results are supported by experiments on synthetic data.

- **High-dimensional linear contextual bandits:** Bandit learning algorithms typically involve the balance of exploration and exploitation. However, in many practical applications, worst-case scenarios needing systematic exploration are seldom encountered. In Chapter 6, we consider a smoothed setting for structured linear contextual bandits where the adversarial contexts are perturbed by Gaussian noise and the unknown parameter θ^* has structure, e.g., sparsity, group sparsity, low rank, etc. We propose simple greedy algorithms for both the single- and multi-parameter (i.e., different parameter for each context) settings and provide a unified regret analysis for θ^* with any assumed structure. The regret bounds are expressed in terms of geometric quantities such as Gaussian widths associated with the structure of θ^*. We also obtain sharper regret bounds compared to earlier work for the unstructured θ^* setting as a consequence of our improved analysis. We show there is implicit exploration in the smoothed setting where a simple greedy algorithm works precluding the exploration for exploration strategies!

Much of the analysis borrows tools and techniques from random matrix analysis, probability tools like generic chaining and, in general, probability results for behavior of random variables, vectors in high-dimensional space. In Chapter 2, we briefly review relevant prior literature in high-dimensional statistics and probability theory.

Chapter 2

Background and Preliminaries

In this chapter we review background material on non-asymptotic analysis of high-dimensional statistical models.

2.1 Random Variables, Vectors and Concentration Inequalities

We briefly review definitions and properties of random variables and vectors. We borrow material from [127] which is a more thorough and easily accessible exposition of the below material.

2.1.1 Sub-Gaussian Random Variables

We define and state properties of sub-Gaussian random variables.

Definition 1 (Sub-Gaussian random variables) *A random variable x is sub-Gaussian if it satisfies any of the following properties for positive constants $\kappa_1, \kappa_2, \kappa_3$,*

1. *Tails:* $P(|x| > t) \leq \exp(1 - t^2/\kappa_1^2), \quad t \geq 0$*;*
2. *Moments:* $(E|x|^p)^{1/p} \leq \kappa_2 \sqrt{p}, \quad \forall p \geq 1$*;*
3. *Super-exponential moment:* $E \exp(x^2/\kappa_3^2) \leq e$*.*

Moreover the sub-Gaussian norm of the random variable, denoted as $\|x\|_{\psi_2}$, is the smallest κ_2 such that,

$$\|x\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (E|x|^p)^{1/p} \ . \tag{2.1}$$

The tail decay, moment growth and growth of moment generating function in the definition are equivalent with each implying the others with the constants $\kappa_1, \kappa_2, \kappa_3$ differing from each other by at most an absolute constant factor. The zero mean, σ^2-variance Gaussian distribution $N(0, \sigma^2)$ is a sub-Gaussian distribution with sub-Gaussian norm $c\sigma$ for some constant c.

We characterize large deviation properties of sums of sub-Gaussian random variables below.

Lemma 1 (Hoeffding-type inequality) *Let $x_1, \dots, x_n$ be independent centered sub-Gaussian random variables. Let $\kappa = \max\limits_{1 \le i \le n} \|x_i\|_{\psi_2}$. Then for any $a \in \mathbb{R}^n$ and $t \geq 0$, we have,*

$$P\left(\left|\sum_{i=1}^{n} a_i x_i\right| \geq t\right) \leq e \cdot \exp\left(-\frac{ct^2}{\kappa^2 \|a\|_2^2}\right) \ , \tag{2.2}$$

where $c > 0$ is an absolute constant.

Sub-Gaussian random variables are rotationally invariant.

Lemma 2 (Rotation invariance) *Consider a finite number of independent centered sub-gaussian random variables x_i. Then $\sum_i x_i$ is also a centered sub-gaussian random variable. Moreover,*

$$\| \sum_i x_i \|_{\psi_2}^2 \leq c \sum_i \|x_i\|_{\psi_2}^2 \tag{2.3}$$

2.1.2 Sub-Exponential Random Variables

We define and state properties of sub-exponential distributions.

Definition 2 (Sub-exponential random variables) *A random variable x is sub-exponential if it satisfies any of the following properties for positive constants $\kappa_1, \kappa_2, \kappa_3$,*

1. *Tails: $P(|x| > t) \leq \exp(1 - t/\kappa_1), \quad t \geq 0$;*
2. *Moments: $(E|x|^p)^{1/p} \leq \kappa_2 p, \quad \forall p \geq 1$;*
3. *Super-exponential moment: $E \exp(x/\kappa_3) \leq e$.*

Moreover the sub-exponential norm of the random variable, denoted as $\|x\|_{\psi_1}$, is the smallest κ_2 such that,

$$\|x\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (E|x|)^{1/p} \ . \tag{2.4}$$

The tail-decay, moment growth and growth of moment generating function for sub-exponential random variables are equivalent with each implying the other two and the constants $\kappa_1, \kappa_2, \kappa_3$ differ from each other by at most an absolute constant factor.

The square of sub-Gaussian random variables is sub-Gaussian.

Lemma 2.1.1 (Sub-exponential is sub-Gaussian squared) *A random variable x is sub-Gaussian if and only if x^2 is sub-exponential. Moreover,*

$$\|x\|_{\psi_2}^2 \leq \|x^2\|_{\psi_1} \leq 2\|x\|_{\psi_2}^2 \ . \tag{2.5}$$

The large deviation bounds on sums of sub-exponential random variables is characterized by a Bernstein-type inequality.

Lemma 3 (Bernstein-type inequality) *Let $x_1, \ldots, x_n$ be independent centered sub-exponential random variables. Let $\kappa = \max_{1 \leq i \leq n} \|x_i\|_{\psi_1}$. Then for any $a \in \mathbb{R}^n$ and $t \geq 0$, we have,*

$$P\left(\left| \sum_{i=1}^{n} a_i x_i \right| \geq t \right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{\kappa^2 \|a\|_2^2}, \frac{t}{\kappa \|a\|_\infty} \right) \right) \ , \tag{2.6}$$

where $c > 0$ is an absolute constant.

2.1.3 Random Vectors

We will work with random vectors $x \in \mathbb{R}^p$ which are samples from a probability distribution in $\mathbb{R}^p$.

Definition 3 *(Isotropic random vectors) A random vector $x \in \mathbb{R}^p$ is isotropic if $\Sigma = E[xx^T] = \mathbb{I}_{p \times p}$. Equivalently, $E[\langle x, u \rangle^2] = \|u\|_2^2$ for any $u \in \mathbb{R}^p$.*

An example of an isotropic random vector is the p-dimensional Gaussian random vector $x \sim N(0, \sigma^2 \mathbb{I}_{p \times p})$. Let $\Sigma = E[xx^T]$ be an invertible matrix, which is true if the probability distribution, from which x is sampled, is not supported in any proper subspace of $\mathbb{R}^p$. Then $\Sigma^{-1/2} x$ is an isotropic random vector.

Definition 4 (Sub-Gaussian random vectors) *A random vector $x \in \mathbb{R}^p$ is sub-Gaussian if the one-dimensional marginals $\langle x, u \rangle$ are sub-Gaussian random variables for all $u \in \mathbb{R}^p$. The sub-Gaussian norm of x is defined as,*

$$\|x\|_{\psi_2} = \sup_{u \in S^{p-1}} \|\langle x, u \rangle\|_{\psi_2} \ . \tag{2.7}$$

Sub-exponential random vectors can be defined similarly.

Definition 5 (Sub-exponential random vectors) *A random vector $x \in \mathbb{R}^p$ is sub-exponential if the one-dimensional marginals $\langle x, u \rangle$ are sub-exponential random variables for all $u \in \mathbb{R}^p$. The sub-exponential norm of x is defined as,*

$$\|x\|_{\psi_1} = \sup_{u \in S^{p-1}} \|\langle x, u \rangle\|_{\psi_1} . \tag{2.8}$$

A random vector with sub-Gaussian elements is a sub-Gaussian random vector.

Lemma 4 (Product of sub-Gaussian distributions) *Let $x_1, \ldots, x_p$ be independent centered sub-gaussian random variables. Then $x = (x_1, \ldots, x_p)$ is a centred sub-gaussian random vector in $\mathbb{R}^p$, and*

$$\|x\|_{\psi_2} \leq c \max_{i \leq p} \|x_i\|_{\psi_2} \tag{2.9}$$

where c is an absolute constant.

Projections of sub-Gaussian random vectors in any direction is a sub-Gaussian random variable.

Lemma 5 *Consider a sub-Gaussian random vector $x \in \mathbb{R}^p$ with sub-Gaussian norm $\kappa = \max_i \|x_i\|_{\psi_2}$, then, $z = \langle x, a \rangle$ is a sub-Gaussian random variable with sub-Gaussian norm $\|z\|_{\psi_2} \leq c\kappa \|a\|_2$ for some absolute constant c.*

2.2 Gaussian and Exponential Widths

Informally speaking, widths of sets [121, 122] can be seen as measures for the complexity of sets. The non-asymptotic estimation error bounds for the estimators we consider will be expressed in terms of the Gaussian/exponential widths of sets related to the norm $R(\cdot)$. For example, the Gaussian width of the unit norm ball $\Omega_R = \{u \in \mathbb{R}^p \mid R(u) \leq 1\}$ is a common term which shows up in all results. We provide informal definitions for the width of sets and state a few properties useful in analysis of high-dimensional estimators. While we will only describe aspects relevant to this work, widths and the associated tools like generic chaining are deep topics to which entire books have been devoted [121, 122].

Definition 6 (Gaussian Width) *Consider any set $T \subseteq \mathbb{R}^p$. Let $\{X_t\}_{t \in T} = \langle g, t \rangle$ be a stochastic process indexed by the set T, where each element $g_i \sim N(0, 1),\ 1 \leq i \leq p$ is i.i.d zero mean variance one Gaussian. The quantity $w_g(T) = E_g \left[\sup_t X_t\right]$ is called the Gaussian width of the set T.*

More generally any stochastic process $\{X_t\}_{t\in T}$ indexed by the set $T \subseteq \mathbb{R}^p$ satisfying the following Hoeffding-type increment condition for some constant κ,

$$\forall u > 0,\; P(|X_s - X_t| \geq u) \leq 2\exp\left(-\frac{u^2}{\kappa^2\|s-t\|_2^2}\right) , \tag{2.10}$$

satisfies the following for some constant c which depends on κ due to the majorizing measures theorem,

$$E\left[\sup_{t\in T} X_t\right] \leq c\cdot w_g(T) . \tag{2.11}$$

The *exponential width* of a set can be similarly defined (although the term exponential width is not explicitly used in prior literature).

Definition 7 (Exponential width) *Consider any set $T \subseteq \mathbb{R}^p$. Let $\{X_t\}_{t\in T} = \langle e, t\rangle$, be a stochastic process indexed by the set T where each element $e_i,\ 1 \leq i \leq p$ is a centered i.i.d exponential random variable satisfying the exponential tail decay $P(|e_i| \geq u) \leq 2\exp(-u),\ u \geq 0$. The quantity $w_e(T) = E_e\left[\sup_t X_t\right]$ is called the exponential width of the set T.*

Again, more generally any stochastic process $\{X_t\}_{t\in T}$ indexed by the set $T \subseteq \mathbb{R}^p$ satisfying the following Bernstein-type increment condition for some constant κ,

$$\forall u > 0,\; P(|X_s - X_t| \geq u) \leq 2\exp\left(-c\left(\frac{u^2}{\kappa^2\|s-t\|_2^2}, \frac{u}{\kappa\|s-t\|_\infty}\right)\right) . \tag{2.12}$$

satisfies the following for some constant c which depends on κ,

$$E\left[\sup_{t\in T} X_t\right] \leq c\cdot w_e(T) . \tag{2.13}$$

Below we state a couple of useful properties of widths of sets.

Proposition 2.2.1 (Properties of width) *Let $w(\cdot)$ denote the Gaussian or exponential width. Consider set $A \subseteq \mathbb{R}^p$.*

1. *Widths are invariant under orthogonal and linear transformations, i.e., for some unitary matrix $Q \in \mathbb{R}^{p\times p}$ and vector $b \in \mathbb{R}^p$*

$$w(A) = w(QA) \quad w(A+b) = w(A) , \tag{2.14}$$

where $QA = \{Qu \mid u \in A\}$ and $A + b = \{u + b \mid u \in A\}$.

2. *Width is invariant under taking the convex hull.*

$$w(\mathit{conv}(A)) = w(A) . \tag{2.15}$$

2.3 Loss Functions

Consider the following linear regression model,

$$E[y|x] = \langle x, \theta^* \rangle \ . \tag{2.16}$$

Given n samples $(x_i, y_i), \ 1 \leq i \leq n$, the goal is to estimate θ^* in formulation (2.16). Taking a step back, consider the problem of finding the mean of n i.i.d. random univariate variables $z_1, \ldots, z_n$ sampled from some distribution. A good estimate is the following empirical mean,

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i \ . \tag{2.17}$$

Posed as an optimization problem $\bar{z}$ is the solution of the following,

$$\bar{z} = \underset{z}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (z_i - z)^2 \ . \tag{2.18}$$

This principle can be extended to the estimation of θ^* for the regression model (2.16),

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle x_i, \theta^* \rangle)^2 \ , \tag{2.19}$$

which is the least squares estimator. Under certain conditions on the distribution of $y|x$, (2.19) is also the maximum likelihood estimator.

Now consider the quantile regression model,

$$F^{-1}_{y_i|x_i}(\tau | x_i) = \langle x_i, \theta^*_\tau \rangle, \ \tau \in (0, 1) \ , \tag{2.20}$$

where $F^{-1}_{y_i|x_i}(\cdot)$ is the inverse of the conditional distribution function of y_i given x_i and τ is the quantile. Consider the computation of the quantiles of n i.i.d. random univariate variables $z_1, \ldots, z_n$ sampled from some distribution. The τth quantile can be shown to be the solution of the following optimization problem. Define $\rho_\tau(u) = (\tau - \mathbb{I}(u \leq 0))u$ the asymmetric absolute deviation function [80],

$$\bar{z}_\tau = \underset{z}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(z_i - z) \ . \tag{2.21}$$

Note that when $\tau = 0.5$ this reduces to the median regression problem. Again extending this principle to the regression model, given n samples $(x_i, y_i), \ 1 \leq i \leq n$ the following is an estimator for θ^*:

$$\hat{\theta}_\tau = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \langle x_i, \theta^* \rangle) \ . \tag{2.22}$$

Table 2.1: Properties of atomic norms

Norm	Atomic set	Number of subspaces **m**	Dimension of each subspace **q**
ℓ_1	$\mathcal{A} = \{\pm e_i\}_{i=1}^p$	p	1
ℓ_1/ℓ_2 non-overlapping group sparse	$\mathcal{A}_i = \{a \in \mathbb{R}^p \mid \|a_{\mathcal{G}_i}\|_2 = 1, a_{\mathcal{G}_i}^c = 0\}$ $\mathcal{A} = \cup_{1 \leq i \leq N_{\mathcal{G}}} \mathcal{A}_i$	$N_{\mathcal{G}}$	l
k-support norm	$\mathcal{A} = \{a \in \mathbb{R}^p \mid \|a\|_0 \leq k, \|a\|_2 = 1\}$	$\binom{p}{k} \leq \left(\frac{ep}{k}\right)^k$	k

2.4 Atomic Norms

We will consider the class of atomic norms for the regularizer. Consider a set $\mathcal{A} \subseteq \mathbb{R}^p$ which is a collection of atoms that is compact, centrally symmetric about the origin (that is, $a \in \mathcal{A} \implies -a \in \mathcal{A}$). Let $\|\theta\|_{\mathcal{A}}$ denote the gauge of $\mathcal{A}$. Then the atomic norm regularizer is defined as follows,

$$R(\theta) = \|\theta\|_{\mathcal{A}} = \inf\{t > 0 : \theta \in t \operatorname{conv}(\mathcal{A})\} \tag{2.23}$$

$$= \inf \left\{ \sum_{a \in \mathcal{A}} c_a \ : \ \theta = \sum_{a \in \mathcal{A}} c_a a, \ c_a \geq 0, \ \forall a \in \mathcal{A} \right\} . \tag{2.24}$$

For example when $\mathcal{A} = \{\pm e_i\}_{i=1}^p$ yields $R(\theta) = \|\theta\|_{\mathcal{A}} = \|\theta\|_1$.

Although the atomic set $\mathcal{A}$ may contain uncountably many elements, for many popular vector norms $\mathcal{A}$ can be expressed as a union of q-dimensional subspaces, $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots \cup \mathcal{A}_m$ [38]. We will consider a few such regularizers as examples throughout the thesis.

1. ℓ_1 **norm:** The atomic set for the ℓ_1 norm is $\mathcal{A} = \{\pm e_i\}_{i=1}^p$, where e_i are the standard basis vectors in p dimensional space. Therefore the atomic set is a union of p 1-dimensional subspaces. Throughout we consider the true parameter θ^* is s-sparse.

2. ℓ_1/ℓ_2 **non-overlapping group sparse norm:** Let $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_{N_{\mathcal{G}}}\}$ denote a collection of non-overlapping sets partitioning the set of natural numbers in $1 \leq i \leq p$, that is using popular set notations, $\mathcal{G}_i \subseteq \{1, \ldots, p\}, \mathcal{G}_i \cap \mathcal{G}_j = \phi \ \forall i, j \in \{1, \ldots, N_{\mathcal{G}}\}, \cup_{1 \leq i \leq N_{\mathcal{G}}} \mathcal{G}_i =$

$\{1, \ldots, p\}$. For any vector $\theta \in \mathbb{R}^p$ define the following,

$$\theta_{\mathcal{G}_i} = \{u \in \mathbb{R}^p \mid u_j = \theta_j \text{ if } j \in \mathcal{G}_i \text{ else } u_j = 0\} \ . \tag{2.25}$$

and let $\theta^c_{\mathcal{G}_i}$ denote its complement, that is,

$$\theta^c_{\mathcal{G}_i} = \{u \in \mathbb{R}^p \mid u_j = \theta_j \text{ if } j \notin \mathcal{G}_i \text{ else } u_j = 0\} \ . \tag{2.26}$$

Any vector θ can be expressed as follows,

$$\theta = \sum_{i=1}^{N_{\mathcal{G}}} \|\theta_{\mathcal{G}_i}\|_2 \cdot \frac{\theta_{\mathcal{G}_i}}{\|\theta_{\mathcal{G}_i}\|_2} \ , \tag{2.27}$$

which has an atomic norm representation. Denote $\mathcal{A}_i = \{a \in \mathbb{R}^p \mid \|a_{\mathcal{G}_i}\|_2 = 1, a^c_{\mathcal{G}_i} = 0\}$ and the atomic set is $\mathcal{A} = \cup_{1 \le i \le N_{\mathcal{G}}} \mathcal{A}_i$. The group-sparse norm can thus be defined as,

$$R(\theta) = \sum_{i=1}^{N_{\mathcal{G}}} \|\theta_{\mathcal{G}_i}\|_2 \ . \tag{2.28}$$

Throughout we will assume that the true parameter θ^* only has $S_{\mathcal{G}} \subseteq \{\mathcal{G}_1, \ldots, \mathcal{G}_{N_{\mathcal{G}}}\}, |S_{\mathcal{G}}| = s_{\mathcal{G}}$ non-zero groups. Also without loss of generality we will assume that the size of any group is less than l. Therefore the atomic set is a union of $N_{\mathcal{G}}$ l-dimensional subspaces.

3. k-support norm: The k-support norm is an infimum convolution norm [7].

$$R(\theta) = \|\theta\|_k^{supp} = \inf_{\sum_i u_i = \theta} \left\{ \sum_i \|u_i\|_2 \,\middle|\, \|u\|_0 \le k \right\} \ . \tag{2.29}$$

The atomic set for the k-support norm can be defined as,

$$\mathcal{A} = \{a \in \mathbb{R}^p \mid \|a\|_0 \le k, \|a\|_2 = 1\} \ . \tag{2.30}$$

Therefore the atomic set is a union of $\binom{p}{k}$ k-dimensional subspaces. Throughout we will assume that the true parameter θ^* is s-sparse.

The above norms will be used as examples throughout the thesis. For convenience we list the required properties in Table 2.1. We also list the Gaussian widths and norm compatibility constants in Table 2.2. The norm compatibility constant comes up often in our analysis and is defined as $\Psi(A) = \sup\limits_{u \in A} \frac{R(u)}{\|u\|_2}$, where A is the error set which we define subsequently in the next section.

Table 2.2: Gaussian widths and norm compatibility constant of atomic norms

Norm	Gaussian width of unit norm ball $\boldsymbol{w(\mathcal{A})}$	Gaussian width of error set $\boldsymbol{w(A)}$	Norm compatibility constant $\boldsymbol{\Psi(A)}$
ℓ_1	$\Theta(\sqrt{\log p})$	$\Theta(\sqrt{s \log p})$	$O(\sqrt{s})$
ℓ_1/ℓ_2 non-overlapping group sparse	$\Theta(\sqrt{l + \log N_{\mathcal{G}}})$	$\Theta(\sqrt{l s_{\mathcal{G}} + s_{\mathcal{G}} \log N_{\mathcal{G}}})$	$O(\sqrt{s_{\mathcal{G}}})$
k-support norm	$\Theta(\sqrt{k + k \log \lceil \frac{p}{k} \rceil})$	$\Theta(\sqrt{s + s \log \lceil \frac{p}{k} \rceil})$	$O(\sqrt{2s/k})$

2.5 Estimation in High-Dimensional Regression

Considerable progress has been made over the past decade on high-dimensional structured estimation using suitable M-estimators or norm-regularized regression [105, 11] of the form:

$$\hat{\theta}_{\lambda_n} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\, \mathcal{L}(\theta; X, y) + \lambda_n R(\theta) \;, \tag{2.31}$$

where $\mathcal{L}(\theta; X, y)$ is an empirical loss function, $R(\theta)$ is a suitable norm, and $\lambda_n > 0$ is the regularization parameter. Early work focused on high-dimensional estimation of sparse vectors using the Lasso and related estimators, where $R(\theta) = \|\theta\|_1$ and $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$ [100, 131, 144]. Sample complexity of such estimators have been rigorously established based on the RIP(restricted isometry property) [29, 30] and the more general RE(restricted eigenvalue) conditions [17, 105, 11]. Several subsequent advances have considered structures beyond ℓ_1, using more general norms such as (overlapping) group sparse norms, k-support norm, nuclear norm, and so on [105, 35, 33]. In recent years, much of the literature has been unified and nonasymptotic estimation error bound analysis techniques have been developed for regularized estimation with any norm [11]. Below we describe the major aspects of the analysis framework in [11, 33, 105].

1. Regularization parameter: In [11, 105] the regularization parameter is assumed to be greater than the dual norm of the gradient of the loss function evaluated at the true parameter,

$$\lambda_n \geq 2R^*(\nabla \mathcal{L}(\theta^*)) \;. \tag{2.32}$$

With $\Omega_R = \{u | R(u) \leq 1\}$ denoting the unit norm ball, [11] prove that with high probability a value $\lambda_n = O(w_g(\Omega_R))$ satisfies the above condition for sub-Gaussian design matrices, noise and the least squares loss.

2. Error set: The assumption on the regularization condition ensures that the error vector $\Delta = \hat{\theta} - \theta^*$ lies in the following error set [11],

$$E_r := \left\{ \Delta \mid R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{2} R(\Delta) \right\} . \tag{2.33}$$

3. The norm compatibility constant: It is defined as [105, 11],

$$\Psi(E_r) = \sup_{u \in E_r} \frac{R(u)}{\|u\|_2} . \tag{2.34}$$

4. Restricted Strong Convexity (RSC): In [11, 105] the loss function is shown to satisfy the following RSC condition with high probability once the number of samples is of the order of the square of the Gaussian width of the error set A, that is, $n = O(w^2(A))$, where $A = \text{cone} \cap S^{p-1}$.

$$\begin{aligned} \inf_{u \in E_r} \delta \mathcal{L}(\theta^*, u) &= \inf_{u \in E_r} \left(\mathcal{L}(\theta^* + u) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), u \rangle \right) \\ &\geq \kappa \|u\|^2 . \end{aligned} \tag{2.35}$$

For the squared loss, the RSC condition simplifies to the Restricted Eigenvalue (RE) condition [17]

$$\inf_{u \in E_r} \frac{1}{n} \|Xu\|_2^2 \geq \kappa \|u\|_2^2 . \tag{2.36}$$

5. Recovery Bounds: When RSC and bounds on the regularization parameter are satisfied [11] prove the following deterministic error bound,

$$\|\Delta\|_2 = \|\hat{\theta} - \theta^*\|_2 \leq c \frac{\Psi(E_r) w(\Omega_R)}{\kappa} . \tag{2.37}$$

where c is any constant, $\Omega_R = \{u \ : \ R(u) \leq 1\}$ is the unit norm ball.

Chapter 3

Beyond Sub-Gaussian Measurements: High- Dimensional Structured Estimation with Sub-Exponential Designs

3.1 Introduction

Considerable progress has been made over the past decade on high-dimensional structured estimation using suitable M-estimators or norm-regularized regression [105, 11] of the form:

$$\hat{\theta}_{\lambda_n} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n R(\theta) \ , \tag{3.1}$$

where $R(\theta)$ is a suitable norm, and $\lambda_n > 0$ is the regularization parameter. Most prior work assumes data is generated according to the linear model $y = X\theta^* + \omega$, where the design matrix $X \in \mathbb{R}^{n \times p}$ has i.i.d. sub-Gaussian rows and the noise $\omega \in \mathbb{R}^n$ has i.i.d. sub-Gaussian elements. Informally, sub-Gaussian distribution is a probability distribution whose tails are dominated by a suitably scaled Gaussian distribution. Early work focused on high-dimensional estimation of sparse vectors using the Lasso and related estimators, where $R(\theta) = \|\theta\|_1$ [100, 131, 144]. Sample complexity of such estimators have been rigorously established based on the RIP (restricted isometry property) [29, 30] and the more general RE (restricted eigenvalue) conditions [17, 105, 11]. Recent papers establish sample complexity to satisfy RE and RIP for various random matrices like time frequency structured [53], partial circulant matrices [53, 49] Toeplitz [72] or general anisotropic measurements [115]. More recent work has established the gap between RE and RIP [48, 85] showing examples of random matrices which satisfy RE but not RIP. Several subsequent advances have considered structures beyond ℓ_1, using more general norms such as (overlapping) group sparse norms, k-support norm [7], nuclear norm, and so on [105, 35, 33]. In recent years, much of the literature has been unified and nonasymptotic estimation error bound analysis techniques have been developed for regularized estimation with any norm [11].

In spite of such advances, most of the existing literature relies on the assumption that the design matrix $X \in \mathbb{R}^{n \times p}$ has i.i.d. sub-Gaussian rows. In particular, recent unified treatments based on decomposable norms, atomic norms, or general norms all rely on concentration properties of sub-Gaussian distributions [105, 33, 11]. Certain estimators, such as the Dantzig selector and variants, consider a constrained problem rather than a regularized problem as in (3.1) but the analysis again relies on entries of X being sub-Gaussian [31, 35]. For the setting of constrained estimation, building on prior work by [81], [124] outlines a possible strategy for such analysis which can work for any distribution, but works out details only for the sub-Gaussian case. In recent work [68] considered sub-Gaussian design matrices but with heavy-tailed noise, and suggested modifying the estimator in (1) via a median-of-means type estimator based on multiple estimates of $\hat{\theta}$ from sub-samples. For the special case of the ℓ_1 norm [85, 3] show results for the RE condition for design matrices with heavier tails than sub-Gaussians.

In this chapter, we analyze the norm-regularized estimation problem as in (3.1) for any atomic norm [33] under the assumption that data is sub-exponential. Informally speaking, sub-exponential distributions are probability distributions whose tails are dominated by a suitably scaled exponential distribution. The motivation behind our work are applications of high-dimensional regression in machine learning and statistics where, unlike many problems in compressed sensing, the design matrix cannot be chosen up front but gets determined by the problem. Variables with heavier than sub-Gaussian tails are frequently encountered in many practical application domains like finance, climate science, ecology, social network analysis, etc. [34, 56, 45]. While high dimensional statistical techniques have been used in practice for such applications, theoretical guarantees on their performance has not been analyzed. Our work, to the best of our knowledge, is the first non-asymptotic analysis of regularized high-dimensional estimation problems of the form (3.1) with sub-exponential design matrices, noise and atomic norms.

In our main result, we obtain non-asymptotic bounds on the estimation error $\|\hat{\Delta}_{\lambda_n}\|_2 = \|\hat{\theta}_{\lambda_n} - \theta^*\|_2$, where θ^* is the optimal structured parameter. The analysis has three major components: 1. A characterization of the error set; 2. Characterization of the regularization parameter λ_n and 3. A characterization of the RE condition on the design matrix. We breifly summarize our results comparing them with the results for sub-Gaussian data from prior literature.

Error Set: Let the estimated parameter $\hat{\theta}_{\lambda_n} = \theta^* + \Delta_{\lambda_n}$. If the regularization parameter is sufficiently large (which we characterize subsequently), the error vector Δ_{λ_n} belongs to the following error set $E_r \subseteq \mathbb{R}^p$ [11],

$$E_r = \left\{ \Delta \in \mathbb{R}^p \;\middle|\; R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{2} R(\Delta) \right\} . \tag{3.2}$$

It is straightforward to observe that for the equivalent constrained formulation [33, 109],

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|y - X\theta\|_2^2 \;\text{ s.t. }\; R(\theta) \leq R(\theta^*) , \tag{3.3}$$

the error vector belongs to the following set,

$$E_c = \{\Delta \in \mathbb{R}^p \mid R(\theta^* + \Delta) \leq R(\theta^*)\} . \tag{3.4}$$

Prior results [33, 11] characterize the sample complexity for recovery (minimum number of samples required to satisfy the RE condition defined below) in terms of the Gaussian widths of $A_r = \text{cone}(E_r) \cap S^{p-1}$ and $A_c = \text{cone}(E_c) \cap S^{p-1}$ denoted as $w_g(A_r)$ and $w_g(A_c)$ respectively. The Gaussian width can be viewed as a measure for the size of a set (see Section 3.2 and 3.3 for a formal definition). It is evident from the definition of E_r and E_c that $E_r \subset E_c$ and consequently $w_g(A_r) > w_c(A_c)$. [11] establish an upper bound of the form $w_g(A_r) \leq c \cdot w_g(A_c)$ where c is a constant, thus showing that the Gaussian widths only differ by a multiplicative factor.

Regularization parameter: The error set characterization holds true only if the regularization parameter satisfies the following assumption [105, 11],

$$\lambda_n \geq \beta R^*(\nabla \mathcal{L}(\theta^*; X, y)) = \beta R^* \left(\frac{1}{n} X^T \omega \right) , \tag{3.5}$$

where $\nabla \mathcal{L}(\theta^*; X, y)$ denotes the gradient of the loss function evaluated at θ^* which for the least squares loss function simplifies to $\frac{1}{n} X^T \omega$ with $\omega \in \mathbb{R}^n$ denoting the noise vector. Note that with the assumption that X, ω are random, we want to establish high probability lower bounds on λ_n. For random sub-Gaussian designs and noise, [11] show that the following relationship holds with certain high probability for some constant c_1,

$$R^* \left(\frac{1}{n} X^T y \right) \leq c_1 \cdot \frac{w_g(\Omega_R)}{\sqrt{n}} , \tag{3.6}$$

where $\Omega_R = \{u \mid R(u) \leq 1\}$ is the unit norm ball and, as previously noted, $w_g(\cdot)$ denotes the Gaussian width of a set. For example, when $R(\cdot)$ is the ℓ_1 norm, $w_g(\Omega_R) \leq c_2 \cdot \sqrt{\log p}$ for some constant c_2. Similar deterministic bounds upto constants on Gaussian widths of unit norm balls for various atomic norms like group sparse norms, k-support norms have been derived in prior literature [38]. When X, ω are sub-exponential, we obtain the following result for some constant c_3,

$$R^*\left(\frac{1}{n}X^T y\right) = c_3 \cdot \frac{w_e(\Omega_R)}{\sqrt{n}} \;, \tag{3.7}$$

where $w_e(\cdot)$ denotes the *exponential* width of a set (defined formally in Section 3.3). The exponential width, similar to the Gaussian width, is another measure for the size of a set. A natural question that arises is if the exponential width for various sets can be computed similar to Gaussian widths. One of the contributions of our work is to establish an upper bound for the exponential width in terms of the Gaussian width. We derive the following result in Section 3.3 using results from generic chaining [121, 122] for any set $T \subseteq \mathbb{R}^p$,

$$w_e(T) \leq c_4 \cdot w_g(T)\sqrt{\log p} \;, \tag{3.8}$$

where c_4 is a constant. Therefore using known results for the Gaussian width of sets, we can obtain bounds on $R^*\left(\frac{1}{n}X^T y\right)$ which hold with high probability when the design and noise are sub-exponential. For example, for the ℓ_1 norm, we get $w_e(\Omega_R) \leq c_5 \cdot \log p$ and hence $R^*\left(\frac{1}{n}X^T y\right) \leq c_6 \cdot \frac{\log p}{\sqrt{n}}$

Restricted eigenvalue (RE) condition: When the loss function is squared loss, the design matrix is required to satisfy the following lower bound on the minimum eigenvalue in the restricted error set $A = \text{cone}(E_r) \cap S^{p-1}$ [17, 105, 11],

$$\inf_{u \in A = \text{cone}(E_r) \cap S^{p-1}} \frac{1}{n}\|Xu\|_2^2 \geq \kappa \;, \tag{3.9}$$

where E_r is the error set and κ is a positive constant. Assuming the design matrix to be isotropic sub-Gaussian, [11, 33] obtain the following two-sided bounds which hold with high probability for some constant c,

$$1 - c \cdot \frac{w_g(A)}{\sqrt{n}} \leq \inf_{u \in A} \frac{1}{n}\|Xu\|_2^2 \leq \sup_{u \in A} \|Xu\|_2^2 \leq 1 + c \cdot \frac{w_g(A)}{\sqrt{n}} \;, \tag{3.10}$$

where $w_g(\cdot)$ denotes the Gaussian width of the error set. Such two-sided bounds are referred to as the Restricted Isometry Property (RIP) in prior literature [30, 31]. For the ℓ_1 norm, $w_g(A) \leq c_1 \cdot \sqrt{s \log p}$ for some constant c_1 [121, 122]. It can be easily seen that the RIP is a stronger condition as it requires the upper bound to be satisfied along with the lower bound which is required by RE. However for sub-Gaussian design matrices both RIP and RE are satisfied with high probability when $n \geq c_2 \cdot w_g^2(A)$ for some suitable positive constant c_2. However the gap between RIP and RE becomes evident when we start using heavier-tailed matrices [48, 85]. For example, for design matrices having independent symmetric exponential random elements for the specific case of the ℓ_1 norm, while [3] show the RIP property requires $n \geq c_3 \cdot s \log^2 p = c_4 \cdot w_g^2(A) \log p$, [85] show that RE holds with high probability when $n \geq c_5 \cdot \max\{s \log p, \cdot \log^3 p\}$. In this work, we establish sample complexity bounds for RE for general atomic norms when the rows of the design matrix are sub-exponential. We show that in the general case RE is satisfied when $n \geq c_6 \cdot w_e^2(A)$ where $w_e(\cdot)$ denotes the exponential width of a set. For a specific class of atomic norms (see Section 3.2), which includes ℓ_1, group sparse, k-support norms etc., we establish sharper bounds. For example, for the ℓ_1 norm we show that RE holds when $n \geq c_7 \cdot \max\{s \log p, \log^2 p\}$.

Estimation error bounds: Assuming the RE condition and lower bound on the regularization parameter hold, [11] derive the following result with high probability,

$$\|\Delta_{\lambda_n}\|_2 = \|\hat{\theta}_{\lambda_n} - \theta^*\|_2 \leq c \cdot \frac{\Psi(E_r)\lambda_n}{\kappa} , \tag{3.11}$$

where $\Psi(E_r) = \sup\limits_{u \in E_r} \frac{R(u)}{\|u\|_2}$ is the norm compatibility constant [105, 11]. For example, when the design matrix is sub-exponential and $R(\cdot)$ is the ℓ_1 norm, $\Psi(E_r) = \sqrt{s}$ and combining with the results for the regularization parameter and restricted eigenvalue condition, we obatin the following result with high probability when $n \geq c_1 \cdot \max\{s \log p, \log^2 p\}$,

$$\|\Delta_{\lambda_n}\|_2 \leq c \cdot \frac{\sqrt{s} \log p}{\kappa \sqrt{n}} . \tag{3.12}$$

The upper bounds on the estimation error are $\sqrt{\log p}$ worse compared to the estimation error with isotropic sub-Gaussian design matrices [11].

The rest of the chapter is organized as follows. In Section 3.2 we describe background and preliminaries related to the problem. In Section 3.3 we establish a key result on the relationship between Gaussian and exponential widths of sets which will be useful in all subsequent analysis. Sections 3.4, 3.5, and 3.6 are devoted to the main technical results of the chapter. We show results from experiments on synthetic data in Section 3.7. Throughout the chapter we denote absolute constants by $c, c_1, c_2, \ldots$ whose values can change from one line to the next.

3.2 Background and Preliminaries

We outline the problem formulation and highlight our contributions introducing notations along the way.

3.2.1 Problem Formulation

We assume the linear model $y = X\theta^* + \omega$. The goal is to estimate the unknown parameter $\theta^* \in \mathbb{R}^p$ given n samples $\{x_i, y_i\}_{i=1}^n$ and unknown random noise ω. The rows of the design matrix $X \in \mathbb{R}^{n\times p}$ are assumed to be sub-exponential (see next two subsections for precise definitions). Each element of the unknown noise vector $\omega \in \mathbb{R}^n$ is assumed to be i.i.d. from a zero mean unit variance sub-exponential distribution and is independent of the design matrix. We operate in the high-dimensional regime where potentially $p > n$, but θ^* has structure captured by some atomic norm. We will assume that $\|\theta^*\|_2 = 1$. We analyze the regularized least squares estimator [105, 11],

$$\hat{\theta}_{\lambda_n} = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n R(\theta) \ , \tag{3.13}$$

where λ_n is the regularization parameter whose value will depend on the number of samples (we will drop the subscripts from λ_n and $\hat{\theta}_{\lambda_n}$ going forward for ease of exposition) and $R(\cdot)$ is an atomic norm regularizer capturing the structure assumed for the true parameter θ^*.

3.2.2 Sub-exponential Random Variables

We define sub-exponential random variables and state a few properties.

Definition 8 (Sub-exponential random variables from Definition 5.13 in [127]): *A random variable $z \in \mathbb{R}$ that satisfies any of the below properties is called a sub-exponential random variable.*

$$P(|z| > t) \leq \exp(1 - t/K_1) \qquad \forall t \geq 0 \tag{3.14}$$

$$(E|z|^a)^{1/a} \leq K_2 a \qquad \forall a \geq 1 \tag{3.15}$$

$$E\exp\left(\frac{z}{K_3}\right) \leq e \ . \tag{3.16}$$

The sub-exponential norm of z, denoted as $\|z\|_{\psi_1}$, is defined as

$$\|z\|_{\psi_1} = \sup_{a\geq 1} a^{-1}(E|X|^a)^{1/a} \ . \tag{3.17}$$

For convenience, we state two properties of sub-exponential random variables which will be useful in establishing a few results. Sub-exponential random variables satisfy a Bernstein-type [127] inequality.

Lemma 6 (Proposition 5.16 in [127]) *Let $x \in \mathbb{R}^p$ be a random vector with independent centered sub-exponential random variables with $K = \max_{1 \leq i \leq p} \|x_i\|_{\psi_1}$. Then for every $u \in \mathbb{R}^p$,*

$$P(|\langle x, u\rangle| \geq \tau) \leq 2\exp\left[-c\min\left(\frac{\tau^2}{K^2\|u\|_2^2}, \frac{\tau}{K\|u\|_\infty}\right)\right] . \tag{3.18}$$

Linear forms of vectors with independent sub-exponential random variables are also sub-exponential random variables.

Lemma 7 *Consider $x \in \mathbb{R}^p$ has i.i.d. sub-exponential random variables with $\max_{1 \leq i \leq p} \|x\|_{\psi_1} \leq K$. Then the following is true with some absolute positive constant c,*

$$\sup_{u \in S^{p-1}} \|\langle x, u\rangle\|_{\psi_1} \leq cK . \tag{3.19}$$

3.2.3 Sub-exponential Design Matrix

We outline the assumptions on the design matrix $X \in \mathbb{R}^{n \times p}$. Let $x \in \mathbb{R}^p$ denote any row of X

1. **Sub-exponential marginals:** The rows of the design matrix have sub-exponential norm K, i.e., for any row x we have $\|x\|_{\psi_1} = \sup_{u \in S^{p-1}} \frac{1}{a}(E|\langle x, u\rangle|^a)^{1/a} \leq K$, $\forall a \geq 1$.

2. **Centering:** The rows have zero mean: $E[x] = 0$.

3. **Nondegeneracy:** There is a positive constant α such that for every $u \in S^{p-1}$: $E[|\langle x, u\rangle|] \geq \alpha > 0$.

4. **Isotropic rows:** The covariance is the identity matrix: $E[xx^T] = \mathbb{I}_{p \times p}$

Throughout the chapter a reference to sub-exponential design matrices implicitly means a design matrix satisfying the above properties.

3.3 Complexity Measures

When the design matrix is sub-Gaussian all results can be succintly expressed in terms of the Gaussian width of suitable sets related to the regularizer [11, 33]. Our results for sub-exponential data requires the definition of another complexity measure of sets, which we call the *exponential width* of sets. In this section we provide a gentle introduction to Gaussian and exponential widths of sets. We will also prove an upper bound for the exponential width in terms of the Gaussian width. With this result, precise bounds can be established for the non-asymptotic estimation error by leveraging a body of literature on the computation of Gaussian widths for various structured sets [33, 124, 38]. All our definitions and analysis are based on results from generic chaining [121, 122].

3.3.1 Gaussian and Exponential Widths

We define the Gaussian and exponential width complexity measures for any set $T \subseteq \mathbb{R}^p$.

Gaussian width: Consider any set $T \subseteq \mathbb{R}^p$. Let $\{X_t\}_{t \in T} = \langle g, t \rangle$ where each element $g_i,\ 1 \leq i \leq p$ is i.i.d. $N(0,1)$. The quantity $w_g(T) = E_g\left[\sup_{t \in T} X_t\right]$ is called the Gaussian width of the set T.

Generalizing to any stochastic process, consider $\{X_t\}_{t \in T}$ to be any stochastic process indexed by the set $T \subseteq \mathbb{R}^p$ satisfying the following increment condition for some constants K, c,

$$\forall u > 0,\ P(|X_s - X_t| \geq u) \leq 2 \exp\left(-\frac{cu^2}{K^2 \|s - t\|_2^2}\right) . \tag{3.20}$$

Then the following is true [121, 122]

$$E\left[\sup_{t \in T} X_t\right] \leq c \cdot w_g(T) . \tag{3.21}$$

Exponential width: Let $\{X_t\}_{t \in T} = \langle e, t \rangle$ where each element $e_i, 1 \leq i \leq p$ is a centered i.i.d. sub-exponential random variable satisfying the exponential tail decay $P\left(|e_i| \geq u\right) \leq 2\exp(-u)$. The quantity $w_e(T) = E\left[\sup_{t \in T} X_t\right]$ is called the exponential width of the set T.

Generalizing to any stochastic process $\{X_t\}_{t \in T}$ indexed by a set $T \subseteq \mathbb{R}^p$, if $\{X_t\}_{t \in T}$ satisfies the following increment condition for some constants K, c,

$$\forall u > 0,\ P(|X_s - X_t| \geq u) \leq 2 \exp\left(-c \min\left(\frac{u^2}{K^2 \|s - t\|_2^2}, \frac{u}{K \|s - t\|_\infty}\right)\right) . \tag{3.22}$$

Then the following is true [121, 122]

$$E\left[\sup_{t \in T} X_t\right] \leq c \cdot w_e(T) . \tag{3.23}$$

3.3.2 An Upper Bound for the Exponential Width

We provide an upper bound for the exponential width in terms of the Gaussian width of the set. This serves two purposes. First, for any given set, we are immediately able to quantify the effect of going from a sub-Gaussian to sub-exponential design matrix. Second, we can establish precise upper bounds for the non-asymptotic estimation error for regularizers for which the Gaussian widths of related sets have been computed [38, 33], without going into the specifics of computing the exponential width from the basic definition.

Theorem 1 *For any set $T \subset \mathbb{R}^p$, for some constant c the following holds:*

$$w_e(T) \leq c \cdot w_g(T)\sqrt{\log p}\ . \tag{3.24}$$

The above result is true for any set $T \subseteq \mathbb{R}^p$. For sets which are convex hulls of union of subspaces of the form $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_m$ and for which $\log m \leq \log p$ a stronger result can be proved. For example, the result is true for the group sparse norm, where the number of groups m is smaller than the ambient dimensionality p.

Theorem 2 *Consider any set $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_m \subseteq \mathbb{R}^p$. Then the following is true,*

$$w_e(\mathcal{A}) \leq \min\left\{c_1 w_g(\mathcal{A})\sqrt{\log p}, c_2\left(\max_{1\leq i\leq m} w_e(\mathcal{A}_i) + 2\sup_{u\in\mathcal{A}}\|u\|_2\sqrt{\log m} + 2\sup_{u\in\mathcal{A}}\|u\|_\infty \log m\right)\right\}\ . \tag{3.25}$$

Proofs for Theorem 1 and 2 are provided in the appendix.

3.4 Restricted Eigenvalue (RE) Condition

The error bound analysis requires the design matrix $X \in \mathbb{R}^{n\times p}$ to satisfy the following RE condition [17] on the error set A,

$$\inf_{u\in A=\text{cone}(E_r)\cap S^{p-1}} \frac{1}{\sqrt{n}}\|Xu\|_2 \geq \kappa\ , \tag{3.26}$$

where $\kappa > 0$ is the RE constant. Design matrices with i.i.d. sub-Gaussian rows satisfy the RE condition with high probability when the number of samples scale as square of the Gaussian width of the error set, i.e., $n \geq c_1 \cdot w_g^2(A)$ [33, 11]. We prove the following result for any norm showing that the RE condition is satisfied by design matrices with i.i.d. sub-exponential rows when the number of samples scales as the square of the exponential width of the set, i.e., $n \geq c_2 \cdot w_e^2(A)$. The proof, provided in the appendix, uses similar arguments as Proposition 5.1 in [124].

Theorem 3 *Consider the regularized least squares problem with atomic norm regularizer* $R(\theta) = \|\theta\|_{\mathcal{A}}$ *and let* $X \in \mathbb{R}^{n \times p}$ *be the design matrix whose rows are i.i.d. sub-exponential with sub-exponential norm* K*. With* E_r *denoting the error set, let the small-ball condition hold in the error set* $A = cone(E_r) \cap S^{p-1}$ *with constant* β*,*

$$\beta = \inf_{u \in A} P(|\langle x, u \rangle| > \xi) \; . \tag{3.27}$$

If $n \geq c \cdot w_e^2(A))$ *then with probability atleast* $1 - \exp(-\eta_0 n \beta^2)$*,*

$$\inf_{u \in A} \frac{1}{n} \|Xu\|_2^2 \geq \frac{\xi^2 \beta^2}{4} = \kappa \; , \tag{3.28}$$

where κ *is the restricted eigenvalue constant.*

When $R(\cdot)$ is the ℓ_1 norm and the true parameter θ^* is s-sparse, the result of Theorem 3 implies that $n \geq c_2 \cdot s \log^2 p$ samples are required to satisfy RE. This follows from the result on the upper bound for exponential width in Section 3.3 and noting that $w_g(A) \leq c_3 \cdot s \log p$ [38, 11, 33]. This also matches the result in [3] where it is shown that even the stronger RIP condition holds when $n \geq c_4 \cdot s \log^2 p$. More recently, for the specific case of the ℓ_1 norm, [85] obtained sharp results for RE with design matrices having rows with even heavier tails than sub-exponentials. Specifically for sub-exponential design matrices considered in this work, [85] show that $n \geq c_5 \cdot \max\{s \log p, \log^3 p\}$ samples are required to satisfy to RE. In our second result, we use similar analysis tools in [85] to establish sharper RE sample complexity results for all atomic norms having the union of subspaces form, like the ℓ_1, ℓ_1/ℓ_2 group sparse and the k-support norms. Specifically the RE sample complexity is the minimum number of samples required to satisfy the following two conditions with high probability:

Condition 1: Let the atomic set $\mathcal{A}$ be a union of m sets $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_m$, where each $\mathcal{A}_i$ is a q-dimensional subspace of $\mathbb{R}^p$. The true parameter θ^* is an $kq = s$-sparse vector that can be expressed as a linear combination of k elements as follows,

$$\theta^* = \sum_{i=1}^{k} c_i a_i, \; a_i \in \mathcal{A}_i, \; c_i \geq 0 \; . \tag{3.29}$$

For any vector in the error set $u \in E_r$, let $\Psi(E_r) = \sup\limits_{u \in E_r} \frac{\|u\|_{\mathcal{A}}}{\|u\|_2}$ denote the norm compatibility constant. Then the RE condition should be satisfied by the design matrix with high probability in every $k_1 q = s_1$ subspace where $k_1 = c(\kappa_1)\Psi^2(E_r)$ and c is a constant depending on κ_1 as defined in condition 2 below.

$$\inf_{u \in S^{p-1}, \|u\|_0 \leq s_1} \frac{1}{\sqrt{n}} \|Xu\|_2 \geq \kappa_1 \ . \tag{3.30}$$

Condition 2: The following should hold for all elements a_i of the atomic set $\mathcal{A}$ with high probability,

$$\sup_{a_i \in \mathcal{A}_i, 1 \leq i \leq m} \frac{1}{\sqrt{n}} \|Xa_i\| \leq \kappa_2 \ . \tag{3.31}$$

A keen reader familiar with the RIP condition [30, 31] will immediately realize that condition 2 is weaker than the following upper bound of RIP,

$$\sup_{u \in A = E_r \cap S^{p-1}} \frac{1}{\sqrt{n}} \|Xu\|_2 \leq \kappa_3 \ . \tag{3.32}$$

This is because the atomic set $\mathcal{A}$ is always, sometimes substantially, smaller than the error set A. This leads to a significant gap between the sample complexities required to satisfy RE and RIP respectively [85, 48] for sub-exponential design matrices.

The result below, whose proof is provided in the appendix, applies to atomic norms having the union of subspaces form as outlined in condition 1 above.

Theorem 4 *Consider the regularized least squares problem with atomic norm regularizer $R(\theta) = \|\theta\|_{\mathcal{A}}$ having the union of subspaces form. Let $X \in \mathbb{R}^{n \times p}$ be the design matrix whose rows are i.i.d. sub-exponential with sub-exponential norm K. With E_r denoting the error set, let the small-ball condition hold in the error set $A = cone(E_r) \cap S^{p-1}$ with constant β,*

$$\beta = \inf_{u \in A} P(|\langle x, u \rangle| > \xi) \ . \tag{3.33}$$

Let $\Psi(E_r) = \sup\limits_{u \in E_r} \frac{\|u\|_{\mathcal{A}}}{\|u\|_2}$ denote the norm compatibility constant. Then with,

$$n \geq c \cdot \max \left(\begin{array}{c} \Psi^2(E_r) q + \Psi^2(E_r) \log(em/\Psi^2(E_r)) \\ \min\left(q + \log m, q \log^2\left(\frac{ep}{q}\right)\right) \end{array} \right) \ , \tag{3.34}$$

with probability atleast $1 - \exp(-\eta_0 n \beta^2) - \exp(-\eta_1 \sqrt{n}/K^2)$,

$$\inf_{u \in E_r} \frac{1}{n} \|Xu\|_2^2 \geq \frac{\xi^2 \beta^2}{8} = \kappa \ , \tag{3.35}$$

where κ is the restricted eigenvalue constant.

The requirement $n \geq c \cdot \Psi^2(E_r)q + \Psi^2(E_r)\log(em/\Psi^2(E_r))$ is due to condition 1 above, while condition 2 is satisfied when $n \geq c \cdot \min\left(q + \log^2 m, q\log^2\left(\frac{ep}{q}\right)\right)$.

Corollary 1 *Let $R(\cdot)$ be the ℓ_1 norm with θ^* s-sparse as defined in Section 2.4. The RE condition is satisfied with high probability when $n \geq c \cdot \max\{s\log p, \log^2 p\}$.*

Proof: The bound follows from the result of Theorem 4 with $q = 1,\ m = p$ and $\Psi(E_r) = O(\sqrt{s})$. ■

Corollary 2 *Let $R(\cdot)$ be the ℓ_1/ℓ_2 non-overlapping group-sparse norm with θ^* having $s_{\mathcal{G}}$ non-zero groups as defined in Section 2.4. The RE condition is satisfied with high probability when $n \geq c \cdot \max\{ls_{\mathcal{G}} + \log N, l + \log^2 N)$.*

Proof: The bound follows from the result of Theorem 4 with $q = 1,\ m = N$ and $\Psi(E_r) = O(\sqrt{s_{\mathcal{G}}})$. ■

Corollary 3 *Let $R(\cdot)$ be the k-support norm with θ^* s-sparse as defined in Section 2.4. The RE condition is staisfied with high probability when $n \geq c \cdot \max\{s\log p, k\log^2 p\}$.*

Proof: The bound follows from the result of Theorem 4 with $q = k,\ m = \binom{p}{s} \leq \left(\frac{ep}{s}\right)^s$ and $\Psi(E_r) = O(\sqrt{2s/k})$. ■

3.5 Regularization Parameter

The non-asymptotic estimation error analysis requires the regularization parameter to satisfy the inquality

$$\lambda \geq 2R^*\left(\frac{1}{n}X^T\left(y - X\theta^*\right)\right) = R^*\left(\frac{1}{n}X^T\omega\right) \ , \tag{3.36}$$

where for the last inequality we use $\omega = y - X\theta^*$ is the noise term. Since both X and ω are random quantities we focus on high probability upper bounds for $R^*\left(\frac{1}{n}X^T\omega\right)$. We prove the below result, whose proof is provided in the appendix, characterizing the expectation and large deviation inequalities of $R^*\left(\frac{1}{n}X^T\omega\right)$

Theorem 5 *Let $\Omega_R = \{u : R(u) \leq 1\}$ be the unite norm ball. If the design matrix $X \in \mathbb{R}^{n\times p}$ has i.i.d. sub-exponential rows and noise ω has i.i.d sub-exponential unit variance entries with*

$\|\omega_i\|_{\psi_1} \leq K_\omega,\ E[\omega_i^2] = 1,\ \forall 1 \leq i \leq n$ respectively, then

$$E\left[R^*\left(\frac{1}{n}X^T\omega\right)\right] \leq \frac{c_1 \cdot w_e(\Omega_R)}{\sqrt{n}}\ , \tag{3.37}$$

for some positive constant c_1 depending on K the sub-exponential norm of the rows of the design matrix. Moreover with probability at least $1 - \exp(-\tau_1) - 2\exp\left(-c\tau_2\sqrt{n}/K_\omega^2\right)$, $\tau_1 > 2$, we have

$$R^*\left(\frac{1}{n}X^T\omega\right) \leq \frac{c_2 \cdot \tau_1\sqrt{1+\tau_2}w_e(\Omega_R)}{\sqrt{n}}\ , \tag{3.38}$$

where c_2 is some constant depending on K.

Therefore the regularization parameter can be chosen to scale as $\lambda \geq c_3 \cdot \frac{w_e(\Omega_R)}{\sqrt{n}}$. Below we instantiate the result for the specific norms we consider in this work.

Corollary 4 *Let $R(\cdot)$ be the ℓ_1 norm with θ^* s-sparse as defined in Section 2.4. Then choosing $\lambda \geq c \cdot \frac{\log p}{\sqrt{n}}$ satisfies inequality (3.36) with high probability.*

Proof: With $\Omega_R = \{u\ :\ \|u\|_1 \leq 1\}$ denoting the unit norm ball, using the result of Theorem 1 for some positive constants c, c_1

$$w_e(\Omega_R) \leq cw_g(\Omega_R)\sqrt{\log p} \leq c_1 \log p\ . \tag{3.39}$$

The bound now follows from the result of Theorem 5 and (3.39). ■

Corollary 5 *Let $R(\cdot)$ be the ℓ_1/ℓ_2 non-overlapping group-sparse norm with θ^* having $s_{\mathcal{G}}$ non-zero groups as defined in Section 2.4. Then choosing $\lambda \geq c \cdot \frac{\sqrt{l}+\log N_{\mathcal{G}}}{\sqrt{n}}$ satisfies inequality (3.36) with high probability.*

Proof: First note that the exponential width of the unit norm ball is equivalent to the unit norm ball of the atomic set $w_e(\Omega_R) = w_e(\mathcal{A})$. We will use the result from Theorem 2. First the atomic set is the union of $N_{\mathcal{G}}$ subspaces of dimension less than l. Therefore by elementary arguments for some constant c,

$$\max_{1 \leq i \leq N_{\mathcal{G}}} w_e(\mathcal{A}_i) \leq c\sqrt{l}\ . \tag{3.40}$$

Also $\sup_{u \in \mathcal{A}} \|u\|_2 = \sup_{u \in \mathcal{A}} \|u\|_\infty = 1$. Therefore using the result of (3.40) we get,

$$w_e(\mathcal{A}) \leq c_2\left(\sqrt{l} + 2\sqrt{\log N_{\mathcal{G}}} + 2\log N_{\mathcal{G}}\right)\ . \tag{3.41}$$

The bound now follows from the result of Theorem 5 and (3.41). ■

Corollary 6 *Let $R(\cdot)$ be the k-support norm with θ^* s-sparse as defined in Section 2.4. Then choosing $\lambda = c \cdot \sqrt{k} \log p$ satisfies inequality (3.36) with high probability.*

Proof: With $\Omega_R = \{u \,:\, \|u\|_k^{sp} \leq 1\}$ denoting the unit norm ball, using the result of Theorem 1 for some positive constants c, c_1

$$w_e(\Omega_R) \leq c w_g(\Omega_R)\sqrt{\log p} \leq c_1 \sqrt{k + k \log \left\lceil \frac{p}{k} \right\rceil} \sqrt{\log p}\,. \tag{3.42}$$

The bound now follows from the result of Theorem 5 and (3.42). ■

3.6 Estimation Error Bounds

When the design matrix satisfies the RE condition on the error set and the regularization parameter satisfies inequality (3.36) then the non-asymptotic estimation error can be bounded using the following result:

Theorem 6 *Assume the regularization parameter satisfies the inequality $\lambda_n \geq 2R^*(\frac{1}{n}X^T(y - X\theta^*))$ with probability atleast $1 - \exp(-\tau_1) - 2\exp\left(-c\tau_2\sqrt{n}/K_\omega^2\right)$, $\tau_1 > 2$ and the RE condition is satisfied on the error set $A = cone(E_r) \cap S^{p-1}$ with RE constant κ with probability atleast $1 - \exp(-\eta_0 n\beta^2) - \exp(-\eta_1\sqrt{n}/K^2)$. Then with $\Delta = \hat{\theta} - \theta^*$, for any norm $R(\cdot)$, we have for some positive constant c with probability atleast $1 - \exp(-\tau_1) - 2\exp\left(-c\tau_2\sqrt{n}/K_\omega^2\right) - \exp(-\eta_0 n\beta^2) - \exp(-\eta_1\sqrt{n}/K^2)$, $\tau_1 > 2$,*

$$\|\Delta\|_2 \leq \frac{c \cdot \Psi(E_r)\lambda}{\kappa}\,, \tag{3.43}$$

where $\Psi(E_r) = \sup\limits_{u \in E_r} \frac{R(u)}{\|u\|_2}$ is the norm compatibility constant and the definition of the various constants are same as Theorems 4 and 5

Proof: The conditions on the regularization parameter and the RE condition follow from the results of Theorems 4 and 5. The result on the ℓ_2 norm of the error follows from Theorem 2 in [11].

We instantiate the result for the ℓ_1, group-sparse and k-support norms below.

Corollary 7 *Let $R(\cdot)$ be the ℓ_1 norm with θ^* s-sparse as defined in Section 2.4. Assume conditions in corollaries 4 and 1 are satisfied then,*

$$\|\Delta\|_2 \leq \frac{c\sqrt{s}\log p}{\sqrt{n}\kappa} \,. \tag{3.44}$$

Proof: Noting that $\Psi(E_r) = O(\sqrt{s})$, the bound follows from the results of corollaries 4, 1 and Theorem 6. ∎

Corollary 8 *Let $R(\cdot)$ be the ℓ_1/ℓ_2 non-overlapping group-sparse norm with θ^* having $s_{\mathcal{G}}$ non-zero groups as defined in Section 2.4. Assume conditions in corollaries 5 and 2 are satisfied then,*

$$\|\Delta\|_2 \leq \frac{c\left(\sqrt{l s_{\mathcal{G}}} + \sqrt{s_{\mathcal{G}}}\log N_{\mathcal{G}}\right)}{\sqrt{n}\kappa} \,. \tag{3.45}$$

Proof: Noting that $\Psi(E_r) = O(\sqrt{s_{\mathcal{G}}})$, the bound follows from the results of corollaries 5, 2 and Theorem 6. ∎

Corollary 9 *Let $R(\cdot)$ be the k-support norm with θ^* s-sparse as defined in Section 2.4. Assume conditions in corollaries 6 and 3 are satisfied then,*

$$\|\Delta\|_2 \leq \frac{c\sqrt{s}\log p}{\sqrt{n}\kappa} \,. \tag{3.46}$$

Proof: Noting that $\Psi(E_r) = O(\sqrt{2s/k})$, the bound follows from the results of corollaries 6, 3 and Theorem 6. ∎

3.7 Experiments

We perform experiments on synthetic data for the ℓ_1 norm and ℓ_1/ℓ_2 non-overlapping group sparse norms with Gaussian and sub-exponential data. Data is generated using the linear model $y = X\theta^* + \omega$, where we first generate the parameter vector $\theta^* = [\underbrace{1, 1, 1, 1, 1, 1,}_{5} \underbrace{0, 0, \dots, 0}_{\text{p-5}}] \in \mathbb{R}^p$ for the ℓ_1 norm and $\theta^* = [\underbrace{1, \dots, 1,}_{3} \underbrace{1, \dots, 1,}_{3} \underbrace{1, \dots, 1,}_{3} \underbrace{0, \dots, 0,}_{3} \dots, \underbrace{0, \dots, 0}_{3 + \text{p mod} 3}]$ for the ℓ_1/ℓ_2 group sparse norm and then normalize so that $\|\theta^*\|_2 = 1$. The dimension $p \in [500, 750, 1000]$ and (p mod 3) denotes the remainder of $p/3$. For Gaussian data, the noise and design matrix are sampled from the Gaussian distribution $\omega_i \sim N(0, 0.25), 1 \leq i \leq n$ and $X_{ij} \sim N(0, 1), 1 \leq$

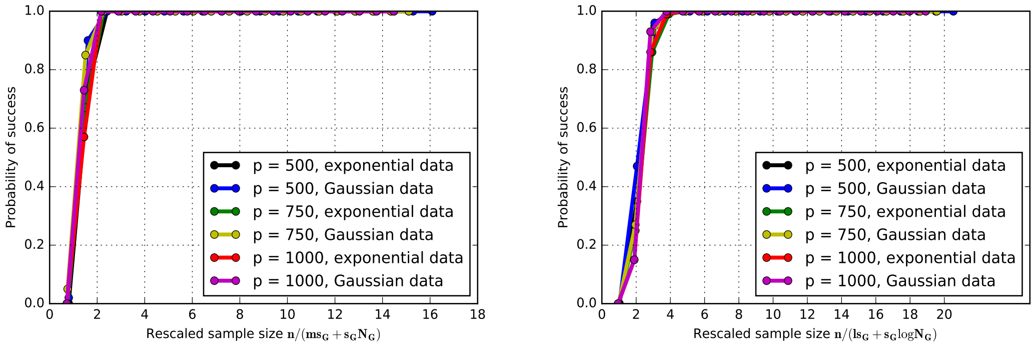

Figure 3.1: Probability of recovering true parameter versus the rescaled sample size for ℓ_1 norm (left) and ℓ_1/ℓ_2 group sparse norm (right). The probability of success rises sharply (phase transition) at the same sample size for both sub-exponential and Gaussian data.

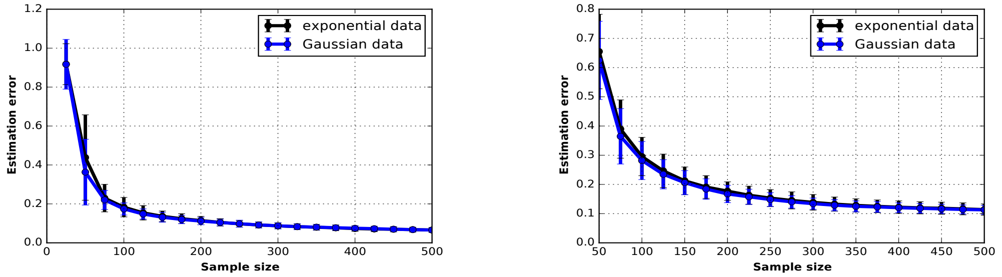

Figure 3.2: Estimation errors for the ℓ_1 and ℓ_1/ℓ_2 group sparse norms for different sample sizes. Estimation error is slightly higher with sub-exponential data.

$i \le n, 1 \le i \le p$ and mean centered. Similarly for sub-exponential designs and noise we sample from the exponential distribution and mean center the data. For each $p \in [500, 750, 1000]$, we generate 100 datasets with sample sizes $n \in [25, 50, \ldots, 500]$. We compute the probability of success out of 100 experiments for each sample size n and dimension p, where success is defined when the truncated estimated parameter correctly estimates the support of the true parameter, e.g., for the ℓ_1 norm the estimated parameter $\hat{\theta}$ is such that $\hat{\theta}_i > 0, 1 \le i \le 5$ and $\hat{\theta}_i > \left|\hat{\theta}_j\right|, \forall i, j, 1 \le i \le 5, 6 \le j \le p$. As can be seen from Figure 3.1 the probability of success increases with n and the probability of success are same for Gaussian and exponential data. We also measure the estimation error for $p = 500$. We plot the mean and standard deviation of the errors across 100 runs for each n in Figure 3.2. We observe that for smaller sample sizes lighter tailed data leads to better estimation performance.

Appendix

3.A Background and Preliminaries

3.A.1 Widths and generic chaining

One of the recurring ideas in our analysis and some previous works in this domain [33, 11, 124] is the concept of width of the set. The manifestation is of the form $E \sup_{t \in T} \langle h, t \rangle$, where $A \subseteq \mathbb{R}^p$ is a set and $h \in \mathbb{R}^p$ is a random vector in p-dimensional space. If h is a standard Gaussian random vector then it is the Gaussian width of the set A. In our analysis, we will also have h to have i.i.d. sub-exponential random variables which leads us to the concept of exponential width of the set. Much work has been done in the field of 'empirical process theory' to understand such quantities for different measures of h. When h has exponentially decaying tails, a number of results have been established using generic chaining [121]. Our endeavour in this section is to give a gentle introduction to the idea of generic chaining. All of the material we present below can be found in [121].

Generic Chaining and Gaussian Width: Much of what we present below will be from [121]. Before we get started we clarify a few notations. We first focus on the Gaussian case. We denote by $\{X_t\}_{t \in T} = \langle g, t \rangle$ a Gaussian process indexed by a certain set $T \subseteq \mathbb{R}^p$ where g has i.i.d $N(0,1)$ entries. By definition the process is centered i.e. $E(X_t) = 0, \ \forall t \in T$. Without any loss of generality, we can assume that the set T is finite. Since the process is centered, we have for some $t_0 \in T$

$$E \sup_{t \in T} X_t = E \sup_{t \in T} (X_t - X_{t_0}) \ . \tag{3.47}$$

The first major observation is that the boundedness of the process $(X_s - X_t)$ is determined by the following distance metric on the metric space T

$$d_2(s,t) = (E(X_s - X_t)^2)^{1/2} = \|s - t\|_2 \ . \tag{3.48}$$

Now for a Gaussian process the following is true

$$P(|X_s - X_t| \geq u) \leq 2 \exp\left(-\frac{u^2}{2 d_2(s,t)^2} \right) \ . \tag{3.49}$$

Therefore from (3.47) and using a union bounding argument, we can write

$$\begin{aligned} P\left(\sup_{t\in T}(X_t - X_{t_0}) > u\right) &\leq \sum_{t\in T} P(X_t - X_{t_0} > u) \\ &\leq \sum_{t\in T} 2\exp\left(-\frac{u^2}{2d_2(t,t_0)^2}\right) \ . \end{aligned}$$

The above bound is very lose if the elements of the set $t \in T$ are clustered in sets. If elements of set T form two different clusters T_1 and T_2, a better bound will be

$$\begin{aligned} P\left(\sup_{t\in T}(X_t - X_{t_0}) > u\right) \leq P(X_{t_1} - X_{t_0} > u_1) + \sum_{t\in T_1} P(X_t - X_{t_1} > u_2) \\ + P(X_{t_2} - X_{t_0} > u_1) + \sum_{t\in T_2} P(X_t - X_{t_2} > u_2) \ . \end{aligned}$$

This is because the distances $d_2(t_1, t), \forall t \in T_1$ and $d_2(t_2, t), \forall t \in T_2$ are much lesser than $d(t, t_0)$ leading to better estimates of the probability on the r.h.s. This is the basic idea behind generic chaining, where the above procedure is repeated successively on the subsets.

Let us formalize this idea. Let $T_0 = \{t_0\}$, and consider a sequence of subsets $\{T_n\}$ such that $T_0 \subseteq T_1 \subseteq T_2 \subseteq \ldots \subseteq T$. For some large enough m, we have $T_n = T$ for $n \geq m$. Let $\pi_n : T \to T_n$ denote a map such that $\pi_n(t)$ denotes the element in T_n which best approximates t. Therefore we can write

$$X_t - X_{t_0} = \sum_{n\geq 1}(X_{\pi_n(t)} - X_{\pi_{n-1}(t)}) \ . \tag{3.50}$$

This decomposition is at the core of generic chaining.

The next main idea is to control the scale of u for a set of a given cardinality. Consider a set T_n of cardinality 2^{2^n}. The question is what should be the scale of u in the below inequality to get exponential tail probabilities.

$$\begin{aligned} P\left(\sup_{t\in T}(X_t - X_{t_0}) \geq u\right) &\leq \sum_{t\in T} 2\exp\left(-\frac{u^2}{2d_2^2(t,t_0)}\right) \\ &\leq 2|T_n|\exp\left(-\frac{u^2}{2d_2^2(t,t_0)}\right) \\ &\leq L(\exp 2^n)\left(-\frac{u^2}{2d_2^2(t,t_0)}\right) \\ &\leq L\left(-\frac{u^2}{2d_2^2(t,t_0)} + 2^n\right) \ . \end{aligned}$$

Obviously $u > 2^{n/2} d_2(t, t_0)$ and hence u should scale as $2^{n/2} d_2(t, t_0)$. The metric $d_2(s, t)$ on the space T and providing a bound based on the cardinality of $T_n, n \geq 0$ are the two central ideas to understand the generic chaining argument. Note that we have shown that the bound depends solely on the following two properties: 1) The concentration inequality of each individual process characterized by a distance metric in the space T and 2) The cardinality of the set T which is a purely geometric quantity. We skip the detailed arguments given in [121] and state the following important result with $d(t, \pi_n(t)) = d(t, T_n) = \inf_{s \in T_n} d_2(t, s)$

$$E \sup_{t \in T} X_t \leq L \sup_t \sum_{n \geq 0} 2^{n/2} d_2(t, T_n) \ . \tag{3.51}$$

We now define the basic complexity parameters used in generic chaining.

Definition 9 ***Admissible sequence:*** *Given a set T an admissible sequence is an increasing sequence $(\mathcal{A}_n)$ of partitions of T such that $card(\mathcal{A}_n) \leq N_n$*

By increasing sequence of partitions we mean that every set of $\mathcal{A}_{n+1}$ is contained in a set of $\mathcal{A}_n$. The elements in the set T_n are picked one from each partition in $\mathcal{A}_n$.

Definition 10 ***γ-functionals:*** *Given $\alpha > 0$, and a metric space (T, d) we define*

$$\gamma_\alpha(T, d) = \inf \sup_t \sum_{n \geq 0} 2^{n/\alpha} \Delta(\mathcal{A}_n(t)) \ . \tag{3.52}$$

where $\Delta(\mathcal{A}_n(t))$ is the diameter of the partition $\mathcal{A}_n$, corresponding to distance metric d.

The below result from [Theorem 2.1.1] [121] gives both an upper and lower bound for Gaussian width in terms of $\gamma_2(T, d_2)$ and signifies that this is the best possible estimate for the Gaussian width. For some universal constant L we have:

$$\frac{1}{L} \gamma_2(T, d_2) \leq E \sup_{t \in T} X_t \leq L \gamma_2(T, d_2) \ . \tag{3.53}$$

3.A.2 Generic Chaining and the Exponential Width

Similar to the previous section, we now denote $\{X_t\}_{t \in T} = \langle e, t \rangle$ is a exponential process where e is a standard exponential random vector. According to Bernstein's inequality [127]

$$P(|X_s - X_t| \geq u) \leq 2 \exp\left(-\frac{1}{K} \min\left(\frac{u^2}{d_2(s,t)^2}, \frac{u}{d_\infty(s,t)} \right) \right) \ . \tag{3.54}$$

where $d_2(s,t) = \|s-t\|_2$ and $d_\infty(s,t) = \|s-t\|_\infty$. We now define the following complexity parameters based on the above concentration inequality:

$$\gamma_2(T,d_2) = \inf \sup_t \sum_{n \geq 0} 2^{n/2} \Delta_2(\mathcal{B}_n(t))$$
$$\gamma_1(T,d_\infty) = \inf \sup_t \sum_{n \geq 0} 2^{n} \Delta_\infty(\mathcal{C}_n(t)) \; .$$

In short we construct increasing sequences of partitions corresponding to each term in the concentration equation. We state the following result for exponential random variables, using results from Theorems 1.2.7 and 5.2.7 in [121] Assume that the random variable X_i are independent, symmetric and satisfy $P(|Y_i| \geq x) = \exp(-x)$. Then we have

$$\frac{1}{L}(\gamma_2(T,d_2) + \gamma_1(T,d_\infty)) \leq E \sup_{t \in T} X_t \leq L(\gamma_2(T,d_2) + \gamma_1(T,d_\infty)) \; . \tag{3.55}$$

The result is applicable to any process that has concentration inequality (3.54) [121].

3.B Relation Between Gaussian and Exponential Width

We prove Theorem 1 and 2.

Theorem 1 *For any set $T \subset \mathbb{R}^p$, for some constant c the following holds:*

$$w_e(T) \leq c \cdot w_g(T)\sqrt{\log p} \; . \tag{3.56}$$

The result depends on geometric results [Lemma 2.6.1] and [Theorem 2.6.2] in [121].

Theorem 7 *[121] Consider a countable set $T \subset \mathbb{R}^p$, and a number $v > 0$. Assume that the Gaussian width is bounded i.e. $S = \gamma_2(T,d_2) \leq \infty$. Then there is a decomposition $T \subset T_1 + T_2$ where $T_1 + T_2 = \{t_1 + t_2 : t_1 \in T_1, t_2 \in T_2\}$, such that*

$$\gamma_2(T_1,d_2) \leq cS \; , \qquad \gamma_1(T_1,d_\infty) \leq cSv \tag{3.57}$$
$$\gamma_2(T_2,d_2) \leq cS \; , \qquad T_2 \subset \frac{cS}{v} B_1 \; , \tag{3.58}$$

where c is some universal constant and B_1 is the unit ℓ_1 norm ball in $\mathbb{R}^p$.

We first examine the exponential widths of the sets T_1 and T_2. For the set T_1:

$$w_e(T_1) \leq c[\gamma_2(T_1, d_2) + \gamma_1(T_1, d_\infty)] \leq c[S + Sv] = c(w_g(T) + w_g(T)v) \ , \tag{3.59}$$

where the first inequality follows from (3.55) and the second inequality follows from (3.57). We will need the following result on bounding the exponential width of an unit ℓ_1-norm ball in p dimensions to compute the exponential width of T_2. The proof is based on the fact $\sup_{t \in B_1} \langle e, t \rangle = \|e\|_\infty$ and then using a simple union bound argument to bound $\|e\|_\infty$.

Lemma 8 *Consider the set $B_1 = \{t \in \mathbb{R}^p : \|t\|_1 \leq 1\}$. Then for some universal positive constant c:*

$$w_e(B_1) = E\left[\sup_{t \in B_1} \langle e, t \rangle\right] \leq c \log p \ . \tag{3.60}$$

Proof: Using the definition of dual norms:

$$E\left[\sup_{t \in B_1} \langle e, t \rangle\right] = E\|e\|_\infty = E\left[\max_{1 \leq i \leq p} |e_i|\right] \ . \tag{3.61}$$

Let $Y = \max\limits_{1 \leq i \leq p} |e_i|$ is a positive random variable. Then,

$$EY = \int_0^\infty P(Y > \tau) d\tau \ . \tag{3.62}$$

Now by definition, $P(|e_i| > \tau \log p) \leq \exp(-\tau \log p)$.

Consider the event Ω_τ is $|e_i| \leq \tau \log p, \forall i \in [1, \ldots, p]$. Therefore, by a simple union bound argument

$$\begin{aligned} P(\Omega_\tau^c) \leq p(\tau) &:= \sum_{i=1}^p \exp(-\tau \log p) \\ &= p \exp(-\tau \log p) \\ &= \exp(-\tau \log p + \log p) \ . \end{aligned}$$

Now if $\tau \geq 2, \log p \geq 2$, then $-\tau \log p \geq -\tau + \log p$, therefore we get,

$$p(\tau) \leq c_1 \exp(-\tau) \ , \tag{3.63}$$

for some constant c_1 when $\tau \geq 2$.

When Ω_τ occurs, we have

$$\max_{1 \leq i \leq p} |e_i| \leq \tau \log p \; . \tag{3.64}$$

Therefore we have,

$$P\left(\max_{1 \leq i \leq p} |e_i| > \tau \log p\right) \leq p(\tau) \Rightarrow P(Y > \tau S) \leq p(\tau) \; , \tag{3.65}$$

where $S = \log p$. Hence, by (3.62) and observing that the integrand is less than 1 we get,

$$w_e(B_1) = E\left[Y\right] \leq c \log p \; , \tag{3.66}$$

for some constant c. This completes the proof. ■

Now returning to the proof of Theorem1, the exponential width of T_2 is:

$$w_e(T_2) = w_e((LS/v)B_1) = (LS/v)w_e(B_1) = (L/v)w_g(T)w_e(B_1) \leq (L/v)w_g(T) \log p \; . \tag{3.67}$$

The first equality follows from (3.58) as T_2 is a subset of a (LS/v)-scaled ℓ_1 norm ball, the second inequality follows from elementary properties of widths of sets and the last inequality follows from Lemma 8. Now as stated in Theorem 7, v in (3.59) and (3.67) is any number greater than 0. We choose $v = \sqrt{\log p}$ and noting that $(1 + \sqrt{\log p}) \leq c\sqrt{\log p}$ for some constant c, the results from (3.67) and (3.59) yields:

$$w_e(T_1) \leq c w_g(T)\sqrt{\log p}, \qquad w_e(T_2) \leq c w_g(T)\sqrt{\log p} \tag{3.68}$$

The final step, following arguments as [Theorem 2.1.6] [121], is to bound exponential width of set T.

$$w_e(T) = E[\sup_{t \in T}\langle e, t\rangle] \leq E[\sup_{t_1 \in T_1}\langle e, t_1\rangle] + E[\sup_{t_2 \in T_2}\langle e, t_2\rangle] \leq w_e(T_1) + w_e(T_2) \leq c w_g(T)\sqrt{\log p} \; .$$

This proves Theorem 1. ■

We now prove Theorem 2.

Theorem 2 *Consider any set $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_m \subseteq \mathbb{R}^p$. Then the following is true,*

$$w_e(\mathcal{A}) \leq \min\left\{c_1 w_g(\mathcal{A})\sqrt{\log p}, c_2\left(\max_{1 \leq i \leq m} w_e(\mathcal{A}_i) + 2\sup_{v \in \mathcal{A}}\|v\|_2\sqrt{\log m} + 2\sup_{v \in \mathcal{A}}\|v\|_\infty \log m\right)\right\} \; . \tag{3.69}$$

Proof: $w_e(\mathcal{A}) \leq c_1 w_g(\mathcal{A})\sqrt{\log p}$ follows from the result of Theorem 1.

For the second result assume $t_i \in \mathcal{A}_i$ are m points from the m subspaces. Let $X_t = \langle e, t\rangle,\ t \in \mathcal{A},\ E[X_t] = 0,\ \forall t \in \mathcal{A}$ be an empirical process indexed by the set $\mathcal{A}$. Using arguments from generic chaining [122] for any $s \in \mathcal{A}_i$,

$$E \sup_{s\in\mathcal{A}_i, 1\leq i\leq m} X_s = E \sup_{s\in\mathcal{A}_i, 1\leq i\leq m} (X_s - X_{t_1}) \ . \tag{3.70}$$

Also,

$$X_s - X_{t_1} = X_s - X_{t_i} + X_{t_i} - X_{t_1} \ . \tag{3.71}$$

Therefore from (3.70) and (3.71),

$$E \sup_{s\in\mathcal{A}_i, 1\leq i\leq m} X_s \leq E \sup_{s\in\mathcal{A}_i, 1\leq i\leq m} (X_s - X_{t_i}) + E \max_{t_i, 1\leq i\leq m} (X_{t_i} - X_{t_1}) \ . \tag{3.72}$$

From the definition of exponential width,

$$\max_{1\leq i\leq m} w_e(\mathcal{A}_i) = E \sup_{s\in\mathcal{A}_i, 1\leq i\leq m} (X_s - X_{t_i}) \ . \tag{3.73}$$

To compute $E \max_{t_i, 1\leq i\leq m} (X_{t_i} - X_{t_1})$, note that since it is a non-negative quantity,

$$E \max_{t_i, 1\leq i\leq m} (X_{t_i} - X_{t_1}) = \int_0^\infty P\left(\max_{t_i, 1\leq i\leq m} (X_{t_i} - X_{t_1}) \geq \tau \right) d\tau \ . \tag{3.74}$$

Let $a = \max_{1\leq i\leq m} \|t_i - t_1\|_2 = 2\sup_{v\in\mathcal{A}} \|v\|_2$ and $b = \max_{1\leq i\leq m} \|t_i - t_1\|_\infty = 2\sup_{v\in\mathcal{A}} \|v\|_\infty$.

Then for any $1 \leq i \leq m$ using Bernstein's inequality result from Lemma 6,

$$P\left(X_{t_i} - X_{t_1} \geq a\tau\sqrt{\log m} + b\tau \log m \right) \leq 2\exp\left[-c\min\left(\tau^2 \log m, \tau \log m \right) \right] \ . \tag{3.75}$$

Since there are m such elements t_i by a simple union bound argument we get,

$$P\left(\max_{1\leq i\leq m} X_{t_i} - X_{t_1} \geq a\tau\sqrt{\log m} + b\tau\log m \right) \leq 2m\exp\left[-c\min\left(\tau^2\log m, \tau\log m \right) \right] \tag{3.76}$$

$$\leq 2\exp\left[-c\min\left(\tau^2\log m, \tau\log m \right) + \log m \right] \ . \tag{3.77}$$

If $c\tau, \log m \geq 2$, then $\tau^2 \log m \geq \tau^2 + \log m$ and $\tau \log m \geq \tau + \log m$. Therefore,

$$P\left(\max_{1\leq i\leq m}(X_{t_i} - X_{t_1}) \geq a\tau\sqrt{\log m} + b\tau\log m\right) \leq 2\exp\left(-c\tau\right) \tag{3.78}$$

$$P\left(\max_{1\leq i\leq m}(X_{t_i} - X_{t_1}) \geq \tau(S_1 + S_2)\right) \leq p(\tau) \tag{3.79}$$

$$\tag{3.80}$$

where $S_1 = a\sqrt{\log m}$, $S_2 = b\log m$ and $p(\tau) \leq c_1\exp(-c\tau)$ for $\tau \geq 2$. Therefore from (3.74), (3.79) and observing that the integrand is ≤ 1,

$$E\max_{t_i,1\leq i\leq m}(X_{t_i} - X_{t_1}) \leq c(2\sup_{v\in\mathcal{A}}\|v\|_2\sqrt{\log m} + 2\sup_{v\in\mathcal{A}}\|v\|_\infty\log m)\ . \tag{3.81}$$

where in the last result we have substituted the values of a and b. The result of Theorem 2 now follows from (3.72), (3.73), (3.81) and the result of Theorem 1. ∎

3.C Regularization Parameter

We give proofs for the bounds on the regularization parameter. The regularization parameter should be the lowest value satisfying the inequality $\lambda \geq 2R^*\left(\frac{1}{n}X^T\omega\right)$.

Theorem 5 *Let $\Omega_R = \{u : R(u) \leq 1\}$. If the design matrix $X \in \mathbb{R}^{n\times p}$ has i.i.d. sub-exponential rows and noise ω has i.i.d sub-exponential unit variance entries with $\|\omega_i\|_{\psi_1} \leq K_\omega,\ E[\omega_i^2] = 1,\ \forall 1 \leq i \leq n$ respectively, then*

$$E\left[R^*\left(\frac{1}{n}X^T\omega\right)\right] \leq \frac{c_1\cdot w_e(\Omega_R)}{\sqrt{n}}\ , \tag{3.82}$$

for some positive constant c_1 depending on K the sub-exponential norm of the rows of the design matrix. Moreover with probability at least $1 - \exp(-\tau_1) - 2\exp\left(-c\tau_2\sqrt{n}/K_\omega^2\right)$, $\tau_1 > 2$, we have

$$R^*\left(\frac{1}{n}X^T\omega\right) \leq \frac{c_2\cdot\tau_1\sqrt{1+\tau_2}w_e(\Omega_R)}{\sqrt{n}}\ , \tag{3.83}$$

where c_2 is some constant depending on K.

Proof: Let us first compute the expected value of $R^*\left(\frac{1}{n}X^T\omega\right)$.

$$E\left[R^*\left(\frac{1}{n}X^T\omega\right)\right] = E\left[\sup_{u\in\mathcal{A}} \|\omega\|_2 \left\langle \frac{1}{n}X^T\frac{\omega}{\|\omega\|_2}, u\right\rangle\right] \tag{3.84}$$

$$= \frac{1}{n}E[\|\omega\|_2]E\left[\sup_{u\in\Omega_R}\langle e, u\rangle\right] \tag{3.85}$$

$$\leq \frac{c_1 \cdot w_e(\Omega_R)}{\sqrt{n}} \; . \tag{3.86}$$

The first equality follows from the definition of dual norm. The second inequality follows from the fact that X and ω are independent of each other. Also from the definition of the design matrix, $e = X^T\frac{\omega}{\|\omega\|_2}$ has i.i.d sub-exponential random variables bounded by $\sup_{\omega\in\mathbb{R}^n}\|\langle x_i, \omega/\|\omega\|_2\rangle\|_{\psi_1} \leq K$, where x_i is any column of the design matrix X. For the last inequality use conditions (3.22) and (3.23) to get the following for some constant c_2 depending on the sub-exponential norm K,

$$E\left[\sup_{u\in\Omega_R}\langle e, u\rangle\right] \leq c_2 \cdot w_e(\Omega_R) \; . \tag{3.87}$$

Also,

$$E[\|\omega\|_2] \leq \sqrt{E[\|\omega\|_2^2]} = \sqrt{E\left[\sum_{i=1}^n \omega_i^2\right]} \leq \sqrt{n} \; , \tag{3.88}$$

Therefore from (3.87) and (3.88), (3.86) follows for some c_1 which depends on K. This proves the result on the expectation.

To establish large deviation bounds around the expectation, consider the quantities $P\left(\sup_{u\in\Omega_R}\langle e, u\rangle \geq \tau_1 c_2 w_e(\Omega_R)\right)$ and $P\left(\left|\frac{1}{n}\sum_i^n \omega_i^2 - 1\right| \geq \tau_2\right)$.

Note that when computing $E\left[\sup_{u\in\Omega_R}\langle e, u\rangle\right] = \int_0^\infty P\left(\sup_{u\in\Omega_R}\langle e, u\rangle \geq \tau_1 c_2 w_e(\Omega_R)\right) d\tau_1$ we prove the following for some $\tau_1 > 2$ (see equation (2.56) in Theorem 2.2.23 in [122]),

$$P\left(\sup_{u\in\Omega_R}\langle e, u\rangle \geq \tau_1 c_2 w_e(\Omega_R)\right) \leq \exp(-\tau_1) \; . \tag{3.89}$$

To establish large deviation inequality $P\left(\left|\frac{1}{n}\sum_i^n \omega_i^2 - 1\right| \geq \tau_2\right)$, we use the following result of Lemma 3.5 in [3].

Lemma 9 *Consider n i.i.d. sub-exponential centered random variables a_i with unit variance and $K = \max_{1 \le i \le n} a_i$. Then with probability atleast $1 - 2\exp\left(-c\sqrt{n}/K^2\right)$ where c is some positive constant,*

$$\left|\frac{1}{n}\sum_{i=1}^{n} a_i^2 - 1\right| \le \tau \ . \tag{3.90}$$

Applying Lemma 9 with sub-exponential random variables ω_i we get with probability atleast $1 - 2\exp\left(-c\tau_2\sqrt{n}/K_\omega^2\right)$,

$$\|\omega\|_2^2 \le (1+\tau_2)n \Rightarrow \|\omega\|_2 \le \sqrt{(1+\tau_2)n} \ . \tag{3.91}$$

The large deviation result now follows from (3.89) and (3.91).

3.D Restricted Eigenvalue Condition

We present the proof for the RE condition result. A few definitions and results from [81, 124] are required before we can prove the main result.

We first define the small-ball condition for random vectors x on subsets $A \subseteq \mathbb{R}^p$ [102], also called the marginal tail function in [124].

Definition 11 ***Small ball condition:*** *Fix a set $A \subseteq S^{p-1}$. Let x be a random vector in $\mathbb{R}^p$. The small-ball property with constants $\xi, \beta > 0$ is defined as follows,*

$$\beta(A; x) = \inf_{u \in A} P(|\langle x, u\rangle| > \xi) \ . \tag{3.92}$$

■

As noted in [81, 124] the small-ball condition reflects the probability that the random variable $|\langle x, u\rangle|$ is close to zero for any fixed vector $u \in A$. When $\beta(A; x)$ is bounded away from zero for some ξ, the nonnegative empirical process is likely to be large. If the rows of the design matrix $x_i \sim x$ are not sufficiently continuous or "spiky", $\beta(A; x)$ will be quite small and recovery of θ will be difficult because we need the row vectors in the design matrix to uniformly explore all possible directions in the error set A to obtain a good estimate of θ. Below we obtain a lower bound on $\beta(A; x)$ when the random vector x is sub-exponential. To ease the notation we will denote $\beta(A; x)$ by β where A, x will be implicit from the context.

Lemma 10 *For random sub-eponential* $x \in \mathbb{R}^p$ *with sub-exponential norm* $K = \sup_{u \in S^{p-1}} \|\langle x, u\rangle\|_{\psi_1}$ *and* $\xi > 0$

$$\beta = \inf_{u \in A} P(|\langle x, u\rangle| > \xi) \geq \frac{(\alpha - \xi)^2}{4K^2} \ . \tag{3.93}$$

Proof: For any non-negative random variable y, the following inequality is true due to the Paley-Zygmund inequality,

$$P(y > t) \geq \frac{(Ey - t)^2}{Ey^2} \ . \tag{3.94}$$

Applying the inequality to the quantity $y = |\langle x, u\rangle|$ and $t = \xi$,

$$P(|\langle x, u\rangle| > \xi) \geq \frac{(E[|\langle x, u\rangle|] - \xi)^2}{E[\langle x, u\rangle^2]} \ . \tag{3.95}$$

From the nondegeneracy assumption on the design matrix $E[|\langle x, u\rangle|] \geq \alpha$. Also by the properties of sub-exponential random variables the denominator is always bounded by $E[\langle x, u\rangle^2]^{1/2} \leq 2\sup_{u \in A} \|\langle x, u\rangle\|_{\psi_1} = 2K$. Therefore $E[\langle x, u\rangle^2] \leq 4K^2$. Substituting the above arguments in (3.95),

$$\beta = \inf_{u \in A} P(|\langle x, u\rangle| > \xi) \geq \frac{(\alpha - \xi)^2}{4K^2} \ . \tag{3.96}$$

■

We now provide the proof of Theorem 3. The proof follows from arguments in [124].

Theorem 3 *Consider the regularized least squares problem with atomic norm regularizer* $R(\theta) = \|\theta\|_{\mathcal{A}}$ *and let* $X \in \mathbb{R}^{n \times p}$ *be the design matrix whose rows are i.i.d. sub-exponential with sub-exponential norm* K*. With* E_r *denoting the error set, let the small-ball condition hold in the error set* $A = cone(E_r) \cap S^{p-1}$ *with constant* β*,*

$$\beta = \inf_{u \in A} P(|\langle x, u\rangle| > \xi) \ . \tag{3.97}$$

Let $w_e(A)$ *denote the exponential width of the error set* A*. If* $n \geq c \cdot w_e^2(A)$ *then with probability atleast* $1 - \exp(-\eta_0 n \beta^2)$*,*

$$\inf_{u \in A} \frac{1}{n}\|Xu\|_2^2 \geq \frac{\xi^2\beta^2}{4} = \kappa \ , \tag{3.98}$$

where κ *is the restricted eigenvalue constant.*

Proof: For any $u \in A$ by Lyapunov's inequality

$$\frac{1}{\sqrt{n}}\|Xu\|_2 = \left(\frac{1}{n}\sum_{i=1}^{n}|\langle x_i, u\rangle|^2\right)^{\frac{1}{2}} \geq \frac{1}{n}\sum_{i=1}^{n}|\langle x_i, u\rangle| \,. \tag{3.99}$$

Now by Markov's inequality for some constant $\xi/2 > 0$

$$\frac{1}{n}\sum_{i=1}^{n}|\langle x_i, u\rangle| \geq \frac{\xi}{2n}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i, u\rangle| \geq \frac{\xi}{2}\right) \,. \tag{3.100}$$

Therefore, we get

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|Xu\|_2 &\geq \frac{\xi}{2n}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i, u\rangle| \geq \frac{\xi}{2}\right) \\
&\geq \xi\Big(P\{|\langle x, u\rangle| \geq \xi\} + \frac{1}{2n}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i, u\rangle| \geq \frac{\xi}{2}\right) - P\{|\langle x, u\rangle| \geq \xi\}\Big) \\
&\geq \xi\Big(\beta + \frac{1}{2n}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i, u\rangle| \geq \frac{\xi}{2}\right) - P\{|\langle x, u\rangle| \geq \xi\}\Big) \,.
\end{aligned} \tag{3.101}$$

where in the second inequality we use $\beta = P\{|\langle x, u\rangle| \geq \xi\}$ to get the third inequality.

Let $\phi_\xi : \mathbb{R}_+ \to [0, 1]$ be a function defined as follows

$$\phi_\xi(t) = \begin{cases} \frac{1}{2} & \text{if } t \geq \xi \\ (t/\xi) - \frac{1}{2} & \text{if } \frac{\xi}{2} \leq t \leq \xi \\ 0 & \text{otherwise} \end{cases} \,, \tag{3.102}$$

so that, for every $t \in \mathbb{R}_+$, $\mathbb{I}_{[\xi/2,\infty)}(t) \geq \phi_\xi(t)$ and $\phi_\xi(t) \geq \mathbb{I}_{[\xi,\infty)}(t)$. Therefore,

$$\frac{1}{2}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i, u\rangle| \geq \frac{\xi}{2}\right) \geq \sum_{i=1}^{n}\phi_\xi(|\langle x_i, u\rangle|) \,. \tag{3.103}$$

$$P\{|\langle x, u\rangle| \geq \xi\} \leq E[\phi_\xi(|\langle x, u\rangle|)] \,. \tag{3.104}$$

Therefore,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|Xu\|_2 &\geq \xi\Big(\beta + \frac{1}{n}\sum_{i=1}^{n}\phi_\xi(|\langle x_i, u\rangle|) - E[\phi_\xi(|\langle x_i, u\rangle|)]\Big) \\
\inf_{u\in A}\frac{1}{\sqrt{n}}\|Xu\|_2 &\geq \xi\Big(\beta - \sup_{u\in A}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_\xi(|\langle x_i, u\rangle|) - E[\phi_\xi(|\langle x, u\rangle|)]\right|\Big) \,.
\end{aligned} \tag{3.105}$$

$\phi_\xi(t)$ is bounded by 1/2. Therefore by bounded differences [20] inequality and the Gene-Zinn symmetrization inequality [126], with probability atleast $1 - 2e^{-\tau}$,

$$\begin{aligned}
&\sup_{u \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \phi_\xi(|\langle x_i, u \rangle|) - E[\phi_\xi(|\langle x, u \rangle|)] \right| \\
&\leq E \sup_{u \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \phi_\xi(|\langle x_i, u \rangle|) - E[\phi_\xi(|\langle x, u \rangle|)] \right| + \sqrt{\frac{\tau}{n}} \\
&\leq \frac{2}{\sqrt{n}} E \sup_{u \in A} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \phi_\xi(|\langle x_i, u \rangle|) \right| + \sqrt{\frac{\tau}{n}} \, . \qquad (3.106)
\end{aligned}$$

ϕ is a Lipschitz function with constant $1/\xi$. Therefore by the contraction inequality for Rademacher sums [86],

$$\begin{aligned}
E \sup_{u \in A} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \phi_\xi(|\langle x_i, u \rangle|) \right| &\leq \frac{1}{\xi} E \sup_{u \in A} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \langle x_i, u \rangle \right| \\
&\leq \frac{1}{\xi} E \sup_{u \in A} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \langle x_i, u \rangle \right| \\
&= \frac{c \cdot w_e(A)}{\xi} \, .
\end{aligned}$$

The last equality follows because $e = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i x_i$ is a sub-exponential random vector with $\|e_i\|_{\psi_1} \leq cK, \ 1 \leq i \leq p$ for some constant c. This follows because of the result of Lemma 7. Therefore combining the above inequalities we get,

$$\inf_{u \in A} \frac{1}{\sqrt{n}} \|Xu\|_2 \geq \xi \left(\beta - \frac{c_1 \cdot w_e(A)}{\xi \sqrt{n}} - \sqrt{\frac{\tau}{n}} \right) \, . \qquad (3.107)$$

The result now follows by letting $n \geq \frac{16 c_1^2 w_e^2(A)}{\beta^2 \xi^2}$ and $\tau = \frac{n \beta^2}{16}$ so that $\frac{c_1 \cdot w_e(A)}{\xi \sqrt{n}} \leq \frac{\beta}{4}$ and $\sqrt{\frac{\tau}{n}} = \frac{\beta}{4}$. ■

Theorem 4 *Consider the regularized least squares problem with atomic norm regularizer $R(\theta) = \|\theta\|_{\mathcal{A}}$ having the union of subspaces form. Let $X \in \mathbb{R}^{n \times p}$ be the design matrix whose rows are i.i.d. sub-exponential with sub-exponential norm K. With E_r denoting the error set, let the small-ball condition hold in the error set $A = cone(E_r) \cap S^{p-1}$ with constant β,*

$$\beta = \inf_{u \in A} P(|\langle x, u \rangle| > \xi) \, . \qquad (3.108)$$

Let $\Psi(E_r) = \sup\limits_{u \in E_r} \frac{\|u\|_{\mathcal{A}}}{\|u\|_2}$ *denote the norm compatibility constant. Then with,*

$$n \geq c \cdot \max \left(\begin{array}{l} \Psi^2(E_r)q + \Psi^2(E_r)\log(em/\Psi^2(E_r)) \\ \min\left(q + \log m, q\log^2\left(\frac{ep}{q}\right)\right) \end{array} \right) , \tag{3.109}$$

with probability atleast $1 - \exp(-\eta_0 n\beta^2) - \exp(-\eta_1\sqrt{n}/K^2)$,

$$\inf_{u \in E_r} \frac{1}{n}\|Xu\|_2^2 \geq \frac{\xi^2\beta^2}{8} = \kappa^2 , \tag{3.110}$$

where κ *is the restricted eigenvalue constant.*

Proof: The proof has several components. We will first need the results from the following Lemma where the arguments are very similar to that of Theorem 3.

Lemma 11 *Let* $X \in \mathbb{R}^{n\times p}$ *be a random matrix with i.i.d subexponential rows* $x_i, 1 \leq i \leq n$. *On a given set* $A \subseteq S^{p-1}$ *define,*

1. *The small-ball condition constant* $\beta = \inf\limits_{u \in A} P(|\langle x, u\rangle| > \xi)$,

2. *The collection of sets* $C_\xi = \{\{\mathbb{I}[|\langle x, u\rangle| > \xi]\}; u \in A\}$ *is a class of* $\{0, 1\}$*-valued functions of VC-dimension at most d.*

Then if $n \geq c \cdot d$, *with probability atleast* $1 - \exp(-\eta_2\beta^2 n)$ *for some positive constant* η_2

$$\inf_{u \in A} \frac{1}{\sqrt{n}}\|Xu\|_2 \geq \frac{\xi\beta}{2} . \tag{3.111}$$

Proof: For any $u \in A$ by Lyapunov's inequality

$$\frac{1}{\sqrt{n}}\|Xu\|_2 = \left(\frac{1}{n}\sum_{i=1}^n |\langle x_i, u\rangle|^2\right)^{\frac{1}{2}} \geq \frac{1}{n}\sum_{i=1}^n |\langle x_i, u\rangle| . \tag{3.112}$$

Now by Markov's inequality for some constant $\xi > 0$

$$\frac{1}{n}\sum_{i=1}^n |\langle x_i, u\rangle| \geq \frac{\xi}{2n}\sum_{i=1}^n \mathbb{I}\left(|\langle x_i, u\rangle| \geq \frac{\xi}{2}\right) . \tag{3.113}$$

Therefore, we get

$$\begin{aligned}\frac{1}{\sqrt{n}}\|Xu\|_2 &\geq \frac{\xi}{2n}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i,u\rangle| \geq \frac{\xi}{2}\right)\\ &\geq \xi\Big(P\{|\langle x,u\rangle| \geq \xi\} + \frac{1}{2n}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i,u\rangle| \geq \frac{\xi}{2}\right) - P\{|\langle x,u\rangle| \geq \xi\}\Big)\\ &\geq \xi\Big(\beta + \frac{1}{2n}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i,u\rangle| \geq \frac{\xi}{2}\right) - P\{|\langle x,u\rangle| \geq \xi\}\Big)\\ &\geq \xi\Big(\beta - \sup_{u\in A}\left|\frac{1}{2n}\sum_{i=1}^{n}\mathbb{I}\left(|\langle x_i,u\rangle| \geq \frac{\xi}{2}\right) - E[\mathbb{I}(|\langle x,u\rangle| \geq \xi)]\right|\Big)\,. \qquad (3.114)\end{aligned}$$

where in the third inequality we use $\beta = P\{|\langle x,u\rangle| \geq \xi\}$. Equation (3.114) above is similar to equation (3.101) in Theorem 3. Therefore using similar arguments as in the proof of 3 we get the following with probability atleast $1 - 2\exp(-\tau)$ with ϵ_i denoting Rademacher random variables,

$$\begin{aligned}\frac{1}{\sqrt{n}}\|Xu\|_2 &\geq \xi\left(\beta - \frac{2}{\sqrt{n}\xi}E\sup_{u\in A}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\langle x_i,u\rangle\right| - \sqrt{\frac{\tau}{n}}\right)\\ &\geq \xi\left(\beta - \frac{2}{\sqrt{n}\xi}E\sup_{u\in A}\left|\sum_{i=1}^{n}\langle e,u\rangle\right| - \sqrt{\frac{\tau}{n}}\right)\\ &\geq \xi\left(\beta - \frac{c_1\sqrt{d}}{\sqrt{n}\xi} - \sqrt{\frac{\tau}{n}}\right)\\ &\geq \xi\left(\beta - \frac{\beta}{4} - \frac{\beta}{4}\right) = \frac{\xi\beta}{2}\,, \qquad (3.115)\end{aligned}$$

where the first inequality follows from (3.105) and (3.106) from Theorem 3, in the second inequality $e = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i x_i$ is a sub-exponential random vector with $\|e_i\|_{\Psi_1} \leq cK, 1 \leq i \leq p$, the third inequality uses the fact that $VC(C) \leq d$ [126] and in the last inequality we use the assumption $n > \frac{16c_1^2 d}{\beta^2\xi^2} = c\cdot d$ and use $\tau = \frac{n\beta^2}{16}$. This proves the result stated in Lemma 11 ■

The following corollary uses the result from Lemma 11 on sets of vectors of the form $v = \sum_{i=1}^{k_1} c_i a_i$ where $a_i \in \mathcal{A}$ and $\mathcal{A}$ is the atomic set. Note that v will be in an $k_1 q = s_1$-sparse subspace of $\mathbb{R}^p$, that is, $\|v\|_0 = s_1$.

Corollary 10 *Let $X \in \mathbb{R}^{n\times p}$ be a random matrix with sub-exponential i.i.d. rows $x_i \in \mathbb{R}^p$. Let*

B be the set,

$$B = \{v \in \mathbb{R}^p \mid v = \sum_{i=1}^{k_1} c_i a_i, c_i > 0, a_i \in \mathcal{A}, \|v\|_2 = 1, \|v\|_0 = k_1 q = s_1\} . \tag{3.116}$$

If $n \geq c \cdot k_1 q + k_1 \log(em/k_1)$, then with probability atleast $1 - \exp(-\eta_0 n \beta^2)$

$$\inf_{v \in B} \frac{1}{\sqrt{n}} \|Xv\|_2 \geq \frac{\xi\beta}{2} . \tag{3.117}$$

Proof: Consider the set D in any s_1-dimensional subspace,

$$D = \{v \in \mathbb{R}^{s_1} \mid \|v\|_2 = 1\} . \tag{3.118}$$

The VC-dimension of a class of half-spaces in $\mathbb{R}^s$ is at most s and thus for every $\xi > 0$ the VC dimension of $C_\xi = \{\{\mathbb{I}_{|\langle x,v\rangle| > \xi}\} : v \in D\}$ is at most $c_2 s_1$. Therefore from the result of Lemma 11,

$$P\left(\inf_{v \in D} \frac{1}{\sqrt{n}} \|Xv\|_2 \geq \frac{\xi\beta}{2}\right) \geq 1 - \exp(-\eta_2 n \beta^2) . \tag{3.119}$$

The set B is the union of such $\binom{m}{k_1}$ spheres D, $B = \cup_{\binom{m}{k_1}} D$.

Therefore using a union bound argument,

$$\begin{aligned} P\left(\inf_{v \in B} \frac{1}{\sqrt{n}} \|Xv\|_2 \geq \frac{\xi\beta}{2}\right) &\geq 1 - \binom{m}{k_1} \exp(-\eta_2 n \beta^2) \\ &\geq 1 - \exp(-\eta_2 n \beta^2 + k_1 \log(em/k_1)) , \end{aligned}$$

where we have used the inequality $\binom{m}{k_1} \leq \left(\frac{em}{k_1}\right)^{k_1} = \exp(k_1 \log(\frac{em}{k_1}))$. Therefore with $n > \frac{c(s_1 + k_1 \log(em/k_1))}{\beta^2}$ for some positive constant c the above inequality is satisfied with probability atleast $1 - \exp(-\eta_0 n \beta^2)$. ■

The Lemma below uses Maurey's empirical approximation argument recently used in [85, 108] to extend the argument from set B to the error set E_r.

Lemma 12 *Consider any vector $u \in \mathbb{R}^p$ which can be expressed as a linear combination of vectors from the atomic set $\mathcal{A}$ as follows,*

$$u = \sum_{i=1}^{m} c_i a_i,\ c_i \geq 0,\ a_i \in \mathcal{A}_i . \tag{3.120}$$

Then with the set B as defined in Corollary 10 and $\kappa_1^2 = \frac{\xi^2\beta^2}{4}$ the following holds,

$$\frac{1}{n}\|Xu\|_2^2 \geq \kappa_1^2\|u\|_2 - \frac{\|u\|_{\mathcal{A}}^2}{k_1 - 1} \sup_{a_i \in \mathcal{A}_i, 1 \leq i \leq m} \frac{1}{n}\langle Xa_i, Xa_i\rangle \ . \tag{3.121}$$

Proof: Let v be a random vector defined as,

$$P(v = \|u\|_{\mathcal{A}} a_i) = \frac{c_i}{\|u\|_{\mathcal{A}}} = \mu_i \ . \tag{3.122}$$

As a result, $E[v] = u$. Let $v_1, v_2, \dots, v_{k_1}$ be independent copies of v and define $z = \frac{1}{k_1}\sum_{i=1}^{k_1} v_i$. Clearly $\frac{z}{\|z\|_2} \in B$, where B is the set as defined in Corollary 10. Therefore from the results of Corollary 10,

$$\frac{1}{n}\|Xz\|_2^2 \geq \frac{\xi^2\beta^2}{4}\|z\|_2^2 = \kappa_1^2\|z\|_2^2 \tag{3.123}$$

$$\Rightarrow \frac{1}{n}E\|Xz\|_2^2 \geq \kappa_1^2 E\|z\|_2^2 \ , \tag{3.124}$$

where the expectation is taken w.r.t. the random variable z. Now from the definition of the random variable z,

$$E\|Xz\|_2^2 = \frac{1}{k_1^2}\sum_{i,j\in[1,\dots,k_1]} E\langle Xv_i, Xv_j\rangle \tag{3.125}$$

$$= \frac{1}{k_1^2}\sum_{i,j\in[1,\dots,k_1], i\neq j} E\langle Xv_i, Xv_j\rangle + \frac{1}{k_1^2}\sum_{i,j\in[1,\dots,k_1], i=j} E\langle Xv_i, Xv_j\rangle \tag{3.126}$$

$$= \frac{k_1(k_1-1)}{k_1^2}\langle Xu, Xu\rangle + \frac{k_1}{k_1^2}\sum_{i=1}^{m}\frac{c_i}{\|u\|_{\mathcal{A}}}\|u\|_{\mathcal{A}}^2\langle Xa_i, Xa_i\rangle \tag{3.127}$$

$$= \frac{k_1(k_1-1)}{k_1^2}\langle Xu, Xu\rangle + \frac{\|u\|_{\mathcal{A}}^2}{k_1}\sum_{i=1}^{m}\mu_i\langle Xa_i, Xa_i\rangle \ . \tag{3.128}$$

With the same argument as above and taking $X = \mathbb{I}_{p\times p}$ the identity matrix in p-dimensions, we get,

$$E\|z\|_2^2 = \frac{k_1(k_1-1)}{k_1^2}\|u\|_2^2 + \frac{\|u\|_{\mathcal{A}}^2}{k_1}\sum_{i=1}^{m}\mu_i\|a_i\|_2^2 \ . \tag{3.129}$$

Combining (3.124), (3.128), (3.129) and with some algebraic manipulation we get the following

$$\frac{1}{n}\|Xu\|_2^2 \geq \kappa_1^2\|u\|_2 - \frac{\|u\|_{\mathcal{A}}^2}{k_1 - 1}\left(\sum_{i=1}^{m}\mu_i\frac{1}{n}\langle Xa_i, Xa_i\rangle - \sum_{i=1}^{m}\mu_i\|a_i\|_2^2\right) \tag{3.130}$$

Noting that $\sum_{i=1}^{m} \mu_i = 1$ we get the following,

$$\frac{1}{n}\|Xu\|_2^2 \geq \kappa_1^2\|u\|_2 - \frac{\|u\|_{\mathcal{A}}^2}{k_1 - 1} \sup_{a_i \in \mathcal{A}_i, 1 \leq i \leq m} \frac{1}{n}\langle Xa_i, Xa_i\rangle \ . \tag{3.131}$$

This completes the proof of the Lemma. ■

Now applying Lemma 12 for any vector in the error set $u \in A = E_r \cap S^{p-1}$ and noting that $\|u\|_{\mathcal{A}}^2 \leq \Psi^2(E_r)\|u\|_2^2 = \Psi^2(E_r)$ we get,

$$\inf_{u \in E_r} \frac{1}{n}\|Xu\|_2^2 \geq \kappa_1^2 - \frac{\Psi^2(E_r)}{k_1 - 1} \sup_{a_i \in \mathcal{A}_i, 1 \leq i \leq m} \frac{1}{n}\langle Xa_i, Xa_i\rangle \ . \tag{3.132}$$

For bounding $\sup_{a_i \in \mathcal{A}_i, 1 \leq i \leq m} \frac{1}{n}\langle Xa_i, Xa_i\rangle$ we will require the result from Theorem 3.3 and Lemma 3.5 from [3]. Their result adapted to our setting is stated below.

Theorem 8 *Let X be a $n \times p$ matrix and $D \subseteq \mathbb{R}^p$ be the set of all q-sparse vectors. Then for every $p \leq \exp(c_1\sqrt{n}/K^2)$,*

$$\sup_{z \in D} \frac{1}{\sqrt{n}}\|Xz\|_2 \leq 1 + c_2K^2\sqrt{\frac{q}{n}} \log\left(\frac{ep}{q\sqrt{q/n}}\right) + \frac{\sqrt{2}-1}{2} \tag{3.133}$$

with probability atleast $1 - c_2\exp(-c_1\sqrt{n}/K^2)$

Aplying the result for any q sparse vector we get when $n > c_3K^2q\log^2\left(\frac{ep}{q}\right)$, with probability atleast $1 - c_2\exp(-c_1\sqrt{n}/K^2)$, we get

$$\sup_{a_i \in \mathcal{A}_i, 1 \leq i \leq m} \frac{1}{n}\langle Xa_i, Xa_i\rangle \leq 2 \ . \tag{3.134}$$

A stronger result is possible when $\log m \leq \log p$. Apply (3.133) to a q dimensional subspace A_i. When $n \geq c_3K^2q$ then with probability atleast $1 - c_2\exp(-c_1\sqrt{n}/K^2)$

$$\sup_{a_i \in \mathcal{A}_i} \frac{1}{n}\langle Xa_i, Xa_i\rangle \leq 2 \ . \tag{3.135}$$

There are m such q dimensional subspaces in the atomic set $\mathcal{A}$. Therefore by a union bound argument when $n \geq c_4(K^2q + \log m)$ with probability atleast $1 - c_2\exp(-c_1\sqrt{n}/K^2)$

$$\sup_{a_i \in \mathcal{A}_i, 1 \leq i \leq m} \frac{1}{n}\langle Xa_i, Xa_i\rangle \leq 2 \ . \tag{3.136}$$

Therefore combining (3.134) and (3.136) we get when n is $c \cdot \min\{q + \log m, q\log^2\left(\frac{ep}{q}\right)\}$ with probability atleast $1 - c_2\exp(-c_1\sqrt{n}/K^2)$,

$$\sup_{a_i \in \mathcal{A}_i, 1\leq i\leq m} \frac{1}{n}\langle Xa_i, Xa_i\rangle \leq 2\ . \tag{3.137}$$

Therefore combining the result from (3.132) and (3.137) with high probability we get,

$$\inf_{u\in E_r} \frac{1}{n}\|Xu\|_2^2 \geq \kappa_1^2 - \frac{2\Psi^2(E_r)}{k_1 - 1}\ . \tag{3.138}$$

Choosing $k_1 \geq \frac{4\Psi^2(E_r)}{\kappa_1^2} + 1$ and using the condition on n from Corollary 10 we get when,

$n \geq c\cdot\max\left\{\Psi^2(E_r)q + \Psi^2(E_r)\log(em/\Psi^2(E_r)), \min\left\{q + \log m, q\log^2\left(\frac{ep}{q}\right)\right\}\right\}$

with probability atleast $1 - \exp(-\eta_0 n\beta^2) - c_2\exp(-c_1\sqrt{n}/K^2) = 1 - \exp(-\eta_0 n\beta^2) - \exp(-\eta_1\sqrt{n}/K^2)$, we get

$$\inf_{u\in E_r} \frac{1}{n}\|Xu\|_2^2 \geq \frac{\kappa_1^2}{2} = \frac{\xi^2\beta^2}{8} = \kappa\ , \tag{3.139}$$

which proves Theorem 4. ■

Chapter 4

High-Dimensional Structured Quantile Regression

4.1 Introduction

In this chapter, we consider the problem of structured quantile regression in high-dimensions, which can be posed as follows: given the response variable y_i and covariates x_i the τth conditional quantile function of y_i given x_i is given by: $F^{-1}_{y_i|x_i}(\tau|x_i) = \langle x_i, \theta^*_\tau \rangle, \tau \in (0,1)$ for some structured θ^*_τ whose structure can be captured by a suitable atomic norm $R(\cdot)$, e.g., ℓ_1-norm for sparsity, ℓ_1/ℓ_2 norm for group sparsity, etc. Here $F^{-1}_{y_i|x_i}(\cdot)$ is the inverse of the conditional distribution function of y_i given x_i. We consider the following regularized estimator for the structured quantile regression problem:

$$\begin{aligned}\hat{\theta}_{\lambda_n} &:= \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\ \mathcal{L}_\tau(\theta; X, y) + \lambda_n R(\theta) \\ &:= \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \langle x_i, \theta \rangle) + \lambda_n R(\theta)\ , \qquad (4.1)\end{aligned}$$

where $\rho_\tau(u) = (\tau - \mathbb{I}(u \leq 0))u$ is the asymmetric absolute deviation function [80], $\mathbb{I}(\cdot)$ is the indicator function. The goal is to get nonasymptotic bounds on the estimation error $\|\hat{\theta}_{\lambda_n} - \theta^*\|_2$.

Many previous papers analyze the asymptotic performance of the estimator in (4.1) [91, 76, 139, 146, 134]. In the non-asymptotic setting of interest in this chapter, special cases of the estimator in (4.1) have been studied in recent literature [16, 78, 54], primarily focusing on specific norms like the ℓ_1-norm and ℓ_1/ℓ_2 non-overlapping group sparse norm. In contrast, our analysis is applicable to any atomic norm $R(\cdot)$ giving considerable flexibility in choosing a suitable structure for real world problems, e.g., hierarchical sparsity, k-support norm, OWL norm etc. More recently [5] consider the more general problem of norm regularized regression with Lipschitz loss functions which includes (4.1) as a special case. They derive similar results to ours for bounds on the estimation error, but their analysis differs significantly and, in our opinion, does not leverage and highlight the key geometric and statistical characteristics of the problem.

In the setting of norm regularized regression with square loss, including the widely studied Lasso estimator [123, 105, 11], the sample complexity n_0 of the estimator gets determined by a certain restricted strong convexity (RSC) property which simplifies to the restricted eigenvalue (RE) condition on the matrix $X^T X$ [17, 105]; in the noiseless setting, i.e., when $\omega_i = 0$, the sample complexity determines a phase transition phenomenon so that the probability of recovering the structured θ^* is minimal when $n < n_0$, but one can exactly recover θ^* with high probability when $n > n_0$. Our work gives an equivalent sample complexity characterization for structured quantile regression, which was not highlighted in prior work. The challenge in characterizing RSC in the context of quantile regression stems partly from the non-smoothness of the objective, so one has to work with sub-gradients. However, the unique aspect stems from the geometry of quantile regression, or as the authoritative book on the topic puts it: "How quantile regression works?" [80][Section 2.2]. In quantile regression, the n samples get divided into three subsets: ν samples which get exactly interpolated, i.e., $y_i = \langle x_i, \hat{\theta} \rangle$, $(n-\nu)\tau$ samples which lie below the curve, i.e., $y_i < \langle x_i, \hat{\theta} \rangle$, and $(n-\nu)(1-\tau)$ samples which lie above the curve, i.e., $y_i > \langle x_i, \hat{\theta} \rangle$. Note that when $\nu = n$ all samples are interpolated, the loss is zero and the same $\hat{\theta}$ is a solution for all quantiles τ. Quantile regression then is clearly not working. The Number of InterPolated Samples (NIPS) ν is an important quantity, inherent to structure in θ^*, and determines the sample complexity for recovery of structured quantile regression estimator (4.1). In fact, we show that when $n > \nu$, the RSC condition associated with the estimator in (4.1) is satisfied. When there is no structure in θ^*, then $\nu \leq c \cdot p$ for some constant c, and hence quantile regression needs $n > c \cdot p$ samples to work. However, when θ^* has structure, such as sparsity or group sparsity, ν can be substantially smaller than p. Specifically we show that ν is of the order of square of Gaussian width of the error set [122, 33] for a class of atomic norms which includes ℓ_1, ℓ_1/ℓ_2 group sparse, k-support [7] and the OWL [19] norms. For example, when θ^* is sparse with s non-zero entries, we show that $\nu \leq c \cdot s \log p$ for some constant c.

When $n > \nu$ and the RSC condition is satisfied, building on recent developments in high-dimensional estimation [105, 11], we derive non-asymptotic bounds on the estimation error $\|\hat{\theta}_n - \theta^*\|_2$ under the assumption that the regularization parameter $\lambda_n \geq 2R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y)))$, where $R^*(\cdot)$ is the dual norm of $R(\cdot)$. While the condition on λ_n looks complex, with its dependency on the dual norm and θ^*, we simplify the inequality and show that it is sufficient to set λ_n based on the Gaussian width [122] of the unit norm ball for $R(\cdot)$ [11, 119]‘. Our analysis and results on the estimation error bound for quantile regression, interestingly, has the same order as that for regularized least squares regression for general norms [11]. In contrast to the least squares loss the quantile loss is more robust as the estimation error is independent of the two norm of the noise. We discuss results for the ℓ_1, ℓ_1/ℓ_2 group sparse and k-support norms as examples, precisely characterizing the sample complexity for recovery and non-asymptotic error bounds. Specifically, our results for the ℓ_1-norm matches those from existing literature on sparse quantile regression [16].

The rest of the chapter is organized as follows. In Section 4.2, we discuss the problem formulation along with assumptions and review the general framework for analyzing regularized estimation problems. In Section 4.3, we analyze the number of interpolated samples and establish precise sample complexities for a class of atomic norms in terms of the Gaussian widths of sets. In Section 4.4 we establish key ingredients of the analysis and provide the main bound. We present experimental results in Section 4.5.

Notations: As for notations, we denote constants by $\nu, \phi, \kappa, \eta, c, c_1, \dots$ whose values can change from one line to the next. The notation $x = O(y)$ denotes $x \leq c \cdot y$ for some constant c. Similarly $x = \Theta(y)$ denotes that there exist constants c, c_1 such that $c \cdot y \leq x \leq c_1 \cdot y$.

4.2 Background, Preliminaries and Contributions

In this section, we formally define the problem, introduce background literature on high-dimensional estimation relevant to our work and highlight key results. We also define quantities and notations used throughout the chapter. We will denote absolute constants by $c, c_1, c_2, \dots$ whose values can change from line to line.

4.2.1 Problem Formulation

We outline assumptions on the data and estimator. We assume the rows of the design matrix $X \in \mathbb{R}^{n \times p}$ are i.i.d. sub-Gaussian satisfying the following properties. Similar assumptions are made in all prior work in high dimensional statistics [124, 84, 119].

1. **Sub-Gaussian marginals:** The rows of the design matrix have sub-Gaussian norm K, i.e., for any row x we have $\|x\|_{\psi_2} = \sup\limits_{u \in S^{p-1}} \frac{1}{a}(E|\langle x, a \rangle|^a)^{1/a} \leq K, \; \forall a \geq 1$

2. **Centering:** The rows have zero mean: $E[x] = 0$.

3. **Nondegeneracy:** There is a positive constant α such that for all $u \in S^{p-1}$: $E[|\langle x, u\rangle|] > \alpha > 0$.

4. **Isotropic rows:** The covariance is the identity matrix: $E[xx^T] = \mathbb{I}_{p\times p}$

We consider a parametric quantile regression model where the τth conditional quantile function of the response variable given covariates is given by,

$$F^{-1}_{y_i|x_i}(\tau|x_i) = \langle x_i, \theta^*_\tau\rangle, \theta^*_\tau \in \mathbb{R}^p, \tau \in (0,1)\ , \tag{4.2}$$

where $F^{-1}_{y_i|x_i}$ is the inverse of the conditional distribution function of y_i given x_i. The conditional density of y_i evaluated at the conditional quantile $\langle x_i, \theta^*_\tau\rangle$ is bounded away from zero uniformly for all τ, that is, $f_{y_i|x_i}(\langle x_i, \theta^*_\tau\rangle) > \underline{f} > 0$ for all τ and all x_i. The goal is to estimate parameter $\hat{\theta}_\tau$ close to θ^*_τ using n observations of the data when $n < p$. The estimator in this chapter belongs to the family of regularized estimators and is of the form:

$$\hat{\theta}_{\lambda_n,\tau} := \operatorname{argmin}_{\theta\in\mathbb{R}^p} \mathcal{L}_\tau(\theta; X, y) + \lambda_n R(\theta)\ , \tag{4.3}$$

where $\mathcal{L}_\tau(\theta; X, y) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \langle x_i, \theta\rangle)$, $\rho_\tau(\cdot)$ is the quantile loss function and $R(\cdot)$ is any atomic norm. Examples of atomic norms we consider in this chapter are the ℓ_1, ℓ_1/ℓ_2 non-overlapping group sparse norm and the k-support norm. The assumptions on the data are much weaker compared to previous work on mean regression with the least squares loss instead of the quantile loss. For example, consider data to be generated according to the linear model $y = X\theta^* + \omega$, where ω is the noise vector. While previous work on mean regression [105, 33, 11, 119, 130] assume the noise vector to have i.i.d. sub-Gaussian or sub-exponential elements, for quantile regression we do not make any other assumption except that the elements are i.i.d. More specifically the noise can be heavy-tailed, bimodal, heteroscedastic as in the location-scale model where $\omega_i = \langle x_i, \eta\rangle\epsilon_i,\ \eta \in \mathbb{R}^p$ and ϵ_i is any noise independent of x_i and so on and so forth.

4.2.2 High-dimensional Estimation and Key Results

Our analysis will be based on the general analysis framework outlined in [11, 105]. We give a brief overview of the main components of the analysis and also summarize the key results in the present work.

Error Set: For any quantile τ, let the estimated parameter $\hat{\theta} = \theta^* + \Delta$. If the regularization parameter is sufficiently large (which we characterize subsequently), for any loss function [11] show that the error vector Δ belongs to the following error set $E_r \subseteq \mathbb{R}^p$,

$$E_r = \left\{ \Delta \in \mathbb{R}^p \;\middle|\; R(\theta^* + \Delta) \leq R(\theta^*) + \frac{1}{2} R(\Delta) \right\} . \tag{4.4}$$

It is straightforward to observe that for the equivalent constrained formulation [33, 109], for any loss function $\mathcal{L}(\cdot)$

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \mathcal{L}(\theta; X, y) \;\text{ s.t. }\; R(\theta) \leq R(\theta^*) , \tag{4.5}$$

the error vector belongs to the following set,

$$E_c = \{ \Delta \in \mathbb{R}^p \mid R(\theta^* + \Delta) \leq R(\theta^*) \} . \tag{4.6}$$

Prior results [33, 11] characterize the sample complexity for recovery for the least squares loss (minimum number of samples required to satisfy the Restricted Eigenvalue (RE) condition defined below) in terms of the Gaussian widths of $A_r = \text{cone}(E_r) \cap S^{p-1}$ and $A_c = \text{cone}(E_c) \cap S^{p-1}$ denoted as $w(A_r)$ and $w(A_c)$ respectively. It is evident from the definition of E_r and E_c that $E_r \subset E_c$ and consequently $w(A_r) > w(A_c)$. [11] establish an upper bound of the form $w(A_r) \leq c \cdot w(A_c)$ where c is a constant, thus showing that the Gaussian widths only differ by a multiplicative factor.

Regularization parameter: For any loss function $\mathcal{L}(\cdot)$, [11, 105] show that the error set characterization holds true only if the regularization parameter satisfies the following assumption,

$$\lambda \geq \beta R^*(\nabla_\theta \mathcal{L}(\theta^*; X, y)) , \tag{4.7}$$

where $\nabla_\theta \mathcal{L}(\theta^*; X, y)$ denotes the gradient of the loss function evaluated at θ^*. Note that since X, y are random quantities the r.h.s. in (4.7) is a random quantity.

For the least squares loss and linear model $y = X\theta^* + \omega$, [11] show the following relationship with high probability,

$$R^*(\nabla_\theta \mathcal{L}(\theta^*; X, y)) = R^*\left(\frac{1}{n} X^T \omega \right) \leq \frac{c \cdot \|\omega\|_2 \cdot w(\Omega_R)}{n} , \tag{4.8}$$

where $\Omega_R = \{u \mid R(u) \leq 1\}$ is the unit norm ball, $w(\cdot)$ denotes the Gaussian width and c is any positive constant. The regularization parameter thus depends on the ℓ_2 norm of the noise vector. In contrast, with the quantile regression loss function we prove the following relationship with high probability,

$$R^*(\nabla_\theta \mathcal{L}(\theta^*; X, y)) \leq \frac{c \cdot \sqrt{n\tau(1-\tau)} \cdot w(\Omega_R)}{n} \, . \tag{4.9}$$

The key difference between (4.8) and (4.9) is the independence of (4.9) from the noise vector. As we will see later, (4.9) can be computed from known quantities X, τ. In [16], the independence of the regularization parameter on unknown quantities like the noise is referred to as the regularization parameter having a *pivotal* distribution.

Restricted Strong Convexity (RSC): In problems like Lasso and basis pursuit [123, 42] a phase transition phenomenon is observed where the probability of parameter estimation increases sharply whenever the number of observed data samples crosses a particular threshold. In the general analysis framework of [11, 105], this is characterized by the loss function satisfying the following RSC condition:

$$\begin{aligned} \inf_{u \in E_r} \delta\mathcal{L}(\theta^*, u; X, y) = \inf_{u \in E_r} \left(\mathcal{L}(\theta^* + u; X, y) - \mathcal{L}(\theta^*; X, y) - \langle \nabla \mathcal{L}(\theta^*; X, y), u \rangle\right) \\ \geq \kappa \|u\|^2 \, . \end{aligned} \tag{4.10}$$

For the least squares loss function the RSC condition is equivalent to the following restricted eigenvalue (RE) condition [17]:

$$\inf_{u \in E_r} \frac{1}{n} \|Xu\|_2^2 \geq \kappa \|u\|_2^2 \, . \tag{4.11}$$

It is well established in prior literature that sub-Gaussian design matrices [31, 30, 17, 105, 33, 11, 53] satisfy (4.11) with high probability when the number of samples $n \geq c \cdot w^2(A_r))$, where $w(\cdot)$ denotes the Gaussian width, $A_r = \text{cone}(E_r) \cap S^{p-1}$ and c is any constant. For example, the RE condition is satisfied for the Lasso problem when the number of samples satisfies $n \geq c \cdot s \log p$.

We show that the RSC condition on the loss function evaluates to the following quantity when the loss is the quantile loss:

$$\inf_{u \in E_r} \frac{1}{n} \sum_{i=1}^{n} \int_0^{\langle x_i, u \rangle} \left(\mathbb{I}(y_i - \langle x_i, \theta^* \rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \theta^* \rangle \leq 0)\right) dz \, . \tag{4.12}$$

Denote the number of interpolated samples (NIPS) by $\nu = \sup_{u \in E_r} |Z| = \sup_{u \in E_r} |\{i \mid y_i = \langle x_i, \theta^* + u \rangle\}|$. We observe that when $\nu = n$ the RSC quantity evaluates to zero. We show in

Section 4.3 that ν gets determined by the structure assumed for θ^*. Specifically we show that $\nu < c \cdot \max\{\Psi^2(E_r)(q + \log(em)), w^2(A)\}$, where $\Psi(E_r) = \sup_{u \in E_r} \frac{R(u)}{\|u\|_2}$ is the norm compatibility constant in the error set, q and m are the subspace dimension and number of subspaces of the atomic norm, E_r, A are the error sets defined earlier and $w(\cdot)$ denotes the Gaussian width. For example for the ℓ_1 norm $\nu = O(s \log p)$ rather than the ambient dimensionality p. Thus, the sum over n points in of the RSC condition simply reduces to the sum over the $(n - \nu)$ points which are not interpolated, and will ensure the RSC condition when $n > \nu$. Once the number of samples is sufficiently larger than ν we show that the following RSC condition is satisfied,

$$\inf_{u \in E_r} \frac{1}{n} \sum_{i=1}^{n} \int_0^{\langle x_i, u \rangle} \left(\mathbb{I}(y_i - \langle x_i, \theta^* \rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \theta^* \rangle \leq 0) \right) dz \geq \underline{f} \kappa \|u\|_2^2 \,, \tag{4.13}$$

where $\underline{f}$ is the lower bound of the conditional density of y_i evaluated at the conditional quantile $\langle x_i, \theta^*_\tau \rangle$.

Estimation error bounds: Assuming the RSC condition and lower bound on the regularization parameter are satisfied, [11] derive the following result with high probability,

$$\|\Delta_{\lambda_n}\|_2 = \|\hat{\theta}_{\lambda_n} - \theta^*\|_2 \leq c \cdot \frac{\Psi(E_r)\lambda_n}{\kappa} \,, \tag{4.14}$$

where $\Psi(E_r) = \sup_{u \in E_r} \frac{R(u)}{\|u\|_2}$ is the norm compatibility constant [105, 11]. For example, when $R(\cdot)$ is the ℓ_1 norm, $\Psi(E_r) = \sqrt{s}$. Combining with the results for the regularization parameter and restricted eigenvalue condition, we show that when $n \geq c \cdot s \log p$ with high probability,

$$\|\Delta\|_2 \leq c \cdot \frac{\sqrt{s \log p}}{\underline{f} \kappa \sqrt{n}} \,. \tag{4.15}$$

4.3 Number of InterPolated Samples (NIPS)

In this section, we reveal some insights on the geometry of the problem. In the high sample, low dimension non penalized setting, when $n >> p$ and $R(\theta) = 0$, the quantile loss is a linear program and hence its solutions are at the vertices, that is, where any p of the n samples are interpolated. Mathematically we define the quantity $Z = \{i : y_i = \langle x_i, \hat{\theta} \rangle = \langle x_i, \theta^* + u \rangle, u \in \mathbb{R}^p\}$ and note that $\nu = \sup_{u \in \mathbb{R}^p} |Z| \leq c \cdot p$ for some constant c. In the high dimensional setting considered in this work, $n < p$ and hence when $R(\theta) = 0$ the number of interpolated samples is $\nu = n$. From an optimization perspective there are multiple such solutions and all solutions are optimal for any τ. But practically quantile regression is not working. Now introducing a regularizer with a suitable choice for the regularization parameter ensures that the error vector lies in a restricted subset of $\mathbb{R}^p$,

$$E_r = \left\{ u \;\middle|\; R(\theta^* + u) \leq R(\theta^*) + \frac{1}{2} R(u) \right\} \subseteq \mathbb{R}^p \; . \tag{4.16}$$

We are now interested in characterizing $\nu = \sup_{u \in E_r} |Z|$ where $Z = \{i : y_i = \langle x_i, \hat{\theta} \rangle = \langle x_i, \theta^* + u \rangle, u \in E_r\}$, that is, the maximum number of interpolated samples with the error restricted to a particular subset of $\mathbb{R}^p$. Again if $\nu = n$, quantile regression is not working. Since there are no restrictions on the number of non-zero elements in the error vector, a first crude estimate will be $\nu \leq \min\{n, p, \|u\|_0\}$, which implies quantile regression will not work unless we have a minimum of p samples. But intuitively the number of interpolated samples should depend on properties of the error set E_r, which the initial crude estimate is failing to take advantage.

Below we state a result which reinforces the intuition of the relation between the number of interpolated samples and the properties of the set E_r. Specifically we show that for the types of atomic norms considered in this work (which includes all popularly known vector norms), the number of interpolated samples does not exceed the product of the square of the norm compatibility constant and the square of the Gaussian width of the unit norm ball. For the norms considered, this is precisely the square of the Gaussian width of the error set E_r. For example for the ℓ_1 norm for an s-sparse parameter vector θ^* this evaluates to an upper bound of $\nu \leq c \cdot s \log p$ with high probability. While the result statement considers sub-Gaussian design matrices, the result can be extended to design matrices sampled from heavy-tailed distributions using arguments similar to [84, 119].

Theorem 9 *Let rows of the design matrix $X \in \mathbb{R}^{n \times p}$ have isotropic sub-Gaussian rows and let θ^* be an s-sparse vector that can be written as a linear combination of k atoms from an atomic set of cardinality m,*

$$\theta^* = \sum_{i=1}^{k} c_i a_i, a_i \in \mathcal{A}, c_i \geq 0, |\mathcal{A}| = m \; . \tag{4.17}$$

Consider the regularized quantile regression problem penalized with the atomic norm $R(\theta) = \|\theta\|_{\mathcal{A}}$,

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \mathcal{L}_\tau(\theta) = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(\theta) + \lambda R(\theta) \; . \tag{4.18}$$

Let $E_r = \left\{ u \;\middle|\; R(\theta^* + u) \leq R(\theta^*) + \frac{1}{2} R(u) \right\}$ denote the error set, let $A = cone(E_r) \cap S^{p-1}$ and let $\lambda \geq R^*(\nabla \mathcal{L}_\tau(\theta^*))$. Then with probability atleast $1 - \exp(-\eta_2 \Psi^2(E_r)(q + \log(em))) -$

$2\exp(-\eta w^2(A))$ the number of interpolated samples,

$$\sup_{u \in E_r} |Z| = \sup_{u \in E_r} |\{i : y_i = \langle x_i, \theta^* + u \rangle,\ u \in E_r\}| \leq \max\{c_1 \Psi^2(E_r)(q + \log(em)), c_2 w^2(A))\}\ , \tag{4.19}$$

where $\Psi(E_r) = \sup_{u \in E_r} \frac{\|u\|_{\mathcal{A}}}{\|u\|_2}$ is the norm compatibility constant, q is the dimension of each of the m subspaces of the atomic norm, and $w(\cdot)$ denotes the Gaussian width of a set and η, η_2, c_1, c_2 are constants.

To understand the intuition behind the proof consider the case of the ℓ_1 norm. We first show that with high probability when the error vector lies in a particular cs-dimensional subspace, with the constant c chosen appropriately, the number of interpolated samples is $c_1 s$. Next we extend this argument to any cs-dimensional subspace by a union bound argument on the $\binom{p}{cs}$ subspaces and show that for an error vector lying in any cs subspace with high probability the number of interpolated samples is upper bounded by $c_2 s \log p$. Finally the argument is extended to all vectors in the error set using the powerful Maurey's empirical approximation argument previously employed in [119, 84, 115].

Suprisingly in prior literature on structured high dimensional quantile regression, the quantity ν and subsequent insights have not been explicitly discussed. This intuition about the importance of ν also shows up in an elegant form in the analysis of the RSC condition in Section 4.2.

Below we provide results for the number of interpolated samples for the ℓ_1, ℓ_1/ℓ_2 non-overlapping group sparse and k-support norms. The results follow from substituting known values for the norm compatibility constant, Gaussian widths of sets, dimension of each subspace and number of subspaces for the different norms from Table 2.2 and 2.1.

Corollary 11 *Let $R(\cdot)$ be the ℓ_1 norm with θ^* being an s-sparse vector. Then for sub-Gaussian design matrices with probability atleast $1 - \exp(-\eta_2(s + s\log p)) - 2\exp(-\eta s \log p)$ the number of interpolated samples,*

$$\nu = \sup_{u \in E_r} |\{i : y_i = \langle x_i, \theta^* + u \rangle\}| \leq c \cdot (s \log p)\ , \tag{4.20}$$

for some constant c.

Proof: The result follows from the result of Theorem 9 after observing that $\Psi(E_r) = \sqrt{s}, q = 1, m = p$ and $w(A) = \Theta(\sqrt{s \log p})$. ■

Corollary 12 *Let $R(\cdot)$ be the ℓ_1/ℓ_2 non-overlapping group sparse norm. Then for sub-Gaussian design matrices with probability atleast* $1-\exp(-\eta_2 s_{\mathcal{G}}(l+\log N_{\mathcal{G}}))-2\exp(-\eta(ls_{\mathcal{G}}+\log N_{\mathcal{G}}))$ *the number of interpolated samples,*

$$\nu = \sup_{u \in E_r} |\{i : y_i = \langle x_i, \theta^* + u\rangle\}| \leq c \cdot s_{\mathcal{G}}(l + \log N_{\mathcal{G}}) \ , \tag{4.21}$$

for some constant c.

Proof: The result follows from the result of Theorem 9 after observing that $\Psi(E_r) = \sqrt{s_{\mathcal{G}}}, q = l, m = N_{\mathcal{G}}$ and $w(A) = \Theta(\sqrt{s_{\mathcal{G}}(l + \log N_{\mathcal{G}})})$. ■

Corollary 13 *Let $R(\cdot)$ be the k-support norm with. Then for sub-Gaussian design matrices with probability atleast* $1 - \exp(-\eta_2(s + s\log\lceil\frac{p}{k}\rceil)) - 2\exp(-\eta(s + s\log\lceil\frac{p}{k}\rceil))$ *the number of interpolated samples,*

$$\nu = \sup_{u \in E_r} |\{i : y_i = \langle x_i, \theta^* + u\rangle\}| \leq c \cdot (s + s\log\lceil p/k\rceil) \ , \tag{4.22}$$

for some constant c.

Proof: The result follows from the result of Theorem 9 after observing that $\Psi(E_r) \leq \sqrt{2s/k}, q = k, m = \binom{p}{k}$ and $w(A) = \Theta(\sqrt{s + s\log\lceil p/k\rceil})$. ■

4.4 Structured Quantile Regression

In this section, we present results for the key components in the general analysis framework of [11] which we briefly described in Section 4.2 of the chapter. We state results on the regularization parameter before establishing sample complexity bounds when the restricted strong convexity condition in is satisfied. Finally we derive an ℓ_2 bound on the estimation error.

4.4.1 Regularization Parameter

The regularization parameter should satisfy the following inquality,

$$\lambda \geq 2R^*(\nabla_\theta \mathcal{L}(\theta^*; X, y)) \ , \tag{4.23}$$

where $\nabla_\theta \mathcal{L}(\theta^*; X, y)$ denotes the gradient of the loss function evaluated at θ^*. Note that since X and y are assumed to be random quantities $R^*(\nabla_\theta \mathcal{L}(\theta^*; X, y))$ is also a random quantity. Theorem 10 gives a result on high probability lower bounds on the quantity $R^*(\nabla_\theta \mathcal{L}(\theta^*; X, y))$.

Theorem 10 *Let $X \in \mathbb{R}^{n \times p}$ be a design matrix with independent isotropic sub-Gaussian rows with sub-Gaussian norm $\|x_i\|_{\psi_2} \leq K$. Define $\Omega_R = \{u \ : \ R(u) \leq 1\}$ the unit norm ball and let $\phi = \sup_u \|u\|_2 / R(u)$. Then the following holds*

$$E\left[R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y))\right] \leq c \frac{K\sqrt{\tau(1-\tau)} w(\Omega_R)}{\sqrt{n}} \ , \tag{4.24}$$

where c is any fixed constant depending only on the sub-Gaussian norm k. Moreover with probability atleast $1 - \eta_1 \exp\left(-\left(\frac{\nu}{\eta_2 \phi K}\right)^2\right)$

$$R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y)) \leq c_1 \frac{K\sqrt{\tau(1-\tau)}(w(\Omega_R) + \nu)}{\sqrt{n}} \ , \tag{4.25}$$

where c_1, η_1, η_2, ν are absolute constants.

A major difference to the least squares loss setting, is the less restrictive assumptions on the noise vector (see for example Theorem 3 and Theorem 4 in [11] where the noise is explicitly assumed to be sub-Gaussian and homoscedastic and the noise enters the analysis through properties of $\|\omega\|_2$). This gives the flexibility of considering, e.g., noise vectors which are heavy tailed or heteroscedastic. Indeed the most interesting applications of quantile regression arise in such settings.

Below we provide bounds for the regularization parameter for different norms by substituting known values of the Gaussian width for the unit norm balls from Table 2.2. Note that the result for the ℓ_1 norm matches the result in Theorem 1 of [16] for the regularization parameter.

Corollary 14 *Let $R(\cdot)$ be the ℓ_1 norm. Then for sub-Gaussian design matrices with rows having sub-Gaussian norm K the following is true with probability atleast $1 - \eta_1 \exp\left(-\left(\frac{\nu}{\eta_2 K}\right)^2\right)$*

$$R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y)) \leq c \frac{K\sqrt{\tau(1-\tau)}(\sqrt{\log p} + \nu)}{\sqrt{n}} \ . \tag{4.26}$$

Proof: The result follows from Theorem 10 after observing that $\phi = 1$ and $w(\Omega_R) = \Theta(\sqrt{\log p})$ ■

Corollary 15 *Let $R(\cdot)$ be the ℓ_1/ℓ_2 non-overlapping group sparse norm. Then for sub-Gaussian design matrices with rows having sub-Gaussian norm K the following is true with* *probability atleast* $1-\eta_1 \exp\left(-\left(\frac{\nu}{\eta_2 K}\right)^2\right)$

$$R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y)) \leq c\frac{K\sqrt{\tau(1-\tau)}(\sqrt{l+\log N_{\mathcal{G}}}+\nu)}{\sqrt{n}} \ . \tag{4.27}$$

Proof: The result follows from Theorem 10 after observing that $\phi = 1$ and $w(\Omega_R) = \Theta(\sqrt{l+N_{\mathcal{G}}})$ ■

Corollary 16 *Let* $R(\cdot)$ *be the* k*-support norm. Then for sub-Gaussian design matrices with rows having sub-Gaussian norm* K *the following is true with probability atleast* $1 - \eta_1 \exp\left(-\left(\frac{\nu}{\eta_2 K}\right)^2\right)$

$$R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y)) \leq c\frac{K\sqrt{\tau(1-\tau)}(\sqrt{k+k\log\lceil\frac{p}{k}\rceil}+\nu)}{\sqrt{n}} \ . \tag{4.28}$$

Proof: The result follows from Theorem 10 after observing that $\phi = 1$ and $w(\Omega_R) = \Theta(\sqrt{\log p})$ ■

4.4.2 Restricted Strong Convexity (RSC)

In the general analysis framework of [11, 105], the loss needs to satisfy the RSC condition:

$$\inf_{u\in E_r} \delta\mathcal{L}(\theta^*, u; X, y) = \inf_{u\in E_r} (\mathcal{L}(\theta^*+u; X, y) - \mathcal{L}(\theta^*; X, y) - \langle\nabla_\theta \mathcal{L}(\theta^*; X, y), u\rangle) \geq \kappa\|u\|^2 \ . \tag{4.29}$$

We begin by providing an intuition for the RSC formulation for the quantile loss. The RSC condition equation (4.29) on the error set E_r evaluates to the following (proof in the appendix),

$$\inf_{u\in E_r} \frac{1}{n}\sum_{i=1}^{n} \int_0^{\langle x_i, u\rangle} (\mathbb{I}(y_i - \langle x_i, \theta^*\rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \theta^*\rangle \leq 0))\, dz \ \geq \kappa\|u\|_2^2. \tag{4.30}$$

Let $\nu = \sup_{u\in E_r} |Z| = \sup_{u\in E_r} |\{i \mid y_i = \langle x_i, \theta^* + u\rangle\}|$ is the number of interpolated samples. For any $n < p$ if the model can interpolate all points, that is, $\nu = n$ then (4.30) evaluates to zero. In general, as shown in Section 4.3, ν gets determined by the structure. For example for the ℓ_1 norm $\nu = O(s\log p)$ rather than the ambient dimensionality p. Thus, the sum over n points in

(4.30) simply reduces to the sum over the $(n - \nu)$ points which are not interpolated, and will ensure the RSC condition when $n > \nu$. The intuition of the NIPS property of Section 4.3 thus shows up elegantly in the RSC condition.

In equation (4.30), let $\xi_i = y_i - \langle x_i, \theta^* \rangle, v_i = \int_0^{\langle x_i, u \rangle} (\mathbb{I}(\xi_i \leq z) - \mathbb{I}(\xi_i \leq 0))$ and consider $\frac{1}{n} \sum_{i=1}^n E[v_i]$, then

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n E[v_i] &= \frac{1}{n} \sum_{i=1}^n \int_0^{\langle x_i, u \rangle} (F_i(\xi_i + z) - F_i(\xi)) \\ &= \frac{1}{n} \sum_{i=1}^n \int_0^{\langle x_i, u \rangle} f_i(\xi_i) z dz + o(1) \\ &= \frac{1}{2n} \sum_{i=1}^n f_i(\xi_i) \langle x_i, u \rangle^2 + o(1) \\ &\geq \frac{\underline{f}}{2n} \|Xu\|_2^2 \geq \frac{\underline{f} \kappa \|u\|_2^2}{2} . \end{aligned}$$

The first line follows from the definition of the cumulative distribution function, the second line by a simple Taylor series expansion, the last line by the assumption that $\underline{f} \leq f_i(\xi_i), \forall i$ and $(1/n) \|Xu\|_2^2 \geq \kappa$, where κ is the restricted eigenvalue (RE) constant. The RE condition is satisfied as the sample complexity bounds for satisfying the NIPS property is of the same order as the RE condition. We can as a result see that RSC is a condition on the minimum eigenvalue of the Jacobian matrix $\frac{1}{n} \sum_{i=1}^n f_i(\xi_i) \langle x_i, u \rangle^2$ restricted to the error set E_r. Prior literature on high dimensional quantile regression has either assumed the RSC condition to be satisfied or has not discussed the RSC condition explicitly, though [55] considers it for the quantile huber loss function.

Theorem 11 *Let $X \in \mathbb{R}^{n \times p}$ have sub-Gaussian rows with sub-Gaussian norm K. Let $0 < \underline{f} < f_i(\langle x_i, \theta^* \rangle)$ be a uniform lower bound on the conditional density around a neighborhood of the τth quantile of the conditional distribution $y_i | x_i$ for all x_i in the support of X. Let the number of samples $n \geq c \cdot w^2(A)$, where $A = cone(E_r) \cap S^{p-1}$ and $E_r = \left\{ u \,| R(\theta^* + u) \leq R(\theta^*) + \frac{1}{2} R(u) \right\}$ is the error set. Also,assume that $\|u\|_2 = \Theta \left(\frac{w(A)}{\underline{f} \sqrt{n}} \right)$. Then, for positive constants ϕ_1, ϕ_2, ν we get with probability atleast $1 - \exp(-\nu^2/2) - \exp\left(-\frac{\phi_1^2 n}{4}\right) - \exp\left(-\eta_2 \phi_2^2 \sqrt{n} w(A)\right)$,*

$$\inf_{u \in E_r} \delta \mathcal{L}_\tau (\theta^*, u; X, y) \geq \kappa \underline{f} \|u\|_2^2 \,, \tag{4.31}$$

for some positive constant κ which depends on the constants ν, ϕ_1, ϕ_2.

Below we instantiate the result for different norms.

Corollary 17 *Let $R(\cdot)$ be the ℓ_1 norm and $X \in \mathbb{R}^{n \times p}$ be a sub-Gaussian design matrix. Let $\underline{f}$ denote a lower bound on the conditional density around a neighborhood of the τth quantile of $y_i|x_i$ for all x_i. Then if $n > c \cdot s \log p$ and the error vector u is such that $\|u\|_2 = \Theta\left(\frac{s \log p}{\underline{f}\sqrt{n}}\right)$ then with probability atleast $1 - \exp(-\nu^2/2) - \exp\left(-\frac{\phi_1^2}{4}\right) - \exp(-\eta_2 \phi_2^2 \sqrt{n}\sqrt{s \log p})$,*

$$\inf_{u \in E_r} \delta\mathcal{L}(\theta^*, u; X, y) \geq \kappa \underline{f} \|u\|_2^2 \,. \tag{4.32}$$

Proof: The result follows from Theorem 11 noting that $w(A) = \Theta(s \log p)$ ■

Corollary 18 *Let $R(\cdot)$ be the ℓ_1/ℓ_2 non-overlapping group sparse norm and $X \in \mathbb{R}^{n \times p}$ be a sub-Gaussian design matrix. Let $\underline{f}$ denote a lower bound on the conditional density around a neighborhood of the τth quantile of $y_i|x_i$ for all x_i. Then if $n > c \cdot (s_{\mathcal{G}} l + s_{\mathcal{G}} \log N_{\mathcal{G}})$ and the error vector u is such that $\|u\|_2 = \Theta\left(\frac{s_{\mathcal{G}} l + s_{\mathcal{G}} \log N_{\mathcal{G}}}{\underline{f}\sqrt{n}}\right)$ then with probability atleast $1 - \exp(-\nu^2/2) - \exp\left(-\frac{\phi_1^2}{4}\right) - \exp(-\eta_2 \phi_2^2 \sqrt{n}\sqrt{s_{\mathcal{G}} l + s_{\mathcal{G}} \log N_{\mathcal{G}}})$,*

$$\inf_{u \in E_r} \delta\mathcal{L}(\theta^*, u; X, y) \geq \kappa \underline{f} \|u\|_2^2 \,. \tag{4.33}$$

Proof: The result follows from Theorem 11 noting that $w(A) = \Theta(s_{\mathcal{G}} l + s_{\mathcal{G}} \log N_{\mathcal{G}})$ ■

Corollary 19 *Let $R(\cdot)$ be the k-support norm and $X \in \mathbb{R}^{n \times p}$ be a sub-Gaussian design matrix. Let $\underline{f}$ denote a lower bound on the conditional density around a neighborhood of the τth quantile of $y_i|x_i$ for all x_i. Then if $n > c \cdot (\sqrt{s + s \log\lceil \frac{p}{k} \rceil})$ and the error vector u is such that $\|u\|_2 = \Theta\left(\frac{\sqrt{s + s \log\lceil \frac{p}{k} \rceil}}{\underline{f}\sqrt{n}}\right)$ then with probability atleast $1 - \exp(-\nu^2/2) - \exp\left(-\frac{\phi_1^2}{4}\right) - \exp(-\eta_2 \phi_2^2 \sqrt{n}\sqrt{s + s \log\lceil \frac{p}{k} \rceil})$,*

$$\inf_{u \in E_r} \delta\mathcal{L}(\theta^*, u; X, y) \geq \kappa \underline{f} \|u\|_2^2 \,. \tag{4.34}$$

Proof: The result follows from Theorem 11 noting that $w(A) = \Theta(\sqrt{s + s \log\lceil \frac{p}{k} \rceil})$ ■

4.4.3 Estimation Error Bounds

When the design matrix satisfies the RSC condition and the regularization parameter satisfies inequality (4.7) then the non-asymptotic estimation error can be bounded using the following result:

Theorem 12 *Let X be a design matrix with i.i.d. isotropic sub-Gaussian rows with sub-Gaussian norm K. Assume the regularization parameter satisfies the inequality $\lambda \geq 2R^*(\nabla_\theta \mathcal{L}(\theta^*; X, y))$ with probability atleast $1 - \eta_1 \exp\left(-\left(\frac{\nu_1}{\eta_2 \phi K}\right)^2\right)$ and the RSC condition is satisfied with probability atleast $1 - \exp(-\nu_2^2/2) - \exp\left(-\frac{\phi_1^2 n}{4}\right) - \exp\left(-\eta_2 \phi_2^2 \sqrt{n} w(A)\right)$. Then with $\Delta = \hat{\theta} - \theta^*$, for any norm $R(\cdot)$, we have for some positive constant c with probability atleast $1 - \eta_1 \exp\left(-\left(\frac{\nu_1}{\eta_2 \phi K}\right)^2\right) - \exp(-\nu_2^2/2) - \exp\left(-\frac{\phi_1^2 n}{4}\right) - \exp\left(-\eta_2 \phi_2^2 \sqrt{n} w(A)\right)$*

$$\|\Delta\|_2 \leq c \cdot \frac{\max\left(\Psi(E_r)\sqrt{\tau(1-\tau)}(w(\Omega_R) + \nu_1), w(A)\right)}{\kappa \underline{f} \sqrt{n}}, \tag{4.35}$$

where all notations are same as Theorem 10 and Theorem 11 and the constant c depends on the sub-Gaussian norm of the rows of the design matrix.

Proof: The conditions on the regularization parameter and the RSC condition follow from the results of Theorems 11 and 10. The result on the ℓ_2 norm of the error follows from Theorem 2 in [11]. The first term inside the $\max$ is from the bounds on the regularization parameter while the second term is from the condition imposed in the proof of the RSC condition. Note that for most norms we consider $w(A)$ is of the same order as $\Psi(E_r)w(\Omega_R)$. ■

The two norm of the error depends on the two terms $\sqrt{\tau(1-\tau)}$ and $\underline{f}$. The $\sqrt{\tau(1-\tau)}$ term is minimized at the tails and hence has the effect of reducing the estimation error. But typically this is dominated by the lower bound on the density $\underline{f}$ term which makes the estimate less precise in regions of low density. This is to be expected as there are very few samples to make a very precise estimate in low density regions. We note that similar observations are made in page 72 of [80]. Another aspect we reiterate here is the independence of the results from the form of the noise. All results make no assumptions on the noise apart from an assumption on the lower bound of the noise density.

Below we provide estimation error bounds for the atomic norms we consider in the work. For the ℓ_1 norm the results match results in [16].

Corollary 20 *Let $R(\cdot)$ be the ℓ_1 norm. Assume conditions in corollaries 14 and 17 are satisfied. Then with probability atleast* $1 - \eta_1 \exp\left(-\left(\frac{\nu_1}{\eta_2 K}\right)^2\right) - \exp(-\nu_2^2/2) - \exp\left(-\frac{\phi_1^2}{4}\right) - \exp(-\eta_2\phi_2^2\sqrt{n}\sqrt{s\log p})$,

$$\|\Delta\|_2 \leq c \cdot \frac{\sqrt{s\log p}}{\underline{f}\kappa\sqrt{n}} \ . \tag{4.36}$$

Proof: The result follows from the result of Theorem 12 and the results in corollaries 14 and 17. ■

Corollary 21 *Let $R(\cdot)$ be the ℓ_1/ℓ_2 non-overlapping group sparse norm. Assume conditions in corollaries 15 and 18 are satisfied. Then with probability atleast* $1 - \eta_1 \exp\left(-\left(\frac{\nu_1}{\eta_2 K}\right)^2\right) - \exp(-\nu_2^2/2) - \exp\left(-\frac{\phi_1^2}{4}\right) - \exp(-\eta_2\phi_2^2\sqrt{n}\sqrt{s_{\mathcal{G}}l + s_{\mathcal{G}}\log N_{\mathcal{G}}})$,

$$\|\Delta\|_2 \leq c \cdot \frac{\sqrt{s_{\mathcal{G}}l + s_{\mathcal{G}}\log N_{\mathcal{G}}}}{\underline{f}\kappa\sqrt{n}} \ . \tag{4.37}$$

Proof: The result follows from the result of Theorem 12 and the results in corollaries 15 and 18. ■

Corollary 22 *Let $R(\cdot)$ be the k-support norm. Assume conditions in corollaries 16 and 19 are satisfied. Then with probability atleast* $1 - \eta_1 \exp\left(-\left(\frac{\nu_1}{\eta_2 K}\right)^2\right) - \exp(-\nu_2^2/2) - \exp\left(-\frac{\phi_1^2}{4}\right) - \exp(-\eta_2\phi_2^2\sqrt{n}\sqrt{s + s\log\lceil\frac{p}{k}\rceil})$,

$$\|\Delta\|_2 \leq c \cdot \frac{\sqrt{s + s\log\lceil\frac{p}{k}\rceil}}{\underline{f}\kappa\sqrt{n}} \ . \tag{4.38}$$

Proof: The result follows from the result of Theorem 12 and the results in corollaries 16 and 19. ■

4.5 Experiments

We perform simulations with synthetic data.

4.5.1 Phase Transition

Data is generated as $y = X\theta^* + \omega$. $\theta^* = [\underbrace{1,1,1,1,1,1}_{6},\underbrace{0,0,\ldots,0}_{\text{p-6}}] \in \mathbb{R}^p$ for the ℓ_1 norm and $\theta^* = [\underbrace{1,\ldots,1}_{5},\underbrace{1,\ldots,1}_{5},\underbrace{1,\ldots,1}_{5},\underbrace{0,\ldots,0}_{5},\ldots,\underbrace{0,\ldots,0}_{5}]$ for the ℓ_1/ℓ_2 group sparse norm

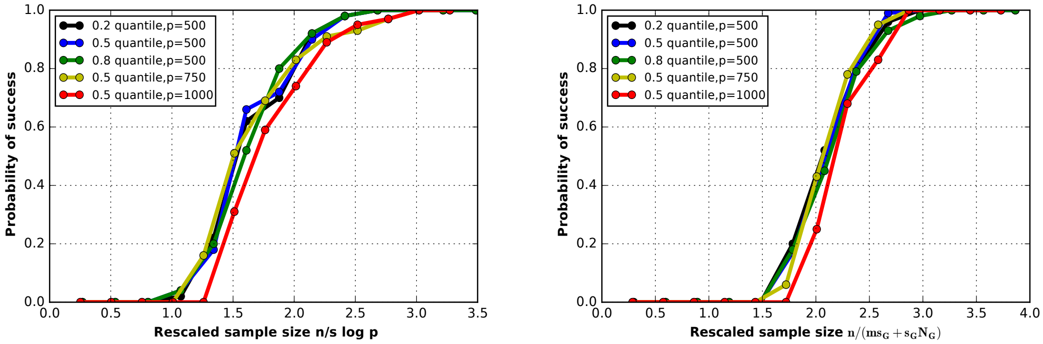

Figure 4.1: Probability of recovering true parameter versus the rescaled sample size for ℓ_1 norm (top) and ℓ_1/ℓ_2 group sparse norm (bottom). There is a sharp phase transition when the number of samples exceeds NIPS

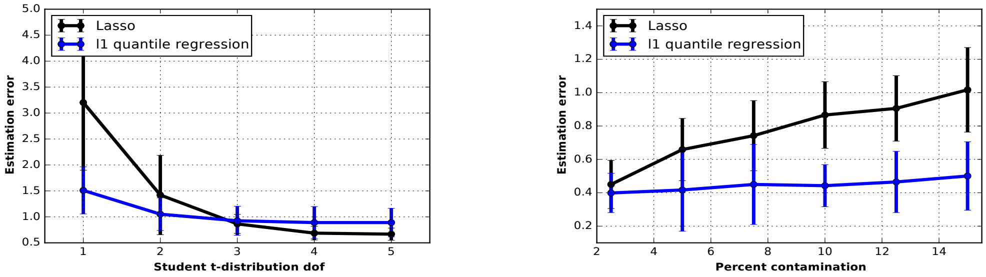

Figure 4.2: Estimation error of Lasso and ℓ_1-penalized quantile regression against different degrees of freedom of the student t-distribution noise (top) and against percentage contamination (bottom). Quantile regression is robust to heavy-tailed noise and outliers

with $p \in [500, 750, 1000]$. The noise $\omega_i \sim N(0, 0.25), \forall i \in [n]$ is Gaussian with zero mean and 0.25 variance. The design matrix $X \sim N(0, \mathbb{I}_{p\times p})$ is multivariate Gaussian with identity covariance. We vary $n = [10, 20, 30, \ldots, 120, 130]$. For each n we generate 100 datasets with the probability of success defined as the fraction of times we are able to faithfully estimate the true parameter. For $p = 500$ we run simulations for $\tau \in [0.1, 0.5, 0.9]$ and for $p \in [750, 1000]$ we run simulations only for $\tau = 0.5$. For the optimization, we use the Alternating Direction Method of Multipliers [22]. The details of the updates can be found in the flare documentation [90]. The code was implemented in Python. The plots in Figure 1 clearly show a phase transition for both the ℓ_1 and ℓ_1/ℓ_2 group sparse norms for all quantiles exemplifying the NIPS property described earlier.

4.5.2 Robustness

We showcase the robustness enjoyed by quantile regression over ordinary least squares estimation against heavy-tailed noise and outliers. We consider the ℓ_1 norm with $y = X\theta^* + \omega$. $\theta^* = [\underbrace{1,1,1,1,1,1}_{6}, \underbrace{0,0,\ldots,0}_{494}]$. For heavy-tailed noise we consider the student t-distribution with different degrees of freedom, with lower degrees of freedom corresponding to heavier tailed data. To show the robustness to outliers we randomly pick a certain percentage of samples from the dataset and multiply the noise by 10, that is, $\omega_i = 10 * \omega_i$ for a certain proportion of the dataset. We vary the proportion of contamination from 2.5% to 15%. We fix $n = 200$ for this simulation. Again for both exercises, we run 100 simulations and plot the mean and standard deviation of the estimation error $\|\hat{\theta} - \theta^*\|_2$. The plots in Figure 2 show both the estimation error against varying degrees of freedom of the student t-distribution and the estimation error against the percent contamination. The observations are in agreement with conventional wisdom on robustness of the quantile regression estimator to heavy-tailed noise and outliers.

Appendix

4.A Proofs for Number of Interpolated Samples

Theorem 9 *Let rows of the design matrix $X \in \mathbb{R}^{n\times p}$ have isotropic sub-Gaussian rows and let θ^* be an s-sparse vector that can be written as a linear combination of k atoms from an atomic set of cardinality m,*

$$\theta^* = \sum_{i=1}^{k} c_i a_i, a_i \in \mathcal{A}, c_i \geq 0, |\mathcal{A}| = m \; . \tag{4.39}$$

Consider the regularized quantile regression problem penalized with the atomic norm $R(\theta) = \|\theta\|_{\mathcal{A}}$,

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \mathcal{L}_\tau(\theta) = \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(\theta) + \lambda R(\theta) \; . \tag{4.40}$$

Let $E_r = \left\{u \;\middle|\; R(\theta^* + u) \leq R(\theta^*) + \frac{1}{2}R(u)\right\}$ denote the error set, let $A = cone(E_r) \cap S^{p-1}$ and let $\lambda \geq R^*(\nabla\mathcal{L}_\tau(\theta^*))$. Then with probability atleast $1 - \exp(-\eta_2\Psi^2(E_r)(q+\log(em))) - 2\exp(-\eta w^2(A))$ the number of interpolated samples,

$$\sup_{u\in E_r} |Z| = \sup_{u\in E_r} |\{i : y_i = \langle x_i, \theta^* + u\rangle, \; u \in E_r\}| \leq \max\{c_1\Psi^2(E_r)(q+\log(em)), c_2 w^2(A))\} \; , \tag{4.41}$$

where $\Psi(E_r) = \sup_{u \in E_r} \frac{\|u\|_{\mathcal{A}}}{\|u\|_2}$ *is the norm compatibility constant,* q *is the dimension of each subspace of the atomic norm and* $w(\cdot)$ *denotes the Gaussian width of a set and* η, η_2, c_1, c_2 *are constants.*

Proof: For simplicity consider the linear model $y = X\theta^* + \omega$. Then, for any $u \in E_r$, if $y_i = \langle x_i, \hat{\theta} \rangle = \langle x_i, \theta^* + u \rangle \implies \omega_i = \langle x_i, u \rangle$. When all samples are interpolated $\omega = Xu$. Hence we will show that when the number of samples n crosses a certain threshold $\inf_{u \in E_r} \|\omega - Xu\|_2^2 > 0$.

Consider a set $D = \{v \mid v = \sum_{i=1}^{k_1} c_i a_i, c_i \geq 0, a_i \in \mathcal{A}, \|v\|_0 \leq k_1 q\}$ be the set of vectors which can be written as a linear combination of k_1 atoms of the atomic set $\mathcal{A}$ such that $\|v\|_0 \leq k_1 q$. Clearly once $n > ck_1 q$ all samples cannot be interpolated. Hence $\inf_{v \in D} \|\omega - Xv\|_2^2$ becomes like a least squares problem and the infimum is achieved when ω is projected on the subspace spanned by $ck_1 q$ columns of X. Let $\Pi_D(\cdot)$ denote the projection operator. Then,

$$\begin{aligned}\inf_{v \in D} \|\omega - Xv\|_2^2 &= \|\omega\|_2^2 - \Pi_D^2(\omega) \\ &= \|\omega\|_2^2 - \|\omega\|_2^2 \Pi_D^2(\omega/\|\omega\|_2) \\ &\leq \|\omega\|_2^2 (1 - \Pi_D^2(\beta)) \, , \end{aligned} \tag{4.42}$$

where $\beta = \omega/\|\omega\|_2$. Below we quote a result from [6] to help get an upper bound for $\Pi_D(\beta)$,

Lemma 13 (Lemma 6.3, Amelunxen et al. 2014) *For each closed convex cone* D *in* $\mathbb{R}^p$ *with* $w(D)$ *denoting the Gaussian width of the cone,*

$$P(n\|\Pi_D(\beta)\|^2 \geq w^2(D) + \delta) \leq \exp\left(\frac{-\delta^2/8}{w^2(D) + \delta}\right) \, . \tag{4.43}$$

■

The lemma leads to the following corollary with cone D the subspace spanned by some s_1 columns of X.

Corollary 23 *Consider a* $k_1 q$ *dimensional subspace and let* $\Pi_D(\cdot)$ *denote the projection operator on the subspace of any unit norm vector. Then the following holds,*

$$P(n\|\Pi_D(\beta)\|^2 \geq ck_1 q + \delta) \leq \exp\left(\frac{-\delta^2/8}{ck_1 q + \delta}\right) \, . \tag{4.44}$$

The result follows directly from Lemma 13 and the fact that the square of the Gaussian width of the subspace is equal to ck_1q. ■

Now consider the set $B = \cup_{\binom{m}{k_1}} D$, that is, the union of all subsets of size k_1 of the atomic set $\mathcal{A}$. Below we establish a result for projection of the vector β to the set B.

Lemma 14 *Consider the set* $B = \cup_{\binom{m}{k_1}} D$. *Then the following holds,*

$$P\{n\|\Pi_B(\beta)\|_2^2 \geq c_2(k_1q + k_1\log(em))\} \leq \exp(-\eta_1 k_1(q + \log(em))) \ . \tag{4.45}$$

The argument follows from a simple union bound,

$$\begin{aligned} P\{n\|\Pi_B(\beta)\| \geq ck_1q + \delta\} &\leq \binom{m}{k_1} \exp\left(\frac{-\delta^2/8}{ck_1q + \delta}\right) \\ &\leq \exp\left(\frac{-\delta^2/8}{ck_1q + \delta} + k_1\log(em)\right) \ . \end{aligned}$$

The result follows from choosing $\delta = c_1(k_1q + k_1\log(em))$ with c_1 large enough so that $c_2 = c + c_1$ and $\frac{\delta^2/8}{ck_1q+\delta} - k_1\log(em) \geq \eta_1 k_1(q + \log(em))$. ■

Now from (4.42), the following can be easily inferred.

$$\inf_{v \in B} \|\omega - Xv\|_2^2 = \|\omega\|_2^2(1 - \Pi_B^2(\beta)) \tag{4.46}$$

Therefore from the result of Lemma 14 we can infer the following when $n \geq c_1(k_1q + k_1\log(em))$, choosing $\kappa_0 = \frac{c_2(k_1q + k_1\log(em))}{n} < 1$ with probability atleast $1 - \exp(-\eta_1 k_1(q + \log(em)))$,

$$\inf_{v \in B} \|\omega - Xv\|_2^2 \geq \|\omega\|_2^2(1 - \kappa_0) > 0 \ . \tag{4.47}$$

Next we move to analyzing conditions under which $\inf_{u \in E_r} \|\omega - Xu\|_2^2 \geq 0$, where E_r is the error set. We need the result from the following lemma which is based on Maurey's empirical approximation argument.

Lemma 15 *Let* $X \in \mathbb{R}^{n \times p}$ *be a random matrix with sub-Gaussian isotropic rows. Let for any* $v \in B$, *where set* B *is as defined above,* $\inf_{v \in B} \|\omega - Xv\|_2^2 > 0$. *Let* $u \in E_r$ *be any non-zero* *vector in the error set whose atomic norm representation is* $u = \sum_{i=1}^{m} c_i a_i, c_i \geq 0, a_i \in \mathcal{A}$*. Then* $\|a\|_{\mathcal{A}} = \sum_{i=1}^{m} c_i$*. Let*

$$\mu_j = c_j / \sum_{j=1}^{m} c_j = c_j / \|u\|_{\mathcal{A}}$$

then

$$\inf_{u \in E_r} \|\omega - Xu\|_2^2 \geq \kappa_1 \|\omega\|_2^2 \,, \tag{4.48}$$

when $n \geq \max\{c_1 \Psi^2(E_r)(q + \log(em)), c_2 w^2(A)\}$ *with probability atleast* $1 - \exp(-\eta_2 \Psi^2(E_r)(q + \log(em))) - 2\exp(-\eta w^2(A))$ *where* $\kappa_1, \eta, \eta_2, c_1, c_2$ *are positive constants.*

Proof: For any $u \in \mathbb{R}^p, u = \sum_{i=1}^{m} c_i a_i, c_i \geq 0, a_i \in \mathcal{A}$, let $W \in \mathbb{R}^p$ be a random vector defined by,

$$P(W = \|u\|_{\mathcal{A}} a_i) = \frac{c_i}{\|u\|_{\mathcal{A}}} \,. \tag{4.49}$$

Hence, $E[W] = u$. Let $W_1, W_2, \ldots, W_{k_1}$ be independent copies of W and set $Z = \frac{1}{k_1} \sum_{i=1}^{k_1} W_i$. Therefore Z belongs to the set B. In the following all expectations are w.r.t Z,

$$\begin{aligned} E\|\omega - XZ\|_2^2 &= \|\omega\|_2^2 + E\|XZ\|_2^2 - E[2\langle \omega, XZ \rangle] \\ &= \|\omega\|_2^2 + \frac{1}{k_1^2} \sum_{i,j \in [1,\ldots,k_1], i \neq j} E\langle XW_i, XW_j \rangle + \frac{1}{k_1^2} \sum_{i \in [1,\ldots,k_1]} E\langle XW_i, XW_i \rangle - 2\langle \omega, Xu \rangle \\ &= \|\omega\|_2^2 + \frac{k_1(k_1-1)}{k_1^2} \langle Xu, Xu \rangle + \frac{k_1}{k_1^2} \sum_{i=1}^{m} \frac{c_i}{\|u\|_{\mathcal{A}}} \|u\|_{\mathcal{A}}^2 \langle Xa_i, Xa_i \rangle - 2\langle \omega, Xu \rangle \\ &= \|\omega\|_2^2 + \left(1 - \frac{1}{k_1}\right) \|Xu\|_2^2 + \frac{\|u\|_{\mathcal{A}}^2}{k_1} \sum_{i=1}^{m} \mu_i \langle Xa_i, Xa_i \rangle - 2\langle \omega, Xu \rangle \\ &\leq \|\omega - Xu\|_2^2 + \frac{\|u\|_{\mathcal{A}}^2}{k_1} \sum_{i=1}^{m} \mu_i \langle Xa_i, Xa_i \rangle \,, \end{aligned}$$

where in the second inequality we use the definition of Z and use $E[Z] = u$, in the third inequality we use the definitions of the W_i's and in the last inequality we use the definitions of μ_i's. Therefore we get,

$$\|\omega - Xu\|_2^2 \geq E\|\omega - XZ\|_2^2 - \frac{\|u\|_{\mathcal{A}}^2}{k_1} \sum_{i=1}^{m} \mu_i \langle Xa_i, Xa_i \rangle \,. \tag{4.50}$$

Therefore for any vector in the error set $u \in E_r$,

$$\inf_{u \in E_r} \|\omega - Xu\|_2^2 \geq E\|\omega - XZ\|_2^2 - \sup_{u \in E_r} \frac{\|u\|_{\mathcal{A}}^2}{k_1} \sum_{i=1}^{m} \mu_i \langle Xa_i, Xa_i \rangle \,. \tag{4.51}$$

From (4.47), since any random vector Z belongs to the set B by design, with probability atleast $1 - \exp(-\eta_1 k_1(q + \log(em)))$ when $n \geq c_1(k_1 q + k_1 \log(em))$,

$$E\|\omega - XZ\|_2^2 \geq \inf_{v \in B} \|\omega - Xv\|_2^2 \geq \|\omega\|_2^2(1 - \kappa_0) \ . \tag{4.52}$$

Now $\|u\|_{\mathcal{A}}^2 \leq \Psi^2(E_r)\|u\|_2^2$ where $\Psi(E_r)$ is the norm compatibility constant. Also from their definition $\mu_i < 1, \sum_{i=1}^m \mu_i = 1$. Therefore we get the following,

$$\sup_{u \in E_r} \frac{\|u\|_{\mathcal{A}}^2}{k_1} \leq \frac{\Psi^2(E_r)\|u\|_2^2}{k_1} \sup_{a_i} \langle Xa_i, Xa_i \rangle \ . \tag{4.53}$$

We need the following result from [11].

Theorem 13 (Theorem 5, Banerjee et al. 2015) *Let X be a design matrix with independent isotropic sub-Gaussian rows, i.e., $\|x_i\|_{\psi_2} \leq K$ and $E[x_i x_i^T] = \mathbb{I}_{p \times p}$. Then for absolute constants $\eta, c > 0$, with probability atleast $1 - 2\exp(-\eta w^2(A))$ we have,*

$$\sup_{u \in A} \left| \frac{1}{n}\|Xu\|_2^2 - 1 \right| = \sup_{u \in A} \left| \frac{1}{n} \sum_{i=1}^n \langle x_i, u \rangle^2 - 1 \right| \leq c\frac{w(A)}{\sqrt{n}} \ , \tag{4.54}$$

or equivalently,

$$1 - c\frac{w(A)}{\sqrt{n}} \leq \inf_{u \in A} \frac{1}{n}\|Xu\|_2^2 \leq \sup_{u \in A} \frac{1}{n}\|Xu\|_2^2 \leq 1 + c\frac{w(A)}{\sqrt{n}} \ . \tag{4.55}$$

Now as a consequence of Theorem 13, since $a_i \in \mathcal{A}$ where $\mathcal{A}$ is the atomic set and $w(\Omega_R)$ is the Gaussian width of the unit norm ball,

$$\sup_{a_i} \langle Xa_i, Xa_i \rangle = \sup_{a_i} \|Xa_i\|_2^2 \leq (n + c\sqrt{n}w(\Omega_R)) \leq 2n \ , \tag{4.56}$$

where we use the result of Theorem 13 for the second inequality and in the third inequality we use that $n > c_2 w^2(A) \geq w(\Omega_R)$, i.e., the Gaussian width of the error set $A = \text{cone}(E_r) \cap \S^{p-1}$ is greater than the Gaussian width of the unit norm ball Ω_R and we choose c_2 large enough so that $\sqrt{n} \geq cw(\Omega_R)$.

Next for a lower bound on $\|u\|_2^2$, note that $\|\omega - Xu\|_2^2$ is minimized when ω and Xu are such that $\langle \omega, Xu \rangle = -\|\omega\|_2^2$ so that $\|\omega\|_2 = \|Xu\|_2^2$. The result of Theorem (13) applies to unit norm vectors $u \in A = \text{cone}(E_r) \cap S^{p-1}$. For $u \in E_r$, the result of Theorem (13) applies to the set $A = \text{cone}(E_r) \cap S^{p-1}$. The following result can be obtained by simple arithmatic manipulation,

$$1 - c\frac{w(A)}{\sqrt{n}} \leq \inf_{u \in A = \text{cone}(E_r) \cap S^{p-1}} \frac{1}{n}\|Xu\|_2^2 \quad \Rightarrow \quad \left(1 - c\frac{w(A)}{\sqrt{n}}\right)\|u\|_2^2 \leq \inf_{u \in E_r} \frac{1}{n}\|Xu\|_2^2 \tag{4.57}$$

Therefore substituting $\|Xu\|_2 = \|\omega\|_2$ in equation (4.57),

$$\|u\|_2^2 \leq \frac{\|\omega\|_2^2}{n\left(1 - c\frac{w(A)}{\sqrt{n}}\right)} \leq \frac{c_4\|\omega\|_2^2}{n}\ , \tag{4.58}$$

where in the last inequality we use $n \geq c_2 w^2(A)$ where c_2 is a constant large enough so that $c_4 = \frac{1}{1-c\frac{w(A)}{\sqrt{n}}}$. Therefore from equations (4.51), (4.52), (4.53), (4.56), (4.58) we can infer the following with probability atleast $1 - \exp(-\eta_1 k_1(q + \log(em))) - 2\exp(-\eta w^2(A))$,

$$\inf_{u \in E_r} \|\omega - Xu\|_2^2 \geq \|\omega\|_2^2\left(1 - \kappa_0 - \frac{2c_4\Psi^2(E_r)}{k_1}\right)\ . \tag{4.59}$$

Now choosing $k_1 \geq \frac{2c_4\Psi^2(E_r)}{1-\kappa_0-\kappa_1}$ for some positive constant $\kappa_1 < 1$, we get with probability atleast $1 - \exp(-\eta_2\Psi^2(E_r)(q + \log(em))) - 2\exp(-\eta w^2(A))$ with $\eta_2 = \frac{\eta_1 2c_4}{1-\kappa_0-\kappa}$,

$$\inf_{u \in E_r} \|\omega - Xu\|_2^2 \geq \kappa_1\|\omega\|_2^2\ . \tag{4.60}$$

As for the condition on the number of samples, it should satisfy the following condition,

$$n \geq \max\{c_1\Psi^2(E_r)(q + \log(em)), c_2 w^2(A)\}\ . \tag{4.61}$$

Hence when $n \geq \max\{c_1\Psi^2(E_r)(q + \log(em)), c_2 w^2(A)\}$ with probability atleast $1 - \exp(-\eta_2\Psi^2(E_r)(q + \log(em))) - 2\exp(-\eta w^2(A))$

$$\inf_{u \in E_r} \|\omega - Xu\|_2^2 \geq \kappa_1\|\omega\|_2^2\ . \tag{4.62}$$

This proves the results of Lemma 15 ■

As outlined in the beginning of proof of Theorem 9 the above result implies that $\max\{c_1\Psi^2(E_r)(q + \log(em)), c_2 w^2(A)\}$ samples will be interpolated with probability atleast $1 - \exp(-\eta_2\Psi^2(E_r)(q + \log(em))) - 2\exp(-\eta w^2(A))$. ■

4.B Proofs for the Regularization Parameter

Theorem 10 *Let $X \in \mathbb{R}^{n \times p}$ be a design matrix with independent isotropic sub-Gaussian rows with sub-Gaussian norm $\|x_i\|_{\psi_2} \leq K$. Define $\Omega_R = \{u \ : \ R(u) \leq 1\}$ the unit norm ball and let $\phi = \sup\limits_u \|u\|_2 / R(u)$. Then the following holds*

$$E\left[R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y))\right] \leq c \frac{K\sqrt{\tau(1-\tau)} w(\Omega_R)}{\sqrt{n}} \ , \tag{4.63}$$

where c is any fixed constant depending only on the sub-Gaussian norm k. Moreover with probability atleast $1 - \eta_1 \exp\left(-\left(\frac{\nu}{\eta_2 \phi K}\right)^2\right)$

$$R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y)) \leq c_1 \frac{K\sqrt{\tau(1-\tau)}(w(\Omega_R) + \nu)}{\sqrt{n}} \ , \tag{4.64}$$

where c_1, η_1, η_2, ν are absolute constants.

Proof: Denote by $\zeta_\tau(u) = (\mathbb{I}(u < 0) - \tau)$ for any $u \in \mathbb{R}$. Also define $\psi_\tau(u) : \mathbb{R}^n \to \mathbb{R}^n$ where the ith element of $(\psi_\tau(u))_i = \zeta_\tau(u_i), 1 \leq i \leq n$. Then the gradient of the loss function evaluated at θ^* is,

$$\begin{aligned}
(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y)) &= \frac{1}{n}\sum_{i=1}^n \langle x_i, \zeta_\tau(y_i - X_i\theta^*)\rangle \\
&= \frac{1}{n} X^T \psi_\tau(y - X\theta^*) \\
&= \frac{1}{n}\|\psi_\tau(y - X\theta^*)\|_2 \frac{X^T\psi_\tau(y - X\theta^*)}{\|\psi_\tau(y - X\theta^*)\|_2} \\
&= \frac{1}{n}\|\psi_\tau(y - X\theta^*)\|_2 X^T\gamma \ ,
\end{aligned}$$

where γ is a unit vector. Hence,

$$\begin{aligned}
R^*(\nabla \mathcal{L}_\tau(\theta^*; X, y)) &= \frac{1}{n} R^*(\|\psi_\tau(y - X\theta^*)\|_2 X^T\gamma) \\
&= \frac{1}{n}\|\psi_\tau(y - X\theta^*)\|_2 R^*(X^T\gamma) \ .
\end{aligned}$$

We will now focus on the random quantity $\|\psi_\tau(y - X\theta^*)\|_2 R^*(X^T\omega)$. We first obtain bounds on $E_{X,y}[\|\psi_\tau(y - X\theta^*)\|_2]$. Note that quantile regression on the τth quantile introduces an ordering where $y_i < \langle x_i, \theta^*\rangle$ for $n\tau$ samples and $y_i > \langle x_i, \theta^*\rangle$ for $n(1-\tau)$ samples (see

Theorem 2.2 in [80]). Therefore for any X, y,

$$\begin{aligned} E_{X,y}[\|\psi_\tau(y - X\theta^*)\|_2^2] &\leq n\tau(1-\tau)^2 + n(1-\tau)\tau^2 \\ &\leq n\tau(1-\tau)(1-\tau+\tau) \\ &\leq n\tau(1-\tau) \end{aligned} \tag{4.65}$$

To bound $E_X[R^*(X^T\gamma)]$ for some unit norm $\gamma \in S^{n-1}$, we use similar arguments as Theorem 3 in [11]. We make the following observation,

$$R^*(X^T\gamma) = \sup_{u:R(u)\leq 1} \langle X^T\gamma, u\rangle = \sup_{u:R(u)\leq 1} \langle h, u\rangle \ , \tag{4.66}$$

where by the definition of sub-Gaussian random vectors, $h = X^T\gamma$ is a sub-Gaussian vector in $\mathbb{R}^p$ for all $\gamma \in S^{n-1}$ and by Lemma 5 the sub-Gaussian norm of each element $\|h_i\|_{\psi_2} \leq cK, 1 \leq i \leq p$ implying $\|h\|_{\psi_2} \leq \sup_i \|h_i\|_{\psi_2} \leq cK$. Now consider the result of Theorem 8 in [11] below.

Lemma 16 (Theorem 8 in Banerjee et. al., 2015) *Let $\Omega_R = \{u \mid R(u) \leq 1\}$ be the unit norm ball of $R(\cdot)$. Assume h is any centered sub-Gaussian random vector with $\|h\|_{\psi_2} \leq cK$, then we have,*

$$E\left[\sup_{u:R(u)\leq 1} \langle h, u\rangle\right] \leq c_1 K w(\Omega_R) \ , \tag{4.67}$$

where c_1 is a constant.

Therefore from equation (4.66) and the result of Lemma 16, we can infer that for all $\gamma \in S^{n-1}$

$$E_X[R^*(X^T\gamma)] \leq c_1 K w(\Omega_R) \ . \tag{4.68}$$

Therefore we can infer the following,

$$\begin{aligned} E_{X,y}[\|\psi_\tau(y - X\theta^*)\|_2 X^T\gamma] &\leq \sup_{X,y} \|\psi_\tau(y - X\theta^*)\|_2 \sup_{\gamma\in S^{n-1}} E_X[X^T\gamma] \\ &\leq c_1 K\sqrt{n\tau(1-\tau)} w(\Omega_R) \ , \end{aligned} \tag{4.69}$$

where the inequality in the second line follows from equations (4.65), (4.68) and a simple application of Jensen's inequality. Therefore it follows that,

$$E\left[R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y))\right] = E\left[\frac{1}{n}\|\psi_\tau(y - X\theta^*)\|_2 R^*(X^T\gamma)\right] \leq c\frac{K\sqrt{\tau(1-\tau)}w(\Omega_R)}{\sqrt{n}} \ , \tag{4.70}$$

which proves the expectation result.

For large deviation bounds of $R^*(X^T\gamma)$, we use the following result from Theorem 9 in [11].

Lemma 17 (Theorem 9 in Banerjee et. al., 2015) *Let $\Omega_R = \{u \mid R(u) \leq 1\|$ be the unit norm ball of $R(\cdot)$. Assuming h is any centered sub-Gaussian random vector with $\|h\|_{\psi_2} \leq cK$, then we have for any $\nu > 0$,*

$$p\left(\sup_{u:R(u)\leq 1} \langle h, u\rangle \geq c_1 K w(\Omega_R) + \nu\right) \leq \eta_1 \exp\left(-\left(\frac{\nu}{\eta_2 K\phi}\right)^2\right) , \tag{4.71}$$

where c_1, η_1, η_2 are constants and $\phi = \sup\limits_{R(u)\leq 1} \|u\|_2$.

Therefore, we can infer that with probability atleast $1 - \eta_1 \exp\left(-\left(\frac{\nu}{\eta_2 K\phi}\right)^2\right)$

$$R^*(\nabla_\theta \mathcal{L}_\tau(\theta^*; X, y)) \leq c\frac{K\sqrt{\tau(1-\tau)}(w(\Omega_R)+\nu)}{\sqrt{n}} . \tag{4.72}$$

This completes the proof. ■

4.C Proofs for the Restricted Strong Convexity (RSC) Condition

Proof of Theorem 11 *Let $X \in \mathbb{R}^{n\times p}$ have sub-Gaussian rows with sub-Gaussian norm K. Let $0 < \underline{f} < f_i(\langle x_i, \theta^*\rangle)$ be a uniform lower bound on the conditional density around a neighborhood of the τth quantile of the conditional distribution $y_i|x_i$ for all x_i in the support of X. Let the number of samples $n \geq c \cdot w^2(A)$, where $A = cone(E_r) \cap S^{p-1}$ and $E_r = \left\{u \,|R(\theta^* + u) \leq R(\theta^*) + \frac{1}{2}R(u)\right\}$ is the error set. Also,assume that $\|u\|_2 = \Theta\left(\frac{w(A)}{\underline{f}\sqrt{n}}\right)$. Then, for positive constants ϕ_1, ϕ_2, ν we get with probability atleast $1 - \exp(-\nu^2/2) - \exp\left(-\frac{\phi_1^2 n}{4}\right) - \exp\left(-c_1\phi_2^2\sqrt{n}w(A)\right)$,*

$$\inf_{u\in E_r} \delta\mathcal{L}_\tau(\theta^*, u; X, y) \geq \kappa\underline{f}\|u\|_2^2 , \tag{4.73}$$

for some positive constant κ.

Proof:

For the quantile loss $\delta\mathcal{L}(\theta^*, u; X, y)$ evaluates to the following,

$$\left[\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \langle x_i, \theta^* + u\rangle) - \rho_\tau(y_i - \langle x_i\theta^*\rangle)\right] - \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, u\rangle)\zeta_\tau(y_i - \langle x_i, \theta^*\rangle) , \tag{4.74}$$

where $\zeta_\tau(u) = (\mathbb{I}(u < 0) - \tau)$ for any $u \in \mathbb{R}$

Now from Equation (4.3) on Pg 121 of [80], for any two scalars w and v we have

$$\begin{aligned}\rho_\tau(w - v) - \rho_\tau(w) &= -v(\tau - \mathbb{I}(w \leq 0)) + \int_0^v (\mathbb{I}(w \leq z) - \mathbb{I}(w \leq 0))dz \\ &= v\zeta_\tau(w) + \int_0^v (\mathbb{I}(w \leq z) - \mathbb{I}(w \leq 0))dz\ .\end{aligned}$$

Now in equation (4.74) for $i = 1, \ldots, n$, let $w_i = y_i - \langle x_i, \theta^* \rangle$ and $v_i = \langle x_i, u \rangle$, so that $\delta\mathcal{L}(\theta^*, u; X, y)$ evaluates to the following,

$$\begin{aligned}\delta\mathcal{L}(\theta^*, u; X, y) &= \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, u\rangle)\psi_\tau(y_i - \langle x_i, \theta^*\rangle) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\int_0^{\langle x_i, u\rangle}(\mathbb{I}(y_i - \langle x_i, \theta^*\rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \theta^*\rangle \leq 0))dz \\ &- \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, u\rangle)\psi_\tau(y_i - \langle x_i, \theta^*\rangle) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\langle x_i, u\rangle}(\mathbb{I}(y_i - \langle x_i, \theta^*\rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \theta^*\rangle \leq 0))dz\ . \qquad (4.75)\end{aligned}$$

Note that for the interpolated observations the integral evaluates to zero. Hence the RSC condition will not be satisfied until $n > \nu$ when we will start getting nonzero positive and negative residuals.

The above formulation can be equivalently written as follows with $\omega_i = y_i - \langle x_i, \theta^* \rangle$,

$$\delta\mathcal{L}(\theta^*, u; X, y) \geq \frac{1}{2n}\sum_{i=1}^{n}|\langle x_i, u\rangle|\,\mathbb{I}\left[|\omega_i| \leq |\langle x_i, u\rangle/2|\right]\,\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]\ . \qquad (4.76)$$

We will consider samples such that $|\langle x_i, u\rangle| \geq \xi\|u\|_2$ for some constant ξ we will define later and assume $\|u\|_2 \geq c\frac{w(A)}{\kappa \underline{f}\sqrt{n}}$. Therefore, we will bound the following quantity,

$$\delta\mathcal{L}(\theta^*, u; X, y) \geq \frac{1}{2n}\sum_{i=1}^{n}\xi\|u\|_2\,\mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2]\,\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]\,\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]\ . \qquad (4.77)$$

Let us first bound the quantity $\frac{1}{n}\sum_{i=1}^{n}\xi\|u\|_2\,\mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2]$

The following result follows from arguments in Proposition 5.1 in [124].

Theorem 14 *Consider the set $A = cone(E_r) \cap S^{p-1}$, where E_r is the error set. Define the marginal tail function as follows for some positive constants β, ξ,*

$$\beta = \inf_{A=cone(E_r)\cap S^{p-1}} P(|\langle x_i, u\rangle| \geq \xi) \ . \tag{4.78}$$

Then the following is true with probability atleast $1 - \exp(-\nu^2/2)$,

$$\inf_{u\in A} \frac{1}{n}\sum_{i=1}^{n} \xi \ \mathbb{I}[|\langle x_i, u\rangle| \geq \xi] \geq \xi\beta - c_2\frac{w(A)}{\sqrt{n}} - \frac{\nu\xi}{\sqrt{n}} \ . \tag{4.79}$$

Moreover, fix the two-norm of the error vector u as $\|u\|_2$. Then the following is true with probability atleast $1 - \exp(-\nu^2/2)$,

$$\inf_{u\in E_r} \frac{1}{n}\sum_{i=1}^{n} \xi\|u\|_2 \ \mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2] \geq \left(\xi\beta - c_2\frac{w(A)}{\sqrt{n}} - \frac{\nu\xi}{\sqrt{n}}\right)\|u\|_2 \ . \tag{4.80}$$

Now due to the conditions $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ and $\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]$ there will only be $m < n$ samples that will satisfy both conditions. We will derive high probability bounds on the number of samples satisfying conditions $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ and $\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]$.

First consider the following event,

$$\sum_{i=1}^{n} \mathbb{I}[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)] \ . \tag{4.81}$$

Since x_i's are symmetric isotropic $\text{sign}(\langle x_i, u\rangle)$ is positive with probability $1/2$ and negative with probability $1/2$. Similarly since we are evaluating the τth quantile ω_i is negative with probability τ and positive with probability $1-\tau$. Also the signs of ω_i and $\langle x_i, u\rangle$ are independent of each other. Hence the variable $\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)$ is a Bernoulli(p) random variable with,

$$\begin{aligned} p = P(\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)) &= P(\langle x_i, u\rangle > 0, \omega_i > 0) + P(\langle x_i, u\rangle < 0, \omega_i < 0) \\ &= (1-\tau) * (1/2) + \tau * (1/2) = 1/2 \ . \end{aligned}$$

We need the result from the below Chernoff bound to establish high probabilitybounds.

Lemma 18 *Let $z_i \sim$ Ber(p) be i.i.d. Bernoulli random variables. Then,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i - p\right| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2)\ . \tag{4.82}$$

By a simple application of Chernoff's bound in Lemma 18 for some constant $0 < \phi_1 < 1$, we get the following,

$$P\left(\sum_{i=1}^{n} \mathbb{I}[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)] \leq \frac{(1-\phi_1)n}{2}\right) \leq \exp\left(-\frac{\phi_1^2 n}{4}\right)\ . \tag{4.83}$$

Now, consider the quantity $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$. Note that $\omega_i = y_i - \langle x_i, \theta^*\rangle$. Let $\underline{f}$ be a uniform lower bound on the conditional density around the probability distribution $y_i|x_i,\ \forall x_i$ around the τth quantile, such that $P(|\omega_i| \leq \xi\|u\|_2/2) \geq \underline{f}\xi\|u\|_2/2$. Clearly $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ is a Bernoulli random variable with success rate $\underline{f}\xi\|u\|_2/2$. Therefore by a simple application of Chernoff's bound in Lemma 18 for some constant $0 \leq \phi_2 \leq 1$, we get the following,

$$P\left(\sum_{i=1}^{n} \mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2] \leq (1-\phi_2)n\underline{f}\xi\|u\|_2/2]\right) \leq \exp\left(-\frac{\phi_2^2 n\underline{f}\xi\|u\|_2}{4}\right)\ . \tag{4.84}$$

Now the events $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ and $\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]$ are independent of each other. Therefore from equations (4.83) and (5.95), we get the following,

$$P\left(\sum_{i=1}^{n} \mathbb{I}[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)]\ \mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2] \leq \frac{(1-\phi_1)(1-\phi_2)n\underline{f}\xi\|u\|_2}{4}\right) \tag{4.85}$$

$$\leq \exp\left(-\frac{\phi_1^2 n}{4}\right) + \exp\left(-\frac{\phi_2^2 n\underline{f}\xi\|u\|_2}{4}\right)$$

Therefore, we have determined that with high probability $m = \frac{(1-\phi_1)(1-\phi_2)n\underline{f}\xi\|u\|_2}{4}$ samples satisfy the conditions $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ and $\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]$. Now applying Theorem 21 for the m samples, we get the following with probability atleast $1 - \exp(-\nu^2/2)$,

$$\inf_{u\in E_r} \sum_{i=1}^{m} \xi\|u\|_2\ \mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2] \geq \left(\xi\beta - c_2\frac{w(A)}{\sqrt{m}} - \frac{\nu\xi}{\sqrt{m}}\right) m\|u\|_2\ . \tag{4.86}$$

Substituting $m = \frac{(1-\phi_1)(1-\phi_2)n\underline{f}\xi\|u\|_2}{4}$ and the assuming $\|u\|_2 = c_3\frac{w(A)}{\underline{f}\sqrt{n}}$ for some big enough positive constant c_3, we get that with probability atleast $1 - \exp(-\nu^2/2) - \exp\left(-\frac{\phi_1^2 n}{4}\right) - \exp\left(-c_1\phi_2^2\sqrt{n}w(A)\right)$

$$\sum_{i=1}^{m}\xi\|u\|_2\,\mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2] \geq \left(\xi\beta - c_2\frac{w(A)^{1/2}}{n^{1/4}} - \frac{\nu\xi}{n^{1/4}w(A)^{1/2}}\right)\frac{(1-\phi_1)(1-\phi_2)n\underline{f}\xi}{4}\|u\|_2^2 \tag{4.87}$$

$$\geq \kappa n\underline{f}\|u\|_2^2\,, \tag{4.88}$$

where we use $\kappa = \left(\xi\beta - c_2\frac{w(A)^{1/2}}{n^{1/4}} - \frac{\tau\xi}{n^{1/4}w(A)^{1/4}}\right)\frac{(1-\phi_1)(1-\phi_2)\xi}{4} > 0$ when $n \geq c \cdot w^2(A)$).

This proves the stated result. ■

Chapter 5

A Unified Analysis of High-Dimensional Single Index Models

5.1 Introduction

Regression analysis [137, 64] is widely used to find relationships between a response variable $y \in \mathbb{R}$ and explanatory variables $x \in \mathbb{R}^p$. Nonparametric regression [135, 125, 129, 65] techniques make very few assumptions on the relationship $y = f(x, \omega)$ with ω denoting the noise in the system. A major limitation is that the number of samples required for efficient estimation increases exponentially with the dimension (curse of dimensionality) [62], making them especially unsuitable for high-dimensional regression problems. At the other end of the spectrum, methods like linear or generalized linear models [13, 25, 98] assume the relationship $y = f(\langle x, \theta^* \rangle, \omega)$ with some known transfer function $f(\cdot)$. The parameter θ^* is then estimated using methods like maximum likelihood estimator, e.g., least squares or quantile regression [80]. However, in many practical applications, the linear model is at best an approximation to the true model. Single index models (SIM) [67, 70, 65] are semiparametric regression models which offer a middle path, assuming the relationship $y = f(\langle x, \theta^* \rangle, \omega)$ where the transfer function $f(\cdot)$ can be unknown, thus offering more flexibility than linear and generalized linear models. On the other hand, the response is assumed to depend on a one-dimensional projection of data on a suitable direction θ^* thus avoiding the computational problems due to curse of dimensionality of nonparametric regression techniques. In this chapter, we consider the problem of estimating the parameter θ^* in high-dimensional SIM [111, 40, 92, 59, 93, 106].

Consider we have n measurements $\{x_i, y_i\}_{i=1}^n$. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be any nonlinear function and $\theta^* \in \mathbb{R}^p$ be a sparse parameter vector whose structure is characterized by atomic norms

[33], e.g., the ℓ_1 norm [123], group-sparse norms [143, 71] or k-support norm [7]. We study two problems characterizing the conditional distribution $y|x$ in the high dimensional setting where possibly $n < p$:

Mean Single Index Models (SIM): Mean SIM make the assumption that the expectation of the response is some unknown nonlinear transformation of a linear function of the covariates [88, 24, 111].

$$E[y|x] = f(\langle x, \theta^* \rangle) \ . \tag{5.1}$$

Special cases of mean SIM's include the linear model with identity transfer function, i.e., $f(u) = u$ and generalized linear models like logistic regression where $f(u) = 1/(1 + e^{-u})$ [98].

Quantile Single Index Models (SIM): Quantile regression is widely used to model the quantiles of the conditional distribution $y|x$ [80]. In quantile SIMs [145, 138], we assume the τ-th quantile of the response for $\tau \in (0, 1)$ is some unknown nonlinear transformation of a linear function of the covariates. Formally we assume the following:

$$F^{-1}_{y|x}(\tau|x) = f_\tau(\langle x, \theta^*_\tau \rangle), \ \tau \in (0, 1) \ , \tag{5.2}$$

where $F_{y|x}(\cdot)$ is the cumulative distribution function of the conditional distribution $y|x$. The parameter θ^*_τ and nonlinear transform $f_\tau(\cdot)$ can vary with $\tau \in (0, 1)$ but satisfy the constraint $f_{\tau_1}(\langle x, \theta^*_{\tau_1} \rangle) \geq f_{\tau_2}(\langle x, \theta^*_{\tau_2} \rangle)$ if $\tau_1 > \tau_2$. Quantile SIM is used in applications to model properties of the conditional distribution $y|x$ other than the mean, e.g., when data is assumed to have the relationship $y = f(\langle x, \theta^* \rangle) + \omega$ and the additive noise ω is asymmetric, heavy-tailed or heteroscedastic.

In this work, we make the assumption that the covariates are non-degenerate elliptically symmetric [27, 57, 79]. An example of an elliptical distribution is the multivariate Gaussian distribution.

Definition 12 Elliptical Distributions: *A random vector $x \in \mathbb{R}^p$ follows an elliptical distribution EC$(\mu, \tilde{\Sigma}, \xi)$ iff x has a stochastic representation:*

$$x \sim \mu + \xi A u \ , \tag{5.3}$$

where $\mu \in \mathbb{R}^p$, $q \triangleq$ rank(A), $A \in \mathbb{R}^{p \times q}$, $\xi \geq 0$ is a random variable independent of u, $u \in S^{q-1}$ uniformly distributed on the unit sphere and $E[AA^T] = \tilde{\Sigma}$. Also $E[x] = \mu$ and cov$[x] = \frac{E[\xi^2]}{q}\tilde{\Sigma}$.

At first glance, it may seem impossible to estimate the parameter direction without some knowledge on the properties of the transfer function. But when the covariates are elliptically symmetric the following result in Theorem 15 can be leveraged to design very simple and efficient estimators without any knowledge of the transfer function [111, 88, 40, 51, 92, 145].

Theorem 15 *(Theorem 2.1 in [51]) Consider a single index model with elliptically symmetric covariates. Then the inverse regression curve $E[x|y]$ falls along a line:*

$$E[x|y] = \mu + \beta(y)\Sigma\theta^* \ , \tag{5.4}$$

where $\mu = E[x]$, $\Sigma = cov(x)$ and β is a scalar function of y:

$$\beta(y) = \frac{E[\langle x - \mu, \theta^* \rangle | y]}{(\theta^*)^T \Sigma \theta^*} \ . \tag{5.5}$$

Based on the elegant result of Theorem 15, we make the following contributions to the study of estimators for SIM models when the covariates are elliptically symmetric:

(1) We establish a unified analysis framework under which we compare existing estimators [111, 93, 92, 40, 59] for high-dimensional mean SIM. Prior work has assumed elliptically symmetric covariates explicitly [59] or implicitly by assuming Gaussian covariates [111, 93, 92, 40]. We show past estimators to be variations of the following constrained formulation:

$$\hat{\theta} := \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2n} \|X\theta\|_2^2 - \langle \eta, \theta \rangle \quad \text{s.t.} \quad R(\theta) \leq \beta R(\theta^*) \ , \tag{5.6}$$

with the parameter $\eta = \frac{1}{n} X^T \tilde{y}$ is a weighted combination of the rows of the design matrix with the weights $\tilde{y}$ computed differently for different estimators and $R(\cdot)$ is any atomic norm constraint. For example, the constrained least squares estimator with $\tilde{y} = y$ is shown to be a consistent estimator of θ^* [111]. Using the elliptical symmetry property of covariates we show $E[\eta] = \beta\Sigma\theta^*$ for all estimators, where $\Sigma = \text{cov}(x)$. The constant β varies with estimator, transfer function and data.

(2) We design a new estimator inspired by ideas from sliced inverse regression [88, 51]. We show both theoretically and empirically through multiple experiments on synthetic data that our new estimator, unlike estimators in [111, 93, 92, 40], is simultaneously robust to heavy-tailed, outlier noise, is sample efficient, and can handle non-monotonic functions.

(3) The following constrained linear quantile estimator (5.7) [80] has been shown to consistently estimate the direction of θ^* in quantile SIM when the covariates are elliptically symmetric [145]:

$$\hat{\theta}_n := \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \langle x_i, \theta \rangle) \quad s.t. \quad R(\theta) \leq \beta R(\theta^*) \; , \tag{5.7}$$

where $\rho_\tau(u) = (\tau - \mathbb{I}(u \leq 0))u$ is the asymmetric absolute deviation function [80], $\mathbb{I}(\cdot)$ is the indicator function and β varies with data. Our main contribution on quantile SIM models is to derive non-asymptotic error bounds for the constrained quantile regression estimator.

We highlight a few important aspects of our work. First, all our bounds are non-asymptotic and are general enough to handle any atomic norm. All results are expressed in terms of easily computable geometric quantities [38, 33], like Gaussian widths [60, 122] and norm compatibility constants [17, 105], on sets related to the norm. Second, the robust techniques can complement techniques like Huber loss function [69] when covariates are elliptically symmetric. Third, although we do not focus on characterizing $f(\cdot)$, practically once θ^* is efficiently estimated standard smoothing techniques from nonparametric regression can be used to model properties of $f(\cdot)$ [104, 136, 41].

The chapter is organized as follows. Section 5.2 and Section 5.3 discuss parameter estimation in mean and quantile SIM respectively. We show results on synthetic datasets in Section 5.4. For both mean and quantile SIM, we will assume for sake of convenience $\mu = E[x] = 0$, $\|\Sigma^{1/2}\theta^*\|_2 = 1$. We note that in general only the direction of θ^* is identifiable and any magnitude for cases when $\|\Sigma^{1/2}\theta^*\|_2 \neq 1$ can be absorbed in the definition of $f(\cdot)$ [67, 111, 88, 40].

5.2 High-dimensional Mean SIMs

We consider the problem of parameter estimation in mean SIM in this section. In Section 5.2.1 we establish a common framework through which we highlight similarities between existing estimators for mean SIM. We also briefly outline the estimator properties w.r.t. sample complexity required for estimation, robustness to heavy-tailed/outlier noise and assumptions made on the transfer functions. In Section 5.2.2, we design a new estimator inspired from sliced inverse regression, establish non-asymptotic estimation error bounds assuming sub-Gaussian elliptically symmetric design matrices and compare its performance to the estimators we consider in Section 5.2.1.

5.2.1 Comparison of Estimators

Existing estimators like constrained least squares [111], sliced inverse regression [92, 93, 106] and estimator based on U-statistics [40] for mean SIM are variations of the constrained estimator (5.6). The quantity $\eta \in \mathbb{R}^p$ satisfies the following properties:

(a) $\eta = \frac{1}{n} X^T \tilde{y}$, with $\tilde{y} \in \mathbb{R}^n$ computed differently for each estimator. The computation of the quantity $\tilde{y}$ is influenced by the response.

(b) Consider the derivative of the loss function in (5.6) without the constraint,

$$\mathcal{F}(\theta) = \frac{1}{n} X^T X \theta - \eta \, . \tag{5.8}$$

Using the result of Theorem 15, we can show η is such that $E_{X,\omega}[\eta] = \beta \Sigma \theta^*$ with β varying with the estimator and transfer function. Therefore the true parameter θ^* satisfies $E_{X,\omega}\left[\mathcal{F}(\beta\theta^*)\right] = 0$. The result further implies that $E_{X,\omega}[\hat{\theta}] = \beta\theta^*$ and hence (5.6) is an unbiased estimator of $\beta\theta^*$.

Theorem 16 below is a formal statement whose proof we provide in the appendix along with a short description for sliced inverse regression [92, 93, 106] and the U-statistics based estimator [40]. The essence of each method is to compute η as a weighted combination of the rows of the design matrix such that $E[\eta] = \beta\Sigma\theta^*$ using property in Theorem 15 for elliptically symmetric covariates. The bounded weights $\tilde{y}$ in sliced inverse regression and U-statistics based estimators ensure robustness to heavy-tailed/outlier response in contrast to the constrained least squares estimator, On the other hand, sliced inverse regression estimator requires considerably more samples for parameter estimation. For example, when $R(\cdot)$ is $\|\cdot\|_1$, [93, 92] show that without making assumptions on the covariance matrix $\Theta(\max(s \log p, \sqrt{p}))$ are required for consistent estimation. In contrast both constrained least squares and the U-statistics based estimator need only $\Theta(s \log p)$ samples for consistent estimation. But the design of the U-statistics based estimator explicitly assumes the transfer function to be monotonic, i.e., $f(u) \geq f(v), \forall u > v$ or $f(u) \leq f(v), \forall u > v$. All estimators require the transfer function to be non-even. Even functions satisfy $f(\langle x, \theta^* \rangle) = f(\langle -x, \theta^* \rangle)$ for all x, e.g., $f(u) = u^2$.

Theorem 16 *The constrained least squares estimator [111], sliced inverse regression [92, 93, 106] and U-statistics based estimator [40] are equivalent to the constrained estimator (5.6).*

1. *For the constrained least squares estimator [111, 23],*

$$\eta = \frac{1}{n}X^T y, \quad \beta = E_{X,\omega}\left[f(\langle x, \theta^*\rangle)\langle x, \theta^*\rangle\right] \tag{5.9}$$

2. *For the sliced inverse regression estimator [92, 93, 106], let $y^{\uparrow}$ denote the response variable sorted by ascending order. Divide the range of $y^{\uparrow}$ into H contiguous slices denoted by I_h. Let $\gamma_h,\ h = 1, \cdots, H$ are bounded weights assigned to each slice such that $\|\eta\|_2 = 1$. Then,*

$$\eta = \frac{1}{n}X^T\tilde{y}, \quad \tilde{y}_i = \sum_{h=1}^{H}\gamma_h\mathbb{I}[y_i \in I_h]; \quad \beta = 1 \tag{5.10}$$

3. *For the U-statistics based estimator [40], let $y^{\uparrow}$ denote the response variable sorted by ascending order. Then $\tilde{y}$ are assigned values between $[-2, 2]$ based on the ordering in $\mathbf{y}^{\uparrow}$. More formally let $i^{\uparrow} \in \{0, \ldots, n-1\}$ denote the position of response variable y_i in $y^{\uparrow}$.*

$$\eta = \frac{1}{n}X^T\tilde{y}, \quad \tilde{y}_i = -2 + \frac{2 * i^{\uparrow}}{(n-1)}$$
$$\beta = E_{X,\omega}\left[\sum_{i,j\in[1,\ldots,n];i\neq j} sign(y_i - y_j)\cdot\langle x_i - x_j, \theta^*\rangle\right] \tag{5.11}$$

Moreover for all estimators $E_{X,\omega}[\eta] = \beta\Sigma\theta^*$ so that $E_{X,\omega}\left[\mathcal{F}(\beta\theta^*)\right] = 0$ where $\mathcal{F}(\cdot)$ is the derivate of the loss function.

5.2.2 A New Sliced Inverse Regression Based Estimator

In this section, we discuss a new estimator based on ideas from sliced inverse regression which we show to be simultaneously sample efficient, robust and can gracefully work with non-monotonic functions. We also derive non-asymptotic estimation error bounds for the estimator.

The first couple of steps in Algorithm 1 are similar to original sliced inverse regression in [88]. Our algorithm differs in the way η is computed in step 3. The result of Lemma 19 helps explain the motivations behind computation of η.

Lemma 19 *Consider the estimate of η in Algorithm 1. For some $\tilde{y}$ with $\tilde{y}_i \in [-1, 1], 1 \leq i \leq n$, the quantity η can be equivalently written as $\eta = \frac{1}{n}X^T\tilde{y}$. Also the following is true,*

$$E_{X,\omega}\left[\frac{\beta}{n}X^T X\theta^* - \eta\right] = 0\ , \tag{5.12}$$

Algorithm 1 Sliced Inverse Regression for High-dimensional Single Index Model (SIR-Hd-SIM)

```
1: Sort $y$ to obtain $y^{\uparrow}$. Divide range of $y^{\uparrow}$ into $H$ slices $I_h,\ h = 1, \ldots, H$ each containing
   $n_h = \lfloor n/H \rfloor$ samples
2: Within each slice compute sample mean $\bar{x}_h$ of $x$
       $$\bar{x}_h = \frac{1}{n_h} \sum_{y_i \in I_h} x_i,\ h = 1, \ldots, H\ . \qquad (5.13)$$
3: Let
       $$\gamma = \underbrace{[\overbrace{-1, \ldots, -1}^{H/2}, \overbrace{1, \ldots, 1}^{H/2}]}_{\textbf{(unweighted SIR)}} \quad \text{OR} \quad \gamma = \underbrace{[-1, -1 + 1/(H-1), -1 + 2/(H-1), \ldots, 1]}_{\textbf{(weighted SIR)}} \qquad (5.14)$$
   and compute
       $$\eta = \frac{1}{H} \sum_{h=1}^{H} \gamma_h \bar{x}_h\ . \qquad (5.15)$$
4: Compute $\hat{\theta}$
       $$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \|X\theta\|_2^2 - \langle \eta, \theta \rangle \quad \text{s.t.} \quad R(\theta) \leq \beta R(\theta^*)\ . \qquad (5.16)$$
```

with $\beta = E_{X,\omega}\left[\frac{1}{H}\sum_{h=1}^{H} \gamma_h \langle x, \theta^* \rangle \;\middle|\; y \in I_h\right]$ *is defined in step 3.*

Unweighted SIR: By Lemma 19, the estimator works when $\beta = E_{X,\omega}[\sum_{h=1}^{H} \gamma_h \langle x, \theta^* \rangle | y \in I_h]$ is not zero, where γ_h is the weights assigned to the slices in Step 3 of the Algorithm. Consider the case when $\gamma_h = 1,\ 1 \leq h \leq H$. Then $E_{X,\omega}[\sum_{h=1}^{H} \langle x, \theta^* \rangle | y \in I_h] = E[\langle x, \theta^* \rangle] = 0$ and hence the estimator will not work since $\beta = 0$ in Lemma 19. Now consider formulation for unweighted SIR. Let the transfer function be monotonically increasing. $y^{\uparrow} \in \mathbb{R}^n$ is the sorted response vector. Divide the range of $y^{\uparrow}$ into two such that each half has equal number of samples. Ignoring the noise, the values of $y_i^{\uparrow}$'s in the bottom half (smaller values) can be argued to be generated as $y_i^{\uparrow} = f(\langle x_i^{\uparrow}, \theta^* \rangle),\ \langle x_i^{\uparrow}, \theta^* \rangle < 0$ and in the top half (larger values) to be generated as $y_i^{\uparrow} = f(\langle x_i^{\uparrow}, \theta^* \rangle),\ \langle x_i^{\uparrow}, \theta^* \rangle > 0$, where we denote $x_i^{\uparrow}$ to be the covariates corresponding to response $y_i^{\uparrow}$. Therefore our choice of γ for unweighted SIR is such that $\gamma_h < 0$ when $\langle x, \theta^* \rangle < 0$ and $\gamma_h > 0$ when $\langle x, \theta^* \rangle > 0$ and hence

$\beta = \sum_{h=1}^{H} E[\gamma_h \langle x, \theta^* \rangle | y \in I_h]$ evaluates to a non-zero value.

Weighted SIR: We find the weighted SIR estimator to perform better than unweighted SIR on most synthetic datasets in Section 5.4. The better performance can be attributed to the following argument. Consider f to be monotonically increasing. For slices with low or high y_i values, $\beta_h = E[\langle x_i, \theta^* \rangle | y_i \in I_h]$ should be expected to have high absolute values. Since the empirical mean in each slice is $\bar{x}_h = E[x | y \in I_h] + \Delta = \beta_h \Sigma \theta^* + \Delta$ [51, 88] by Theorem 15, where Δ is a random sub-Gaussian noise, the high values of β_h ensure the signal to noise ratio $\|\beta_h \Sigma \theta^*\|_2 / \|\Delta\|_2$ is high. In contrast the mid values of y_i correspond to slices with smaller values for $\langle x_i, \theta^* \rangle$ and consequently β_h. Low values of β_h in the middle slices mean a lower signal to noise ratio. The weighted SIR formulation downweights estimates from the middle slices, i.e., corresponding γ_h values are close to 0, which correspond to low signal to noise ratio leading to better estimator performance.

Number of slices (H), weights (γ): As shown in [88], the number of slices can vary from as low as 2 to as high as $n/2$. In general, the method works as long the choice of γ_h's is such that $\beta \neq 0$.

Connections to the U-statistics estimator: On closer inspection, the U-statistics estimator [40] is similar to weighted sliced inverse regression in Algorithm 1. The U-statistics method divides the data into n slices based on sorted response values, assigns increasing weights to each slice in the range $[-1, 1]$ and computes η as a weighted combination of the covariates belonging to each slice. Based on this interpretation the U-statistics estimator should work for non-monotonic transfer functions for which Algorithm 1 works, a hypothesis which we positively validate on synthetic data in Section 5.4.

Before we establish non-asymptotic estimation error bounds for the estimator in Algorithm 1, we highlight a few properties of the estimator. Due to bounded $\tilde{y}$, the estimator is robust to heavy-tailed/outlier noise. The algorithm can be used with any transfer function except even functions. For even functions, the estimate $\beta_h = E[\langle x_i, \theta^* \rangle | y \in I_h] = 0, \ \forall h$, since $-x$ and x generate the same y value. As we show in the result below the estimator is sample efficient, e.g., $O(s \log p)$ samples are sufficient to estimate the parameter direction with the ℓ_1 norm.

We use tools from the uniform analysis framework established in [11, 105, 33] for high-dimensional linear regression to derive non-asymptotic estimation error bounds when the design matrix is sub-Gaussian elliptically symmetric. We first define a few important notations. The constraint in estimator (5.16) ensures the error vector $\Delta = \hat{\theta} - \theta^*$ belongs to the set $E_c = \{u \in$

$\mathbb{R}^p \mid R(\beta\theta^* + u) \leq R(\beta\theta^*)\}$. Define the error set as $A = \text{cone}(E_c) \cap S^{p-1}$. Also define the unit norm ball $\Omega_R = \{u \in \mathbb{R}^p \mid R(u) \leq 1\}$. The estimation error can be expressed using geometric quantities related to the sets A, Ω_R. The first is the norm compatibility constant [17, 105] defined as $\Psi(A) = \sup\limits_{u \in A} \frac{R(u)}{\|u\|_2}$. For example, when $R(\cdot)$ is $\|\cdot\|_1$, $\Psi(A) = \sqrt{s}$ for s-sparse θ^*. The second geometric quantity is the Gaussian width of a set [121, 122], which informally speaking is a measure of the size/complexity of a set. Formally, for any set $A \subseteq \mathbb{R}^p$, the Gaussian width denoted by $w(\cdot)$ is defined as $w(A) = E\left[\sup\limits_{u \in A} \langle g, u \rangle\right]$ where $g \sim N(0, \mathbb{I}_{p \times p})$ is a standard Gaussian random vector. The number of samples required for estimation, the estimation error depend on the Gaussian widths of the error set and unit norm ball. Gaussian width values for popular norms can be found in [33, 38]. For example, for the ℓ_1 norm $w(A) = O(\sqrt{s \log p})$, $w(\Omega_R) = O(\sqrt{\log p})$ when θ^* is s-sparse.

Theorem 17 *Let $X \in \mathbb{R}^{n \times p}$ have independent sub-Gaussian elliptically symmetric rows x_i with $E[x_i] = 0$, $cov(x) = \Sigma$ and sub-Gaussian norm $\|x_i \Sigma^{-1/2}\|_{\psi_2} \leq \kappa$. Let $\Lambda_{\max}(\Sigma)$ denote the largest eigenvalue of Σ and $\Lambda_{\max}(\Sigma|A) = \sup\limits_{u \in A} u^T \Sigma u$. Then for the sliced inverse regression estimator in Algorithm 1 when $n > \Theta(w^2(A))$ the following estimation error bound holds with probability atleast $1 - 2\exp(-\nu_1 w^2(A)) - \nu_2 \exp\left(-\left(\frac{\tau}{\nu_3 \kappa \phi}\right)^2\right)$:*

$$\|\Delta\|_2 = \|\hat{\theta} - \theta^*\|_2 \leq O\left(\frac{\Psi(A)\kappa\sqrt{\Lambda_{\max}(\Sigma)}w(\Omega_R) + \beta\Lambda_{\max}(\Sigma|A)w(A)}{\sqrt{n}}\right), \quad (5.17)$$

where β is the value defined in Lemma 19 and $\eta, \nu_1, \nu_2, \nu_3, \tau$ are absolute constants.

For example, for the ℓ_1 norm with s-sparse parameter the result of Theorem 17 implies $n = \Theta(s \log p)$ samples are sufficient for consistent estimation and the estimation error is $\|\Delta\|_2 = O(\sqrt{s \log p / n})$ matching the bounds for the linear model [17, 33, 105, 11]. Also our analysis is general enough to handle most atomic norms [33]. For example, we obtain the same bounds as ℓ_1 norm when $R(\cdot)$ is the k-support norm [7] with s-sparse parameter, matching the results for the linear model [7, 38].

5.3 High-dimensional Quantile SIMs

In this section, we discuss parameter estimation in high-dimensional quantile SIM [16, 118]. An important consequence when covariates are elliptical symmetric is the linear quantile regression estimator (5.7) is consistent for the direction of θ^* [145] when the transfer function is not even.

Lemma 20 *(Theorem 1 in [145]) Assume the quantile single index model with elliptically symmetric covariates and function $f(\cdot)$ which is not even. Then the following is true,*

$$\beta\theta^*_\tau = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \, E_{x,\omega}\left[\rho_\tau(y - \langle x, \theta \rangle)\right] \ , \tag{5.18}$$

for some β depending on the data.

Before establishing non-asymptotic estimation error bounds we remark on a couple of interesting properties of the quantile SIM model. Quantile regression is robust in the sense that the estimator is less sensitive to outliers in the data compared to least squares regression. Quantile regression is also equivariant to monotone transformations [80]. Let $F_y^{-1}(\tau)$ denote the quantile of some univariate random variable y. Then equivariance to monotone transformations implies that $F_{f(y)}^{-1}(\tau) = f(F_y^{-1}(\tau))$ when the function $f(\cdot)$ is monotone. Applied to quantile regression, this implies that $F_{y_i|x_i}^{-1}(\tau|x_i) = f(\langle x, \theta^* \rangle)$ is equivalent to assuming the linear model on the transformed variable $f^{-1}(y)$, i.e., $F_{f^{-1}(y)}^{-1}(\tau) = \langle x, \theta^* \rangle$. This is not true for mean estimation as $E(f(y)) \neq f(E(y))$.

The result below characterizes the non-asymptotic estimation error bounds for the estimator for quantile τ using notations outlined in Section 5.2.2 before Theorem 17.

Theorem 18 *Let $X \in \mathbb{R}^{n \times p}$ have independent sub-Gaussian elliptically symmetric rows x_i with $E[x_i] = 0$, $cov(x) = \Sigma$ and sub-Gaussian norm $\|x_i \Sigma^{-1/2}\|_{\psi_2} \leq \kappa$. Let $\Lambda_{\max}(\Sigma)$ denote the largest eigenvalue of Σ and $\Lambda_{\max}(\Sigma|A) = \sup_{u \in A} u^T \Sigma u$. Then when $n = \Theta(w^2(A))$ the estimation error for the estimator (5.7) applied to the quantile SIM model satisfies the following bound with probability atleast $1 - \exp(-\tau_1^2/2) - \exp(-\nu_1 \phi_2^2 \sqrt{n} w(A)) - \nu_2 \exp\left(-\left(\frac{\tau_2}{c_2 \phi k}\right)^2\right)$ when $n > \Theta(w^2(A))$*

$$\|\Delta\|_2 = \|\hat{\theta} - \theta^*_\tau\|_2 = O\left(\frac{\sqrt{\Lambda_{\max}(\Sigma)}\max\{\max\{\tau, 1-\tau\}\Psi(A) w(\Omega_R), w(A)\}}{\underline{f}\sqrt{n}}\right) \ , \tag{5.19}$$

where $0 < \underline{f}$ is a uniform lower bound on the conditional density of the distribution $y_i|x_i$ around the noise term $y_i - \langle x_i, \beta\theta^* \rangle$ and $\tau_1, \tau_2, \nu_1, \nu_2$ are absolute constants.

Consider the ℓ_1 norm as an example with s-sparse parameter. Application of Theorem 18 implies the estimation error is $O\left(\sqrt{s \log p}/\underline{f}\sqrt{n}\right)$ matching bounds in [118, 16] for linear quantile regression. Note the extra $\underline{f}$ in the denominator which means the estimation error will be higher in low density regions. This is to be expected since very few samples are available in low density regions to make precise estimates [80]. Again the result is general enough to handle any atomic norm. For example, we obtain the same bounds as the ℓ_1 norm for the k-support norm matching results in [118].

5.4 Synthetic Experiments

In this section, we empirically compare the performance of different estimators on synthetic data generated using five functions: (a) $f(a) = a$, the linear function; (b) $f(a) = 1/(1 + \exp(-a))$, a bounded monotonic function; (c) $f(a) = a^3$, an unbounded monotonic function; (d) $f(a) =$ $\sin(a)$ a non-monotonic function; and (e) $f(a) = a$ if $|a| < 1$ and $1/a$ otherwise a non-monotonic function. For each function, we experiment on four synthetic datasets generated as follows:

Gaussian design, Gaussian noise: The rows of the design matrix are sampled i.i.d. from a $N(0, \mathbb{I}_{1000\times 1000})$ distribution. The noise is i.i.d. $N(0, 0.25 * v)$ where $v = \text{var}(f(\langle x_i, \theta^*\rangle))$.

Gaussian design, student-t distribution noise: The design matrix is Gaussian as above, but the noise sampled i.i.d. from a 1-dof student-t distribution.

Gaussian design, outlier noise: The design matrix and noise are initially i.i.d. Gaussian. Then for 5% of the datapoints we multiply the noise by 5, i.e., $\omega_i = 5 * \omega_i$.

Heavy tailed design, noise: Finally we generate heavy tailed design and noise as described in Section 4 of [59]. The rows of the design matrix $x_i = \mu_i U_i$, where $U_i \in \mathbb{R}^{1000}$ are i.i.d. with uniform distribution over sphere of radius $\sqrt{1000}$, the random variables $\mu_i \in \mathbb{R}$ are i.i.d., independent of U_i and such that $\mu_i = \frac{1}{2c(q)}(\xi_{i,1} - \xi_{i,2})$, where $\xi_{i,1}, \xi_{i,2}$ are i.i.d. with Pareto distribution, so that their probability density function is given by $p(t;q) = \frac{q}{(1+t)^{1+q}}\mathbb{I}_{\{t>0\}}$, $c(q) = \text{variance}(\xi) = \frac{q}{(q-1)(q-2)}$, and $q = 2.1$. The noise is i.i.d. with Pareto distribution such that $\omega_i = \frac{1}{\sqrt{20c(q)}}(\xi_{i,1} - \xi_{i,2})$ where $\xi_{i,1}, \xi_{i,2}$ and $c(q)$ are Pareto distribution variables and variance respectively as described earlier.

The response vector is $y_i = f(\langle x_i, \theta^*\rangle) + \omega_i$, where $f(\cdot)$'s, x_i's and ω_i's are generated using the procedure outlined above. In each case, $\theta^* =$ $[\underbrace{1/\sqrt{10}, \ldots, 1/\sqrt{10}}_{10}, \underbrace{0, \ldots, 0}_{990}]$. We measure the estimation error with different sample sizes $n = [250, 500, 750, 1000, 1250, 1500, 1750, 2000]$. For each sample size, we compute the mean estimation error $\|\hat{\theta} - \theta^*\|_2$ over 50 different datasets.

For datasets with light tailed Gaussian designs and noise we perform numerical experiments

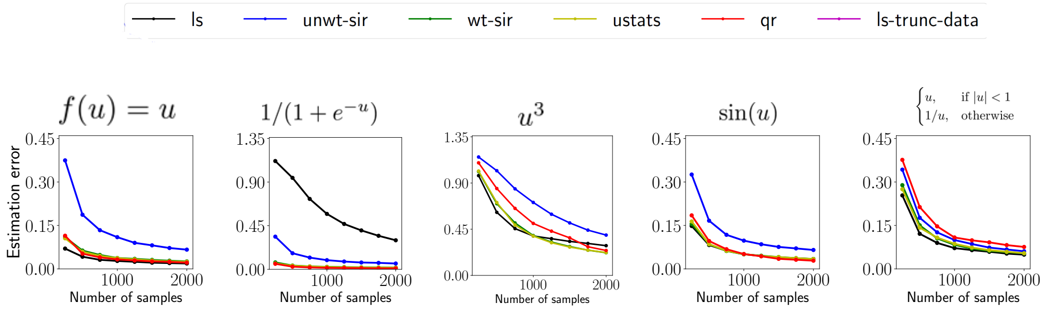

Figure 5.1: Estimation error vs sample size for Gaussian design, Gaussian noise. Weighted SIR, Ustats are equal or better than other mean estimators.

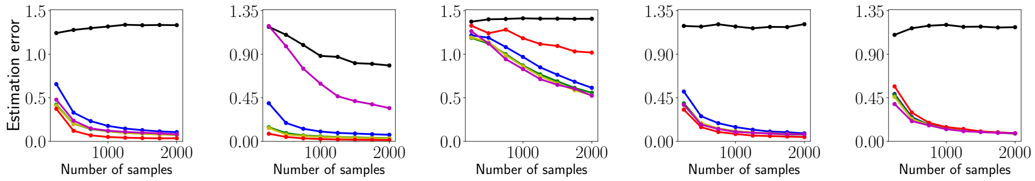

Figure 5.2: Performance with heavy-tailed noise. Weighted SIR, Ustats perform best. Least squares does not converge. Quantile regression is robust.

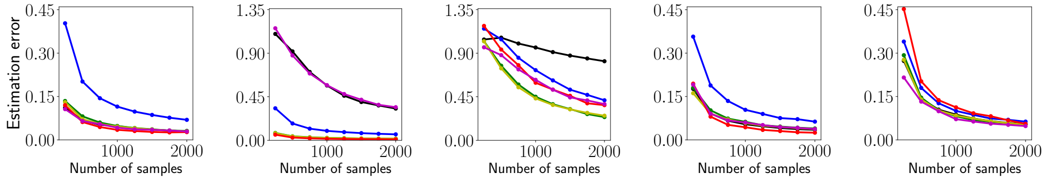

Figure 5.3: Performance with outlier noise. Weighted SIR, Ustats perform best. Quantile regression is robust.

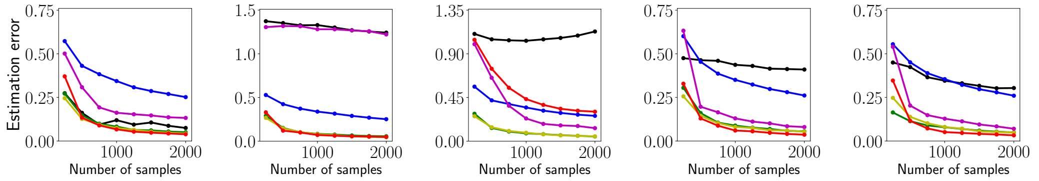

Figure 5.4: Performance with heavy tailed design, noise. Weighted SIR and Ustats again perform best and quantile regression is robust.

with five different ℓ_1 regularized estimators: (i) Least squares estimator with $\eta = \frac{1}{n}X^T y$ (ls); (ii) unweighted sliced inverse regression estimator (unwt-sir); (iii) weighted sliced inverse regression estimator with $n/2$ slices (wt-sir); (iv) U-statistics based estimator proposed in [40] (ustats); and (v) quantile regression estimator for $\tau = 0.5$ quantile (qr). For the other datasets we also compare with the ℓ_1 regularized least squares estimator on truncated data as proposed in [59] (ls-trunc).

Three observations summarizing the results of the numerical experiments are immediately evident. First, for mean SIM, the U-statistics estimator [40] and weighted SIR estimator perform equally or better than other estimators for each data, function combination. Both vanilla least squares and least squares applied on truncated data [59] fail to estimate the true parameter in certain settings. Second, the U-statistics based estimator, in contrast to observations made in [40], works for non-monotonic functions too like the weighted SIR estimator. Third, the performance of the quantile regression estimator supports the observations made in Section 5.3.

Appendix

5.A Background and Preliminaries

We assume the covariates to satisfy the following condition,

$$E\left[x|\langle x, \theta^*\rangle\right] = \mu + \frac{\langle x-\mu, \theta^*\rangle}{\theta^{*T}\Sigma\theta^*}\Sigma\theta^* \; , \tag{5.20}$$

where $\mu = E[x]$. The class of elliptical distributions [27, 57, 79] are known to satisfy condition (5.20). An example of an elliptical distribution is the multivariate Gaussian distribution.

We give the proof for Theorem 15 below.

Theorem 15(Theorem 2.1 in [51]) *Consider a single index model with elliptically symmetric covariates. Then the inverse regression curve $E[x|y]$ falls along a line:*

$$E[x|y] = \mu + \beta(y)\Sigma\theta^* \; , \tag{5.21}$$

where $\mu = E[x]$, $\Sigma = cov(x)$ and β is a scalar function of y:

$$\beta(y) = \frac{E[\langle x-\mu, \theta^*\rangle|y]}{(\theta^*)^T\Sigma\theta^*} \; . \tag{5.22}$$

Proof: The result follows from the following observations,

$$E[x|y] = E[E[x|\langle x, \theta^*\rangle]|y] \tag{5.23}$$

$$= \mu + \frac{E[\langle x-\mu, \theta^*\rangle|y]}{(\theta^*)^T\Sigma\theta^*}\Sigma\theta^* \tag{5.24}$$

$$= \mu + \beta(y)\Sigma\theta^* , \tag{5.25}$$

where in the second line we use condition 5.20.

Throughout we will assume the covariates to be centered so that $\mu = 0$ and $(\theta^*)^T\Sigma\theta^* = \|\Sigma^{1/2}\theta^*\|_2^2 = 1$.

5.B Comparison of Estimators

We show past estimators [111, 92, 106, 40] are variants of the following constrained estimator:

$$\hat{\theta} := \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n}\|X\theta\|_2^2 - \langle \eta, \theta\rangle \quad \text{s.t.} \quad R(\theta) \leq \beta R(\theta^*) \ , \ , \tag{5.26}$$

under the assumption of elliptically symmetric covariates satisfying the following property:

$$E\left[x|\langle x, \theta^*\rangle\right] = \frac{\langle x, \theta^*\rangle}{\theta^{*T}\Sigma\theta^*}\Sigma\theta^* = \langle x, \theta^*\rangle\Sigma\theta^* \ , \tag{5.27}$$

where $\Sigma = E[xx^T]$ is the covariance matrix of the rows of the design matrix and in the second inequality above we have used the assumption that $\|\Sigma^{1/2}\theta^*\|_2 = 1$. The variable $\eta \in \mathbb{R}^p$ in equation (5.26) satisfies the following properties:

(a) $\eta = \frac{1}{n}X^T\tilde{y}$, with $\tilde{y} \in \mathbb{R}^n$ computed differently for each estimator. The computation of the quantity $\tilde{y}$ is influenced by the response.

(b) Consider the derivative of the loss function in (5.26) without the constraint,

$$\mathcal{F}(\theta) = \frac{1}{n}X^TX\theta - \eta \ . \tag{5.28}$$

We show η is such that $E_{X,\omega}[\eta] = \beta\Sigma\theta^*$ with β varying with the estimator and transfer function. Therefore the true parameter θ^* satisfies $E_{X,\omega}\left[\mathcal{F}(\beta\theta^*)\right] = 0$. The result further implies that $E_{X,\omega}[\hat{\theta}] = \beta\theta^*$ and hence (5.26) is an unbiased estimator of $\beta\theta^*$.

Theorem 16 *The constrained least squares estimator [111], sliced inverse regression [92, 93, 106] and U-statistics based estimator [40] are equivalent to the constrained estimator (5.26).*

1. *For the constrained least squares estimator [111, 23],*

$$\eta = \frac{1}{n} X^T y, \quad \beta = E_{X,\omega}\left[f(\langle x, \theta^* \rangle)\langle x, \theta^* \rangle\right] \tag{5.29}$$

2. *For the sliced inverse regression estimator [92, 93, 106], let* $y^\uparrow$ *denote the response variable sorted by ascending order. Divide the range of* $y^\uparrow$ *into* H *contiguous slices denoted by* I_h*. Let* $\gamma_h,\ h = 1, \cdots, H$ *are bounded weights assigned to each slice such that* $\|\eta\|_2 = 1$*. Then,*

$$\eta = \frac{1}{n} X^T \tilde{y}, \quad \tilde{y}_i = \sum_{h=1}^{H} \gamma_h \mathbb{I}[y_i \in I_h]; \quad \beta = 1 \tag{5.30}$$

3. *For the U-statistics based estimator [40], let* $y^\uparrow$ *denote the response variable sorted by ascending order. Then* $\tilde{y}$ *are assigned values between* $[-2, 2]$ *based on the ordering in* $\mathbf{y}^\uparrow$*. More formally let* $i^\uparrow \in \{0, \ldots, n-1\}$ *denote the position of response variable* y_i *in* $y^\uparrow$*.*

$$\begin{aligned} \eta &= \frac{1}{n} X^T \tilde{y}, \quad \tilde{y}_i = -2 + \frac{2 * i^\uparrow}{(n-1)} \\ \beta &= E_{X,\omega}\left[\sum_{i,j \in [1,\ldots,n]; i \neq j} sign(y_i - y_j) \cdot \langle x_i - x_j, \theta^* \rangle\right] \end{aligned} \tag{5.31}$$

Moreover for all estimators $E_{X,\omega}[\eta] = \beta \Sigma \theta^*$ *so that* $E_{X,\omega}\left[\mathcal{F}(\beta\theta^*)\right] = 0$ *where* $\mathcal{F}(\cdot)$ *is the derivate of the loss function.*

Proof:

1. Constrained least squares estimator

Let $\eta = \frac{1}{n} X^T y$. First consider the quantity $E\left[\eta\right]$. The result below is from [24, 23].

$$\begin{aligned} E_{X,\omega}[\eta] = \text{cov}(y, x) &= \text{cov}\left(f(\langle x, \theta^* \rangle), E_x(x \mid \langle x, \theta^* \rangle)\right) \\ &= E_{X,\omega}(f(\langle x, \theta^* \rangle) \cdot \langle x, \theta^* \rangle)\Sigma\theta^* \\ &= \beta\Sigma\theta^* . \end{aligned}$$

The first line above follows from the assumed SIM model and the assumption that the noise ω is independent of x. The second line follows from the assumption that x is elliptically symmetric and application of (5.27). Now $E_{X,\omega}\left[\mathcal{F}(\beta\theta^*)\right] = 0$ follows by observing that $E_X\left[\frac{1}{n}X^TX\theta^*\right] = \Sigma\theta^*$.

2. Sliced Inverse Regression

In sliced inverse regression, the response is sorted in ascending order to obtain $y^{\uparrow}$, the range of $y^{\uparrow}$ is divided into H slices $I_1, \ldots, I_H$. The mean of the corresponding x's in each slice is computed.

$$\bar{x}_h = \frac{1}{n_h}\sum_{i:y_i\in I_h} x_i \quad \text{for} \quad h \in [1, \ldots, H]\ , \tag{5.32}$$

where n_h is the number of samples in slice h. Then the matrix Γ is constructed as follows,

$$\Gamma = \sum_{h=1}^{H} \hat{p}_h \bar{x}_h \bar{x}_h^T\ , \tag{5.33}$$

where $\hat{p}_h = n_h/n$ is the proportion of samples in the hth slice. The quantity η is the eigenvector corresponding to the largest eigenvalue of Γ. Let λ here denote the largest eigenvalue. Therefore the following is true,

$$\begin{aligned}\lambda\eta &= \sum_{h=1}^{H}\frac{n_h}{n}\bar{x}_h\langle\bar{x}_h, \eta\rangle \\ &= \frac{1}{n}\sum_{h=1}^{H} n_h \cdot \frac{1}{n_h}\sum_{i:y_i\in I_h} x_i\gamma_h' \\ \Rightarrow\eta &= \frac{1}{n}\sum_{i=1}^{n} x_i\tilde{y}_i\ .\end{aligned}$$

In the first line λ is the largest eigenvalue of Γ and η is the corresponding eigenvector. In the second line we substitute the value of $\bar{x}_h$ and assign $\gamma_h' = \langle\bar{x}_h, \eta\rangle = \frac{1}{n_h}\sum_{i:y_i\in I_h}\langle x_i, \eta\rangle$. In the third line we use $\tilde{y}_i = \frac{1}{\lambda}\sum_{h=1}^{H}\gamma_h'\mathbb{I}[y_i \in I_h] = \sum_{h=1}^{H}\gamma_h\mathbb{I}[y_i \in I_h]$ which proves claim (5.30).

For the second results we first prove that $E[\eta] = \Sigma\theta^*$. Note the following,

$$\begin{aligned}E_{X,\omega}[\bar{x}_h] &= E_{X,\omega}[x|y \in I_h] \\ &= E_{X,\omega}[E_x[x|\langle x, \theta^*\rangle]|y \in I_h] \\ &= E_{X,\omega}[\langle x, \theta^*\rangle|y \in I_h]\Sigma\theta^* \\ &= \beta_h\Sigma\theta^*\ .\end{aligned}$$

The equality on the second line follows from the law of total expectation. The equality on the third line follows from (5.27). Therefore the following is true,

$$\begin{aligned} E_{X,\omega}[\Gamma] &= E_{X,\omega}[\sum_{h=1}^{H} \hat{p}_h \bar{x}_h \bar{x}_h^T] \\ &= \sum_{h=1}^{H} p_h \beta_h^2 (\Sigma\theta^*)(\Sigma\theta^*)^T \ . \end{aligned}$$

The equality on the second line uses $E_{X,\omega}[\hat{p}_h] = p_h$. Therefore in expectation $\eta = \Sigma\theta^*$ is the eigenvector of Γ with eigenvalue $\lambda = \sum_{h=1}^{H} p_h \beta_h^2$, which proves the that $E_{X,\omega}[\eta] = \Sigma\theta^*$. This now implies the following,

$$E_{X,\omega}\left[\frac{1}{n} X^T X \theta^* - \eta\right] = \Sigma\theta^* - \Sigma\theta^* = 0 \ , \tag{5.34}$$

thus proving $E_{X,\omega}\left[\mathcal{F}(\beta\theta^*)\right] = 0$ with $\beta = 1$.

3. U-statistics based estimator

The general algorithm based on U-statistics [40] first computes the quantity $u((x_1, y_1), \ldots, (x_m, y_m)) = \sum_{i=1}^{m} q_i(y_1, \ldots, y_m) \cdot x_i$ from any m i.i.d. observations $(x_1, y_1), \ldots, (x_m, y_m)$ with bounded functions $q_i : \mathbb{R}^m \mapsto [-1, 1]$.

$$u((x_1, y_1), \ldots, (x_m, y_m)) = \sum_{i=1}^{m} q_i(y_1, \ldots, y_m) \cdot x_i \ . \tag{5.35}$$

Noting that there are $\frac{n!}{(n-m)!}$ ways of choosing m samples from n samples, the quantity η in (5.26) is computed as follows (equation (7) in [40]):

$$\eta = \frac{(n-m)!}{n!} \sum_{\substack{1 \leq i_1, \ldots, i_m \leq n, \\ i_1 \neq \ldots \neq i_m}} u((x_{i_1}, y_{i_1}), \ldots, (x_{i_m}, y_{i_m})) \ . \tag{5.36}$$

An example in [40] with $m = 2$ for monotonic functions which we use in the statement of Theorem 2 is the following:

$$\eta = \frac{(n-2)!}{n!} \sum_{\substack{1 \leq i,j \leq n, \\ i \neq j}} \text{sign}(y_i - y_j) \cdot (x_i - x_j) \tag{5.37}$$

We first prove a general result starting from formulation (5.36). The claim (5.31) in Theorem 2 will then be an application of the general result for estimator (5.37).

Consider formulation (5.36). Note that out of the $\frac{n!}{(n-m)!}$ sample subsets of size m chosen from n samples each sample $i,\ 1 \leq i \leq n$ is contained in $\frac{(n-1)!m}{(n-m)!}$ subsets. This is because assuming sample i is chosen in a given subset, the remaining $m-1$ samples are chosen from $n-1$ samples. Additionally within each m-size subset containing the sample i there are m different ways of arranging the samples amongst themselves. Additionally, each $|q_i| \leq 1$. Therefore the following is true,

$$
\begin{aligned}
-\frac{(n-m)!}{n!}\frac{(n-1)!m}{(n-m)!} = -\frac{m}{n} \leq \frac{(n-m)!}{n!} & \sum_{\substack{i,1\leq i_1,\ldots,i_{m-1}\leq n,\\ i_1\neq\ldots i_{m-1}\neq i}} q_i(y_i, y_{i_1}, \ldots, y_{i_{m-1}}) \\
& \leq \frac{(n-m)!}{n!}\frac{(n-1)!m}{(n-m)!} = \frac{m}{n} \ . \qquad (5.38)
\end{aligned}
$$

It therefore follows that $\eta = \frac{1}{n}X^T\tilde{y}$ with $\tilde{y}_i \in [-m, m]$. Also consider the quantity $E_{X,\omega}[\eta]$

$$
\begin{aligned}
E_{X,\omega}[\eta] &= E_{X,\omega}[u((x_1, y_1), \ldots, (x_m, y_m))] \\
&= E_{X,\omega}[\sum_{i=1}^{m} q_i(y_1, \ldots, y_m) \cdot x_i] \\
&= E_{X,\omega}[E_{X,\omega}[\sum_{i}^{m} q_i(y_1, \ldots, y_m) \cdot x_i | \langle x_1, \theta^* \rangle, \ldots, \langle x_m, \theta^* \rangle] \\
&= E_{X,\omega}[\sum_{i}^{m} q_i(y_1, \ldots, y_m) \cdot E_x[x_i | \langle x_i, \theta^* \rangle]] \\
&= \beta\Sigma\theta^* \ . \qquad (5.39)
\end{aligned}
$$

The first two equalities follow from definition of η. The third equality is because of the law of total expectation. The fourth inequality follows because given $\langle x_i, \theta^* \rangle$, y_i is independent of x_i and also due to our assumption that the x_i's are i.i.d. The equality on the fifth line follows from (5.27) with $\beta = E_{X,\omega}[\sum_{i=1}^{m} q_i(y_1, \ldots, y_m)\langle x_i, \theta^* \rangle]$.

Now consider the formulation for monotonic function in (5.37). We have $m = 2$, therefore using result (5.38), $\eta = \frac{1}{n}X^T\tilde{y}$ with $\tilde{y}_i \in [-2, 2]$. Also consider $y^{\uparrow}$ to be the vector obtained by sorting the values of the response vector y. Denote element i of $y_i^{\uparrow}$, indexed starting at 0. Let $x_i^{\uparrow}$ denote the corresponding covariates and $\tilde{y}_i^{\uparrow}$ denote the corresponding weights. Consider the zeroth element in $y^{\uparrow}$. It is evident from (5.37) that $x_1^{\uparrow}$ will always get a -1 weight in the summation. So it follows that $\tilde{y}_1^{\uparrow} = -2$. Now consider $y_1^{\uparrow}$. In formulation (5.37), $x_1^{\uparrow}$ will receive weight $\frac{-2*(n-1)!}{(n-2)!}+2$ times since $x^{\uparrow}_1$ will always receive a negative weight except for two times when i,j correspond to elements $y^{\uparrow}_0$ and $y^{\uparrow}_1$. Therefore effectively the weight $\tilde{y'}^{\uparrow}_1$ can be computed as follows:

$$\tilde{y'}^{\uparrow}_1 = \frac{\frac{-2*(n-1)!}{(n-2)!}+2}{\frac{n!}{(n-2)!}} = \frac{-2*(n-1)!+2*(n-2)!}{n!} = -\frac{2}{n}+\frac{2}{(n-1)*n} \tag{5.40}$$

Therefore $\tilde{y}^{\uparrow}_1 = n * \tilde{y'}^{\uparrow}_1 = -2+\frac{2}{n-1}$. Inductively similar arguments can be made for $\tilde{y}^{\uparrow}_i$ for $i \in [2,3,\ldots,n-1]$. Also from argument (5.39) it can be easily deduced that,

$$\beta = E_{X,\omega}\left[\sum_{i,j\in[1,\ldots,n];i\neq j} \text{sign}(y_i-y_j)\cdot\langle x_i-x_j,\theta^*\rangle\right] , \tag{5.41}$$

and

$$E_{X,\omega}[\eta] = \beta\Sigma\theta^* . \tag{5.42}$$

We have thus proved claim (5.31) and $E_{X,\omega}\left[\mathcal{F}(\beta\theta^*)\right]=0$. ■

5.C Robust Sliced Inverse Regression Based Estimators

We provide the proof for Lemma 19 and Theorem 17.

Lemma 19 *Consider the estimate of η in Algorithm 1. For some $\tilde{y}$ with $\tilde{y}_i \in [-1,1], 1 \leq i \leq n$, the quantity η can be equivalently written as $\eta = \frac{1}{n}X^T\tilde{y}$. Also the following is true,*

$$E_{X,\omega}\left[\frac{\beta}{n}X^TX\theta^* - \eta\right] = 0 , \tag{5.43}$$

with $\beta = E_{X,\omega}\left[\frac{1}{H}\sum_{h=1}^{H}\gamma_h\langle x,\theta^*\rangle \;\middle|\; y\in I_h\right]$ where γ is defined in step 3 of Algorithm 1.

Proof: It is evident from Algorithm 1 that $\tilde{y}_i \in [-1,1]$. Also condition 5.20 for centered covariates implies the following for all slices $1 \leq h \leq H$.

$$\bar{x}_h = E[\langle x,\theta^*\rangle | y \in I_h]\Sigma\theta^* , \tag{5.44}$$

where we have used the assumption that $(\theta^*)^T\Sigma\theta^* = 1$. From the computation of η in Algorithm 1 it immediately follows that,

$$E_{X,\omega}[\eta] = \beta\Sigma\theta^* , \tag{5.45}$$

where

$$\beta = E_{X,\omega}\left[\frac{1}{H}\sum_{h=1}^{H}\gamma_h\langle x, \theta^*\rangle \;\middle|\; y \in I_h\right] . \tag{5.46}$$

Also,

$$E_{X,\omega}\left[\frac{\beta}{n}X^T X\theta^* - \eta\right] = 0 . \tag{5.47}$$

This proves the stated result. ■

Theorem 17 *Let $X \in \mathbb{R}^{n\times p}$ have independent sub-Gaussian elliptically symmetric rows x_i with $E[x_i] = 0$, $cov(x) = \Sigma$ and sub-Gaussian norm $\|x_i\Sigma^{-1/2}\|_{\psi_2} \leq \kappa$. Let $\Lambda_{\max}(\Sigma)$ denote the largest eigenvalue of Σ and $\Lambda_{\max}(\Sigma|A) = \sup_{u\in A} u^T\Sigma u$. Then for the sliced inverse regression estimator in Algorithm 1 when $n > \Theta(w^2(A))$ the following estimation error bound holds with probability atleast $1 - 2\exp(-\nu_1 w^2(A)) - \nu_2\exp\left(-\left(\frac{\tau}{\nu_3\kappa\phi}\right)^2\right)$:*

$$\|\Delta\|_2 = \|\hat{\theta} - \theta^*\|_2 \leq O\left(\frac{\Psi(A)\kappa\sqrt{\Lambda_{\max}(\Sigma)}w(\Omega_R) + \beta\Lambda_{\max}(\Sigma|A)w(A)}{\sqrt{n}}\right) , \tag{5.48}$$

where β is the value defined in Lemma 1 and $\eta, \nu_1, \nu_2, \nu_3, \tau$ are absolute constants.

Proof: The loss function is minimized at $\beta\theta^*$ in expectation. Let the empirical loss be minimized at $\hat{\theta} = \beta\theta^* + \Delta$. Because of the constraint the error vector lies in the following set,

$$E_c = \{\Delta \mid R(\beta\theta^* + \Delta) \leq R(\beta\theta^*)\} . \tag{5.49}$$

Since the loss function is minimized at $\beta\theta^* + \Delta$ the following is true,

$$\begin{aligned}
0 &\geq \left(\frac{1}{2n}\|X(\beta\theta^* + \Delta)\|_2^2 - \langle\eta, \beta\theta^* + \Delta\rangle\right) - \left(\frac{1}{2n}\|X\beta\theta^*\|_2^2 - \langle\eta, \beta\theta^*\rangle\right) \\
&\geq \frac{1}{2n}\|X\Delta\|_2^2 + \frac{\beta}{n}\langle X\theta^*, X\Delta\rangle - \langle\eta, \Delta\rangle \\
&\geq \frac{1}{2n}\|X\Delta\|_2^2 + \langle\Delta, \frac{\beta}{n}X^T X\theta^* - \eta\rangle \\
&\geq \frac{1}{2n}\|X\Delta\|_2^2 + \langle\Delta, \frac{\beta}{n}X^T X\theta^* - \beta\Sigma\theta^* - \eta + \beta\Sigma\theta^*\rangle
\end{aligned}$$

Therefore the following is true,

$$\begin{aligned}
\frac{1}{2n}\|X\Delta\|_2^2 &\leq |\langle\Delta, \eta - \beta\Sigma\theta^*| + \left|\langle\Delta, \frac{\beta}{n}X^T X\theta^* - \beta\Sigma\theta^*\rangle\right| \\
&\leq R(\Delta)R^*(\eta - \beta\Sigma\theta^*) + \left|\langle\Delta, \frac{\beta}{n}X^T X\theta^* - \beta\Sigma\theta^*\rangle\right| .
\end{aligned} \tag{5.50}$$

We will first provide upper bounds for $R^*(\eta - \beta\Sigma\theta^*)$. For that we first prove that $\eta - \beta\Sigma\theta^*$ is a sub-Gaussian random vector with sub-Gaussian norm $\frac{\kappa\sqrt{\Lambda_{\max}(\Sigma)}}{\sqrt{n}}$ where $\Lambda_{\max}(\Sigma)$ is the maximum eigenvalue of the covariance matrix Σ. Remember that $\eta = \frac{1}{H}\sum_{h=1}^{H}\gamma_h\bar{x}_h$. Also $\bar{x}_h = \frac{1}{\lfloor n/H \rfloor}\sum_{y_i \in I_h} x_i$ where the range of $y^{\uparrow}$, the response sorted in ascending order, is divided into H intervals $[I_1, \dots, I_H]$ such that the number of samples in the hth slice is $\lfloor n/H \rfloor$. For ease of exposition, without loss of generality, we assume that the number of samples is such that n/H is an integer. Note that $E[\bar{x}_h] = E[\langle x_i, \theta^*\rangle | y_i \in I_h]\Sigma\theta^*$ and $\frac{1}{(n/H)}\sum_{y_i \in I_h} x_i$ is an empirical estimator of $\bar{x}_h$. Also x_i are sub-Gaussian random vectors having sub-Gaussian norm,

$$\begin{aligned} \|x_i\|_{\psi_2} &= \sup_{u \in S^{p-1}} \|\langle x_i, u\rangle\|_{\psi_2} \\ &= \sup_{u \in S^{p-1}} \|\langle \Sigma^{1/2} g, u\rangle\|_{\psi_2} \le \sqrt{\Lambda_{\max}(\Sigma)}\kappa\,, \end{aligned} \tag{5.51}$$

where $g \in \mathbb{R}^p$ above is a sub-Gaussian random vector with i.i.d. entries and having sub-Gaussian norm $\|g\|_{\psi_2} \le \kappa$. Therefore by the rotation invariance property in Lemma 2 and result of Lemma 4 $\frac{1}{(n/H)}\sum_{y_i \in I_h}(x_i - E[\langle x_i, \theta^*\rangle | y_i \in I_h]\Sigma\theta^*)$ is a centered sub-Gaussian random vector with sub-Gaussian norm,

$$\begin{aligned} \|\bar{x}_h - E[\langle x_i, \theta^*\rangle | y_i \in I_h]\Sigma\theta^*\|_{\psi_2}^2 &= \left\| \frac{1}{(n/H)} \sum_{y_i \in I_h} x_i - E[\langle x_i, \theta^*\rangle | y_i \in I_h]\Sigma\theta^* \right\|_{\psi_2}^2 \\ &\le \frac{c}{(n/H)^2} \sum_{y_i \in I_h} \|x_i - E[\langle x_i, \theta^*\rangle | y_i \in I_h]\Sigma\theta^*\|_{\psi_2}^2 \\ &\le \frac{2c}{(n/H)^2} \sum_{y_i \in I_h} \|x_i\|_{\psi_2}^2 \\ &\le \frac{c_1\kappa\sqrt{\Lambda_{\max}(\Sigma)}(n/H)}{(n/H)^2}\,, \end{aligned} \tag{5.52}$$

where the second line following from the application of Lemma 4 followed by roational invariance property in Lemma 2, third line is due to Remark 5.18 in [127]. Therefore, we get,

$$\|\bar{x}_h - E[\langle x_i, \theta^*\rangle | y_i \in I_h]\Sigma\theta^*\|_{\psi_2} \le \frac{c_1\kappa\sqrt{\Lambda_{\max}(\Sigma)}}{\sqrt{n/H}} \tag{5.53}$$

Again note by construction, $\bar{x}_h,\ h \in [1, \dots, H]$ are independent of each other. Therefore rotational invariance property of Lemma 2 and result in Lemma 4 as above $\frac{1}{H}\sum_{h=1}^{H}\gamma_h(\bar{x}_h -$

$E[\langle x_i, \theta^* \rangle | y_i \in I_h] \Sigma\theta^*)$ is a centered sub-Gaussian random variable with sub-Gaussian norm,

$$\begin{aligned} \left\| \frac{1}{H} \sum_{h=1}^{H} \gamma_h (\bar{x}_h - E[\langle x_i, \theta^* \rangle | y_i \in I_h] \Sigma\theta^*) \right\|_{\psi_2} &\leq \frac{c\kappa\sqrt{\Lambda_{\max}}}{\sqrt{n/H}} \sqrt{\sum_{h=1}^{H} \frac{\gamma_h^2}{H^2}} \\ &\leq \frac{c\kappa\sqrt{\Lambda_{\max}}}{\sqrt{n}} , \end{aligned} \tag{5.54}$$

where in the last equation $\sqrt{\sum_{h=1}^{H} \frac{\gamma_h^2}{H^2}} \leq \sqrt{\frac{1}{H}}$ since $\gamma_h \leq 1,\ h \in [1, \ldots, H]$. Also,

$$\frac{1}{H} \sum_{h=1}^{H} \gamma_h (\bar{x}_h - E[\langle x_i, \theta^* \rangle | y_i \in I_h] \Sigma\theta^*) = \eta - \beta\Sigma\theta^* , \tag{5.55}$$

where $\beta = \frac{1}{H} \sum_{h=1}^{H} \gamma_h E[\langle x_i, \theta^* \rangle | y_i \in I_h]$. Hence it follows from (5.54) and (5.55) that $\eta - \beta\Sigma\theta^*$ is a sub-Gaussian random vector with sub-Gaussian norm $\kappa\sqrt{\Lambda_{\max}(\Sigma)}/\sqrt{n}$. To get high probability bounds on $R^*(\eta - \beta\Sigma\theta^*)$, note the following,

$$R^*(\eta - \beta\Sigma\theta^*) \leq \sup_{u \in \Omega_R} \langle \eta - \beta\Sigma\theta^*, u \rangle = \sup_{u \in \Omega_R} \langle h, u \rangle , \tag{5.56}$$

where Ω_R is the unit norm ball and h is a sub-Gaussian random vector with sub-Gaussian norm $\|h\|_{\psi_2} \leq \frac{c\kappa\sqrt{\Lambda_{\max}(\Sigma)}}{\sqrt{n}}$. The bound on $R^*(\eta - \beta\Sigma\theta^*)$ now follows directly from the following Theorem 9 in [11].

Theorem 19 *Let $\Omega_R = \{u : R(u) \leq 1\}$ be the unit norm ball of $R(\cdot)$. Assuming h is any centered sub-Gaussian random vector with $\|h\|_{\psi_2} \leq \kappa$, then we have for any $t > 0$,*

$$P\left(\sup_{u \in \Omega_R} \langle h, u \rangle \geq \nu_0 \kappa w(\Omega_R) + t \right) \leq \nu_1 \exp\left(-\left(\frac{t}{\nu_2 \kappa \phi} \right)^2 \right) , \tag{5.57}$$

where ν_0, ν_1, ν_2 are universal constants, and $\phi = \sup_{R(u) \leq 1} \|u\|_2$.

We therefore have established the following,

$$P\left(R^*\left(\eta - \frac{\beta}{n} X^T X \theta^* \right) \geq \frac{c\kappa\sqrt{\Lambda_{\max}} w(\Omega_R) + t}{\sqrt{n}} \right) \leq \nu_2 \exp\left(-\left(\frac{t}{\nu_3 \kappa \sqrt{\Lambda_{\max}(\Sigma))} \phi} \right)^2 \right) . \tag{5.58}$$

Now to bound the value of $R(\Delta)$ in terms of $\|\Delta\|_2$, let $\Psi(A) = \sup_{u \in A} \frac{R(u)}{\|u\|_2}$ denote the norm compatibility constant [105, 11] in the error set. We get,

$$R(\Delta) \leq \|\Delta\|_2 \Psi(A) . \tag{5.59}$$

Next we bound give lower bounds on $\inf_{u \in A = \text{cone}(E_c) \cap S^{p-1}} \|Xu\|_2^2$. This is equivalent to the Restricted Eigenvalue condition popular in prior literature on high dimensional estimation [17, 105, 11]. We use the following Theorem 12 from [11].

Theorem 20 *Let X be a design matrix with independent anisotropic rows, i.e., $E[x_i^T x_i] = \Sigma$ and $\|x_i \Sigma^{-1/2}\|_{\psi_2} \leq \kappa$. Then, for absolute constants $\nu_1, c > 0$, with probability atleast $(1 - 2\exp(-\nu_1 w^2(A)))$, we have*

$$\sup_{u \in A} \left| \frac{1}{n} \frac{1}{u^T \Sigma u} \|Xu\|_2^2 - 1 \right| \leq c \frac{w(A)}{\sqrt{n}} . \tag{5.60}$$

Further,

$$\Lambda_{\min}(\Sigma|A) \left(1 - c\frac{w(A)}{\sqrt{n}}\right) \leq \inf_{u \in A} \frac{1}{n} \|Xu\|_2^2 \leq \sup_{u \in A} \frac{1}{n} \|Xu\|_2^2 \leq \Lambda_{\max}(\Sigma|A) \left(1 + c\frac{w(A)}{\sqrt{n}}\right) , \tag{5.61}$$

where $\Lambda_{\min}(\Sigma|A) = \inf_{u \in A} u^T \Sigma u$, and $\Lambda_{\max}(\Sigma|A) = \sup_{u \in A} u^T \Sigma u$ are the restricted minimum and maximum eigenvalues of Σ restricted to $A \subseteq S^{p-1}$.

Applying the above result for the set $A = \mathit{cone}(E_c) \cap S^{p-1}$ we get the following,

$$P\left(\inf_{u \in A} \frac{1}{n} \|Xu\|_2^2 \leq \Lambda_{\min}(\Sigma|A) \left(1 - c\frac{w(A)}{\sqrt{n}}\right)\right) \leq 2\exp(-\nu_1 w^2(A)) . \tag{5.62}$$

Now since $\Delta \in E_c$ from the above result we conclude that for $k = \Lambda_{\min}(\Sigma|A) \left(1 - c\frac{w(A)}{\sqrt{n}}\right) > 0$ when $n = \Theta(w^2(A))$, where k is called the RE constant,

$$P\left(\inf_{u \in A} \frac{1}{n} \|X\tilde{\Delta}\|_2^2 \leq k\right) \leq 2\exp(-\nu_1 w^2(A)) , \tag{5.63}$$

where $\tilde{\Delta} = \Delta / \|\Delta\|_2 \in A$ is a unit norm vector in set A.

Next we derive upper bounds on $\left|\langle \Delta, \frac{\beta}{n} X^T X \theta^* - \beta \Sigma \theta^* \rangle\right|$. We note the following,

$$\begin{aligned} \left|\left\langle \Delta, \frac{\beta}{n} X^T X \theta^* - \beta \Sigma \theta^* \right\rangle\right| &= \left|\left\langle \Delta, \left(\frac{1}{n} X^T X - \Sigma\right) \beta \theta^* \right\rangle\right| \\ &= \|\Delta\|_2 \beta \|\theta^*\|_2 \left|\left\langle \tilde{\Delta}, \left(\frac{1}{n} X^T X - \Sigma\right) \frac{\theta^*}{\|\theta^*\|_2} \right\rangle\right| , \end{aligned} \tag{5.64}$$

where in the second equality we use $\tilde{\Delta} = \frac{\Delta}{\|\Delta\|_2}$. Now note that when $n = \Theta(w^2(A))$ then from Theorem 20 $\frac{1}{n}X^TX - \Sigma$ is positive definite restricted to vectors in the set A. Also the vector $-\theta^*$ lies in the error set. Hence we have the following result,

$$P\left(\left|\left\langle \Delta, \frac{\beta}{n}X^TX\theta^* - \beta\Sigma\theta^* \right\rangle\right| \geq \frac{c\beta\|\Delta\|_2\Lambda_{\max}(\Sigma|A)w(A)}{\sqrt{n}}\right) \leq 2\exp(-\nu_1 w^2(A)) \,. \tag{5.65}$$

Now combining the results of equations (5.50), (5.58), (5.59), (5.63), (5.65) we get the following result with probability atleast $1 - 2\exp(-\nu_1 w^2(A)) - \nu_2 \exp\left(-\left(\frac{\tau}{\nu_3\kappa\phi}\right)^2\right)$

$$\begin{aligned} k\|\Delta\|_2^2 &\leq \frac{1}{n}\|X\Delta\|_2^2 \\ &\leq c\frac{\|\Delta\|_2\Psi(A)\kappa\sqrt{\Lambda_{\max}}w(\Omega_R) + \beta\|\Delta\|_2\Lambda_{\max}(\Sigma|A)w(A)}{\sqrt{n}} \,. \end{aligned} \tag{5.66}$$

Dividing by $\|\Delta\|_2$ throughout we get,

$$\|\Delta\|_2 \leq c\frac{\Psi(A)\kappa\sqrt{\Lambda_{\max}}w(\Omega_R) + \beta\Lambda_{\max}(\Sigma|A)w(A)}{k\sqrt{n}} \,, \tag{5.67}$$

which is the result we set out to prove. ■

Corollary 24 *Consider the mean single index model with sub-Gaussian elliptically symmetric design matrix.*

1. *Let $R(\cdot)$ be the ℓ_1 norm and θ^* be s-sparse. Then with probability atleast $1 - 2\exp(-\nu_1 s\log p) - \nu_2\exp\left(-\left(\frac{\tau}{\nu_3\kappa\phi}\right)^2\right)$ the estimation error for the estimator in Algorithm 1 satisfies,*

$$\|\Delta\|_2 \leq O\left(\frac{\kappa\sqrt{\Lambda_{\max}}\sqrt{s\log p} + \beta\Lambda_{\max}(\Sigma|A)\sqrt{s\log p}}{\sqrt{n}}\right) \,. \tag{5.68}$$

2. *Let $R(\cdot)$ be the ℓ_1/ℓ_2 nonoverlapping group sparse norm. Then with probability atleast $1 - 2\exp(-\nu_1(ls_{\mathcal{G}} + s_{\mathcal{G}}\log N_{\mathcal{G}})) - \nu_2\exp\left(-\left(\frac{\tau}{\nu_3\kappa\phi}\right)^2\right)$ the estimation error for the estimator in Algorithm 1 satisfies,*

$$\|\Delta\|_2 \leq O\left(\frac{\kappa\sqrt{\Lambda_{\max}}\sqrt{ls_{\mathcal{G}} + s_{\mathcal{G}}\log N_{\mathcal{G}}} + \beta\Lambda_{\max}(\Sigma|A)\sqrt{ls_{\mathcal{G}} + s_{\mathcal{G}}\log N_{\mathcal{G}}}}{\sqrt{n}}\right) \,. \tag{5.69}$$

3. *Let $R(\cdot)$ be the k-support norm with $k < s$. Then with probability atleast* $1 - 2\exp(-\nu_1 s \log p) - \nu_2 \exp\left(-\left(\frac{\tau}{\nu_3\kappa\phi}\right)^2\right)$ *the estimation error for the estimator in Algorithm 1 satisfies,*

$$\|\Delta\|_2 \leq O\left(\frac{\kappa\sqrt{\Lambda_{\max}}\sqrt{s\log p} + \beta\Lambda_{\max}(\Sigma|A)\sqrt{s\log p}}{\sqrt{n}}\right) . \tag{5.70}$$

Proof: The claims follow from the result of Theorem 3 and values of various quantities for the respective norms from Table 2.2 ■

5.D Quantile Single Index Models

We give proofs for claims on high-dimensional quantile single index models in Section 3.

Lemma 20: (Theorem 1 in [145]) *Assume the quantile single index model with elliptically symmetric covariates and function $f(\cdot)$ which is not even. Then the following is true,*

$$\beta\theta^*_\tau = \underset{\theta\in\mathbb{R}^p}{\text{argmin}}\, E_{x,\omega}\left[\rho_\tau(y - \langle x, \theta\rangle)\right] , \tag{5.71}$$

for some β depending on the data.

Proof: Consider the quantity $E_{x,\omega}\left[\rho_\tau(y - \langle x, \theta\rangle)\right]$.

$$\begin{aligned} E_{x,\omega}\left[\rho_\tau(y - \langle x, \theta\rangle)\right] &= E_{x,\omega}\left[E_x\left[\rho_\tau(y - \langle x, \theta\rangle)|\langle x, \theta^*_\tau\rangle, \omega\right]\right] \\ &\geq E_{x,\omega}\left[\rho_\tau(E_x[y|\langle x, \theta^*_\tau\rangle, \omega] - E_x[\langle x, \theta\rangle|\langle x, \theta^*_\tau\rangle, \omega])\right] \\ &\geq E_{x,\omega}\left[\rho_\tau(y - \beta\langle x, \theta^*_\tau\rangle)\right] . \end{aligned} \tag{5.72}$$

The first equality above follows from the iterative law of conditional expectation; the second inequality from Jensen's inequality since the quantile loss is convex; the third inequality follows since x is independent of the noise and from (5.27) we get $\beta = \theta^{*T}_\tau\Sigma\theta^*_\tau$. ■

Theorem 18 *Let $X \in \mathbb{R}^{n\times p}$ have independent sub-Gaussian elliptically symmetric rows x_i with $E[x_i] = 0$, $cov(x) = \Sigma$ and sub-Gaussian norm $\|x_i\Sigma^{-1/2}\|_{\psi_2} \leq \kappa$. Let $\Lambda_{\max}(\Sigma)$ denote the largest eigenvalue of Σ and $\Lambda_{\max}(\Sigma|A) = \sup_{u\in A} u^T\Sigma u$. Then when $n = \Theta(w^2(A))$ the estimation error for the constrained linear quantile regression estimator applied to the quantile SIM model satisfies the following bound with probability atleast* $1 - \exp(-\tau_1^2/2) - \exp(-\nu_1\phi_2^2\sqrt{n}w(A)) - \nu_2\exp\left(-\left(\frac{\tau_2}{c_2\phi\kappa}\right)^2\right)$,

$$\|\Delta\|_2 = \|\hat{\theta} - \theta^*_\tau\|_2 = O\left(\frac{\sqrt{\Lambda_{\max}(\Sigma)}\max\{\max\{\tau, 1-\tau\}\Psi(A)w(\Omega_R), w(A)\}}{\underline{f}\sqrt{n}}\right) , \tag{5.73}$$

where $0 < \underline{f}$ is a uniform lower bound on the conditional density of the distribution $y_i|x_i$ around the noise term $y_i - \langle x_i, \beta\theta^* \rangle$ and $\tau_1, \tau_2, \nu_1, \nu_2$ are absolute constants.

Proof: The proof has similar arguments as made in [118], although we analyze the constrained estimator while [118] analyzed the regularized estimator. We first prove the result on the number of interpolated samples (NIPS) which implies $n = \Theta(w^2(A))$ samples are sufficient for estimation. Note that if the NIPS property is not satisfied quantile regression does not work as there exists a vector in the error set when the empirical loss function evaluates to 0 [118]. Consider $\omega = y - \beta X\theta^*$, where β is the constant from Lemma 2, is the noise. Note that this is different than the actual noise $y_i - f(\langle x_i, \theta^* \rangle),\ 1 \leq i \leq n$. Consider the error set $E_c := \{\Delta = \hat{\theta} - \theta^* \mid R(\beta\theta^* + \Delta) \leq R(\beta\theta^*)\}$. If all samples are interpolated then there exists a $u \in E_c$ such that $\omega = Xu$. Therefore we will show that $\|\omega - Xu\|_2^2 > 0$ when the number of samples crosses a certain threshold.

Consider the subspace $D = \{v \mid v = \sum_{i=1}^{k_1} c_i a_i,\ c_i \geq 0,\ a_i \in \mathcal{A},\ \|v\|_0 \leq s_1\}$ be the set of vectors which are a linear combination of some k_1 atoms from the atomic set $\mathcal{A}$ such that all vectors have the same support and the ℓ_0 norm is less than s_1. Clearly, once $n > cs_1$, for some constant c, it will be not be possible to find a vector $v \in D$ such that $\omega = Xv$. Therefore $\inf\limits_{v \in D} \|\omega - Xv\|_2^2$ becomes a least squares problem and the infimum is achieved when ω is projected on the subspace spanned by the s_1 columns of X corresponding to the support of D. Now, let $\Pi_D(\cdot)$ denote the projection operator. Then,

$$\begin{aligned}\inf_{v \in D} \|\omega - Xv\|_2^2 &= \|\omega\|_2^2 - \Pi_D^2(\omega) \\ &= \|\omega\|_2^2 - \|\omega\|_2^2 \Pi_D^2(\omega/\|\omega\|_2) \\ &\leq \|\omega\|_2^2(1 - \Pi_D^2(\gamma))\ ,\end{aligned}$$

where $\gamma = \omega/\|\omega\|_2$ is a unit norm vector. Below we state a result from [6] to obtain an upper bound for $\Pi_D(\gamma)$,

Lemma 21 (Lemma 6.3, Amelunxen et al. 2014) *For each closed convex cone E in $\mathbb{R}^p$ with $w(E)$ denoting the Gaussian width of the cone,*

$$P(n\|\Pi_E(\gamma)\|^2 \geq w^2(E) + \delta) \leq \exp\left(\frac{-\delta^2/8}{w^2(E) + \delta}\right)\ . \tag{5.74}$$

This leads to the following corollary with subspace D as the cone E in the result above.

Corollary 25 *Consider a s_1 dimensional subspace and let $\Pi_D(\cdot)$ denote the projection operator on the subspace of any unit norm vector. Then the following holds,*

$$P(n\|\Pi_D(\gamma)\|^2 \geq s_1 + \delta) \leq \exp\left(\frac{-\delta^2/8}{s_1+\delta}\right) . \tag{5.75}$$

The result follows directly from Lemma 1 and the fact that the square of the Gaussian width of the subspace is equal to $O(s_1)$.

Now consider the set $B = \cup_{\binom{m}{k_1}} D$, that is, the union of all subsets of size k_1 of the atomic set $\mathcal{A}$. Below we establish a result for $\Pi_B(\cdot)$.

Lemma 22 *Consider the set $B = \cup_{\binom{m}{k_1}} C$. Then the following holds,*

$$P\{n\|\Pi_B(\gamma)\|_2^2 \geq s_1 + c_1 k_1 \log(em)\} \leq \exp(-\nu_1 k_1 \log(em)) . \tag{5.76}$$

The argument follows from a simple union bound,

$$\begin{aligned} P\{n\|\Pi_B(\gamma)\| \geq s_1 + \delta\} &\leq \binom{m}{k_1} \exp\left(\frac{-\delta^2/8}{s_1+\delta}\right) \\ &\leq \exp\left(\frac{-\delta^2/8}{s_1+\delta} + k_1\log(em)\right) . \end{aligned}$$

The result follows from choosing $\delta = c_1 k_1 \log(em)$ and assume $\nu_1 = \left(\frac{c_1^2/8k_1\log(em)}{s_1+c_1k_1\log(em)} - 1\right)$. We have as a result proved that once $n > c(s_1 + c_1 k_1 \log(em))$ choosing $\kappa_0 = (1/c) < 1$, with probability atleast $1 - \exp(-\nu_1 k_1 \log(em))$,

$$\inf_{v\in B} \|\omega - Xv\|_2^2 \geq \|\omega\|_2^2(1-\kappa_0) > 0 . \tag{5.77}$$

■

Next we move to analyzing conditions under which $\inf_{u\in E_c} \|\omega - Xu\|_2^2 \geq 0$, where E_c is the error set. For that we need the result from the following lemma which is based on Maurey's empirical approximation argument.

Lemma 23 *Let $X \in \mathbb{R}^{n\times p}$ be a random matrix with subGaussian isotropic rows. Let for any $v \in B$, where B is as defined above, $\inf_{v\in B} \|\omega - Xv\|_2^2 > 0$. Take any vector in the error set, $u \in E_c, u = \sum_{i=1}^m c_i a_i, c_i \geq 0, a_i \in \mathcal{A}$ is a nonzero vector and $\mu_j = c_j / \sum_{j=1}^m c_j = c_j/\|u\|_{\mathcal{A}}$, then, when $n > (s_1 + c\Psi^2(A)\log(em))$ with probability atleast $1 - \exp(-\nu_1 k_1 \log(em)) - 2\exp(-\nu_2 w^2(A))$,*

$$\inf_{u\in E_c} \|\omega - Xu\|_2^2 > 0 . \tag{5.78}$$

Proof: First consider any vector $w \in \mathbb{R}^p, w = \sum_{i=1}^m c_i a_i, c_i \geq 0, a_i \in \mathcal{A}$, let $W \in \mathbb{R}^p$ be a random vector defined by,

$$P(W = \|w\|_{\mathcal{A}} a_i) = \frac{c_i}{\|w\|_{\mathcal{A}}} \ . \tag{5.79}$$

Hence, $E[W] = w$. Let $W_1, W_2, \dots, W_{k_1}$ be independent copies of W and set $Z = \frac{1}{k_1}\sum_{i=1}^{k_1} W_i$. Therefore Z belongs to the set B. In the following all expectations are w.r.t Z,

$$\begin{aligned}
E\|\omega - XZ\|_2^2 &= \|\omega\|_2^2 + E\|XZ\|_2^2 - E[2\langle \omega, XZ\rangle] \\
&= \|\omega\|_2^2 + \frac{1}{k_1^2}\sum_{i,j\in[1,\dots,k_1], i\neq j} E\langle XW_i, XW_j\rangle + \frac{1}{k_1^2}\sum_{i\in[1,\dots,k_1]} E\langle XW_i, XW_i\rangle - 2\langle \omega, Xw\rangle \\
&= \|\omega\|_2^2 + \frac{k_1(k_1-1)}{k_1^2}\langle Xw, Xw\rangle + \frac{\|w\|_{\mathcal{A}}}{k_1}\sum_{i=1}^m c_i\langle Xa_i, Xa_i\rangle - 2\langle \omega, Xw\rangle \\
&= \|\omega\|_2^2 + \left(1 - \frac{1}{k_1}\right)\|Xu\|_2^2 + \frac{\|w\|_{\mathcal{A}}^2}{k_1}\sum_{i=1}^m \mu_i\langle Xa_i, Xa_i\rangle - 2\langle \omega, Xw\rangle \\
&\leq \|\omega - Xw\|_2^2 + \frac{\|w\|_{\mathcal{A}}^2}{k_1}\sum_{i=1}^m \mu_i\langle Xa_i, Xa_i\rangle \ .
\end{aligned}$$

Therefore we get,

$$\|\omega - Xw\|_2^2 \geq E\|\omega - XZ\|_2^2 - \frac{\|w\|_{\mathcal{A}}^2}{k_1}\sum_{i=1}^m \mu_i\langle Xa_i, Xa_i\rangle \ . \tag{5.80}$$

The above results holds for any general vector $w \in \mathbb{R}^p$. The following conclusion follows by considering the above result for all vectors in the error set $u \in E_c$,

$$\inf_{u\in E_c}\|\omega - Xu\|_2^2 \geq E\|\omega - XZ\|_2^2 - \sup_{u\in E_c}\frac{\|u\|_{\mathcal{A}}^2}{k_1}\sum_{i=1}^m \mu_i\langle Xa_i, Xa_i\rangle \ . \tag{5.81}$$

As established before each vector $Z \in \mathbb{R}^p$ belongs to the set B and we have already established that $\inf_{v\in B}\|\omega - Xv\|_2^2 \geq \|\omega\|_2^2(1-\kappa_0)$ with high probability when $n > (s_1 + c_1 k_1 \log(em))$. Hence it follows that $E\|\omega - XZ\|_2^2 \geq \|\omega\|_2^2(1-\kappa_0)$.

Now $\|u\|_{\mathcal{A}}^2 \leq \Psi^2(A)\|u\|_2^2$ where $\Psi(A)$ is the norm compatibility constant in the error set. Also $\mu_i < 0, \sum_{i=1}^m \mu_i = 1$. Therefore by standard arguments using Gaussian widths if $n > cw^2(\mathcal{A})$, that is, n is of the order of the square of the Gaussian width of the unit norm ball with probability atleast $1 - 2\exp(-\nu_2 w^2(\mathcal{A}))$ for some constant η [11] (see Theorem 5, Theorem 6 in [11]),

$$\sum_i \mu_i\langle Xa_i, Xa_i\rangle \leq \sup_i\langle Xa_i, Xa_i\rangle \leq n + c_1 w^2(\mathcal{A}) < c_2 n \ .$$

Therefore combining the previous results, we get with high probability when $n > (s_1 + c_1 k_1 \log(em))$,

$$\inf_{u \in E_c} \|\omega - Xu\|_2^2 \geq \|\omega\|_2^2 (1 - \kappa_0) - \frac{c_2 n \Psi^2(A)}{k_1} \, .$$

If we assume $\|\omega\|_2^2 > c_3 n$, which is not unreasonable since ω is the noise, and if $k_1 > \frac{c_2 \Psi^2(A)}{c_3(1-\kappa_0)}$ then $\|\omega - Xu\|_2^2 > 0$.

Hence it follows that when $n > (s_1 + c\Psi^2(A) \log(em))$ with probability atleast $1 - \exp(-\nu_1 k_1 \log(em)) - 2\exp(-\nu_2 w^2(E_c))$,

$$\inf_{u \in E_c} \|\omega - Xu\|_2^2 > 0 \, . \tag{5.82}$$

■

As stated in the beginning of the proof the above result leads to the conclusion that when $n > c(s_1 + c\Psi^2(A) \log(em)) = c_1 w^2(A)$, with high probability all samples cannot be interpolated.

We now turn our focus on obtaining high probability estimation error bounds. Suppose the empirical loss function is minimized at some $\beta\theta^* + \Delta, \ \Delta \in E_c$. Then the following inequality holds true,

$$\frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \langle x_i, \beta\theta^* + \Delta \rangle) - \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \langle x_i, \beta\theta^* \rangle) \leq 0 \, . \tag{5.83}$$

Now for any two scalars w, v, we have the following relationship from Equation (4.3) on Pg. 121 of [80],

$$\begin{aligned} \rho_\tau(w - v) - \rho_\tau(w) &= -v(\tau - \mathbb{I}(w \leq 0)) + \int_0^v (\mathbb{I}(w \leq z) - \mathbb{I}(w \leq 0)) dz \\ &= v\psi_\tau(w) + \int_0^v (\mathbb{I}(w \leq z) - \mathbb{I}(w \leq 0)) dz \, . \end{aligned}$$

Substitute $w_i = y_i - \langle x_i, \beta\theta^* \rangle$ and $v_i = \langle x_i, \Delta \rangle$ for $1 \leq i \leq n$ in the above inequality, with a little manipulation we get the following inequality,

$$\frac{1}{n} \sum_{i=1}^{n} \int_0^{\langle x_i, \Delta \rangle} (\mathbb{I}(y_i - \langle x_i, \beta\theta^* \rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \beta\theta^* \rangle \leq 0)) dz \leq \frac{1}{n} \sum_{i=1}^{n} \langle x_i, \Delta \rangle (\tau - \mathbb{I}(y_i - \langle x_i, \beta\theta^* \rangle \leq 0)) \, . \tag{5.84}$$

We will provide upper bounds on the rhs and lower bounds on the lhs of the inequality in (5.84). For upper bounds for the rhs quantity we need to obtain bounds on $\sup\limits_{u \in E_c} \frac{1}{n} \sum_{i=1}^{n} \langle x_i, u \rangle (\tau - \mathbb{I}(y_i - \langle x_i, \beta\theta^* \rangle \leq 0))$. We make the following observations,

$$\begin{aligned} \sum_{i=1}^{n} \frac{1}{n} \langle x_i, u \rangle (\tau - \mathbb{I}(y_i - \langle x_i, \beta\theta^* \rangle \leq 0)) &= \frac{1}{n} \langle X^T \psi_\tau(y - X\theta^*), u \rangle \\ &= \frac{1}{n} \|u\|_{\mathcal{A}} \|\psi_\tau(y - X\theta^*)\|_2 \left\langle X^T \gamma, \frac{u}{\|u\|_{\mathcal{A}}} \right\rangle \\ &\leq \frac{1}{n} \Psi(A) \|u\|_2 \|\psi_\tau(y - X\theta^*)\|_2 R^*(X^T\gamma) \ , \quad (5.85) \end{aligned}$$

where in the first inequality $\psi_\tau(y - X\theta^*) \in \mathbb{R}^n$ is an n-dimensional vector whose components are $(\psi_\tau(y - X\theta^*))_i = \tau - \mathbb{I}(y_i - \langle x_i, \beta\theta^* \rangle < 0)$; in the second inequality $\gamma = \psi_\tau(y - X\theta^*)/\|\psi_\tau(y - X\theta^*)\|_2$ is a unit norm vector; the last inequality follows from the definition of dual norm where $R^*(\cdot)$ is the dual norm corresponding to the atomic norm $R(\cdot)$ and also that $\Psi(A) = \sup\limits_{u \in A} \frac{\|u\|_{\mathcal{A}}}{\|u\|_2}$ is the norm compatibility constant.

We will now focus on the random quantity $R^*(X^T\omega)$. By arguments proposed in [11, 119][For example see proof of Theorem 3 in [11]], $E_X[R^*(X^T\gamma)] = c \cdot w(\Omega_R)$ where $w(\Omega_R)$ is the Gaussian width of the unit norm ball for subGaussian design matrices and c depends on the sub-Gaussian norm of X. For large deviation bounds around $E_X[R^*(X^T\gamma)]$ we refer to the proof from Theorem 4 in [11] and note that with probability atleast $1 - \nu_2 \exp\left(-\left(\frac{\tau_2}{c_2\phi\kappa}\right)^2\right)$

$$R^*(X^T\gamma) \leq c \cdot \sqrt{\Lambda_{\max}(\Sigma)} w(\Omega_R) + \tau \ . \quad (5.86)$$

Also the following is true,

$$\frac{1}{\sqrt{n}} \|\psi_\tau(y - X\theta^*)\|_2 \leq \max\{\tau, 1 - \tau\} \ . \quad (5.87)$$

Therefore, combining the results in (5.85), (5.86) and (5.87) we get that with probability atleast $1 - \nu_2 \exp\left(-\left(\frac{\tau_2}{c_2\phi k}\right)^2\right)$,

$$\sup_{u \in E_c} \frac{1}{n} \sum_{i=1}^{n} \langle x_i, u \rangle (\tau - \mathbb{I}(y_i - \langle x_i, \beta\theta^* \rangle \leq 0)) \leq \frac{c\sqrt{\Lambda_{\max}(\Sigma)} \max\{\tau, 1 - \tau\} \Psi(A) w(\Omega_R) \|u\|_2 + \tau}{\sqrt{n}} \ . \quad (5.88)$$

Next we provide a lower bound on the lhs quantity in (5.84).

Consider the quantity $\inf_{u \in E_c} \frac{1}{n}\sum_{i=1}^{n} \int_0^{\langle x_i,u\rangle} (\mathbb{I}(y_i - \langle x_i, \beta\theta^*\rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \beta\theta^*\rangle \leq 0))dz$. First note that if all samples are interpolated the quantity evaluates to 0.

We will instead lower bound the quantity on the rhs of the inequality below with $\omega_i = y_i - \langle x_i, \beta\theta^*\rangle$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \int_0^{\langle x_i,u\rangle} (\mathbb{I}(y_i - \langle x_i, \beta\theta^*\rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \beta\theta^*\rangle \leq 0))dz \\ &\geq \frac{1}{n}\sum_{i=1}^{n} |\langle x_i, u\rangle| \, \mathbb{I}\left[|\omega_i| \leq |\langle x_i, u\rangle/2|\right] \, \mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right] \ .\end{aligned}$$

We will consider samples such that $|\langle x_i, u\rangle| \geq \xi \|u\|_2$ for some constant ξ we will define later. Therefore, we will lower bound the following quantity,

$$\frac{1}{n}\sum_{i=1}^{n} \xi\|u\|_2 \, \mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2] \, \mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2] \mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right] \ . \quad (5.89)$$

We first focus on lower bounds for the quantity $\frac{1}{n}\sum_{i=1}^{n} \xi\|u\|_2 \, \mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2]$

The following result is derived from Proposition 5.1 in [124].

Theorem 21 *Consider the set* $A = cone(E_c) \cap S^{p-1}$*, where* E_c *is the error set. Define the marginal tail function as follows for some positive constants* β, ξ*,*

$$\beta = \inf_{u \in A} P(|\langle x_i, u\rangle| \geq \xi) \ . \quad (5.90)$$

Then the following is true with probability atleast $1 - \exp(-\tau_1^2/2)$*,*

$$\frac{1}{n}\sum_{i=1}^{n} \xi \, \mathbb{I}[|\langle x_i, u\rangle| \geq \xi] \geq \xi\beta - c_2\frac{w(A)}{\sqrt{n}} - \frac{\tau\xi}{\sqrt{n}} \ . \quad (5.91)$$

Moreover, fix the two-norm of the error vector u *as* $\|u\|_2$*. Then the following is true with probability atleast* $1 - \exp(-\tau_1^2/2)$*,*

$$\frac{1}{n}\sum_{i=1}^{n} \xi\|u\|_2 \, \mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2] \geq \left(\xi\beta - c_2\frac{w(A)}{\sqrt{n}} - \frac{\tau\xi}{\sqrt{n}}\right)\|u\|_2 \ . \quad (5.92)$$

Now due to the conditions $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ and $\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]$ there will only be $m < n$ samples that will satisfy all the conditions. We will derive high probability bounds on the number of samples satisfying conditions $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ and $\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]$.

First consider the following event,

$$\sum_{i=1}^{n} \mathbb{I}[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)] \,. \tag{5.93}$$

Since x_i's are symmetric isotropic $\text{sign}(\langle x_i, u\rangle)$ is positive with probability $1/2$ and negative with probability $1/2$. Hence, we can reasonably assume that $\sum_{i=1}^{n} \mathbb{I}[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)] \sim \frac{n}{2}$ with high probability.

Now, consider the quantity $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$. Note that $\omega_i = y_i - \langle x_i, \beta\theta^*\rangle$. Let $\underline{f}$ be a uniform lower bound on the conditional density around the probability distribution $y_i|x_i,\ \forall x_i$ around the noise term $y_i - \langle x_i, \beta\theta^*\rangle$, such that $P(|\omega_i| \leq \xi\|u\|_2/2) \geq \underline{f}\xi\|u\|_2/2$. Clearly $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ is a Bernoulli random variable with success rate $\underline{f}\xi\|u\|_2/2$. Therefore by a simple application of Chernoff's bound for some constant $0 \leq \phi_2 \leq 1$, we get the following,

$$P\left(\sum_{i=1}^{n} \mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2] \leq (1-\phi_2)n\underline{f}\xi\|u\|_2/2]\right) \leq \exp\left(-\frac{\phi_2^2 n\underline{f}\xi\|u\|_2}{4}\right) \,. \tag{5.94}$$

Note that assuming that $\|u\|_2 \geq c\frac{w(A)}{\underline{f}\sqrt{n}}$ where the set $A = \text{cone}(E_c) \cap S^{p-1}$ for some big enough positive constant c, the above equation evaluates to the following with some positive constant c_1,

$$P\left(\sum_{i=1}^{n} \mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2] \leq (1-\phi_2)n\underline{f}\xi\|u\|_2/2]\right) \leq \exp\left(-\nu_1\phi_2^2\sqrt{n}w(A)\right) \,. \tag{5.95}$$

Therefore, we have determined that with high probability $m = \frac{(1-\phi_2)n\underline{f}\xi\|u\|_2}{4}$ samples satisfy the conditions $\mathbb{I}[|\omega_i| \leq \xi\|u\|_2/2]$ and $\mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right]$. Now applying Theorem 21 for the m samples, we get the following with probability atleast $1 - \exp(-\tau_1^2/2)$,

$$\sum_{i=1}^{m} \xi\|u\|_2\, \mathbb{I}[|\langle x_i, u\rangle| \geq \xi\|u\|_2] \geq \left(\xi\beta - c_2\frac{w(A)}{\sqrt{m}} - \frac{\tau\xi}{\sqrt{m}}\right) m\|u\|_2 \,. \tag{5.96}$$

Substituting $m = \frac{(1-\phi_2)n\underline{f}\xi\|u\|_2}{4}$ and assuming $\|u\|_2 = c_3\frac{w(A)}{\underline{f}\sqrt{n}}$ for some big enough positive constant c_3, we get,

$$\sum_{i=1}^{m} \xi \|u\|_2 \, \mathbb{I}[|\langle x_i, u\rangle| \geq \xi \|u\|_2] \geq \left(\xi\beta - c_2 \frac{w(A)^{1/2}}{n^{1/4}} - \frac{\tau\xi}{n^{1/4} w(A)^{1/2}} \right) \frac{(1-\phi_2) n \underline{f} \xi}{4} \|u\|_2^2 \tag{5.97}$$

$$\geq \kappa n \underline{f} \|u\|_2^2 \,, \tag{5.98}$$

where we use $\kappa = \left(\xi\beta - c_2 \frac{w(A)^{1/2}}{n^{1/4}} - \frac{\tau\xi}{n^{1/4} w(A)^{1/4}} \right) \frac{(1-\phi_2)\xi}{4} > 0$ when $n = \Theta(w^2(A))$. We therefore conclude that with probability atleast $1 - \exp(-\tau_1^2/2) - \exp(-\nu_1 \phi_2^2 \sqrt{n} w(A))$ when the number of samples $n \geq \Theta(w^2(A))$,

$$\begin{aligned} & \inf_{u \in E_c} \frac{1}{n} \sum_{i=1}^{n} \int_0^{\langle x_i, u\rangle} (\mathbb{I}(y_i - \langle x_i, \beta\theta^*\rangle \leq z) - \mathbb{I}(y_i - \langle x_i, \beta\theta^*\rangle \leq 0)) dz \\ & \geq \inf_{u \in E_r} \frac{1}{n} \sum_{i=1}^{n} |\langle x_i, u\rangle| \, \mathbb{I}\left[|\omega_i| \leq |\langle x_i, u\rangle/2|\right] \, \mathbb{I}\left[\text{sign}(\omega_i) = \text{sign}(\langle x_i, u\rangle)\right] \\ & \geq \kappa n \underline{f} \|u\|_2^2 \,. \end{aligned} \tag{5.99}$$

Now combining the results in (5.84), (5.88) and (5.99) we get the following result with probability atleast $1 - \exp(-\tau_1^2/2) - \exp(-\nu_1 \phi_2^2 \sqrt{n} w(A)) - \nu_2 \exp\left(- \left(\frac{\tau_2}{c_2 \phi \kappa} \right)^2 \right)$ when $n > \Theta(w^2(A))$,

$$\|\Delta\|_2 \leq c \frac{\sqrt{\Lambda_{\max}(\Sigma)} \max\{\max\{\tau, 1-\tau\} \Psi(A) w(\Omega_R), w(A)\}}{\underline{f} \kappa \sqrt{n}} \,. \tag{5.100}$$

■

Corollary 26 *Consider the mean quantile index model with sub-Gaussian elliptically symmetric design matrix.*

1. *Let $R(\cdot)$ be the ℓ_1 norm and θ^* be s-sparse. Then with probability atleast $1 - \exp(-\tau_1^2/2) - \exp(-\nu_1 \phi_2^2 \sqrt{n} \sqrt{s \log p}) - \nu_2 \exp\left(- \left(\frac{\tau_2}{c_2 \phi \kappa} \right)^2 \right)$ the estimation error for the constrained quantile estimator satisfies,*

$$\|\Delta\|_2 \leq O\left(\frac{\sqrt{\Lambda_{\max}(\Sigma)} \max\{\tau, 1-\tau\} \sqrt{s \log p}}{\underline{f} \sqrt{n}} \right) \,. \tag{5.101}$$

2. *Let* $R(\cdot)$ *be the* ℓ_1 *norm and* θ^* *be s-sparse. Then with probability atleast* $1 - \exp(-\tau_1^2/2) - \exp(-\nu_1 \phi_2^2 \sqrt{n}\sqrt{l s_{\mathcal{G}} + s_{\mathcal{G}} \log N_{\mathcal{G}}}) - \nu_2 \exp\left(-\left(\frac{\tau_2}{c_2 \phi \kappa}\right)^2\right)$ *the estimation error for the constrained quantile estimator satisfies,*

$$\|\Delta\|_2 \leq O\left(\frac{\sqrt{\Lambda_{\max}(\Sigma)}\max\{\tau, 1-\tau\}\sqrt{l s_{\mathcal{G}} + s_{\mathcal{G}} \log N_{\mathcal{G}}}}{\underline{f}\sqrt{n}}\right) . \tag{5.102}$$

3. *Let* $R(\cdot)$ *be the* ℓ_1 *norm and* θ^* *be s-sparse. Then with probability atleast* $1 - \exp(-\tau_1^2/2) - \exp(-\nu_1 \phi_2^2 \sqrt{n}\sqrt{s \log p}) - \nu_2 \exp\left(-\left(\frac{\tau_2}{c_2 \phi \kappa}\right)^2\right)$ *the estimation error for the constrained quantile estimator satisfies,*

$$\|\Delta\|_2 \leq O\left(\frac{\sqrt{\Lambda_{\max}(\Sigma)}\max\{\tau, 1-\tau\}\sqrt{s \log p}}{\underline{f}\sqrt{n}}\right) . \tag{5.103}$$

Proof: The claims follow from the result of Theorem 4 and values of various quantities for the respective norms from Table 2.2 ■

Chapter 6

Structured Linear Contextual Bandits: A Sharp and Geometric Smoothed Analysis

6.1 Introduction

Contextual bandits [83] is a powerful framework for sequential decision-making, with many applications to clinical trials, web search and content optimization. In a typical scenario, users arrive over time, and the algorithm chooses among various content (e.g., news articles) to present to each user and observes the outcome (e.g. clicks). A popular parametric formulation for this problem is the linear contextual bandit setting [43, 89]: in rounds $t = 1, \dots, T$, the algorithm selects a context $x^t_{i^t}$ from k available contexts $x^t_1, \dots, x^t_k$ and receives a noisy reward $r^t(x^t_{i^t}) = \langle x^t_{i^t}, \theta^* \rangle + \omega^t$ where θ^*, ω^t are the unknown parameter and noise respectively. The goal of the algorithm is to select arms to maximize rewards over time observing only the contexts and the reward associated with the selected context in each round. Such algorithms typically need to balance *exploration*, making potentially sub-optimal decisions for the sake of information acquisition, and *exploitation*, selecting decisions that are optimal based on the estimate. Particularly, purely exploitative algorithms like greedy which myopically select contexts maximizing rewards based on the current parameter estimate $\hat{\theta}$, i.e., choosing $x^t_{i^t} = \underset{x^t_i : 1 \leq i \leq k}{\operatorname{argmax}} \langle x^t_i, \hat{\theta} \rangle$ are known to be sub-optimal in the worst case (see [97] for an example). At the same time, the greedy algorithm offers several appealing features, including its simplicity in computation and its best-effort treatments to every user.

Given the advantages of the greedy algorithm, there has been recent work that investigates when the greedy algorithms perform well. On the practical side, Bietti et al. [18] shows that there is strong empirical evidence that exploration free algorithms perform well on real data sets. On the theoretical side, a line of work [14, 77, 112] analyzed conditions under which inherent diversity in the data makes explicit exploration unnecessary. In particular, the work of [77, 112] provide a *smoothed analysis* on the greedy algorithm under the following setting: in each round the contexts $x_i^t, 1 \leq i \leq k$ are of the form $\mu_i^t + g_i^t, 1 \leq i \leq k$, where the $\mu_i^t \in \mathbb{R}^p$'s are possibly selected adverserially with the constraint $\|\mu_i^t\|_2 \leq 1$ and $g_i^t \sim N(0, \sigma^2 \mathbb{I}_{p \times p})$ are random Gaussian perturbations independent of the μ_i^t's. The algorithm in each round selects a context $x_{i^t}^t$ and receives noisy reward $r^t = \langle x_{i^t}^t, \theta_{i^t}^* \rangle + \omega^t$ where the parameter $\theta_{i^t}^*$ is unknown and there can be a different parameter corresponding to each context.

Our work further expands on the the smoothed analysis framework for linear contextual bandits considered in [77, 112]. In contrast to [77, 112], we assume the unknown parameters have structure characterized by low values according to some atomic norm $R(\cdot)$ like the ℓ_1 norm, group-sparse norms, nuclear norms, k-support norm [71, 7, 143, 123, 28, 114] among many others. We consider two variants of the problem: the multi parameter setting when there is a separate parameter corresponding to each context, i.e., $\theta_1^*, \ldots, \theta_k^*$ and the single parameter setting when there is a single unknown parameter, i.e., $\theta^* = \theta_1^* = \theta_2^* = \ldots = \theta_k^*$. In any round t the greedy algorithm maintains estimates of the true parameters $\hat{\theta}_1^t, \ldots, \hat{\theta}_k^t$ using the constrained least squares estimator:

$$\hat{\theta}_i^t = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \quad \mathcal{L}(\theta; Z_i^t, y_i^t) \quad \text{s.t.} \quad R(\theta) \leq R(\theta_i^*) \ , \tag{6.1}$$

where $\mathcal{L}(\theta; Z_i^t, y_i^t)$ is the least squares loss, Z_i^t is the design matrix in round t whose rows are contexts chosen in the rounds prior to t and y_i^t is a vector with the corresponding rewards for context i. The greedy algorithm then selects the arm corresponding to the highest reward w.r.t. to the current parameter estimate, i.e., $x_{i^t}^t = \underset{x_i^t : 1 \leq i \leq k}{\operatorname{argmax}} \langle x_i^t, \hat{\theta}_i^t \rangle$. We analyze the performance of the greedy algorithm w.r.t. the regret which compares the performance with a clairvoyant learner having knowledge of the optimal parameter θ_i^*,

$$\text{Reg}(T) = \sum_{t=1}^{T} \left(\max_i \langle x_i^t, \theta_i^* \rangle - \langle x_{i^t}^t, \theta_{i^t}^* \rangle \right) \ , \tag{6.2}$$

where in the single parameter setting $\theta_i^* = \theta_{i^t}^* = \theta^*$.

In our main results we derive worst case regret bounds for the single and multi parameter settings. Consider first the single parameter problem setting. In any round t, denote the error vector $\Delta^t = \hat{\theta}^t - \theta^*$. It is evident from equation (6.1) that the error vector lies in the error set:

$$E_c = \{\Delta \mid R(\theta^* + \Delta) \leq R(\theta^*)\}$$

. Now consider the set $A = \text{cone}(E_c) \cap S^{p-1}$ [17, 105] and define by $w(A)$ the Gaussian width of set A [121, 122, 60]. The Gaussian width is a metric for the complexity/size of a set [121, 122, 60] widely used in literature on high-dimensional statistics [11, 38, 33, 119]. For example, Gaussian width of the error set for $R(\cdot) = \|\cdot\|_1$ and s-sparse θ^* is $\Theta(s \log p)$. Note that the notations $y = \Theta(x)$ (respectively $y = O(x)$, $y = \Omega(x)$) implies there exists absolute constants c_1, c_2, c_3, c_4 such that $c_1 \cdot x \leq y \leq c_2 \cdot x$ (respectively $y \leq c_3 \cdot x$, $y \geq c_4 \cdot x$) and $\tilde{\Theta}(\cdot)$, $\tilde{\Omega}(\cdot)$ and $\tilde{O}(\cdot)$ notations hide the dependence on logarithm terms and noise variance. We show that the single parameter setting requires a warm start phase of $t_{\min} = \tilde{\Theta}(w^2(A))$ rounds when the contexts are chosen randomly or in round robin fashion.After the first $t_{\min}$ rounds when the algorithm accrues linear regret, we obtain worst case regret bounds of the form:

$$\text{Reg}(T) = \tilde{O}\left(\frac{w(A)\sqrt{T}}{\sigma}\right) , \tag{6.3}$$

where σ^2 is the variance of the Gaussian perturbations on the contexts. We make the following observations comparing our results to prior work.

1. For the unconstrained problem $w(A) = \Theta(\sqrt{p})$ and $\text{Reg}(T) = \tilde{O}\left(\frac{\sqrt{pT}}{\sigma}\right)$. When $\sigma^2 = \frac{1}{p}$ as considered in [77], ignoring logarithmic factors, the regret bounds are sharper compared to the results in [77] by a factor $\sqrt{p}$. Moreover when $\sigma^2 = \frac{1}{p}$, the regret upper bound is of the same order as the regret upper bounds obtained for UCB [44, 1] and better than the regret upper bounds for Thompson sampling [4]. With more smoothing when $\sigma^2 > \frac{1}{p}$ the greedy algorithm performs better than the UCB algorithm whereas less smoothing has the reverse effect.

2. For $R(\cdot) = \|\cdot\|_1$ and s-sparse θ^*, $w(A) = \Theta(\sqrt{s \log p})$ leading to the regret bounds, $\text{Reg}(T) = \tilde{O}\left(\frac{\sqrt{s \log p \cdot T}}{\sigma}\right)$. Again when $\sigma^2 = \frac{1}{p}$, the regret upper bounds are of the same order as [2] where a UCB algorithm was proposed for the ℓ_1 regularized problem. Note that the algorithm proposed in [2] is computationally involved and difficult to optimize.

The multi parameter setting requires a warm start phase of $\tilde{\Theta}\left(\frac{kw^2(A)}{\sigma^4}\right)$, where k is the number of contexts. When $\sigma^2 = \frac{1}{p}$, in the worst case, we require the length of the initial warm start phase to be $\tilde{\Theta}(k \cdot p^2 \cdot w^2(A))$ In the unregularized setting, $w^2(A) = p$ which translates to $\tilde{\Theta}(kp^3)$ rounds in the warm start phase which improves over the $\tilde{\Theta}(kp^6)$ rounds in [77] (see Theorem 4.2). The algorithm achieves $\tilde{O}\left(\frac{w(A)\sqrt{Tk}}{\sigma}\right)$ regret after the first initial warm start rounds which is $\sqrt{k}$ times worse compared to the single parameter setting.

We briefly summarize the organization and notations used throughout the chapter. We concisely present the main ideas and technical results in the chapter in Section 6.2. Section 6.2 also highlights new tools we use for analysis of high-dimensional estimation problems which maybe of independent interest to practitioners in the field. Results for the single parameter and multi parameter settings are presented in Section 6.3 and 6.4 respectively. All proofs are pushed to the supplementary section.

Notations: Throughout the chapter we use constants like $c, c_1, c_2, \ldots$ whose definition may change from one line to the next. In certain places we use the terms contexts and arms interchangeably.

6.2 Overview of Main Results

We briefly summarize the major ideas and results in this chapter.

Episodic algorithm: The bandit algorithm we propose has an episodic theme [73] with the episode lengths increasing geometrically with time. Let T denote the total number of rounds. In the single parameter setting, denote the episode number by e and let T_e denote the total number of rounds in episode e. The number of rounds in each episode increases geometrically with time, i.e., $T_1 = 2T_0$, $T_2 = 2T_1$ and so on. The total number of rounds $T = \sum_e T_e$. It follows that the number of episodes scales as $\log T$. The regression parameter is estimated at the beginning of each episode using only the contexts and rewards observed in the T_{e-1} rounds in the immediately preceding episode. In any round belonging to a particular episode, the contexts are chosen greedily using the parameter estimated at the beginning of the episode. The episodic algorithm is primarily used due to its computational simplicity and efficiency [73]. As we will show later, an episodic algorithm only contributes a multiplicative $\log T$ term to the regret bound compared to an algorithm where the parameters are estimated in every round. For the multi parameter setting, we maintain separate estimates for the parameters corresponding to each context. Similar to the single parameter setting the algorithm has multiple episodes, but there are separate episodes for the different arms. Denote episode numbers for each arm i as e_i and let the maximum number of episodes for arm i is denoted by $e_{i,\max}$. Let out of T total rounds, arm i is played in T_i rounds so that $T = \sum_{i=1}^{k} T_i$. The T_i rounds for each arm are divided into multiple episodes with the total number of rounds in episode e_i denoted as T_{i,e_i} so that $T_i = \sum_{e_i=0}^{e_{i,\max}} T_{i,e_i}$. The number of rounds in any episode is twice the number of rounds in the previous episode, i.e., $T_{i,1} = 2T_{i,0}$, $T_{i,2} = 2T_{i,1}$ and so on. The parameters for each arm are again estimated only at the beginning of each episode.

Estimation error: The regret depends on the estimation error for the parameter estimated using the constrained least squares estimator at the beginning of each episode. In the single parameter setting, let $Z^{(e)} \in \mathbb{R}^{T_e \times p}$ denote the design matrix with the observed contexts in episode e as rows and $y^{(e)} \in \mathbb{R}^{T_e}$ be the corresponding observed rewards. We precondition the data before parameter estimation using the Puffer transformation [75]. The Puffer transformation computes the SVD of the design matrix as $\frac{1}{\sqrt{T_e}} Z^{(e)} = U^{(e)} D^{(e)} (V^{(e)})^\intercal$ followed by transforming the data as $\tilde{Z}^{(e)} = F^{(e)} Z^{(e)}$, $\tilde{y}^{(e)} = F^{(e)} y^{(e)}$ where $F^{(e)} = U^{(e)} (D^{(e)})^{-1} (U^{(e)})^\intercal$. The parameter at the beginning of episode $e+1$ is then estimated using the following constrained estimator,

$$\hat{\theta}^{(e+1)} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{T_e} \|\tilde{y}^{(e)} - \tilde{Z}^{(e)}\theta\|_2^2 \quad s.t. \quad R(\theta) \leq R(\theta^*) \ . \tag{6.4}$$

One of the technical results in this work is the upper bounds on the parameter estimation error using the Puffer transformed data. In the worst case Puffer transformed data gives better estimation bounds compared to the bounds obtained using raw data in previous literature [33, 105, 11]. Our analysis borrows tools and techniques from the existing vast literature on high-dimensional estimation [130, 128]. Specifically, following the analysis framework established in [11], we need three main results. First, note that to satisfy the constraint in (6.4) the error vector Δ with $\hat{\theta} = \theta^* + \Delta$ lies in the following set,

$$E_r = \{\Delta \mid R(\theta^* + \Delta) \leq R(\theta^*)\} \ . \tag{6.5}$$

Second, for consistent estimation the design matrix should satisfy the following restricted eigenvalue (RE) condition on the error set $A = \text{cone}(E_r) \cap S^{p-1}$ [17, 105],

$$\inf_{u \in A} \frac{1}{T_e} \|\tilde{Z}^{(e)} u\|_2^2 \geq \kappa \ , \tag{6.6}$$

where κ is a positive constant. In another main technical result, we show that with high probability, in the single parameter setting the RE condition is satisfied once $T > t_{\min} = \Omega\left((w(A) + \sqrt{\log\log T})^2 \log^2 k\right)$ across all episodes. Here $w(A)$ denotes the Gaussian width of set A [121, 122, 60]. Third, assuming $T > t_{\min}$ we show that estimation with the transformed data leads to the following upper bound on the estimation error across all episodes with high probability,

$$\max_e \|\hat{\theta}^{(e)} - \theta^*\|_2 \leq O\left(\frac{\kappa_\omega(w(A) + \sqrt{\log\log T})\sqrt{\log k}}{\sigma\sqrt{T_{e-1}}}\right) , \tag{6.7}$$

where κ_ω denotes the sub-Gaussian norm of the noise and $\sigma = \frac{c}{\sqrt{p}}$ is the variance of the Gaussian perturbations present in the smoothed contexts. The results on parameter estimation errors also holds in the multi parameter setting except we maintain separate parameter estimates for each context.

Regret: We show that the regret formulation (6.2), in the single parameter setting, can be upper bounded as follows:

$$\text{Reg}(T) \leq 4\beta t_{\min} + \sum_e \sum_1^{T_e} 2\beta\|\hat{\theta}^{(e)} - \theta^*\|_2 , \tag{6.8}$$

where $\theta^{(e)}$ is the parameter estimated at the beginning of episode e and $\beta = \max_{\substack{1\leq i\leq k, 1\leq t\leq T \\ a\in R^p:\|a\|=1}} \langle x_i^t, a\rangle$. We show with high probability that $\beta \leq O(1 + \sigma\sqrt{\log(Tk)})$. The first $t_{\min}$ rounds are random exploration rounds required to satisfy the RE condition for parameter estimation across all episodes. Once $T > t_{\min}$, the regret depends on the ℓ_2 norm of the estimation error. Together with result (6.7), noting that by algorithm design $T_e = 2T_{e-1}$, $e \leq \log T$ and simple algebraic manipulations lead to the following sublinear regret bounds when $T >> t_{\min}$,

$$\text{Reg}(T) \leq O\left(\frac{\beta\kappa_\omega(w(A) + \sqrt{\log\log T})\sqrt{\log k}\log(T)\sqrt{T}}{\sigma}\right) . \tag{6.9}$$

With the choice $\sigma = \frac{c}{\sqrt{p}}$ and $\kappa_\omega = 1$, ignoring logarithmic terms, we get $\text{Reg}(T) \leq O\left(\sqrt{p\cdot w^2(A)\cdot T}\right)$.

The upper bound for the regret in the multi parameter setting evaluates to the following:

$$\text{Reg}(T) \leq 2\beta t_{\min} + \sum_{i=1}^{k}\sum_{e_i=1}^{e_{i,\max}}\left(\sum_1^{T_{i,e_i}} \beta\|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2 + \sum_1^{T^*_{i,e_i}} \beta\|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2\right) , \tag{6.10}$$

where, as described above, e_i indexes episode numbers for context i, $e_{i,\max}$ is the maximum number of episodes for context i, T_{i,e_i} denotes the total rounds in episode e_i and T^*_{i,e_i} denotes total rounds when context i was optimal but not chosen by the greedy algorithm, i.e., before arm i was chosen in T_{i,e_i} rounds after the end of episode $e_i - 1$, the greedy algorithm did not choose context i in theT^*_{i,e_i} rounds when $x_i^t = \underset{x_j^t : 1 \leq j \leq k}{\text{argmax}} \langle x_j^t, \theta_j^* \rangle$. It is evident that the regret will suffer if T^*_{i,e_i} is large compared to T_{i,e_i}. Therefore our present analysis requires that the random exploration rounds $t_{\min} = \Omega\left(\frac{kw^2(A)\beta^2}{\sigma^4}\right)$, ignoring logarithmic terms, so that there is a non-zero fixed probability that the greedy algorithm chooses context i when it is optimal. When $t_{\min} = \Omega\left(\frac{kw^2(A)\beta^2}{\sigma^4}\right)$, we show that $T^*_{i,e_i} \leq 40 T_{i,e_i}$ with high probability leading to the following high probability regret bounds,

$$\text{Reg}(T) \leq 2\beta t_{\min} + \sum_{i=1}^{k} \sum_{e_i=1}^{e_{i,\max}} \sum_{1}^{T_{i,e_i}} 41\beta \|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2 \tag{6.11}$$

Note that with $\sigma = \frac{c}{\sqrt{p}}$ the condition on $t_{\min}$ implies each context should be chosen randomly in more than p^2 rounds before the greedy algorithm can be used. This maybe prohibitive for many practical high-dimensional applications where p can be very large. Since $t_{\min}$ is greater than $\Omega((w(A) + \sqrt{\log\log T})^2 \log^2 k)$ the parameter estimation error rates are same as the single parameter setting. Noting that $e_{i,\max} \leq \log(T)$ and with simple algebraic manipulations we get get the following upper bounds on the regret when $T >> t_{\min}$,

$$\text{Reg}(T) \leq O\left(\frac{\kappa_\omega \beta (w(A) + \sqrt{\log\log T})\sqrt{\log k}\log(T)\sqrt{kT}}{\sigma}\right) . \tag{6.12}$$

The regret is $\sqrt{k}$ worse compared to the single parameter setting.

6.3 Single Parameter Regret Analysis

We present results for the single parameter setting in this section. We first present the algorithm and derive regret bounds in terms of the parameter estimation error rates. We then present analysis for the case without the adversary, i.e., in each round we observe k contexts which are sampled as $g_i^t \sim N(0, \sigma^2 \mathbb{I}_{p\times p}),\ 1 \leq i \leq k$ before deriving bounds for the adverserial setting.

The greedy algorithm proceeds in multiple episodes with the length of each episode increasing geometrically with time [73]. We index episode numbers by e, time steps by t and arms by i. We denote by T the total number of rounds and by T_e the number of rounds played in episode e. In each round, the algorithm observes contexts $x_i^t, 1 \leq i \leq k$ and greedily selects the optimal arm based on the current parameter estimate, i.e., $z^t = \underset{x_i^t : 1 \leq i \leq k}{\text{argmax}} \langle x_i^t, \hat{\theta}^{(e)} \rangle$ and receives noisy

Algorithm 2 Structured Greedy (single parameter)

```
Initialize empty design matrix and reward vector $Z^{(0)} = [], y^{(0)} = []$
for $e = 1, 2, 3, \dots, \lfloor \log_2 T \rfloor$ do
    Compute SVD as $\frac{1}{\sqrt{T_{e-1}}} Z^{(e-1)} = U^{(e-1)} D^{(e-1)} (V^{(e-1)})^\intercal$
    Compute the Puffer transformation $F^{(e-1)} = U^{(e-1)} (D^{(e-1)})^{-1} (U^{(e-1)})^\intercal$ and define $\tilde{Z}^{(e-1)} = F^{(e-1)} Z^{(e-1)}$ and $\tilde{y}^{(e-1)} = F^{(e-1)} y^{(e-1)}$
    Estimate parameter using constrained least squares estimator breaking ties arbitrarily when necessary
        $\hat{\theta}^{(e)} = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2T_{e-1}} \|\tilde{y}^{(e-1)} - \tilde{Z}^{(e-1)} \theta\|_2^2 \quad \text{s.t.} \quad R(\theta) \leq R(\theta^*) \, , \qquad (6.13)$
    where $T_{e-1}$ is the number of observations in the previous episode.
    Initialize empty design matrix and reward vector $Z^{(e)} = [], y^{(e)} = []$. Set $T_e = 2^{e-1}$
    for $t = 2^{(e-1)} + 1$ to $2^e$ do
        Observe contexts $x_1^t, \dots, x_k^t \in \mathbb{R}^p$
        Choose arm $z^t = \underset{x_i^t : 1 \leq i \leq k}{\text{argmax}} \langle x_i^t, \hat{\theta}^{(e)} \rangle$ and observe reward $y^t = \langle z^t, \theta^* \rangle + \omega^t$ where $\omega^t$ is zero mean $\kappa_\omega$-sub-Gaussian noise
        Append observations $(z^t, y^t)$ to $(Z^{(e)}, y^{(e)})$
    end for
end for
```

reward $\langle z^t, \theta^* \rangle + \omega^t$ with ω^t denoting the noise at time t. The parameter is estimated at the beginning of each episode using the arms and rewards observed in the previous episode. The parameter is estimated using the constrained least squares estimator on the Puffer transformed design matrix and response [75].

Lemma 24 provides an upper bound for the regret for Algorithm 2. The greedy algorithm accrues linear regret in the first $t_{\min}$ rounds when the design matrix is rank deficient for parameter estimation, i.e., it does not satisfy the restricted eigenvalue condition. Subsequent rounds are played in an episodic fashion with the regret in any round depending on the accuracy of parameter estimation at the beginning of the episode.

Lemma 24 (Single Parameter Regret Bounds) *Denote by* $\beta = \max_{\substack{1 \leq i \leq k, 1 \leq t \leq T \\ a \in R^p : \|a\| = 1}} \langle x_i^t, a \rangle$. *Assume*

$T > t_{\min}$, where $t_{\min}$ depends on properties of the true parameter θ^* and the regularizer $R(\cdot)$. Then,

$$Reg(T) \leq 4\beta t_{\min} + \sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{1}^{T_e} 2\beta \|\hat{\theta}^{(e)} - \theta^*\|_2 \ . \tag{6.14}$$

6.3.1 Gaussian Arms

We start with an analysis of the performance of the Greedy algorithm when the contexts are completely stochastic, i.e., we derive regret bounds when the contexts are sampled independently from a Gaussian distribution, , $x_i^t \sim N(0, \sigma^2 \mathbb{I}_{p\times p}), 1 \leq i \leq k, t \leq T$ in step 9 of Algorithm 2. Our primary goal in this section is to establish the analysis framework for the smoothed adversary setting. As we will subsequently see, with a few additional arguments we obtain the same parameter estimation error bounds in the smoothed adversary setting as the completely stochastic setting.

Since the regret bound is influenced by the efficiency of parameter estimation, the focus will be on characterizing the non-asymptotic estimation error bounds of the form $\|\hat{\theta}^{(e)} - \theta^*\|_2$. The major difference from traditional literature on high-dimensional structured estimation [31, 105, 17, 11, 33] is the rows of the design matrix. While most high-dimensional estimation literature assumes the rows to be independent (sub)-Gaussian, here we have independent Gaussian arms chosen as the maximum of the dot product with the estimated parameter vector, i.e., $z^t = \underset{x_i^t : 1 \leq i \leq k}{\operatorname{argmax}} \langle x_i^T, \hat{\theta}^{(e)} \rangle$. The Lemma below describes relevant properties of the design matrix useful to bound the estimation error. In particular, we observe that the minimum eigenvalue condition is satisfied in expectation and has an inverse relationship with the logarithm of the number of arms.

Lemma 25 (Single Parameter Gaussian Arms Design Matrix Properties) *The rows of the design matrix $Z^{(e)} \in \mathbb{R}^{T_e \times p}$ in any episode e satisfy $\kappa_z = \|z^t\|_{\psi_2} \leq c_2 \sigma \sqrt{\log k}$ for c_2 some positive constant. Moreover the minimum eigenvalue of the matrix $E_{z^t}[z^t (z^T)^T]$ satisfies,*

$$\lambda_{\min}(E_{z^t}[z^t (z^t)^\intercal]) \geq c_1 \frac{\sigma^2}{\log k} \ , \tag{6.15}$$

where c_1 is some positive constant and the expectation is over the random draws of contexts.

The estimation error bounds can now be obtained using the results of Lemma 25 using standard analysis tools established in prior literature in high-dimensional statistics [105, 11, 33]. A slight departure from more traditional estimation is the use of the Puffer transformation on the design matrix and response before applying the constrained least squares estimator [75]. The Puffer transformation is a preconditioning technique analyzed in [75] and was practically found have better performance in estimating the sparsity pattern with the Lasso estimator when the design matrix had heavily correlated rows. Our analysis also leads us to conclude that preconditioning with the Puffer transformation before estimation gives better worst case estimation error bounds.

Theorem 22 (Single Parameter Gaussian Arms Estimation Error Bounds) *Let* $T_e \geq c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$*. Then with probability atleast* $1 - 2\delta$ *assuming ,*

$$\|\hat{\theta}^{(e+1)} - \theta^*\|_2 \leq O\left(\frac{\gamma}{\sigma\sqrt{T_e}}\right) \ , \tag{6.16}$$

where $\gamma = c\kappa_\omega\sqrt{\log k}\left(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)}\right)$*,* $E_c = \{\Delta \mid R(\theta^* + \Delta) \leq R(\theta^*)\}$*,* $A = cone(E_c) \cap S^{p-1}$ *is the error set and* $w(\cdot)$ *denotes the Gaussian width of a set.*

Now high probability regret bounds can be obtained combining the results of Lemma 24 together with the high probability estimation error bounds of Theorem 22.

Theorem 23 (Gaussian Arms Regret Bounds) *Consider Gaussian arms. Then with probability atleast* $1 - 2\delta$

$$\beta = \max_{\substack{1 \leq i \leq k, 1 \leq t \leq T \\ a \in R^p : \|a\| = 1}} \langle x_i^t, a\rangle \leq c_1\sigma(\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)}) \ . \tag{6.17}$$

Also with $T >> t_{\min} \geq c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ *with probability atleast* $1 - 4\delta$ *the following is an upper bound on the regret for the Greedy algorithm,*

$$\textit{Reg}(T) \leq O\left(\frac{\gamma \cdot \beta \cdot \log(T) \cdot \sqrt{T}}{\sigma}\right) \tag{6.18}$$

where $\gamma = c\kappa_\omega\sqrt{\log k}(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$

6.3.2 Smoothed Perturbed Adversary

We now derive regret bounds when the contexts are $x_i^t = \mu_i^t + g_i^t, 1 \leq i \leq k, \forall 1 \leq t \leq T$. Remember that an adversary can choose $\mu_i^t, \|\mu_i^t\|_2 = 1, \forall 1 \leq i \leq k$ based on the observed contexts and rewards in the previous rounds. The adversary can potentially influence the properties of the design matrix (primarily messing up the minimum eigenvalue) and the subsequent non-asymptotic estimation error bounds. Our first main result is Lemma 26 where we show that even in the adverserial setting in the smoothed analysis framework the minimum eigenvalue of the design matrix in expectation is no worse than the Gaussian setting. In particular, adding small random perturbations to adverserially selected contexts leads to implicit exploration where the greedy algorithm works well. Another departure from the Gaussian setting is that the contexts are no longer independent since an adversary having access to all data before a particular round chooses the μ_i^t's. Technically the non-asymptotic analysis for the minimum eigenvalue of the design matrix requires application of new results analyzing quadratic random variables with dependence from [12].

Lemma 26 (Design matrix properties for smoothed adversary) *The rows of the design matrix $Z^{(e)} \in \mathbb{R}^{T_e \times p}$ in any episode e are $z^t = \mu^t + g^t$ where $\mu^t, g^t = \operatorname*{argmax}_{\mu_i^t, g_i^t : 1 \le i \le k} \langle \mu_i^t + g_i^t, \hat{\theta}^{(e-1)} \rangle$, $g_i^t \sim N(0, \sigma^2 \mathbb{I}_{p \times p})$ with the sub-Gaussian norm of g^t satisfying $\|g^t\|_{\psi_2} \le c_2 \sigma \sqrt{\log k}$ for some constant c_2. Moreover we have the following lower bound on the expected minimum eigenvalue for any μ_i^t's:*

$$\lambda_{\min}(E_{z^t}[z^t (z^t)^\intercal]) \ge c_1 \frac{\sigma^2}{\log k} , \tag{6.19}$$

where c_1 is some constant.

Note that the minimum eigenvalue does not get worse compared to the purely stochastic case. Also note that we obtain sharper lower bounds on the minimum eigenvalue than the $c_1 \frac{\sigma^2}{\log T}$ obtained in [77] (see Lemma 3.7, Lemma 3.6). The estimation error bounds can now be obtained similar to the purely stochastic setting.

Theorem 24 (Estimation Error Bounds) *The design matrix $Z^{(e)} \in \mathbb{R}^{T_e \times p}$ in all episode where $T_e \ge c_1 (w(A) + \sqrt{\log \log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ satisfies the following minimum eigenvalue condition with probability atleast $1 - 3\delta$,*

$$\inf_{u \in A} \frac{1}{T_e} \|Z^{(e)} u\|_2^2 \ge c \frac{\sigma^2}{\log k} . \tag{6.20}$$

Moreover, for all episodes when $T_e \ge c_1 (w(A) + \sqrt{\log \log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ with probability atleast $1 - 4\delta$,

$$\|\hat{\theta}^{(e+1)} - \theta^*\|_2 \le O\left(\frac{\gamma}{\sigma \sqrt{T_e}} \right) , \tag{6.21}$$

where $\gamma = c\kappa_\omega \sqrt{\log k}(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$.

High probability regret bounds can now be obtained combining results from Lemma 24 and Theorem 24.

Theorem 25 (Smoothed Perturbed Adversary Regret Bounds) *In the smoothed adversary setting with probability atleast* $1 - 2\delta$

$$\beta = \max_{\substack{1 \le i \le k, 1 \le t \le T \\ a \in R^p : \|a\| = 1}} \langle x_i^t, a \rangle \le (1 + c_1 \sigma(\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)}))\ . \tag{6.22}$$

Also with $T >> t_{\min} \ge c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ *with probability atleast* $1 - 6\delta$ *the following is an upper bound on the regret,*

$$\textit{Reg(T)} \le O\left(\frac{\gamma \cdot \beta \cdot \log(T) \cdot \sqrt{T}}{\sigma}\right)\ , \tag{6.23}$$

where $\gamma = c\kappa_\omega \sqrt{\log k}(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$.

6.3.3 Examples

We instantiate the regret bounds for a few norms assuming very mild conditions on σ. Note that for ℓ_2^2 regularization the setting is similar to [77]. If θ^* is sparse exploiting structure, e.g. using the ℓ_1 norm, the regret bounds depend on $\sqrt{s \log p}$ instead of $\sqrt{p}$.

Corollary 27 (Regret Bounds for ℓ_1 norm) *Consider the smoothed adversary setting. Let* θ^* *be an s-sparse vector,* $R(\cdot)$ *be the* ℓ_1 *norm and* $\sigma = \frac{1}{\sqrt{p}}$*. Then with* $T >> c_1(\sqrt{s \log p} + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ *with probability atleast* $1 - 6\delta$*,*

$$\textit{Reg}(T) \le O\left(\sqrt{p} \cdot \gamma \cdot \beta \cdot \log(T) \cdot \sqrt{T}\right)\ , \tag{6.24}$$

where $\gamma = c\kappa_\omega \sqrt{\log k}(\sqrt{s \log p} + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$

Proof: For the ℓ_1 norm and s-sparse θ^*, $w(A) \le c_2\sqrt{s \log p}$ [33, 11, 38]. Hence the result follows from the result in Theorem 25. ■

Corollary 28 (Regret Bounds for low rank matrix recovery) *Consider the smoothed adversary setting. Let* $\theta^* \in \mathbb{R}^{m \times p}$ *be a matrix with rank* $r < \min\{m, p\}$*,* $R(\cdot)$ *be the nuclear* *norm and* $\sigma = \frac{1}{p}$*. Then with* $T >> c_7(\sqrt{r(m+p)} + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ *with probability atleast* $1 - 6\delta$

$$Reg(T) \leq O\left(p \cdot \gamma \cdot \beta \cdot \log(T) \cdot \sqrt{T}\right) \ , \tag{6.25}$$

where $\gamma = c\kappa_\omega \sqrt{\log k}(\sqrt{r(m+p)} + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$

Proof: In the low rank matrix considered $w(A) \leq c_2\sqrt{r(m+p)}$ [33, 39]. Hence the result follows from the result in Theorem 25. ■

Corollary 29 (Regret Bounds for ridge regularization) *Consider the smoothed adversary setting with ridge regularization. Then with* $T >> c_1(\sqrt{p} + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ *with probability atleast* $1 - 6\delta$

$$Reg(T) \leq O\left(\sqrt{p} \cdot \gamma \cdot \beta \cdot \log(T) \cdot \sqrt{T}\right) \ , \tag{6.26}$$

where $\gamma = c\kappa_\omega \sqrt{\log k}(\sqrt{p} + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$

Proof: In the ridge regularized setting $w(A) = \sqrt{p}$. Hence the result follows from the result in Theorem 25. ■

6.4 Multi Parameter Regret Analysis

We present results for the multi parameter setting in this section. The multi parameter setting has a separate parameter corresponding to each context. The multi parameter setting requires a warm start phase of T_0 rounds where the contexts are chosen in a round robin fashion before employing the greedy algorithm. As we show later, the warm start phase has dependence on the variance of the perturbations and is required to obtain sublinear regret. Similar to the single parameter setting, after the warm start phase the greedy algorithm proceeds in an episodic fashion, except that there are separate episodes for each context. We denote the episode numbers for context i by e_i and the maximum number of episodes for context i as $e_{i,\max}$. For each context, the parameter estimates are updated at the beginning of an episode. In episode e_i, context i is chosen by the greedy algorithm in T_{i,e_i} rounds. During episode e_i, before context i is chosen in T_{i,e_i} rounds by the greedy algorithm, there can also be rounds when context i was optimal but was not chosen by the algorithm, i.e., $x_i^t = \underset{x_j^t: 1\leq j \leq k}{\text{argmax}} \langle x_j^t, \theta_j^* \rangle$ but $x_i^t \neq \underset{x_j^t: 1\leq j \leq k}{\text{argmax}} \langle x_j^t, \hat{\theta}_j^{(e_j)} \rangle$. We denote the number of rounds this happens in episode e_i by T^*_{i,e_i}.

Lemma 27 below gives an upper bound on the regret for Algorithm 3.

Lemma 27 (Multi Parameter Regret Bounds) *The greedy algorithm plays the contexts in an episodic fashion with the maximum episode number for each context* $e_i \leq e_{i,\max} \leq \lfloor \log T \rfloor$*. Denote by* $\beta = \max_{\substack{1\leq i\leq k, 1\leq t\leq T \\ a\in R^p : \|a\|=1}} \langle x_i^t, a\rangle$*. Let* $t_{\min} < T$*, where* $t_{\min}$ *depends on properties of the true parameters* θ_i^**, the regularizer* $R(\cdot)$*, the noise properties, the number of contexts* k *and the quantity* β*. Then,*

$$\begin{aligned} Reg(T) &\leq 2\beta t_{\min} + \\ &+ \beta \sum_{i=1}^{k} \sum_{e_i=1}^{e_{i,\max}} \left(T_{i,e_i} \|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2 + T_{i,e_i}^* \|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2 \right) \end{aligned} \tag{6.27}$$

The regret thus depends on the following: a) the accuracy of estimating θ_i^* in each episode for all contexts; b) the number of rounds when context i is optimal but not chosen,i.e., the quantities T_{i,e_i}^*, and c) the number of episodes each context is chosen, i.e., the quantities $e_{i,\max}$. Note that the estimate of any context parameter improves with the number of times the particular context is chosen. The quantities T_{i,e_i}^*, while contributing to the regret, represent rounds when the context is not chosen and hence do not contribute to improvement of the parameter estimate. We need the warm start to ensure the greedy algorithm chooses contexts with constant probability when they are optimal, i.e., to limit the quantities T_{i,e_i}^*.

We focus on regret bounds when the contexts are $x_i^t = \mu_i^t + g_i^t, 1 \leq i \leq k, 1 \leq t \leq T$, where μ_i^t's are adverserially chosen and g_i^t's are the Gaussian perturbations. We start with a characterization of the number of rounds required in the warm start phase. The warm start phase is required to ensure that conditioned on a context being optimal in a given round, there is a constant non-zero probability the greedy algorithm also perceives it to be optimal which in turn limits the quantities T_{i,e_i}^*. Intuitively, this is the essence of Lemma 28. We build towards the result Lemma 28 with a couple of propositions. Proposition 1 is a straightforward observation on the relationship between the first and second optimal contexts where we introduce the quantity r.

Proposition 1 *Consider any round* t *when the episode numbers of the* k *contexts are* $e_1, \ldots, e_k$*. Let* i^* *denote the context with the maximum reward, i.e.,* $i^* = \operatorname*{argmax}_{1\leq l\leq k} \langle \mu_l^t + g_l^t, \theta_l^* \rangle$*. Let* j *denote the context having the second largest reward, i.e.,* $j = \operatorname*{argmax}_{1\leq l\leq k; l\neq i^*} \langle \mu_l^t + g_l^t, \theta_l^* \rangle$*. Define*

$r = \langle \mu_j^t + g_j^t, \theta_j^* \rangle - \langle \mu_{i^*}^t, \theta_{i^*}^* \rangle$. *Then the following condition is satisfied,*

$$\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r \ . \tag{6.28}$$

Proposition 2 states conditions when there is a definite match between the actual optimal context and the context perceived to be optimal by the greedy algorithm.

Proposition 2 *Assume context j' such that $j' = \underset{1 \leq l \leq k, l \neq i^*}{\operatorname{argmax}} \langle \mu_l^t + g_l^t, \hat{\theta}_l^{(e_l)} \rangle$, i.e., the context other than i^* which has the highest estimated reward. Also assume the parameter estimate for context i^* to be $\hat{\theta}_{i^*}^{(e_{i^*})} = \theta_{i^*}^* + \Delta_{i^*}^{(e_{i^*})}$. Then the greedy algorithm selects context i^* if the following condition is satisfied,*

$$\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \ . \tag{6.31}$$

The greedy algorithm always picks the optimal context if the condition in equation (6.31) is satisfied. Intuitively, to limit T_{i,e_i}^* condition (6.31) should be satisfied with some constant probability. Lemma 28 formalizes this notion.

Lemma 28 **(Margin Condition)** *Consider good events as when $r \leq c_3 \sigma \sqrt{\log(Tk)}$ and consider errors $\Delta_{i^*}^{(e_{i^*})}$ and $\Delta_{j'}^{(e_{j'})}$ to be small enough such that $\langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \leq \frac{\sigma^2}{r}$. Then the following holds,*

$$P\left(\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \frac{\sigma^2}{r} \;\middle|\; \langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r \right) \geq \frac{1}{20} \ , \tag{6.32}$$

for all $r \leq c_3 \sigma \sqrt{\log(Tk)}$.

The length of the warm start phase is influenced by the condition that $\|\Delta_{j'}^{(e_{j'})}\|_2$ and $\|\Delta_{i^*}^{(e_{i^*})}\|_2$ are small enough so that $\langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \leq \frac{\sigma^2}{r}$ in Lemma 28. Let $\beta = \max_{\substack{1 \leq i \leq k, 1 \leq t \leq T \\ a \in R^p : \|a\| = 1}} \langle \mu_i^t + g_i^t, a \rangle$. We show that $\beta = \tilde{O}(1)$ with high probability in Theorem 25. Then with some simple algebraic manipulations we show in the worst case that $\langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \leq \frac{\sigma^2}{r}$ is satisfied when:

$$\|\Delta_i^{(e_i)}\|_2 = \|\hat{\theta}_i^{(e_i)} - \theta_i^*\|_2 \leq \tilde{O}\left(\sigma\right) \ . \tag{6.33}$$

Algorithm 3 High-dimensional Greedy (multi parameter)

```
Set $e_1 = \ldots = e_k = 0$. Initialize empty design matrices and rewards $Z_1^{(0)}, \ldots, Z_k^{(0)} = []$,$y_1^{(0)}, \ldots, y_k^{(0)} = []$
for $t = 1$ to $T_0$ do
    Observe contexts $x_1^t, \ldots, x_k^t \in \mathbb{R}^p$
    Pick context $i^t$ from $\{1, \ldots, k\}$ in round robin fashion and observe reward $r_{i^t}^t = \langle x_{i^t}^t, \theta_{i^t}^* \rangle + \omega^t$ where $\omega^t$ is zero mean $\kappa_\omega$-sub-Gaussian noise
    Append observations $(x_{i^t}^t, r_{i^t}^t)$ to $(Z_{i^t}^{(0)}, y_{i^t}^{(0)})$
end for
Compute SVD of $\frac{1}{\sqrt{T_{i,0}}} Z_i^{(0)} = U_i^{(0)} D_i^{(0)} (V_i^{(0)})^\intercal$
Define the Puffer transformation $F_i^{(0)} = U_i^{(0)} (D_i^{(0)})^{-1} (U_i^{(0)})^\intercal$ and compute $\tilde{y}_i^{(0)} = F_i^{(0)} y_i^{(0)}$ and $\tilde{Z}_i^{(0)} = F_i^{(0)} Z_i^{(0)}$
Estimate parameters using constrained least squares estimator for each context with $T_{1,0} = \ldots = T_{i,0} = \ldots = T_{k,0} = T_0/k$
    $$\hat{\theta}_i^{(1)} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2T_{i,0}} \|\tilde{y}_i^{(0)} - \tilde{Z}_i^{(0)} \theta\|_2^2 \quad \text{s.t.} \quad R(\theta) \leq R(\theta_i^*) \,, \qquad (6.29)$$
Increment all $e_i = e_i + 1, 1 \leq i \leq k$. Initialize empty design matrices and rewards $Z_1^{(e_1)}, \ldots, Z_k^{(e_k)} = []$,$y_1^{(e_1)}, \ldots, y_k^{(e_2)} = []$. Also initialize $t_1 = \ldots = t_k = 0$.
for $t = T_0$ to $T$ do
    Observe contexts $x_1^t, \ldots, x_k^t \in \mathbb{R}^p$
    Pick context $i^t$ such that $i^t = \underset{1 \leq i \leq k}{\operatorname{argmax}} \langle x_i^t, \hat{\theta}_i^{(e_i)} \rangle$, receive reward $r_{i^t}^t = \langle x_{i^t}^t, \theta_{i^t}^* \rangle + \omega^t$ and increment $t_{i^t} = t_{i^t} + 1$
    Append observations $(x_{i^t}^t, r_{i^t}^t)$ to $(Z_{i^t}^{(e_{i^t})}, y_{i^t}^{(e_{i^t})})$
    if $t_{i^t} = 2T_{i^t, e_{i^t}-1} = T_{i^t, e_{i^t}}$ then
        Compute SVD of $\frac{1}{\sqrt{T_{i,e_{i^t}}}} Z_{i^t}^{(e_{i^t})} = U_{i^t}^{(e_{i^t})} D_{i^t}^{(e_{i^t})} (V_{i^t}^{(e_{i^t})})^\intercal$
        Compute the Puffer transformation $F_{i^t}^{(e_{i^t})} = U_{i^t}^{(e_{i^t})} (D_{i^t}^{(e_{i^t})})^{-1} (U_{i^t}^{(e_{i^t})})^\intercal$ and compute $\tilde{Z}_{i^t}^{(e_{i^t})} = F_{i^t}^{(e_{i^t})} Z_{i^t}^{(e_{i^t})}$ and $\tilde{y}_{i^t}^{(e_{i^t})} = F_{i^t}^{(e_{i^t})} y_{i^t}^{(e_{i^t})}$
        Estimate parameter using constrained least squares estimator
            $$\hat{\theta}_{i^t}^{(e_{i^t}+1)} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2T_{i^t, e_{i^t}}} \|\tilde{y}_{i^t}^{(e_{i^t})} - \tilde{Z}_{i^t}^{(e_{i^t})} \theta\|_2^2 \quad \text{s.t.} \quad R(\theta) \leq R(\theta_{i^t}^*) \,, \qquad (6.30)$$
        where $T_{i^t, e_{i^t}} = 2T_{i^t, e_{i^t}-1}$.
        Increment $e_{i^t} = e_{i^t} + 1$. Initialize empty design matrix $Z_{i^t}^{(e_{i^t})} = []$ and reward $y_{i^t}^{(e_{i^t})} = []$. Initialize $t_{i^t} = 0$.
    end if
end for
```

The estimation error bounds are influenced by the properties of the design matrices after the warm start phase.

Lemma 29 (Multi parameter Design Matrix Properties) *Consider any context i and a particular episode e_i. The rows of the design matrix $Z_i^{(e_i)} \in \mathbb{R}^{T_{i,e_i} \times p}$ are $z_i^t = \mu_i^t + g_i^t$ where in round t context i is chosen by the Greedy algorithm, i.e., $i = \underset{1 \leq l \leq k}{\operatorname{argmax}} \langle x_l^t, \hat{\theta}_l^{(e_l)} \rangle$ where $x_l^t = \mu_l^t + g_l^t$, $g_l^t \sim N(0, \sigma^2 \mathbb{I}_{p \times p})$. Then under the condition $\langle g_i^t, \theta_i^* \rangle \geq r$ for some $r \leq c_3 \sigma \sqrt{\log(Tk)}$,*

$$\lambda_{\min}\left(E_{z^t}\left[z_i^t (z_i^t)^\intercal \mid z_i^t \textit{ satisfies } \zeta\right]\right) \geq c_2 \frac{\sigma^2}{\log(Tk)} ,$$

where ζ is the condition $z_i^t = \underset{g_l^t : 1 \leq l \leq k}{\operatorname{argmax}} \langle x_l^t, \hat{\theta}_l^{(e_l)} \rangle; \langle g_i^t, \theta_i^* \rangle \geq r; r \leq c_3 \sigma \sqrt{\log(Tk)}$.

The only difference in the properties of the design matrix compared to the single parameter setting are the sub-Gaussian norm and expected minimum eigenvalue of the covariance matrix. Therefore following similar proof arguments as in the single parameter setting we obtain the following upper bound on the maximum estimation error across all contexts and episodes:

$$\sup_{1 \leq i \leq k} \sup_{e_i \leq e_{i,\max}} \|\hat{\theta}_i^{(e_i+1)} - \theta_i^*\|_2 \leq \tilde{O}\left(\frac{w(A)}{\sigma \sqrt{T_{i,e_i}}}\right) . \tag{6.34}$$

Comparing equations (6.33) and (6.34) it can be easily inferred that $T_{i,e_i} = \tilde{\Theta}\left(\frac{w^2(A)}{\sigma^4}\right)$ to satisfy the margin condition and hence the length of the warm start phase $T_0 = \tilde{\Theta}\left(\frac{kw^2(A)}{\sigma^4}\right)$. In the interest of space all proofs are provided in the supplement.

If the length of the warm start phase satisfies $T_0 = \tilde{\Theta}\left(\frac{kw^2(A)}{\sigma^4}\right)$, the margin condition of Lemma 28 holds which ensures that the greedy algorithm chooses any context i with probability greater than $1/20$ if context i is optimal in a given round. In other words in expectation $T_{i,e_i}^* \leq 20 T_{i,e_i}$, i.e., in any particular episode for any context the number of rounds when the context is optimal but not perceived to be optimal by the greedy algorithm is upper bounded by 20 times the length of the episode. With the result on T_{i,e_i}^*'s the upper bound for the regret in the multi parameter setting can be derived from the result of Lemma 27 to be $\tilde{O}\left(\frac{w(A)\sqrt{Tk}}{\sigma}\right)$ which is $\sqrt{k}$ times worse compared to the single parameter setting.

Theorem 26 (Multi parameter Smoothed Adversary Regret Bounds) *Consider computation of regret for the Greedy algorithm in the multi parameter setting following Lemma 4. Define the following quantities* $r \leq c_3\sigma\sqrt{\log(Tk)}$, $\gamma = \frac{c_{12}\kappa_\omega(w(A)+\sqrt{\log\log T}+\sqrt{\log k}+\sqrt{\log(1/\delta)})\sqrt{\log(Tk)}}{\sigma}$ *and* $\beta = \max\limits_{\substack{1\leq i\leq k, 1\leq t\leq T\\ a\in R^p:\|a\|=1}} \langle x_i^t, a\rangle$. *The margin condition in Lemma 5 is satisfied with probability atleast* $1-5\delta$ *when,*

$$t_{\min} \geq \frac{4k\gamma^2r^2\beta^2}{\sigma^4} + 1 + \sqrt{\frac{1}{2}\log(1/\delta)}\ . \tag{6.35}$$

Under the margin condition, the regret is maximized when in each round each context has equal probability to be selected by the Greedy algorithm. The equal probability implies that in expectation $T_1 = T_2 = \ldots = T_k = \frac{T}{k}$. *Also the regret is upper bounded as follows,*

$$\mathit{Reg}(T) \leq 2\beta t_{\min} + 82\beta\gamma\sqrt{Tk}\log(T)\ . \tag{6.36}$$

Moreover $\beta = \max\limits_{\substack{1\leq i\leq k, 1\leq t\leq T\\ a\in R^p:\|a\|=1}} \langle x_i^t, a\rangle \leq (1 + c_1\sigma(\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)}))$ *with probability atleast* $1-2\delta$. *Therefore with probability atleast* $1-8\delta$ *when* $T >> t_{\min}$

$$\mathit{Reg}(T) \leq O\left(\gamma\cdot\beta\cdot\log(T)\cdot\sqrt{Tk}\right) \tag{6.37}$$

6.4.1 Examples

As examples we instantiate the regret bounds for the ℓ_1 norm, nuclear norm and ridge regularization.

Corollary 30 (Regret Bounds for ℓ_1-norm) *Let* θ^* *be an s-sparse vector,* $R(\cdot)$ *be the* ℓ_1 *norm and* $\sigma = \frac{1}{\sqrt{p}}$. *Then with* $T >> c_4kp^2\gamma^2r^2\beta^2$ *with probability atleast* $1-8\delta$,

$$\mathit{Reg}(T) \leq O\left(\sqrt{p}\cdot\gamma\cdot\beta\cdot\log(T)\cdot\sqrt{Tk}\right)\ , \tag{6.38}$$

where $\gamma = c_2\kappa_\omega\sqrt{\log(Tk)}(\sqrt{s\log p} + \sqrt{\log\log T} + \sqrt{\log k} + \sqrt{\log(1/\delta)})$.

Proof: In the setting considered $w(A) \leq c_5\sqrt{s\log p}$. The result follows from the result in Theorem 26. ■

Corollary 31 **(Regret Bounds for nuclear norm)** *Let $\theta^* \in \mathbb{R}^{m\times p}$ be a rank r matrix, $R(\cdot)$ be the nuclear norm and $\sigma = \frac{1}{p}$. Then with $T >> c_4 k n^2 \gamma^2 r^2 \beta^2$ with probability atleast $1 - 8\delta$,*

$$\textit{Reg}(T) \leq O\left(p \cdot \gamma \cdot \beta \cdot \log(T) \cdot \sqrt{Tk}\right) \ , \tag{6.39}$$

where $\gamma = c_2 \kappa_\omega \sqrt{\log(Tk)}(\sqrt{r(m+p)} + \sqrt{\log\log T} + \sqrt{\log k} + \sqrt{\log(1/\delta)})$.

Proof: In the setting considered $w(A) \leq c_5\sqrt{r(m+p)}$. The result follows from the result in Theorem 26. ■

Corollary 32 **(Regret Bounds for ridge regularization)** *Let $R(\cdot)$ be the ridge regularizer and $\sigma = \frac{1}{\sqrt{p}}$. Then with $T >> c_4 k p^2 \gamma^2 r^2 \beta^2$ with probability atleast $1 - 8\delta$,*

$$\textit{Reg}(T) \leq O\left(p \cdot \gamma \cdot \beta \cdot \log(T) \cdot \sqrt{Tk}\right) \ , \tag{6.40}$$

where $\gamma = c_2 \kappa_\omega \sqrt{\log(Tk)}(\sqrt{p} + \sqrt{\log\log T} + \sqrt{\log k} + \sqrt{\log(1/\delta)})$.

Proof: In the setting considered $w(A) \leq c_5\sqrt{p}$. The result follows from the result in Theorem 26. ■

Appendix

6.A Background and Preliminaries

We present a variant of the Hoeffding bound where the coefficients can depend on the randomness of prior random variables.

Lemma 30 *Let $\{Z_t\}$ be a sub-Gaussian martingale difference sequence (MDS) and let $z_{1:t}$ denote a realization of $Z_{1:t}$. Let $\{a_t\}$ be a sequence of random variables such that $a_t = f_t(z_{1:(t-1)})$ for some sequence function f_t with $|a_t| \leq \alpha_t$ a.s. for suitable constants $\alpha_1, \ldots, \alpha_T$. Then, for any $\tau > 0$, we have*

$$P\left(\left|\sum_{t=1}^{T} a_t z_t\right| \geq \tau\right) \leq 2\exp\left\{-\frac{\tau^2}{4c\kappa^2 \sum_{t=1}^{T}\alpha_t^2}\right\} \ , \tag{6.41}$$

for absolute constants $c > 0$ and where κ is the ψ_2-norm of the conditional subGaussian random variables.

Proof: For any realization $z_{1:(t-1)}$ since $Z_t|z_{1:(t-1)}$ is a sub-Gaussian random variable with zero mean, then the conditional moment-generating function (MGF) satisfies: for all $s > 0$

$$E[\exp(sZ_t) \mid z_{1:(t-1)}] \leq \exp(cs^2\kappa^2) \ , \tag{6.42}$$

where κ is ψ_2-norm of Z_t conditioned on any realization $z_{1:(t-1)}$ and $c > 0$ is an absolute constant. Further, for $a_t = f(z_{1:(t-1)})$ with $|a_t| \leq \alpha_t$, we have

$$E[\exp(sa_tZ_t) \mid z_{1:(t-1)}] \leq \exp(ca_t^2s^2\kappa^2) \leq \exp(c\alpha_t^2s^2\kappa^2) \ , \tag{6.43}$$

where the last inequality holds for all realiztions $z_{1:(t-1)}$.

For any $s > 0$, note that

$$\begin{aligned} P\left(\sum_{t=1}^{T} a_tZ_t \geq \tau\right) &= P\left(\exp\left(s\sum_{t=1}^{T} a_tZ_t\right) \geq \exp(s\tau)\right) \\ &\leq \exp(-s\tau)E\left[\exp\left(s\sum_{t=1}^{T} a_tZ_t\right)\right] \ . \end{aligned} \tag{6.44}$$

Now, using (6.42), we have

$$\begin{aligned} E\left[\exp\left(s\sum_{t=1}^{T} a_tZ_t\right)\right] &= E_{(Z_1,\ldots,Z_T)}\left[\prod_{t=1}^{T}\exp(sa_tZ_t)\right] \\ &= E_{(Z_1,\ldots,Z_{T-1})}\left[E_{Z_T|Z_1,\ldots,Z_{T-1}}\left[\exp(sa_TZ_T)\right]\prod_{t=1}^{T-1}\exp(sa_tZ_t)\right] \\ &\leq \exp(cs^2\alpha_T^2\kappa^2)E_{(Z_1,\ldots,Z_{T-1})}\left[\prod_{t=1}^{T-1}\exp(sa_tZ_t)\right] \\ &\leq \exp(cs^2\alpha_T^2\kappa^2)\exp(cs^2\alpha_{T-1}^2\kappa^2)E_{(Z_1,\ldots,Z_{T-2})}\left[\prod_{t=1}^{T-2}\exp(sa_tZ_t)\right] \\ &\ldots \\ &\leq \exp\left(cs^2\kappa^2\sum_{t=1}^{T}\alpha_t^2\right) \ . \end{aligned}$$

Plugging this back to (6.44), we have

$$P\left(\sum_{t=1}^{T} a_tZ_t \geq \tau\right) \leq \exp\left(-s\tau + cs^2\kappa^2\sum_{t=1}^{T}\alpha_t^2\right) \ . \tag{6.45}$$

Choosing $s = \frac{\tau}{2c\kappa^2 \sum_{t=1}^{T} \alpha_t^2}$, we obtain

$$P\left(\sum_{t=1}^{T} a_t Z_t \geq \tau\right) \leq \exp\left\{-\frac{\tau^2}{4c\kappa^2 \sum_{t=1}^{T} \alpha_t^2}\right\} . \tag{6.46}$$

Repeating the same argument with $-Z_t$ instead of X_t, we obtain the same bound for $P(-\sum_t a_t Z_t \geq \tau)$. Combining the two results gives us (6.41). ■

6.B Results on Gaussian Random Variables

Lemma 31 *Consider k Gaussians $g_1, \ldots, g_k$ sampled from a $N(0, \sigma^2)$ distribution. Let $g_{(1)} = \max_{g_i : 1 \leq i \leq k} g_i$. Then for some constant c_2,*

$$P(g_{(1)} \leq \sqrt{2}\sigma(\sqrt{\log k} + \sqrt{\log(1/\delta)})) \geq 1 - 2\delta \tag{6.47}$$

Proof: We first obtain upper bounds on $g_{(1)}$. We make the following observations,

$$\begin{aligned} \exp(tE[g_{(1)}]) &\leq E[\exp(tg_{(1)})] \\ &\leq E[\max \exp(tg_i)] \\ &\leq \sum_{i=1}^{k} E[\exp(tg_i)] \\ &\leq n \exp(t^2\sigma^2/2) , \end{aligned} \tag{6.48}$$

where the first inequality is Jensen's inequality, the second is the union bound, and the final inequality follows from the definition of the moment generating function. Taking logarithm of both sides of the inequality, we get

$$E[g_{(1)}] \leq \frac{\log k}{t} + \frac{t\sigma^2}{2} . \tag{6.49}$$

This can be minimized by setting $t = \frac{\sqrt{2 \log k}}{\sigma}$ to give,

$$E[g_{(1)}] \leq \sigma\sqrt{2}\sqrt{\log k} . \tag{6.50}$$

We now use the following result from [121] to provide large deviation bounds around $E[g_{(1)}]$.

Lemma 32 (Lemma 2.1.3 in [121]) *Consider a Gaussian process $(Z_t)_{t \in U}$, where U is finite and a number σ such that $\sigma \geq \sup_{t \in U} (E[Z_t^2])^{1/2}$. Then for $u > 0$ we have,*

$$P\left(\left|\sup_{t \in U} Z_t - E \sup_{t \in U} Z_t\right| \geq u\right) \leq 2 \exp\left(-\frac{u^2}{2\sigma^2}\right) . \tag{6.51}$$

In the context of our setting we have $g_i = Z_i$, $g_{(1)} = \max_{t \in U} Z_t$ and $\max_{t \in U}(E[Z_t^2])^{1/2} = E[g_i^2]^{1/2} = \sigma$. Therefore using $u = \sigma\sqrt{2}\sqrt{\log(1/\delta)}$ we get,

$$P(g_{(1)} \geq E[g_{(1)}] + \sigma\sqrt{2}\sqrt{\log(1/\delta)}) \leq 2\delta , \tag{6.52}$$

Now the stated result can be derived from (6.50) and (6.52). ■

Proof: Now from the proof in Lemma 31 $E[g_{(1)}] \leq \sigma\sqrt{2}\sqrt{\log k}$. The result then follows using Lemma A.4 and using $c_2 = 1$.

Lemma 33 *Let $x_1, \ldots, x_k$ be k independent Gaussian random variables with variance σ^2 and let $z = \operatorname{argmax}_{1 \leq i \leq k} x_i$. Then,*

$$\mathit{Var}(z) \geq c_1 \frac{\sigma^2}{\log k} , \tag{6.53}$$

where c_1 is some positive constant.

Proof: For each i, let e_i denote the event of $i = \operatorname{argmax}_{i'} x_{i'}$. Then the variance of z can be written as

$$\begin{aligned} \mathrm{Var}(z) &\geq \sum_i P[e_i]\,\mathrm{Var}[z \mid e_i] && (6.54) \\ &= \frac{1}{k} \sum_i \mathrm{Var}[z \mid e_i] = \mathrm{Var}[z \mid e_1] && (6.55) \end{aligned}$$

where the last two steps follow from the fact that the distributions among arms are identical. Furthermore,

$$\begin{aligned} &\mathrm{Var}[z \mid e_1] \\ &= \mathrm{Var}\left[x_1 \mid e_1\right] \\ &> P\left[x_1 \geq \sqrt{\log(k)} \wedge \max_{i>1} x_i < \sqrt{\log(k)} \mid e_1\right] \mathrm{Var}\left[x_1 \mid x_1 \geq \sqrt{\log(k)} \wedge \max_{i>1} x_i < \sqrt{\log(k)}\right] \\ &> \frac{P\left[x_1 \geq \sqrt{\log(k)} \wedge \max_{i>1} x_i < \sqrt{\log(k)}\right]}{P[e_1]} \mathrm{Var}\left[x_1 \mid x_1 \geq \sqrt{\log(k)}\right] \\ &= k\,P\left[x_1 \geq \sqrt{\log(k)}\right] P\left[\max_{i>1} x_i < \sqrt{\log(k)}\right] \mathrm{Var}\left[x_1 \mid x_1 \geq \sqrt{\log(k)}\right] \end{aligned}$$

where the last step follows from $P[e_1] = 1/k$ and that x_1 is independent from all other draws.

We use the following known inequality of the Gaussian distribution [63].

$$P[x_1 \geq \sqrt{\log(k)}] > 1/k.$$

Furthermore,

$$\begin{aligned} P\left[\max_{i>1} x_i < \sqrt{\log(k)}\right] &= \prod_{i=2}^{k} P\left[x_i < \sqrt{\log(k)}\right] && (6.56) \\ &\geq (1 - C/k)^k \geq 1/e^C && (6.57) \end{aligned}$$

where C is an absolute constant. Finally, by [77], we have $\text{Var}[x_1 \mid x_1 \geq \sqrt{\log(k)}] \geq \Omega(1/\log(k))$

Putting everything together, we have

$$\text{Var}[z] \geq \text{Var}[z \mid e_1] \geq k(1/k)(1/e^C)\text{Var}[x_1 \mid x_1 \geq \sqrt{\log(k)}] \geq 1/\log(k)$$

■

Lemma 34 *Let an adversary pick $\mu_i \in \mathbb{R}, 1 \leq i \leq k$ and then consider k random draws from a Gaussian distribution $g_i \sim N(0, \sigma^2), 1 \leq i \leq k$. Then the following is true for any adversary,*

$$\textit{Var}[g \mid g = \operatorname*{argmax}_{g_i : 1 \leq i \leq k} g_i + \mu_i] \geq \textit{Var}[g \mid g = \operatorname*{argmax}_{g_i : 1 \leq i \leq k} g_i] \,. \quad (6.58)$$

Proof: Without loss of generality assume $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_k$. Also let $g_{(1)} \geq g_{(2)} \geq \ldots g_{(k)}$ denote the order statistics of the Gaussian variables. Now any μ_i can be mapped to any $g_{(j)}$ to give $k!$ possibilities. Lets divide the $k!$ events into k disjoint sets $A_1, \ldots, A_k$ in the following way. Consider one mapping $\{(\mu_{i_l}, g_{(1)}), \ldots, (\mu_i, g_{(j)}), \ldots, (\mu_{i_h}, g_{(k)})\}$ where there are indices $i, j = \operatorname*{argmax}_{1 \leq i,j \leq k} \mu_i + g_{(j)}$. If $j = k$ then we put the mapping in the bin A_k. Otherwise assuming $j < k$, let $\mu_{i_1}, \ldots, \mu_{i_h}$ be mapped to $g_{(j+1)}, \ldots, g_{(j+h)}$ such that $1 \leq j \leq j + h \leq k$. We then find an index i_m such that after swapping μ_i and μ_{i_m} such that the new mapping is $\{(\mu_{i_l}, g_{(1)}), \ldots, (\mu_{i_m}, g_{(j)}), \ldots, (\mu_i, g_{(j+m)}), \ldots (\mu_{i_h}, g_{(k)})\}$ we find that $i, j + m = \operatorname*{argmax}_{1 \leq i_m, j+m \leq k} \mu_i + g_{(j+m)}$ but when we swap μ_i with $\mu_{i_{m+1}}$ for the mapping $\{(\mu_{i_l}, g_{(1)}), \ldots, (\mu_{i_{m+1}}, g_{(j)}), \ldots, (\mu_i, g_{(j+m+1)}), \ldots (\mu_{i_h}, g_{(k)})\}$ then $\mu_i + g_{j+m+1}$ is no longer the maximum. We then put the mapping $\{(\mu_{i_l}, g_{(1)}), \ldots, (\mu_i, g_{(j)}), \ldots, (\mu_{i_h}, g_{(k)})\}$ in the bin A_{j+m}. Note that the bin A_{j+m} will also have the mappings $\{(\mu_{i_l}, g_{(1)}), \ldots, (\mu_{i_n}, g_{(j)}), \ldots, (\mu_i, g_{(j+n)}), \ldots (\mu_{i_h}, g_{(k)})\}$ for all $1 \leq n \leq m$ with μ_{i_n} swapped with μ_i where $\mu_i + g_{(j+n)}$ is the maximum as also the mappings $\{(\mu_{i_l}, g_{(1)}), \ldots, (\mu_{i_n}, g_{(j-n)}), \ldots, (\mu_{i_n}, g_{(j)}), \ldots (\mu_{i_h}, g_{(k)})\}, 1 \leq n \leq j-1$ where $\mu_i + g_{(j-n)}$ is the maximum. Since all these mappings are equally probable, bin A_{j+m} is a set of events out of $k!$ where any $g_{(o)}, 1 \leq o \leq j+m$ are equally probably such that $i, o = \underset{1 \leq o, i \leq k}{\operatorname{argmax}} \mu_i + g_{(o)}$ for some $1 \leq i \leq k$. Moreover from the construction we see that the sets $A_i \cap A_j = \phi, 1 \leq i, j \leq k$ are disjoint and $\cup_{1 \leq i \leq k} A_i$ contains all $k!$ events. Therefore with this construction we make the following observations,

$$\text{Var}[g \mid g = \underset{g_i : 1 \leq i \leq k}{\operatorname{argmax}} g_i + \mu_i] = \sum_{i=1}^{k} \text{Var}(g \mid g \sim \{g_{(1)}, \ldots, g_{(i)}\}) P(A_i) \ . \tag{6.59}$$

We note that the minimum variance is achieved when $P(A_1) = 1$ and Var$(g \mid g \sim g_{(1)}) =$ Var$[g \mid g = \underset{g_i : 1 \leq i \leq k}{\operatorname{argmax}} g_i]$ which is the desired result. ■

6.C Proof for Single Parameter Setting with Gaussian Contexts

We give the proof for Lemma 24.

Lemma 24 *Denote by* $\beta = \underset{\substack{1 \leq i \leq k, 1 \leq t \leq T \\ a \in R^p : \|a\| = 1}}{\max} \langle x_i^t, a \rangle$*. Assume* $T > t_{\min}$*, where* $t_{\min}$ *depends on properties of the true parameter* θ^* *and the regularizer* $R(\cdot)$*. Then,*

$$\textit{Reg}(T) \leq 4\beta t_{\min} + \sum_{e = \lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} 2 T_e \beta \|\hat{\theta}^{(e)} - \theta^*\|_2 \ . \tag{6.60}$$

Proof: Let the episodes be indexed by e, let T_e denote the number of rounds in episode e and let T denote the total number of rounds. Let S_e denote the rounds in episode e. If context i^t is selected in round t and i^* denotes the optimal context then the regret can be computed as follows,

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T} \langle \theta^*, x_{i^*}^t - x_{i^t}^t \rangle \\
&= \sum_{e=1}^{\lfloor \log T \rfloor} \sum_{t \in S_e} \langle \theta^*, x_{i^*}^t - x_{i^t}^t \rangle \\
&= \sum_{e=1}^{\lceil \log t_{\min} \rceil} \sum_{t \in S_e} \langle \theta^*, x_{i^*}^t - x_{i^t}^t \rangle + \sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{t \in S_e} \langle \theta^*, x_{i^*}^t - x_{i^t}^t \rangle \ . \qquad (6.61)
\end{aligned}$$

The first term on the r.h.s. of (6.61) can be upper bounded as follows,

$$\begin{aligned}
\sum_{e=1}^{\lceil \log t_{\min} \rceil} \sum_{t \in S_e} \langle \theta^*, x_{i^*}^t - x_{i^t}^t \rangle &\leq \sum_{e=1}^{\lceil \log t_{\min} \rceil} \sum_{t \in S_e} |\langle \theta^*, x_{i^*}^t \rangle| + |\langle \theta^*, x_{i^t}^t \rangle| \\
&\leq \sum_{e=1}^{\lceil \log t_{\min} \rceil} \sum_{t \in S_e} 2\beta \\
&\leq 4\beta t_{\min} \ , \qquad (6.62)
\end{aligned}$$

where in the third line we use from the algorithm $T_e = 2^{e-1}$ and hence $\sum_{e=1}^{\lceil \log t_{\min} \rceil} \sum_{t=1}^{T_e} 1 \leq 2t_{\min}$.

We make the following observations to bound the second term on the r.h.s. in equation (6.61),

$$\begin{aligned}
\sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{t \in S_e} \langle \theta^*, x_{i^*}^t - x_{i^t}^t \rangle &= \sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{t \in S_e} \langle \theta^* - \hat{\theta}^{(e)}, x_{i^*}^t \rangle - \langle \theta^* - \hat{\theta}^{(e)}, x_i^t \rangle + \langle \hat{\theta}^{(e)}, x_{i^*}^t - x_i^t \rangle \\
&\leq \sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{t \in S_e} |\langle \theta^* - \hat{\theta}^{(e)}, x_{i^*}^t \rangle| + |\langle \theta^* - \hat{\theta}^{(e)}, x_i^t \rangle| \\
&\leq \sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{t \in S_e} 2\beta \|\theta^* - \hat{\theta}^{(e)}\|_2 \ , \qquad (6.63)
\end{aligned}$$

where in the second line we use $\langle \hat{\theta}^e, x_{i^*}^t \rangle \leq \langle \hat{\theta}^e, x_i^t \rangle$ as $x_{i^t}^t$ was chosen ahead of $x_{i^*}^t$ in round t.

The stated result now follows from (6.61), (6.62) and (6.63). ■

We give the proof for Lemma 25.

Lemma 25 *The rows of the design matrix $Z^{(e)} \in \mathbb{R}^{T_e \times p}$ in any episode e satisfy $\kappa_z = \|z^t\|_{\psi_2} \leq c_2 \sigma \sqrt{\log k}$ for c_2 some positive constant. Moreover the minimum eigenvalue of the* *matrix $E_{z^t}[z^t(z^T)^T]$ satisfies,*

$$\lambda_{\min}(E_{z^t}[z^t(z^t)^\intercal]) \geq c_1 \frac{\sigma^2}{\log k} \ , \tag{6.64}$$

where c_1 is some positive constant and the expectation is over the random draws of contexts.

Proof: The rows of the design matrix satisfy,

$$z^t = \underset{x_i^t : 1 \leq i \leq k}{\operatorname{argmax}} \langle x_i^t, \hat{\theta}^{(e)} \rangle \ , \tag{6.65}$$

where $\hat{\theta^{(e)}}$ is the estimated parametrer in episode e. We first prove the result on the sub-Gaussian norm of z^t. Let Q be an orthogonal matrix such that $Q\hat{\theta^{(e)}} = (\|\hat{\theta}^{(e)}\|_2, 0, \ldots, 0)$. Also for any round t, let $(x_1^t, \ldots, x_k^t) = (Q^\intercal \epsilon_1^t, \ldots, Q^\intercal \epsilon_k^t)$. Due to rotational invariance $\epsilon_i^t \sim N(0, \sigma^2 \mathbb{I}_{p \times p}), 1 \leq i \leq k$. Therefore,

$$\begin{aligned} z^t &= \underset{x_i^t : 1 \leq i \leq k}{\operatorname{argmax}} \langle x_i^t, \hat{\theta^{(e)}} \rangle \\ Qz^t = \epsilon^t &= \underset{\epsilon_i^t : 1 \leq i \leq k}{\operatorname{argmax}} \langle \epsilon_i^t, Q\hat{\theta^{(e)}} \rangle \end{aligned} \tag{6.66}$$

Therefore $\epsilon^t \in \mathbb{R}^p$ is a p-dimensional random vector such that elements $(\epsilon^t)_j, 1 \leq j \leq p$ are random $N(0, \sigma^2)$ elements with $\|(\epsilon^t)_j\|_{\psi_2} \leq c_3 \sigma$ for some constant c_3. For the element at the first position,

$$(\epsilon^t)_1 = \underset{1 \leq i \leq k}{\operatorname{argmax}} (\epsilon_i^t)_1 \ , \tag{6.67}$$

where $(\epsilon_i^t)_1$ are $N(0, \sigma^2)$ elements. The following Lemma bounds the sub-Gaussian norm of $(\epsilon_i^t)_1$:

Lemma 35 *Let $g_1, \cdots, \mathbf{g}_k$ be k Gaussian $N(0, \sigma^2)$ elements and let $h = \underset{1 \leq i \leq k}{\operatorname{argmax}} g_i$. Then the sub-Gaussian norm of h satisfies the following:*

$$\|h\|_{\psi_2} \leq c_6 \sigma \sqrt{\log k} \ . \tag{6.68}$$

Proof: The maximum of k-Gaussian elements can be expressed as follows with vector $g = [g_1, \ldots, g_k] \in \mathbb{R}^k$:

$$\|g\|_\infty = \sup_{u : \|u\|_1 \leq 1} \langle g, u \rangle \ . \tag{6.69}$$

Therefore,

$$E\left[\sup_{1\leq i\leq k} g_i\right] = E\left[\sup_{u:\|u\|_1\leq 1} \langle g, u\rangle\right] \leq c_4\sigma\sqrt{\log k}\ , \tag{6.70}$$

where the last inequality is because the Gaussian width of the unit ℓ_1 norm ball is $\sqrt{\log k}$ [121, 122, 38] and by the majorizing measure theorem (see Theorem 2.1.1 in [121]). Now from the result of Lemma 2.1.3 in [121],

$$P(|\sup_{1\leq i\leq k} \mathbf{g}_i - E\sup_{1\leq i\leq k} \mathbf{g}_i| \geq u) \leq 2\exp\left(-\frac{u^2}{2\sigma^2}\right)\ . \tag{6.71}$$

Note that any random variable ξ is a sub-Gaussian random variable with sub-Gaussian norm c_5K is it satisfies the following tail decay [127],

$$P(|\xi| \geq u) \leq 2\exp\left(-\frac{u^2}{2K^2}\right)\ . \tag{6.72}$$

Therefore $\left(\sup_{1\leq i\leq k} \mathbf{g}_i - E\sup_{1\leq i\leq k} \mathbf{g}_i\right)$ is a $c_5\sigma$-sub-Gaussian random variable. Therefore,

$$\begin{aligned}\|h\|_{\psi_2} &= \|\sup_{1\leq i\leq k} g_i - E\sup_{1\leq i\leq k} g_i + E\sup_{1\leq i\leq k} g_i\|_{\psi_2}\\ &\leq \|\sup_{1\leq i\leq k} g_i - E\sup_{1\leq i\leq k} g_i\|_{\psi_2} + \|E\sup_{1\leq i\leq k} g_i\|_{\psi_2}\\ &\leq c_5\sigma + c_4\sigma\sqrt{\log k}\\ &\leq c_6\sigma\sqrt{\log k}\ .\end{aligned} \tag{6.73}$$

Therefore by the definition of sub-Gaussian random variables $(\epsilon^t)_1$ is a sub-Gaussian random variable with $\|(\epsilon^t)_1\|_{\psi_2} \leq c_6\sigma\sqrt{\log k}$ for some constant c_6. Therefore Qz^t is a random vector with independent sub-Gaussian random elements. Therefore from the result of Lemma 4 the elements of $z^t = Q^TQz^t$ are also independent sub-Gaussian random variables with sub-Gaussian norm of each element $\|(z^t)_i\|_{\psi_2} \leq c_7\sigma\sqrt{\log k}$. Also from the result of Lemma 5, z^t is a sub-Gaussian random variable with $\|z^t\|_{\psi_2} \leq c_2\sigma\sqrt{\log k}$ for some constant c_2 which proves the first result.

In order to prove the minimum eigenvalue condition, let Q be an orthogonal matrix such that $Q\hat{\theta}^{(e)} = (\|\hat{\theta}^{(e)}\|_2, 0, \ldots, 0)$ as outlined earlier. Again for any round t, let $(x_1^t, \ldots, x_k^t) = (Q^\intercal\epsilon_1^t, \ldots, Q^\intercal\epsilon_k^t)$. Due to rotational invariance $\epsilon_i^t \sim N(0, \sigma^2\mathbb{I}_{p\times p}), 1 \leq i \leq k$. Now with

$z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t, \hat{\theta}\rangle = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle Qx_i^t, Q\hat{\theta}\rangle$ and let $\epsilon^t = Qz^t$

$$\begin{aligned}
&\lambda_{\min}\left(E\left[z^t(z^t)^\intercal \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t, \hat{\theta}^{(e)}\rangle\right]\right) \\
&= \min_{w:\|w\|_2=1} w^\intercal\left(E\left[z^t(z^t)^\intercal \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t, \hat{\theta}^{(e)}\rangle\right]\right) w \\
&= \min_{w:\|w\|_2=1} \left(E\left[w^\intercal z^t(z^t)^\intercal w \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t, \hat{\theta}^{(e)}\rangle\right]\right) \\
&= \min_{w:\|w\|_2=1} \left(E\left[\langle w, z^t\rangle^2 \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t, \hat{\theta}^{(e)}\rangle\right]\right) \\
&\ge \min_{w:\|w\|_2=1} \text{Var}\left(\langle w, z^t\rangle \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t, \hat{\theta}^{(e)}\rangle\right) \\
&= \min_{w:\|w\|_2=1} \text{Var}\left(\langle Qw, Qz^t\rangle \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle Qx_i^t, Q\hat{\theta}^{(e)}\rangle\right) \\
&= \min_{w:\|w\|_2=1} \text{Var}\left(\langle Qw, Qz^t\rangle \,\middle|\, z_t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}(Qx^t)_1\|\hat{\theta}^{(e)}\|_2\right) \\
&= \min_{w:\|w\|_2=1} \text{Var}\left(\langle w, Qz^t\rangle \,\middle|\, z_t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}(Qx^t)_1\|\hat{\theta}^{(e)}\|_2\right)\ , \qquad (6.74)
\end{aligned}$$

where the last line uses that minimizing over w and over Qw yield the same result. Now $\epsilon^t = Qz^t$ is a $N(0, \sigma^2\mathbb{I}_{p\times p})$ random vector. Therefore it follows,

$$\begin{aligned}
&\lambda_{\min}\left(E\left[z^t(z^t)^\intercal \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t, \hat{\theta}^{(e)}\rangle\right]\right) \\
&\ge \min_{w:\|w\|_2=1} \left(\text{Var}\left[\langle w, \epsilon^t\rangle \,\middle|\, \epsilon^t = \underset{\epsilon_i^t:1\le i\le k}{\operatorname{argmax}}(\epsilon_i^t)_1\|\hat{\theta}^{(e)}\|\right]\right) \\
&\ge \min_{w:\|w\|_2=1} \left(w_1^2\text{Var}((\epsilon^t)_1) + \sum_{j=2}^{p} w_j^2\text{Var}((\epsilon^t)_j) \,\middle|\, \epsilon^t = \underset{\epsilon_i^t:1\le i\le k}{\operatorname{argmax}}(\epsilon_i^t)_1\|\hat{\theta}^{(e)}\|)\right) \\
&\ge c_1\frac{\sigma^2}{\log k}\ , \qquad (6.75)
\end{aligned}$$

where second line follows as the coordinates of ϵ^t are independent and the third line follows as from the result of Lemma 33 where $\text{Var}((\epsilon^t)_1|\epsilon^t = \underset{\epsilon_i^t:1\le i\le k}{\operatorname{argmax}}(\epsilon_i^t)_1\|\hat{\theta}^{(e)}\|) \ge c_1\frac{\sigma^2}{\log k}$ and

$\mathrm{Var}((\epsilon^t)_j | \epsilon^t = \underset{\epsilon_i^t : 1 \leq i \leq k}{\mathrm{argmax}} (\epsilon_i^t)_1 \|\hat{\theta}^{(e)}\|) = \sigma^2.$ ■

We give the proof for the estimation error in each episode for the Gaussian contexts setting.

Theorem 22 *Let* $T_e \geq c_7(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$. *Then with probability atleast* $1 - \delta \exp(-\eta_2 w^2(A)) - \delta$,

$$\|\hat{\theta}^{(e+1)} - \theta^*\|_2 \leq O\left(\frac{\gamma}{\sigma\sqrt{T_e}}\right) , \tag{6.76}$$

where $\gamma = c\kappa_\omega\sqrt{\log k}(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$, $E_c = \{\Delta \mid R(\theta^* + \Delta) \leq R(\theta^*)\}$, $A = cone(E_c) \cap S^{p-1}$ *is the error set,* $w(\cdot)$ *denotes the Gaussian width of a set.*

Proof: Consider parameter estimation at the beginning of episode $e+1$. Assume the design matrix has the SVD decomposition $\frac{1}{\sqrt{T_e}} Z^{(e)} = UDV^\intercal$ where $U \in \mathbb{R}^{T_e \times d}$, $D \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{p \times d}$, where d is the rank of $Z^{(e)}$. Also let $\Sigma^{1/2} = VDV^\intercal$. Define the Puffer transformation $F = UD^{-1}U^\intercal$ [75] and consider the preconditioned design matrix $\tilde{Z}^{(e)} = FZ^{(e)}$ and response $\tilde{y}^{(e)} = Fy^{(e)}$. Since $y^{(e)} = Z^{(e)}\theta^* + \omega^{(e)}$, it follows that $Fy^{(e)} = FZ^{(e)}\theta^* + F\omega^{(e)}$, i.e. $\tilde{y}^{(e)} = \tilde{Z}^{(e)}\theta^* + \tilde{\omega}^{(e)}$ where $\tilde{\omega}^{(e)} = F\omega^{(e)}$ We then compute the constrained regression estimator $\hat{\theta}^{(e)} = \underset{\theta \in \mathbb{R}^p}{\mathrm{argmin}} \frac{1}{2T_e} \|\tilde{y}^{(e)} - \tilde{Z}^{(e)}\theta\|_2^2 \quad \text{s.t.} \quad R(\theta) \leq R(\theta^*)$. Since $\hat{\theta}^{(e)}$ minimizes the loss function the following observation is straightforward,

$$\frac{1}{2T_e}\|\tilde{y}^{(e)} - \tilde{Z}^{(e)}\hat{\theta}^{(e)}\|_2^2 - \frac{1}{2T_e}\|\tilde{y}^{(e)} - \tilde{Z}^{(e)}\theta^*\|_2^2 \leq 0 \tag{6.77}$$

Let $\hat{\theta}^{(e)} = \theta^* + \Delta^{(e)}$ where $\Delta^{(e)}$ satisfies $R(\theta^* + \Delta^{(e)}) \leq R(\theta^*)$. Substituting it in (6.77) and subsequent simplification using $u = \frac{\Delta^{(e)}}{\|\Delta^{(e)}\|_2}$ yields the following,

$$
\begin{aligned}
\frac{1}{2T_e}\|\tilde{Z}^{(e)}\Delta^{(e)}\|_2^2 &= \frac{1}{2T_e}\|\tilde{Z}^{(e)}u\|_2^2\|\Delta^{(e)}\|_2^2 &(6.78)\\
&\leq \frac{1}{T_e}\left\langle \tilde{y}^{(e)} - \tilde{Z}^{(e)}\theta^*, \tilde{Z}^{(e)}\Delta^{(e)}\right\rangle\\
&\leq \frac{1}{T_e}\left\langle (\tilde{Z}^{(e)})^\intercal\tilde{\omega}^{(e)}, \Delta^{(e)}\right\rangle\\
&\leq \frac{1}{\sqrt{T_e}}\left\langle \frac{1}{\sqrt{T_e}}\Sigma^{1/2}(\tilde{Z}^{(e)})^\intercal\tilde{\omega}^{(e)}, \Sigma^{-1/2}\Delta^{(e)}\right\rangle\\
&\leq \frac{1}{\sqrt{T_e}}\left\langle \frac{1}{\sqrt{T_e}}\Sigma^{1/2}(FZ^{(e)})^\intercal F\omega^{(e)}, \Sigma^{-1/2}\Delta^{(e)}\right\rangle\\
&\leq \frac{1}{\sqrt{T_e}}\left\langle VDV^\intercal VDU^\intercal UD^{-1}U^\intercal UD^{-1}U^\intercal\omega^{(e)}, \Sigma^{-1/2}\Delta^{(e)}\right\rangle\\
&\leq \frac{1}{\sqrt{T_e}}\left\langle VU^\intercal\omega^{(e)}, \Sigma^{-1/2}\Delta^{(e)}\right\rangle\\
&\leq \frac{1}{\sqrt{T_e}}\langle h, \Sigma^{-1/2}\Delta^{(e)}\rangle\\
&\leq \frac{1}{\sqrt{T_e}}\langle h, \Sigma^{-1/2}u\rangle\|\Delta^{(e)}\|_2\,, &(6.79)
\end{aligned}
$$

where in the fourth line we use $\tilde{Z}^{(e)} = FZ^{(e)}$, $\tilde{\omega}^{(e)} = F\omega^{(e)}$; in the fifth line we use that $\Sigma^{1/2} = VDV^\intercal$, $\frac{1}{\sqrt{T_e}}Z^{(e)} = UDV^\intercal$ and $F^{(e)} = UD^{-1}U^\intercal$. In the second last line we observe that $h \in \mathbb{R}^p$ is a sub-Gaussian random vector with $\|h\|_{\psi_2} \leq c_3\kappa_\omega$. This is because using results from Lemma 5 it can be inferred that $U^\intercal\omega^{(e)} \in \mathbb{R}^d$ is sub-Gaussian with $\|U^\intercal\omega^{(e)}\|_{\psi_2} \leq c_4\kappa_\omega$ and again from Lemma 5, $h = VU^\intercal\omega^{(e)} \in \mathbb{R}^p$ is sub-Gaussian with $\|h\|_{\psi_2} \leq c_3\kappa_\omega$. We will now focus on lower bounds for $\inf\limits_{u\in A}\frac{1}{T_e}\|Z^{(e)}u\|_2^2$ and upper bounds for $\sup\limits_{u\in A}\frac{1}{\sqrt{T_e}}\langle h, \Sigma^{-1/2}u\rangle$

1. Minimum eigenvalue condition: Lower bounds for $\inf\limits_{u\in A}\frac{1}{T_e}\|Z^{(e)}u\|_2^2$

We obtain high probability lower bounds on the quantity $\inf\limits_{u\in A}\frac{1}{T_e}\|Z^{(e)}u\|_2^2$. Remember that $Z^{(e)} \in \mathbb{R}^{T_e\times p}$ is the design matrix before the Puffer transformation. We make the following observations:

$$\begin{aligned}\frac{1}{T_e}\|Z^{(e)}u\|_2^2 &= \frac{1}{T_e}\sum_{t=1}^{T_e}\langle z^t, u\rangle^2 \\ &= \frac{1}{T_e}\sum_{t=1}^{T_e}\langle z^t - E[z^t] + E[z^t], u\rangle^2 \\ &= \frac{1}{T_e}\sum_{t=1}^{T_e}\langle z^t - E[z^t], u\rangle^2 + \frac{1}{T_e}\sum_{t=1}^{T_e}\langle E[z^t], u\rangle^2 - \frac{2}{T_e}\sum_{t=1}^{T_e}\langle z^t - E[z^t], u\rangle\langle E[z^t], u\rangle\end{aligned} \tag{6.80}$$

We first analyze the quantity $\frac{1}{T_e}\sum_{t=1}^{T_e}\langle z^t - E[z^t], u\rangle^2$. Let $G \in \mathbb{R}^{T_e \times p}$ be the design matrix with rows as $z^t - E[z^t]$. Using the results of Lemma 25 and the episodic algorithm, we make the observation that the rows of the matrix G are i.i.d. σ-sub-Gaussian. We obtain lower bounds on the quantity $\frac{1}{T_e}\|Gu\|_2^2$. We use the following result [11, 103].

Theorem 27 (Mendelson, Pajor, Tomczak-Jaegermann [103]) *There exist absolute constants c_2, c_3, c_4 for which the following holds. Let (Ω, μ) be a probability space, set F be a subset of the unit sphere of $L_2(\mu)$, i.e., $F \subseteq S_{L_2} = \{f : \|f\|_{L_2} = 1\}$, and assume that $\sup_{f\in F}\ \|f\|_{\psi_2} \leq \kappa$. Then, for any $\theta > 0$ and $n \geq 1$ satisfying*

$$c_2\kappa\gamma_2(F, \|\cdot\|_{\psi_2}) \leq \theta\sqrt{n}\ , \tag{6.81}$$

with probability at least $1 - \exp(-c_3\theta^2 n/\kappa^4)$,

$$\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f^2(X_i) - E\left[f^2\right]\right| \leq \theta\ . \tag{6.82}$$

Further, if F is symmetric, then

$$\begin{aligned}&E\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f^2(X_i) - E\left[f^2\right]\right|\right] \\ &\leq c_4 \max\left\{2\kappa\frac{\gamma_2(F, \|\cdot\|_{\psi_2})}{\sqrt{n}}, \frac{\gamma_2^2(F, \|\cdot\|_{\psi_2})}{n}\right\}\end{aligned} \tag{6.83}$$

For convenience let z_0 have the same distribution as the rows of the design matrix G. Consider the following class of functions:

$$F = \{f_u, u \in A : f_u(\cdot) = \frac{1}{\sqrt{E[\langle\cdot, u\rangle^2]}}\langle\cdot, u\rangle\}\ . \tag{6.84}$$

Then, $f_u(z_0) = \frac{1}{\sqrt{E[\langle z_0, u\rangle^2]}}\langle z_0, u\rangle$ and F is a subset of the unit sphere, i.e., $F \subseteq S_{L_2}$, since $\|f\|_{L_2} = E[f_u^2] = 1$.

Next, we get an upper bound on $\sup_{f_u \in F} \|f_u\|_{\psi_2} = \sup_{u \in A} \left\| \frac{1}{\sqrt{E[\langle z_0, u\rangle^2]}}\langle z_0, u\rangle \right\|_{\psi_2}$. Note that $\kappa_z = \|z_0\|_{\psi_2} = \sup_{v \in S^{p-1}} \|\langle z_0, v\rangle\|_{\psi_2} \leq c_2\sigma$. Also from the result of Lemma 25, $E[\langle z_0, u\rangle^2] \geq \frac{\sigma^2}{\log k}$ Therefore,

$$\begin{aligned} \sup_{f_u \in F} \|f_u\|_{\psi_2} &= \sup_{u \in A} \left\| \frac{1}{\sqrt{E[\langle z_0, u\rangle^2]}}\langle z_0, u\rangle \right\|_{\psi_2} && (6.85) \\ &\leq \frac{c_2\sigma\sqrt{\log k}}{c_3\sigma} && (6.86) \\ &\leq c_4\sqrt{\log k}\,. && (6.87) \end{aligned}$$

As a result we have,

$$\gamma_2(F \cap S_{L_2}, \|\cdot\|_{\psi_2}) \leq c_4\gamma_2(F \cap S_{L_2}, \|\cdot\|_{L_2}) \leq c_4c_5w(A)\sqrt{\log k}\,, \qquad (6.88)$$

where the last line follows from generic chaining [122, 121], for some constant $c_5 > 0$. Therefore, in the context of Theorem 27, we choose,

$$\begin{aligned} \theta &= c_4^2\frac{(c_6c_5w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})\log k}{\sqrt{T_e}} \\ &\geq c_6c_4\sqrt{\log k}\frac{\gamma_2(F \cap S_{L_2}, \|\cdot\|_{\psi_2})}{\sqrt{T_e}}\,, && (6.89) \end{aligned}$$

for some constant $0 < \delta < 1$, so that the condition on θ is satisfied. With this choice of θ, we have,

$$\begin{aligned} \frac{\theta^2 T_e}{c_4^4\log^2 k} &\geq c_6^2c_5^2w^2(A) + \log\log T + \log(1/\delta) \\ &= \eta_2 w^2(A) + \log\log T + \log(1/\delta)\,. && (6.90) \end{aligned}$$

Then, from Theorem 27 it follows that with probability atleast $1 - \exp(-\eta_2 w^2(A) - \log\log T -$

$\log(1/\delta))$ with $z^t, 1 \leq t \leq T_e$ denoting the rows of Z^e , we have,

$$\sup_{u \in A} \left| \frac{1}{T_e} \frac{1}{E[\langle z^t - E[z^t], u\rangle^2]} \sum_{t \in [T_e]} \langle z^t - E[z^t], u\rangle^2 - 1 \right|$$
$$\leq c_4^2 \frac{(c_6 c_5 w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)}) \log k}{\sqrt{T_e}}$$
$$\Rightarrow \inf_{u \in A} \frac{1}{T_e} \|Gu\|_2^2 \geq E[\langle z^t - E[z^t], u\rangle^2]$$
$$\left(1 - c_4^2 \frac{(c_6 c_5 w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)}) \log k}{\sqrt{T_e}}\right) .$$

Substituting $T_e \geq c_7(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ so that $1 - c_4^2 \frac{(c_6 c_5 w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)}) \log k}{\sqrt{T_e}} \geq c_9$ and noting from Lemma 25 that $E[\langle z^t - E[z^t], u\rangle^2] = \text{Var}[\langle z^t, u\rangle]) \geq c_3 \frac{\sigma^2}{\log k}$ it follows with probability atleast $1 - \exp(-\eta_2 w^2(A) - \log\log T - \sqrt{\log(1/\delta)})$,

$$\inf_{u \in A} \frac{1}{T_e} \|Gu\|_2^2 \geq c_3 \frac{\sigma^2}{\log k} . \tag{6.91}$$

Now by a union bound argument for all episodes $e \leq \lfloor \log T \rfloor$ with probability atleast $1 - \exp(-\eta_2 w^2(A) - \log(1/\delta)) = 1 - \delta \exp(-\eta_2 w^2(A))$,

$$\inf_e \inf_{u \in A} \frac{1}{T_e} \|Gu\|_2^2 = \inf_e \inf_{u \in A} \frac{1}{T_e} \sum_{t=1}^{T_e} \langle z^t - E[z^t], u\rangle^2 \geq c_3 \frac{\sigma^2}{\log k} . \tag{6.92}$$

We now derive upper bounds for the quantity $\frac{1}{\sqrt{T_e}} \sum_{t=1}^{T_e} \langle z^t - E[z^t], u\rangle \langle E[z^t], u\rangle$. Let $\alpha \in \mathbb{R}^{T_e}$ be the vector whose elements $\alpha_i = \frac{1}{\sqrt{T_e}} \langle E[z^i], u\rangle$ and therefore $\|\alpha\|_2 = \frac{1}{\sqrt{T_e}} \sqrt{\sum_{t=1}^{T_e} \langle E[z^t], u\rangle^2}$. Note that it follows from Lemma 5 that $\langle z^t - E[z^t], u\rangle$ is a $c_2\sigma$-sub-Gaussian random variables, i.e., $\|\langle z^t - E[z^t], u\rangle\|_{\psi_2} \leq c_2\sigma$. Therefore from the Hoeffding inequality of Lemma 1:

$$P\left(\left|\sum_{t=1}^{T_e} \alpha_t \langle z^t - E[z^t], u\rangle\right| \geq \tau\right) \leq 2\exp\left(-\frac{\tau^2}{c_3 \sigma^2 \|\alpha\|_2^2}\right) . \tag{6.93}$$

Now for any $u, v \in A$, $\langle z^t - E[z^t], u - v\rangle$ is a $c_2\sigma\|u - v\|_2$-sub-Gaussian random variable. Therefore by an application of Lemma 1:

$$P\left(\left|\sum_{t=1}^{T_e} \alpha_t \langle z^t - E[z^t], u - v\rangle\right| \geq \tau\right) \leq 2\exp\left(-\frac{\tau^2}{c_3 \sigma^2 \|u - v\|_2^2 \|\alpha\|_2^2}\right) . \tag{6.94}$$

Therefore substituting $\sigma_1 = \sqrt{c_3}\sigma\|\alpha\|_2$, we get,

$$P\left(\left|\sum_{t=1}^{T_e}\alpha_t\langle z^t - E[z^t], u - v\rangle\right| \geq \tau\right) \leq 2\exp\left(-\frac{\tau^2}{\sigma_1^2\|u-v\|_2^2}\right) . \tag{6.95}$$

Therefore using properties of the Gaussian width defined in Section 2.2,

$$E\left[\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha_t\langle z^t - E[z^t], u\rangle\right|\right] \leq c_4\sigma_1 w(A) = c_5\sigma\|\alpha\|_2 w(A) . \tag{6.96}$$

Now for the high probability bounds we refer Theorem 2.2.27 in [122]. Applying the result of Theorem 2.2.27 [122] leads to the following result :

$$P\left(\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle z^t - E[z^t], u\rangle\right| \geq E\left[\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle z^t - E[z^t], u\rangle\right|\right] + c_6\sigma_1\tau\right) \leq c_7{\cdot}\exp(-\tau^2) . \tag{6.97}$$

Let $\tau = c_8(\sqrt{\log(1/\delta)} + \sqrt{\log\log T})$ choosing c_8 large enough so that $c_7 \cdot \exp(-\tau^2) \geq c_7 \cdot \exp(-c_8^2(\log\log T + \log(1/\delta))) \geq \exp(-\log\log T - \log(1/\delta))$. Also substituting the value of $E\left[\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right|\right]$ from equation (6.138) and choosing constant c_9 large enough, we get the following:

$$P\left(A \geq c_9\sigma\|\alpha\|_2(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})\right) \leq \exp\left(-\log(1/\delta) - \log\log T\right) . \tag{6.98}$$

This above is true for any single episode e. Taking a union bound over all $\lfloor\log\log T\rfloor$ episodes, we get:

$$\begin{aligned} P\left(A \geq c_8\sigma\|\alpha\|_2(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})\right) &\leq \exp\left(-\log(1/\delta) - \log\log T + \log\log T\right) \\ &\leq \delta . \end{aligned} \tag{6.99}$$

Now from equations (6.80), (6.92) and (6.99) we get,

$$\frac{1}{T_e}\|Z^{(e)}u\|_2^2 \geq c_3\frac{\sigma^2}{\log k} + \|\alpha\|_2^2 - \frac{2c_9\sigma\|\alpha\|_2(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})}{\sqrt{T_e}} . \tag{6.100}$$

Equation (6.100) is minimized when $\|\alpha\|_2 = \frac{c_9\sigma\|\alpha\|_2(w(A)+\sqrt{\log(1/\delta)}+\sqrt{\log\log T})}{\sqrt{T_e}}$. Substituting the minimum value in equation (6.100) and by simple algebraic manipulations we get:

$$\frac{1}{T_e}\|Z^{(e)}u\|_2^2 \geq \frac{\sigma^2}{\log k}\left(c_3 - \frac{c_9^2(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})^2\log k}{T_e}\right) \tag{6.101}$$

Then with $T_e \geq c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ and choosing c_1 large enough so that $c = c_3 - \frac{c_9^2(w(A)+\sqrt{\log\log T}+\sqrt{\log(1/\delta)})^2 \log k}{T_e} > 0$, we get:

$$\frac{1}{T_e}\|Z^{(e)}u\|_2^2 \geq c\frac{\sigma^2}{\log k} \; . \tag{6.102}$$

2. Upper Bounds for $\frac{1}{\sqrt{T_e}}\langle h, \Sigma^{-1/2}u\rangle$:

h is a sub-Gaussian random vector with $\|h\|_{\psi_2} \leq c_3\kappa_\omega$. We use the following result from generic chaining [121, 122] (also Theorem 9 in [11])

Theorem 28 *Let set $B \subseteq \mathbb{R}^p$. Assuming h is any centered sub-Gaussian random vector with $\|h\|_{\psi_2} \leq \kappa$, then we have for any $\tau_1 > 0$,*

$$P\left(\sup_{u\in B}\langle h, u\rangle \geq c_6\kappa w(B) + \tau_1\right) \leq \eta_4 \exp\left(-\left(\frac{\tau_1}{c_7\phi\kappa}\right)^2\right) \; , \tag{6.103}$$

where c_6, η_4, c_7 are positive constants and $\phi = \sup_{u\in B}\|u\|_2$.

Therefore applying Theorem 28 on the set $B = \{v \in \mathbb{R}^p \mid v = \Sigma^{-1/2}u, u \in A\}$ with A denoting the error set, we get the following noting that $w(B) \leq \sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}w(A)$ where $\sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}$ denotes the restricted maximum eigenvalue of the matrix, i.e. $\Lambda_{\max}(\Sigma^{-1}|A) = \sup_{u\in A} u^T\Sigma^{-1}u$ and $\phi = \sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}$,

$$P\left(\sup_{v\in B}\langle h, v\rangle \geq c_6c_3\kappa_\omega\sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}w(A) + \tau_1\right) \leq \eta_3 \exp\left(-\left(\frac{\tau_1}{c_7c_3\sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}\kappa_\omega}\right)^2\right) \tag{6.104}$$

Substituting $\tau_1 = c_3\kappa_\omega\sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}(c_7\sqrt{\log\log T} + c_8\sqrt{\log(1/\delta)})$, where we choose c_8 such that $\left(\frac{c_8}{c_7}\right)^2 \log(1/\delta) + \log\eta_3 > \log(1/\delta)$ we get:

$$\begin{aligned} &P\left(\sup_{v\in B}\langle h, u\rangle \geq c_3\kappa_\omega\sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}(c_6w(A) + c_7\sqrt{\log\log T} + c_8\sqrt{\log(1/\delta)})\right) \\ &\leq \exp(-\log(1/\delta) - \log\log T) \; . \end{aligned} \tag{6.105}$$

Inequality (6.105) is true for any episode e, taking a union bound over all $\lfloor \log T \rfloor$ episodes, we get for all episodes,

$$\begin{aligned} &P\left(\sup_{e}\sup_{v\in B}\langle h, u\rangle \geq c_3\kappa_\omega\sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}(c_6w(A) + c_7\sqrt{\log\log T} + c_8\sqrt{\log(1/\delta)})\right) \\ &\leq \exp(-\log(1/\delta)) = \delta \; . \end{aligned} \tag{6.106}$$

3. Estimation Error: Putting it all Together

Now consider the l.h.s of equation (6.79). Using the result equation (6.102), it is nonzero with probability atleast $1 - \delta \exp(-\eta_2 w^2(A))$ when $T_e \geq c_7(w(A) + \sqrt{\log \log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$. Moreover due to the preconditioning all eigenvalues are unit length and hence,

$$\inf_e \inf_{u \in A} \frac{1}{2T_e} \|\tilde{Z}^{(e)} u\|_2^2 \geq c_5 \|u\|_2^2 \geq c_5 \; . \tag{6.107}$$

Therefore from equations (6.79),(6.144) and (6.107), we get that with probability atleast $1 - \delta \exp(-\eta_2 w^2(A)) - \delta$

$$\begin{aligned} &\sup_e \|\hat{\theta}^{(e)} - \theta^*\|_2 = \|\Delta^{(e)}\|_2 \\ &\leq \frac{c_9 \kappa_\omega \sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}(c_6 w(A) + c_7\sqrt{\log\log T} + c_8\sqrt{\log(1/\delta)})}{\sqrt{T_e}} \; , \end{aligned} \tag{6.108}$$

where $c_9 = \frac{c_3}{c_5}$. Now from equation (6.102), $\Lambda_{\max}(\Sigma^{-1}|A) \leq \frac{\log k}{c\sigma^2}$. We have thus proved the advertised result. ■

We now prove regret bounds in the Gaussian contexts setting.

Theorem 25 *Consider Gaussian contexts. Then with probability atleast* $1 - 2\delta$

$$\beta = \max_{\substack{1 \leq i \leq k, 1 \leq t \leq T \\ a \in R^p : \|a\| = 1}} \langle x_i^t, a \rangle \leq c_1 \sigma (\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)}) \; . \tag{6.109}$$

Also with $T >> t_{\min} = c_7(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ *with probability atleast* $1 - \delta \exp(-\eta_1 w^2(A)) - 3\delta$ *the following is an upper bound on the regret for the Greedy algorithm,*

$$\textit{Reg}(T) \leq O\left(\frac{\gamma \cdot \beta \cdot \log(T) \cdot \sqrt{T}}{\sigma}\right) \tag{6.110}$$

where $\gamma = c\kappa_\omega \sqrt{\log k}(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$ *and*

Proof: From the result of Lemma 24 we have,

$$\text{Reg}(T) \leq 4\beta t_{\min} + \sum_{e = \lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{1}^{T_e} 2\beta \|\hat{\theta}^{(e)} - \theta^*\|_2 \; . \tag{6.111}$$

From the result in Theorem 1, we need $T_e > t_{\min} = c_7(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ for the RE condition to be satisfied. Moreover in each episode e we use the

$\hat{\theta}^{(e)}$ estimated using rounds played in the previous episode $e-1$ with $T_e = 2T_{e-1}$. Therefore substituting from the result of Theorem 3 the value of $\|\hat{\theta}^{(e)} - \theta^*\|_2$ in (6.111) we get,

$$\begin{aligned}
\text{Regret}(T) &\leq 4\beta t_{\min} + \sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{1}^{T_e} 2\beta \|\hat{\theta}^{(e)} - \theta^*\|_2 \\
&\leq 4\beta c_7 (w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k + \sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \sum_{1}^{T_e} \frac{2c\beta\gamma}{\sigma\sqrt{T_{e-1}}} \\
&\leq 4\beta c_7 (w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k + \sum_{e=\lceil \log t_{\min} \rceil}^{\lfloor \log T \rfloor} \frac{4c\beta\gamma T_{e-1}}{\sigma\sqrt{T_{e-1}}} \\
&\leq 4\beta c_7 (w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k + \frac{4c\beta\gamma\sqrt{T}\log T}{\sigma} \ , \qquad (6.112)
\end{aligned}$$

where in the second line we use that in the eth episode we play with $\hat{\theta}^{(e)}$ estimated using T_{e-1} rounds played in the previous episode, in the third line we use $T_e = 2T_{e-1}$ and in the last line we use $T > T_e$ for all e.

Also from the result of Lemma 31 we get with probability atleast $1 - 2\delta$, $\beta \leq c_1\sigma(\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)})$. ■

6.D Proofs for Single Parameter Setting with Adversary

We give proof for Lemma 26.

Lemma 26 *The rows of the design matrix $Z^{(e)} \in \mathbb{R}^{T_e \times p}$ in any episode e are $z^t = \mu^t + g^t$ where $\mu^t, g^t = \underset{\mu_i^t, g_i^t : 1 \leq i \leq k}{\text{argmax}} \langle \mu_i^t + g_i^t, \hat{\theta}^{(e-1)} \rangle$, $g_i^t \sim N(0, \sigma^2 \mathbb{I}_{p\times p})$ with the sub-Gaussian norm of g^t satisfying $\|g^t\|_{\psi_2} \leq c_2\sigma\sqrt{\log k}$ for some constant c_2. Moreover we have the following lower bound on the expected minimum eigenvalue for any μ_i^t's:*

$$\lambda_{\min}(E_{z^t}[z^t(z^t)^\intercal]) \geq c_1 \frac{\sigma^2}{\log k} \ , \qquad (6.113)$$

where c_1 is some constant.

Proof: For convenience we drop the superscript from $\hat{\theta}^{(e-1)}$. To bound the minimum eigenvalue we make the following observations.

$$\begin{aligned}\lambda_{\min}\left(E\left[z^t(z^t)^\intercal \,\middle|\, x^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t,\hat{\theta}\rangle\right]\right) &= \min_{w:\|w\|_2=1} w^T\left(E\left[z^t(z^t)^\intercal \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t,\hat{\theta}\rangle\right]\right)w\\ &= \min_{w:\|w\|_2=1}\left(E\left[w^\intercal z^t(z^t)^\intercal w \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t,\hat{\theta}\rangle\right]\right)\\ &\ge \min_{w:\|w\|_2=1}\text{Var}\left(\left[\langle w,z^t\rangle \,\middle|\, z^t = \underset{x_i^t:1\le i\le k}{\operatorname{argmax}}\langle x_i^t,\hat{\theta}\rangle\right]\right)\\ &\ge \min_{w:\|w\|_2=1}\text{Var}\left(\left[\langle w,g^t\rangle \,\middle|\, g^t = \underset{g_i^t:1\le i\le k}{\operatorname{argmax}}\langle \mu_i^t+g_i^t,\hat{\theta}\rangle\right]\right)\ ,\end{aligned} \tag{6.114}$$

where the last line follows because $\langle w,z^t\rangle = \langle w,\mu^t\rangle + \langle w,g^t\rangle$.

We will now prove that

$$\min_{w:\|w\|=1}\text{Var}\left[\langle g^t,w\rangle \,\middle|\, g^t = \underset{g_i^t:1\le i\le k}{\operatorname{argmax}}\langle \mu_i^t+g_i^t,\hat{\theta}\rangle\right] \ge \min_{w:\|w\|=1}\text{Var}\left[\langle g^t,w\rangle \,\middle|\, g^t = \underset{g_i^t:1\le i\le k}{\operatorname{argmax}}\langle g_i^t,\hat{\theta}\rangle\right]\ . \tag{6.115}$$

Therefore the worst any adversary can do is to ensure that the context corresponding to $g^t = \underset{g_i^t:1\le i\le k}{\operatorname{argmax}}\langle g_i^t,\hat{\theta}\rangle$ is chosen in each round. In fact this can be achieved by choosing $\mu_1^t = \mu_2^t = \ldots = \mu_k^t$ in any round.

We make the following observations. Let Q be an orthogonal matrix such that $Q\hat{\theta} = (\|\hat{\theta}\|_2, 0, \ldots, 0)$. Also let $(g_1^t,\ldots,g_k^t) = (Q^T\epsilon_1^t,\ldots,Q^T\epsilon_k^t)$. Due to rotational invariance

$\epsilon_i^t \sim N(0, \sigma^2 \mathbb{I}_{p\times p}), 1 \leq i \leq k$. Therefore,

$$
\begin{aligned}
&\min_{w:\|w\|=1} \text{Var}\left[\langle g^t, w\rangle \;\middle|\; g^t = \underset{g_i^t:1\leq i\leq k}{\text{argmax}} \langle \mu_i^t + g_i^t, \hat{\theta}\rangle\right] \\
&= \min_{w:\|w\|=1} \text{Var}\left[\langle Qg^t, Qw\rangle \;\middle|\; g^t = \underset{g_i^t:1\leq i\leq k}{\text{argmax}} \langle Q\mu_i^t + Qg_i^t, Q\hat{\theta}\rangle\right] \\
&= \min_{w:\|w\|=1} \text{Var}\left[\langle \epsilon^t, Qw\rangle \;\middle|\; \epsilon^t = \underset{\epsilon_i^t:1\leq i\leq k}{\text{argmax}} \langle Q\mu_i^t + \epsilon_i^t, Q\hat{\theta}\rangle\right] \\
&= \min_{w:\|w\|=1} \text{Var}\left[\langle \epsilon^t, w\rangle \;\middle|\; \epsilon^t = \underset{\epsilon_i^t:1\leq i\leq k}{\text{argmax}} (Q\mu_i^t + \epsilon_i^t)_1\right] \\
&= \min_{w:\|w\|=1} \left(w_1^2 \text{Var}((\epsilon^t)_1) + \sum_{j=2}^{p} w_j^2 \text{Var}((\epsilon^t)_j) \;\middle|\; \epsilon^t = \underset{\epsilon_i^t:1\leq i\leq k}{\text{argmax}} (Q\mu_i^t + \epsilon_i^t)_1\right) \\
&\geq c\frac{\sigma^2}{\log k} \qquad (6.116)
\end{aligned}
$$

where the last line is because the coordinates of ϵ^t are independent and from Lemma 34 and 33 we have

$$
\left(\text{Var}((\epsilon^t)_1 \;\middle|\; \epsilon^t = \underset{\epsilon_i^t:1\leq i\leq k}{\text{argmax}} (Q\mu_i^t + \epsilon_i^t)_1)\right) \geq \left(\text{Var}((\epsilon^t)_1 \;\middle|\; \epsilon^t = \underset{\epsilon_i^t:1\leq i\leq k}{\text{argmax}} (\epsilon_i^t)_1\right) \geq \frac{\sigma^2}{\log k}
$$

and $\text{Var}\left((\epsilon^t)_j \mid \epsilon^t = \underset{\epsilon_i^t:1\leq i\leq k}{\text{argmax}} (Q\mu_i^t + \epsilon_i^t)_1)\right) = \sigma^2$. That completes the proof. ■

Theorem 24 *The design matrix $Z^{(e)} \in \mathbb{R}^{T_e \times p}$ in all episode where $T_e \geq c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ satisfies the following minimum eigenvalue condition with probability atleast $1 - \delta\exp(-\eta_1 w^2(A)) - 2\delta$,*

$$
\inf_{u\in A} \frac{1}{T_e}\|Z^{(e)}u\|_2^2 \geq c\frac{\sigma^2}{\log k} \; . \qquad (6.117)
$$

Moreover, for all episodes when $T_e \geq c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ with probability atleast $1 - \delta\exp(-\eta_1 w^2(A)) - 3\delta$,

$$
\|\hat{\theta}^{(e+1)} - \theta^*\|_2 \leq O\left(\frac{\gamma}{\sigma\sqrt{T_e}}\right) \; , \qquad (6.118)
$$

where $\gamma = c\kappa_\omega\sqrt{\log k}(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$.

Proof: Using similar arguments as Theorem 22, we get:

$$\frac{1}{2T_e}\|\tilde{Z}^{(e)}u\|_2^2\|\Delta^{(e)}\|_2^2 \leq \frac{1}{\sqrt{T_e}}\langle h, \Sigma^{-1/2}u\rangle\|\Delta^{(e)}\|_2 \,. \tag{6.119}$$

Note that $h = VU^{\intercal}\omega^{(e)}$ is a sub-Gaussian random vector $\|h\|_{\psi_2} \leq c_1\kappa_\omega$ by direct application of Lemma 30. We obtain lower bounds for $\inf\limits_{u\in A}\frac{1}{T_e}\|Z^{(e)}u\|_2^2$ and upper bounds for $\sup\limits_{u\in A}\frac{1}{\sqrt{T_e}}\langle h, \Sigma^{-1/2}u\rangle$

1. Lower bounds for $\inf\limits_{u\in A}\frac{1}{T_e}\|Z^{(e)}u\|_2^2$

We first prove that $\inf\limits_{u\in A}\frac{1}{T_e}\|Z^{(e)}u\|_2^2 \geq c\frac{\sigma^2}{\log k}$ with high probability when $T_e \geq c_1(w^2(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2\log^2 k$. We make the following observations for some $u \in A$,

$$\begin{aligned}
\frac{1}{T_e}\|Z^{(e)}u\|_2^2 &= \frac{1}{T_e}\sum_{t=1}^{T_e}\langle z^t, u\rangle^2 \\
&= \frac{1}{T_e}\sum_{t=1}^{T_e}\left\langle g^t - E[g^t] + E[g^t] + \mu^t, u\right\rangle^2 \\
&= \frac{1}{T_e}\sum_{t=1}^{T_e}\left\langle g^t - E[g^t], u\right\rangle^2 + \frac{1}{T_e}\sum_{t=1}^{T_e}\left\langle E[g^t] + \mu^t, u\right\rangle^2 + \\
&\qquad \frac{2}{T_e}\sum_{t=1}^{T_e}\left\langle g^t - E[g^t], u\right\rangle\left\langle E[g^t] + \mu^t, u\right\rangle \qquad (6.120)\\
&= \frac{1}{T_e}\sum_{t=1}^{T_e}\left\langle g^t - E[g^t], u\right\rangle^2 + \|\alpha\|_2^2 + \frac{2\|\alpha\|_2}{\sqrt{T_e}}\sum_{t=1}^{T_e}\alpha^t\left\langle g^t - E[g^t], u\right\rangle \qquad (6.121)
\end{aligned}$$

where we denote $\alpha = \frac{1}{\sqrt{T_e}}[\langle E[g^1] + \mu^1, u\rangle, \ldots, \langle E[g^{T_e}] + \mu^{T_e}, u\rangle] \in \mathbb{R}^{T_e}$.

We will first obtain lower bounds for the quantity $\inf\limits_{u\in A}\frac{1}{T_e}\sum_{t=1}^{T_e}\left\langle g^t - E[g^t], u\right\rangle^2$ where A is the error set. Compared to the Gaussian context setting, the g^t's can no longer be assumed to be independent. The g^t's are adaptively generated based on observing the history of contexts chosen in earlier rounds and the corresponding rewards. We adopt the nomenclature in [12] to use their Theorem 5. Let $\xi^t = g^t - E[g^t]$ denote the centered random smoothing vector with $\|\xi^t\|_{\psi_2} \leq \sigma$ (see result before equation (6.73)) and $\xi = [(\xi^1)^{\intercal}, \ldots, (\xi^{T_e})^{\intercal}]^{\intercal} \in \mathbb{R}^{T_e p\times 1}$ be a random vector formed by concatenating the rows of the centered random smoothed component.

Also let $V \in R^{T_e \times T_e p}$ denote the following matrix indexed by vectors $u \in A$:

$$V(u) = \frac{1}{\sqrt{T_e}} \begin{bmatrix} u^T & 0 & \cdots & 0 \\ 0 & u^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u^T \end{bmatrix} . \tag{6.122}$$

Then by simple algebraic manipulations the following is a straightforward observation with $\Xi \in \mathbb{R}^{T_e \times p}$ denoting the random matrix obtained by stacking the g^t as rows:

$$\frac{1}{T_e} \sum_{t=1}^{T_e} \langle \xi^t, u \rangle^2 = \frac{1}{T_e} \|\Xi u\|_2^2 = \|V(u)\xi\|_2^2 . \tag{6.123}$$

To obtain lower bounds on $\inf_{u \in A} \|V(u)\xi|_2^2$ we focus on lower bounding $\inf_{u \in A} \left(\|V(u)\xi\|_2^2 - E\|V(u)\xi\|_2^2 \right)$ which can be obtained using the result of Theorem 5 in [12]. To apply Theorem 5, we first show that the random quantity satisfies the conditions required to apply the result of Theorem 5. Application of Theorem 5 in [12] requires the data generated to satisfy conditions (SP-1) and (SP-2) manifested by three graphical models. We first show that the data generation in the contextual bandit problem can be modelled using graphical model GM3 in [12]. We make the following observations:

1. Let $\mathcal{H}_{t-1}$ denote historical data observed until time $t-1$. In time step $t-1$ an adaptive adversary $\mathcal{A}_{t-1}$ maps the histories to k contexts $\mu_1^t, \ldots, \mu_k^t$ in $\mathbb{R}^p$ with $\|\mu_i^t\|_2 \leq 1$, i.e., $\mathcal{A}_{t-1} : \mathcal{H}_{t-1} \to (B_2^p)^k$ where B_2^p represents the unit ball in p dimensions. Nature perturbs the contexts with random Gaussian noise, i.e., $x_i^t = \mu_i^t + g_i^t$ with $g_i^t \sim N(0, \sigma^2 \mathbb{I}_{p \times p})$. Now, in the context of graphical model GM3, $\mathcal{H}_{t-1} \cup \{x_1^t, \ldots, x_k^t\}$ represents $F_{1:t-1}$.

2. In time step t, a learner chooses one among k contexts $\{x_1^t, \ldots, x_k^t\}$ based on historical data $\mathcal{H}_{t-1}$. Let z^t denote the selected context and g_t denote the corresponding Gaussian perturbation. In the context of GM3, we denote the centered Gaussian perturbation $g^t - \mathbf{E}[g^t]$ by ξ^t. The learner receives the noisy reward $y^t = \langle z^t, \theta^* \rangle + \omega^t$ where ω^t is an unknown sub-Gaussian noise. History at time step t is now augmented with the new data, i.e., $\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{\{x_1^t, \ldots, x_k^t\}, z^t, y^t\}$.

3. Now similar to step 1, the contexts in time step t, $\{x_1^{t+1}, \ldots, x_k^{t+1}\}$, are generated by an adversary $\mathcal{A}_t : \mathcal{H}_t \to (B_2^p)^k$ perturbed with Gaussian noise and $\mathcal{H}_t \cup \{x_1^{t+1}, \ldots, x_k^{t+1}\}$ represents $F_{1:t}$.

Lemma 36 *Let $G, \xi, V(u), \nu$ be constructed as above. Define the set $\mathcal{A} = \{V(u) \mid u \in A\}$. Then with probability atleast $1 - \exp(-c_9\epsilon^2 T_e)$*

$$\inf_{V(u)\in\mathcal{A}} \left(\|V(u)\xi\|_2^2 - E\|V(u)\xi\|_2^2\right) \geq \sigma^2 \left(-c_{10}\frac{w(A)}{T_e} - \epsilon\right) . \tag{6.124}$$

Proof: We start with the result of Theorem 5 in [12]. Let ξ' be a random vector constructed similar to ξ but with 1-sub-Gaussian norm. Therefore $\xi_i = c_4\sigma\xi'_i$ for some constant c_4. Also,

$$\begin{aligned} \|V(u)\xi\|_2^2 &= c_4^2\sigma^2\|V(u)\xi'\|_2^2 \\ E\|V(u)\xi\|_2^2 &= c_4^2\sigma^2\|V(u)\xi'\|_2^2 . \end{aligned} \tag{6.125}$$

We now apply Theorem 5 and Corollary 4 to obtain bounds on $\inf\limits_{u\in A}\|V(u)\xi'\|_2^2$. The values of the quantities in Theorem 5 of [12] are $\|V(u)\|_F = \|u\|_2 = 1$, $d_F(\mathcal{A}) = 1$, $\|V(u)\|_{2\to 2} = \frac{1}{\sqrt{T_e}}\|u\|_2 = \frac{1}{\sqrt{T_e}}$ and $d_{2\to 2}(\mathcal{A}) = \frac{1}{\sqrt{T_e}}$. Also the Gaussian width of the set $\mathcal{A}$:

$$\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) \leq c_5\frac{w(A)}{\sqrt{T_e}} . \tag{6.126}$$

Therefore we have,

$$M \leq c_6\left(\frac{w(A)}{T_e}\right), \quad V = O\left(\frac{1}{\sqrt{T_e}}\right), \quad U = \frac{1}{T_e} . \tag{6.127}$$

Then by application of result in Theorem 5 in [12], with $0 < \epsilon' < 1$ with probability atleast $1 - \exp(-c_7(\epsilon')^2 T_e)$,

$$\inf_{V(u)\in\mathcal{A}} \left(\|V(u)\xi'\|_2^2 - E\|V(u)\xi'\|_2^2\right) \geq -c_8\frac{w(A)}{T_e} - \epsilon' . \tag{6.128}$$

Now from the relationship (6.125), we get with probability atleast $1 - \exp(-c_9\epsilon^2 T_e)$,

$$\inf_{V(u)\in\mathcal{A}} \left(\|V(u)\xi\|_2^2 - E\|V(u)\xi\|_2^2\right) \geq \sigma^2 \left(-c_{10}\frac{w(A)}{T_e} - \epsilon\right) , \tag{6.129}$$

where $\epsilon = c_4^2\epsilon'$. $c_{10} = c_8 c_4^2$ and $c_9 = c_7/c_4^4$. This proves the stated result. ■

From the result of Lemma 26 we have,

$$\mathrm{Var}[\langle g^t, u\rangle] = E[\langle g^t - E[g^t], u\rangle^2] \geq c_1\frac{\sigma^2}{\log k} . \tag{6.130}$$

Therefore by simple algebraic manipulations we get,

$$E\|V(u)\xi\|_2^2 \geq c_{11}\frac{\sigma^2}{\log k} \; . \tag{6.131}$$

Therefore using the result of Lemma 36 with probability atleast $1 - \exp(-c_9\epsilon^2 T_e)$, we get

$$\inf_{V(u)\in\mathcal{A}} \|V(u)\xi\|_2^2 \geq \frac{\sigma^2}{\log k}\left(c_{11} - c_{10}\frac{w(A)\log k}{T_e} - \epsilon\log k\right) \tag{6.132}$$

Now choosing $T_e \geq c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2\log^2 k$ with $c_1 > 1$ large enough so that $0 < c_3 \leq \left(c_{11} - c_{10}\frac{w(A)\log k}{T_e} - \epsilon\log k\right)$, choosing $\epsilon \leq \frac{\epsilon'}{\log k}$ with $0 < \epsilon' < 1$ and choosing $\eta_1 = c_7(\epsilon')^2 c_1$, we get with probability atleast $1 - \delta\exp(-\eta_1 w^2(A) - \log\log T)$,

$$\inf_{V(u)\in\mathcal{A}} \|V(u)\xi\|_2^2 \geq c_3\frac{\sigma^2}{\log k} \; . \tag{6.133}$$

This is the bound for estimation in episode e. Taking a union bound over all episodes $e < \log\log T$, we get that with probability atleast $1 - \delta\exp(-\eta_1 w^2(A))$ over all rounds,

$$\inf_{u\in A}\sum_{t=1}^{T_e}\frac{1}{T_e}\langle g^t - E[g^t], u\rangle^2 = \inf_{V(u)\in\mathcal{A}}\|V(u)\xi\|_2^2 \geq c_3\frac{\sigma^2}{\log k} \; . \tag{6.134}$$

We now obtain upper bounds for $\sup\limits_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right|$.

Note that $g^t - E[g^t]$ is a $c_1\sigma$-sub-Gaussian random vector and hence $\beta^t = \langle g^t - E[g^t], u\rangle$ is a centered $c_1\sigma\|u\|_2 = c_1\sigma$ sub-Gaussian random variable by Lemma 5 in Chapter 2. Also β^t's are MDS with $E[\beta^i|\beta^1,\ldots,\beta^{i-1}] = 0$ and the coefficients $\alpha^1,\ldots,\alpha^t$ are adaptive, i.e., $\alpha_i = f_i((x_1^1,\ldots,x_k^1,z^t,y^1),\ldots,(x_1^{i-1},\ldots,x_k^{i-1},z^{i-1},y^{i-1}))$ depends on the history of the previously seen contexts and rewards. By an application of Lemma 30 for some $u\in A$, we get,

$$P\left(\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right| \geq \tau\right) \leq 2\exp\left(-\frac{\tau^2}{c_2\sigma^2\|\alpha\|_2^2}\right) \tag{6.135}$$

Now for any $u, v \in A$, $\langle g^t - E[g^t], u - v\rangle$ is a $c_1\sigma\|u-v\|_2$-sub-Gaussian random variable. Therefore by the application of Lemma 30 we get,

$$P\left(\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u - v\rangle\right| \geq \tau\right) \leq 2\exp\left(-\frac{\tau^2}{c_2\sigma^2\|u-v\|_2^2\|\alpha\|_2^2}\right) \; . \tag{6.136}$$

Therefore substituting $\sigma_1 = \sqrt{c_2}\sigma\|\alpha\|_2$, we get,

$$P\left(\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u - v\rangle\right| \geq \tau\right) \leq 2\exp\left(-\frac{\tau^2}{\sigma_1^2\|u-v\|_2^2}\right) . \tag{6.137}$$

Therefore, from the definition of Gaussian width and the majorizing measures theorem (put references),

$$E\left[\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right|\right] \leq c_3\sigma_1 w(A) = c_4\sigma\|\alpha\|_2 w(A) . \tag{6.138}$$

Now for the high probability bounds we refer Theorem 2.2.27 in [122]. Applying the result of Theorem 2.2.27 [122] leads to the following,

$$P\left(\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right| \geq E\left[\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right|\right] + c_5\sigma_1\tau\right) \leq c_6\cdot\exp(-\tau^2) . \tag{6.139}$$

Let $\tau = c_7(\sqrt{\log(1/\delta)} + \sqrt{\log\log T})$ choosing c_7 large enough so that $c_6 \cdot \exp(-\tau^2) \geq c_6 \cdot \exp(-c_7^2(\log\log T + \log(1/\delta))) \geq \exp(-\log\log T - \log(1/\delta))$. Also substituting the value of $E\left[\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right|\right]$ from equation (6.138) and choosing constant c_8 large enough, we get the following:

$$P\left(\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right| \geq c_8\sigma\|\alpha\|_2(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})\right)$$
$$\leq \exp\left(-\log(1/\delta) - \log\log T\right) . \tag{6.140}$$

This above is true for any episode e. Taking a union bound over all $\lfloor\log\log T\rfloor$ episodes, we get,

$$P\left(\sup_{u\in A}\left|\sum_{t=1}^{T_e}\alpha^t\langle g^t - E[g^t], u\rangle\right| \geq c_8\sigma\|\alpha\|_2(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})\right)$$
$$\leq \exp\left(-\log(1/\delta) - \log\log T + \log\log T\right) = \delta . \tag{6.141}$$

From equations (6.121), (6.134) and (6.141), we get with probability atleast $1 - \delta\exp(-\eta_1 w^2(A)) - \delta$,

$$\frac{1}{T_e}\|Z^{(e)}u\|_2^2 \geq c_3\frac{\sigma^2}{\log k} + \|\alpha\|_2^2 - \frac{2c_8\sigma\|\alpha\|_2}{\sqrt{T_e}}(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T}) . \tag{6.142}$$

Equation (6.142) is minimized when $\|\alpha\|_2 = \frac{c_8\sigma}{\sqrt{T_e}}(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})$. Substituting the minimum value in equation (6.142) and by simple algebraic manipulations, we get,

$$\frac{1}{T_e}\|Z^{(e)}u\|_2^2 \geq \frac{\sigma^2}{\log k}\left(c_3 - \frac{c_8^2(w(A) + \sqrt{\log(1/\delta)} + \sqrt{\log\log T})\log k}{\sqrt{T_e}}\right) \tag{6.143}$$

Then with $T_e \geq c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2 \log^2 k$ and choosing c_1 large enough so that $c = c_3 - \frac{c_8^2(w(A)+\sqrt{\log\log T}+\sqrt{\log(1/\delta)})^2\log k}{\sqrt{T_e}} > 0$, we get the advertised result on the minimum eigenvalue.

2. Upper Bounds for $\frac{1}{\sqrt{T_e}}\langle h, \Sigma^{-1/2}u\rangle$:

The following upper bound can be obtained using similar arguments as Theorem 22:

$$P\left(\sup_e \sup_{v\in B}\langle h, u\rangle \geq c_3\kappa_\omega\sqrt{\Lambda_{\max}(\Sigma^{-1}|A)}(c_6 w(A) + c_7\sqrt{\log\log T} + c_8\sqrt{\log(1/\delta)})\right)$$
$$\leq \exp(-\log(1/\delta)) = \delta\ . \tag{6.144}$$

3. Estimation Error: Putting it all Together Again by following similar arguments as Theorem 22, we obtain the following estimation error bounds with probability atleast $1 - \delta\exp(-\eta_2 w^2(A)) - \delta$:

$$\sup_e \|\hat{\theta}^{(e)} - \theta^*\|_2 = \|\Delta^{(e)}\|_2 \leq \frac{c_9\kappa_\omega(c_6 w(A) + c_7\sqrt{\log\log T} + c_8\sqrt{\log(1/\delta)})}{\sigma\sqrt{T_e}}\ . \tag{6.145}$$

Theorem 25 *In the smoothed adversary setting with probability atleast* $1 - 2\delta$

$$\beta = \max_{\substack{1\leq i\leq k, 1\leq t\leq T\\ a\in R^p:\|a\|=1}}\langle x_i^t, a\rangle \leq (1 + c_1\sigma(\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)})) \tag{6.146}$$

Also with $T >> t_{\min} \geq c_1(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2\log^2 k$ *with probability atleast* $1 - \delta\exp(-\eta_1 w^2(A)) - 5\delta$ *the following is an upper bound on the regret,*

$$\textit{Reg(T)} \leq O\left(\frac{\gamma\cdot\beta\cdot\log(T)\cdot\sqrt{T}}{\sigma}\right)\ , \tag{6.147}$$

where $\gamma = c\kappa_\omega\sqrt{\log k}(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})$.

Proof: We argue similarly as Theorem 25 to get bounds,

$$\text{Reg}(T) \leq 4\beta c(w(A) + \sqrt{\log\log T} + \sqrt{\log(1/\delta)})^2\log^2 k + \frac{4c\beta\gamma\sqrt{T}\log T}{\sigma}\ . \tag{6.148}$$

Now $\beta = \max_{\substack{1\leq i\leq k, 1\leq t\leq T\\ a\in R^p:\|a\|=1}}\langle x_i^t, a\rangle \leq \max_{\substack{1\leq i\leq k, 1\leq t\leq T\\ a\in R^p:\|a\|=1}}\left(\langle \mu_i^t, a\rangle + \langle g_i^t, a\rangle\right) \leq (1 + c_1\sigma(\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)}))$ with probability atleast $1 - 2\delta$ follows from Lemma (31). ■

6.E Proofs for Multi Parameter Setting

Lemma 27 *The greedy algorithm plays the contexts in an episodic fashion with the maximum episode number for each context $e_i \leq e_{i,\max} \leq \lfloor \log T \rfloor$. Denote by $\beta = \max_{\substack{1 \leq i \leq k, 1 \leq t \leq T \\ a \in R^p : \|a\| = 1}} \langle x_i^t, a \rangle$. Let $t_{\min} < T$, where $t_{\min}$ depends on properties of the true parameters θ_i^*, the regularizer $R(\cdot)$, the noise properties, the number of contexts k and the quantity β. Then,*

$$Reg(T) \leq 2\beta t_{\min} + \sum_{i=1}^{k} \sum_{e_i=1}^{e_{i,\max}} \left(\sum_{1}^{T_{i,e_i}} \beta \|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2 + \sum_{1}^{T_{i,e_i}^*} \beta \|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2 \right) \tag{6.149}$$

Proof: Let $i^*(t) = \underset{1 \leq i \leq k}{\operatorname{argmax}} \langle x_i^t, \theta_i^* \rangle$ denote the optimal context in any round t. Its context, for shorthand, is $\theta_{i^*}^*$ and let $x_{i^*}^t$ denote the context. Let i^t denote the context chosen in round t. The regret can be computed as follows,

$$\begin{aligned} \text{Reg}(T) &\leq \sum_{t=1}^{T} \langle \theta_{i^*}^*, x_{i^*}^t \rangle - \langle \theta_{i^t}^*, x_{i^t}^t \rangle \\ &\leq \sum_{t=1}^{t_{\min}} \langle \theta_{i^*}^*, x_{i^*}^t \rangle - \langle \theta_{i^t}^*, x_{i^t}^t \rangle + \sum_{t=t_{\min}+1}^{T} \langle \theta_{i^*}^*, x_{i^*}^t \rangle - \langle \theta_{i^t}^*, x_{i^t}^t \rangle \end{aligned} \tag{6.150}$$

The first term on the r.h.s. of (6.150) can be upper bounded as follows,

$$\begin{aligned} \sum_{t=1}^{t_{\min}} \langle \theta_{i^*}^*, x_{i^*}^t \rangle - \langle \theta_{i^t}^*, x_{i^t}^t \rangle &\leq \sum_{t=1}^{t_{\min}} |\langle \theta_{i^*}^*, x_{i^*}^t \rangle| + |\langle \theta_{i^t}^*, x_{i^t}^t \rangle| \\ &\leq \sum_{t=1}^{t_{\min}} 2\beta \qquad (6.151) \\ &\leq 2\beta t_{\min} \,. \qquad (6.152) \end{aligned}$$

To bound the second term on the r.h.s. in (6.150), assume in round t let $e_{i^*(t)}$ denote the episode number corresponding to the optimal context i^* and e_{i^t} denote the episode number corresponding to the selected context i^t. Again for shorthand we denote $e_{i^*(t)}$ by e_{i^*}. Let

$T_1, \ldots, T_k$ be the total number of rounds where contexts $1, \ldots, k$ are played respectively.

$$\begin{aligned}\sum_{t=t_{\min}+1}^{T} \langle \theta_{i^*}^*, x_{i^*}^t \rangle - \langle \theta_{i^t}^*, x_{i^t}^t \rangle = & \sum_{t=t_{\min}+1}^{T} \langle \theta_{i^*}^* - \hat{\theta}_{i^*}^{(e_{i^*})}, x_{i^*}^t \rangle - \langle \theta_{i^t}^* - \hat{\theta}_{i^t}^{(e_{i^t})}, x_{i^t}^t \rangle \qquad (6.153) \\ & + \langle \hat{\theta}_{i^*}^{(e_{i^*})}, x_{i^*}^t \rangle - \langle \hat{\theta}_{i^t}^{(e_{i^t})}, x_{i^t}^t \rangle \\ \leq & \sum_{t=t_{\min}+1}^{T} |\langle \theta_{i^*}^* - \hat{\theta}_{i^*}^{(e_{i^*})}, x_{i^*}^t \rangle| + |\langle \theta_{i^t}^* - \hat{\theta}_{i^t}^{(e_{i^t})}, x_{i^t}^t \rangle| \\ \leq & \sum_{i=1}^{k} \sum_{e_i=1}^{e_{i,\max}} \left(\sum_{1}^{T_{i,e_i}} \beta \|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2 + \sum_{1}^{T_{i,e_i}^*} \beta \|\theta_i^* - \hat{\theta}_i^{(e_i)}\|_2 \right) , \qquad (6.154)\end{aligned}$$

where the second inequality follows because $\langle \hat{\theta}_{i^*}^{(e_{i^*})}, x_{i^*}^t \rangle \leq \langle \hat{\theta}_{i^t}^{(e_{i^t})}, x_{i^t}^t \rangle$ as context i^t was chosen ahead of i^* in round t and the third inequality directly follows from the definitions of the various quantities in the Lemma.

The stated result now follows from (6.150), (6.152) and (6.154). ■

Proposition 1 *Consider any round t when the episode numbers of the k contexts are $e_1, \ldots, e_k$. Let i^* denote the context with the maximum reward, i.e., $i^* = \underset{1 \leq l \leq k}{\operatorname{argmax}} \langle \mu_l^t + g_l^t, \theta_l^* \rangle$. Let j denote the context having the second largest reward, i.e., $j = \underset{1 \leq l \leq k; l \neq i^*}{\operatorname{argmax}} \langle \mu_l^t + g_l^t, \theta_l^* \rangle$. Define $r = \langle \mu_j^t + g_j^t, \theta_j^* \rangle - \langle \mu_{i^*}^t, \theta_{i^*}^* \rangle$. Then the following condition is satisfied,*

$$\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r \; . \qquad (6.155)$$

Proof: Since context i^* is optimal in round t, we have

$$\begin{aligned} & \langle \mu_{i^*}^t + g_{i^*}^t, \theta_{i^*}^* \rangle \geq \langle \mu_j^t + g_j^t, \theta_j^* \rangle \\ \Rightarrow & \langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq \langle \mu_j^t + g_j^t, \theta_j^* \rangle - \langle \mu_{i^*}^t, \theta_{i^*}^* \rangle \; , \qquad (6.156)\end{aligned}$$

which proves the stated result. ■

Proposition 2 *Assume context j' such that $j' = \underset{1 \leq l \leq k, l \neq i^*}{\operatorname{argmax}} \langle \mu_l^t + g_l^t, \hat{\theta}_l^{(e_l)} \rangle = \underset{1 \leq l \leq k, l \neq i^*}{\operatorname{argmax}} \langle \mu_l^t + g_l^t, \theta_l^* + \Delta_l^{(e_l)} \rangle$, i.e., the context other than i^* which has the highest estimated reward. Also assume the parameter estimate for context i^* to be $\hat{\theta}_{i^*}^{(e_{i^*})} = \theta_{i^*}^* + \Delta_{i^*}^{(e_{i^*})}$. Then the greedy algorithm selects context i^* if the following condition is satisfied,*

$$\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \; . \qquad (6.157)$$

Proof: Now for context i^* to be optimal according to the Greedy algorithm the following condition should be satisfied,

$$\begin{aligned}
&\langle \mu_{i^*}^t + g_{i^*}^t, \theta_{i^*}^* + \Delta_{i^*}^{(e_{i^*})} \rangle \geq \langle \mu_{j'}^t + g_{j'}^t, \theta_{j'}^* + \Delta_{j'}^{(e_{j'})} \rangle \\
\Rightarrow &\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq \langle \mu_{j'}^t + g_{j'}^t, \theta_{j'}^* \rangle - \langle \mu_{i^*}^t, \theta_{i^*}^* \rangle + \langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \\
\Rightarrow &\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \,, \qquad (6.158)
\end{aligned}$$

where in the third line we use the assumption that $j = \underset{1 \leq l \leq k; l \neq i^*}{\operatorname{argmax}} \langle \mu_l^t + g_l^t, \theta_l^* \rangle$ and hence $\langle \mu_{j'}^t + g_{j'}^t, \theta_{j'}^* \rangle - \langle \mu_{i^*}^t, \theta_{i^*}^* \rangle \leq \langle \mu_j^t + g_j^t, \theta_j^* \rangle - \langle \mu_{i^*}^t, \theta_{i^*}^* \rangle = r$. ■

Lemma 28 (Margin Condition) *Consider good events as when $r \leq c_3 \sigma \sqrt{\log(Tk)}$ and consider errors $\Delta_{i^*}^{(e_{i^*})}$ and $\Delta_{j'}^{(e_{j'})}$ to be small enough such that $\langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \leq \frac{\sigma^2}{r}$. Then the following holds,*

$$P\left(\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \frac{\sigma^2}{r} \,\middle|\, \langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r \right) \geq \frac{1}{20} \,, \qquad (6.159)$$

for all $r \leq c_3 \sigma \sqrt{\log(Tk)}$.

Proof: We prove that assuming $\langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{(e_{j'})} \rangle - \langle \mu_{i^*}^t + g_{i^*}^t, \Delta_{i^*}^{(e_{i^*})} \rangle \leq \frac{\sigma^2}{r}$, conditioned on context i^* being optimal in round t implies that it will be played by Greedy with some constant non-zero probability, i.e., we prove the following,

$$P\left(\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \frac{\sigma^2}{r} \,\middle|\, \langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r \right) \geq \frac{1}{20} \,,$$

We use the result from Lemma 4.11 in [77] to lower bound $P\left(\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \frac{\sigma^2}{r} \,\middle|\, \langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r \right)$. We repeat the arguments for the sake of completeness. Denote by $\eta = \langle g_{i^*}^t, \theta_{i^*}^* \rangle$ and $\alpha = \frac{\sigma^2}{r}$. Then,

$$\begin{aligned}
P\left(\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \frac{\sigma^2}{r} \,\middle|\, \langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r \right) &= P[\eta \geq r + \alpha \mid \eta \geq r] \\
&= \frac{P[\eta \geq r + \alpha]}{P[\eta \geq r]} \\
&= \frac{1 - \Phi\left(\frac{r+\alpha}{\sigma}\right)}{1 - \Phi\left(\frac{r}{\alpha}\right)} \qquad (6.160)
\end{aligned}$$

Using Gaussian tail bounds (Lemma A.6 in [77]),

$$\frac{\phi(z)}{2z} \leq 1 - \Phi(z) \leq \frac{\phi(z)}{z} \ .$$

This gives,

$$\begin{aligned}
\frac{1 - \Phi\left(\frac{r+\alpha}{\sigma}\right)}{1 - \Phi\left(\frac{r}{\alpha}\right)} &\geq \frac{\phi\left(\frac{r+\alpha}{\sigma}\right)}{\phi\left(\frac{r}{\alpha}\right)} \frac{r}{r+\alpha} \frac{1}{2} \\
&\geq \exp\left[-\frac{(r+\alpha)^2 - r^2}{2\sigma^2}\right] \frac{r}{2(r+\alpha)} \\
&\geq \exp\left[-\frac{2r\alpha + \alpha^2}{2\sigma^2}\right] \frac{r}{2(r+\alpha)} \ .
\end{aligned}$$

Using $\alpha \leq r$ we get,

$$\exp\left[-\frac{2r\alpha + \alpha^2}{2\sigma^2}\right] \frac{r}{2(r+\alpha)} \geq \frac{1}{4} \exp\left[-\frac{3r\alpha}{2\sigma^2}\right] \tag{6.161}$$

$$\geq \frac{1}{4} e^{-\frac{3}{2}} \approx 0.05578 \ , \tag{6.162}$$

where in the second inequality we use $\alpha = \frac{\sigma^2}{r}$. Therefore we obtain,

$$P\left(\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r + \frac{\sigma^2}{r} \ \middle| \ \langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r\right) \approx 0.05578 \geq \frac{1}{20} \ , \tag{6.163}$$

which proves the third result.

Finally $P\left(\langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r' + \frac{\sigma^2}{r} \ \middle| \ \langle g_{i^*}^t, \theta_{i^*}^* \rangle \geq r'\right) \approx 0.05578 \geq \frac{1}{20}$ holds for all $r' < r$ due to the following result from [77].

Lemma 37 (Lemma A.10 in [77]) *Let $\eta \sim N(0, \sigma^2)$. Then for any $\alpha > 0$, the conditional "margin probability",*

$$P[\eta \geq b + \alpha \mid \eta \geq b] \ , \tag{6.164}$$

is decreasing in b.

We have thus proved all the stated results. ■

Lemma 29 (Properties of Design Matrices) *Consider any context i and a particular episode e_i. The rows of the design matrix $Z_i^{(e_i)} \in \mathbb{R}^{T_{i,e_i} \times p}$ are $z_i^t = \mu_i^t + g_i^t$ with t indexing the rounds in episode e_i where context i is chosen by the Greedy algorithm, i.e.,*

$z_i^t = \underset{x_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle x_l^t, \hat{\theta}_l^{(e_l)}\rangle$ *where* $x_l^t = \mu_l^t + g_l^t,\ g_l^t \sim N(0, \sigma^2\mathbb{I}_{p\times p})$. *Then under the condition* $\langle g_i^t, \theta_i^*\rangle \geq r$ *for some* $r \leq c_3\sigma\sqrt{\log(Tk)}$,

$$\lambda_{\min}\left(E\left[z_i^t(z_i^t)^\intercal \;\middle|\; z_i^t = \underset{g_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle x_l^t, \hat{\theta}_l^{(e_l)}\rangle; \langle g_i^t, \theta_i^*\rangle \geq r; r \leq c_3\sigma\sqrt{\log(Tk)}\right]\right) \geq c_2\frac{\sigma^2}{\log(Tk)}\,.$$

Proof: Using similar argument as used in Lemma 26 we get the following,

$$\begin{aligned}
&\lambda_{\min}\left(E\left[z_i^t(z_i^t)^\intercal \;\middle|\; z_i^t = \underset{x_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle x_l^t, \hat{\theta}_l^{(e_l)}\rangle\right]\right)\\
&= \min_{w:\|w\|_2=1} w^T\left(E\left[z_i^t(z_i^t)^\intercal \;\middle|\; z_i^t = \underset{x_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle x_l^t, \hat{\theta}_l^{(e_l)}\rangle\right]\right)w\\
&= \min_{w:\|w\|_2=1}\left(E\left[w^\intercal z_i^t(z_i^t)^\intercal w \;\middle|\; z_i^t = \underset{x_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle x_l^t, \hat{\theta}_l^{(e_l)}\rangle\right]\right)\\
&\geq \min_{w:\|w\|_2=1}\left(\text{Var}\left[\langle w, z_i^t\rangle \;\middle|\; z_i^t = \underset{x_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle x_l^t, \hat{\theta}_l^{(e_l)}\rangle\right]\right)\\
&\geq \min_{w:\|w\|_2=1}\left(\text{Var}\left[\langle w, g_i^t\rangle \;\middle|\; g_i^t = \underset{g_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle \mu_l^t + g_l^t, \hat{\theta}_l^{(e_l)}\rangle\right]\right) \qquad (6.165)
\end{aligned}$$

$$\lambda_{\min}\left(E\left[z_i^t(z_i^t)^\intercal \;\middle|\; x_i^t = \underset{x_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle x_l^t, \hat{\theta}_l^{(e_l)}\rangle\right]\right) \geq \min_{w:\|w\|=1}\text{Var}\left[\langle g_i^t, w\rangle \;\middle|\; g_i^t = \underset{g_l^t:1\leq l\leq k}{\operatorname{argmax}}\langle \mu_l^t + g_l^t, \hat{\theta}_l^{(e_l)}\rangle\right]. \qquad (6.166)$$

Let $j = \underset{1\leq m\leq k; m\neq i}{\operatorname{argmax}}\langle x_m^t, \hat{\theta}_m^{(e_m)}\rangle$ denote the context which has second maximum reward in round t and let $x_j^t = \mu_j^t + g_j^t,\ g_j^t \sim N(0, \sigma^2\mathbb{I}_{p\times p})$. Also let $\hat{\theta}_i^{(e_i)} = \theta_i^* + \Delta_i^{(e_i)}$ and $\hat{\theta}_j^{(e_j)} = \theta_j^* + \Delta_j^{(e_j)}$. Since context i is selected over context j in round t, we have the following,

$$\begin{aligned}
&\langle x_i^t, \hat{\theta}_i^{(e_i)}\rangle \geq \langle x_j^t, \hat{\theta}_j^{(e_j)}\rangle\\
\Rightarrow&\langle x_i^t, \theta_i^* + \Delta_i^{(e_i)}\rangle \geq \langle x_j^t, \theta_j^* + \Delta_j^{(e_j)}\rangle\\
\Rightarrow&\langle \mu_i^t + g_i^t, \theta_i^*\rangle + \langle x_i^t, \Delta_i^{(e_i)}\rangle \geq \langle x_j^t, \theta_j^* + \Delta_j^{(e_j)}\rangle\\
\Rightarrow&\langle g_i^t, \theta_i^*\rangle \geq \langle x_j^t, \theta_j^* + \Delta_j^{(e_j)}\rangle - \langle \mu_i^t, \theta_i^*\rangle - \langle x_i^t, \Delta_i^{(e_i)}\rangle\,. \qquad (6.167)
\end{aligned}$$

We now characterize the good events by the condition that $\langle x_j^t, \theta_j^* + \Delta_j^{(e_j)}\rangle - \langle \mu_i^t, \theta_i^*\rangle - \langle x_i^t, \Delta_i^{(e_i)}\rangle \leq c_3\sigma\sqrt{\log(Tk)}$. Note that there is very less probability on the complementary event $\langle x_j^t, \theta_j^* + \Delta_j^{(e_j)} \rangle - \langle \mu_i^t, \theta_i^* \rangle - \langle x_i^t, \Delta_i^{(e_i)} \rangle \geq c_3 \sigma \sqrt{\log(Tk)}$. Therefore,

$$\begin{aligned} &\text{Var}\left[\langle g_i^t, w \rangle \;\middle|\; g_i^t = \underset{g_l^t : 1 \leq l \leq k}{\text{argmax}} \langle \mu_l^t + g_l^t, \hat{\theta}_l^{(e_l)} \rangle \right] \\ &= \text{Var}\left[\langle g_i^t, w \rangle \;\middle|\; \langle g_i^t, \theta_i^* \rangle \geq \langle x_j^t, \theta_j^* + \Delta_j^{(e_j)} \rangle - \langle \mu_i^t, \theta_i^* \rangle - \langle x_i^t, \Delta_i^{(e_i)} \rangle \right] \\ &\geq \text{Var}\left[\langle g_i^t, w \rangle \;\middle|\; \langle g_i^t, \theta_i^* \rangle \geq c_3 \sigma \sqrt{\log(Tk)} \right] \\ &\geq c_2 \frac{\sigma^2}{\log(Tk)} \, , \end{aligned} \tag{6.168}$$

where in the second line we condition on the good events when $\langle x_j^t, \theta_j^* + \Delta_j^{(e_j)} \rangle - \langle \mu_i^t, \theta_i^* \rangle - \langle x_i^t, \Delta_i^{(e_i)} \rangle \leq c_3 \sigma \sqrt{\log(Tk)}$ and then use the fact $\text{Var}\left[\langle g_i^t, w \rangle \mid \langle g_i^t, \theta_i^* \rangle \geq a \right] \geq \Omega(1/a^2)$ is a decreasing function of a [77] so we condition on the maximum value of $a = \langle x_j^t, \theta_j^* + \Delta_j^{(e_j)} \rangle - \langle \mu_i^t, \theta_i^* \rangle - \langle x_i^t, \Delta_i^{(e_i)} \rangle = c_3 \sigma \sqrt{\log(Tk)}$. Again in the third line we use $\text{Var}\left[\langle g_i^t, w \rangle \mid \langle g_i^t, \theta_i^* \rangle \geq a \right] \geq \Omega(1/a^2)$ [77]. ■

Theorem 29 *Consider contexts to be indexed by i and the episode numbers to be indexed by e_i. Let S_{i,e_i} denote the set of rounds when context i was selected by the Greedy algorithm in episode e_i with $T_{i,e_i} = |S_{i,e_i}|$. Also assume all rounds satisfy the conditions of Lemma 26. Then when $T_{i,e_i} \geq c_9(w(A) + \sqrt{\log\log T} + \sqrt{\log k} + \sqrt{\log(1/\delta)})^2 \log^2(Tk)$, with probability atleast $1 - \delta \exp(-\eta_2 w^2(A)) - \delta$ the following RE condition holds for all contexts $1 \leq i \leq k$,*

$$\inf_{1 \leq i \leq k} \inf_{e_i \leq e_{i,\max}} \inf_{u \in A} \frac{1}{T_{i,e_i}} \| Z_i^{(e_i)} u \|_2^2 \geq c_4 \frac{\sigma^2}{\log(Tk)} \, . \tag{6.169}$$

Also consider parameter estimation using the constrained least squares estimator. Define the following quantities $r \leq c_3 \sigma \sqrt{\log(Tk)}$, $\gamma = \frac{c_{12} \kappa_\omega (w(A) + \sqrt{\log\log T} + \sqrt{\log k} + \sqrt{\log(1/\delta)}) \sqrt{\log(Tk)}}{\sigma}$ and $\beta = \max_{\substack{1 \leq i \leq k, 1 \leq t \leq T \\ a \in R^p : \|a\| = 1}} \langle x_i^t, a \rangle$. Then if $T_{i,e_i} \geq \frac{4\gamma^2 r^2 \beta^2}{\sigma^4}$, then with probability atleast $1 - \delta \exp(-\eta_1 w^2(A)) - 2\delta$,

$$\sup_{1 \leq i \leq k} \sup_{e_i \leq e_{i,\max}} \| \hat{\theta}_i^{(e_i+1)} - \theta_i^* \|_2 \leq \frac{\sigma^2}{2\beta r} \, . \tag{6.170}$$

Proof: The following result can be proved with probability atleast $1 - \delta \exp(-\eta_1 w^2(A)) - 2\delta$, using same arguments as Theorem 24.

$$\| \hat{\theta}_i^{(e_i+1)} - \theta_i^* \|_2 = \| \Delta_i^{(e_i)} \|_2 \leq \frac{c_{11} \kappa_\omega (c_7 w(A) + c_8 \sqrt{\log\log T} + c_8 \sqrt{\log k} + c_9 \sqrt{\log(1/\delta)}) \sqrt{\log(Tk)}}{\sigma \sqrt{T_{i,e_i}}} \, . \tag{6.171}$$

Now with $\gamma = \frac{c_{12}\kappa_\omega(w(A)+\sqrt{\log k}+\sqrt{\log(1/\delta)})\sqrt{\log(Tk)}}{\sigma}$ and $\beta = \max_{\substack{1\leq i\leq k, 1\leq t\leq T\\ a\in R^p:\|a\|=1}} \langle x_i^t, a\rangle$, it follows from (6.171) that when $T_{i,e_i} \geq \frac{4\gamma^2 r^2\beta^2}{\sigma^4}$ then,

$$\|\hat{\theta}_i^{(e_i+1)} - \theta^*\|_2 = \|\Delta_i^{(e_i)}\|_2 \leq \frac{\sigma^2}{2\beta r} \ , \tag{6.172}$$

which is the desired result. ■

Theorem 26 *Consider computation of regret for the Greedy algorithm in the multi parameter setting following Lemma 27. Define the following quantities* $r \leq c_3\sigma\sqrt{\log(Tk)}$, $\gamma = \frac{c_{12}\kappa_\omega(w(A)+\sqrt{\log T}+\sqrt{\log k})\sqrt{\log(Tk)}}{\sigma}$ *and* $\beta = \max_{\substack{1\leq i\leq k, 1\leq t\leq T\\ a\in R^p:\|a\|=1}} \langle x_i^t, a\rangle$. *The margin condition in Lemma 28 is satisfied with probability atleast* $1-\delta\exp(-\eta_1 w^2(A)) - 3\delta$ *when,*

$$t_{\min} \geq \frac{4k\gamma^2 r^2\beta^2}{\sigma^4} + \sqrt{\frac{1}{2}\log(1/\delta)} \ . \tag{6.173}$$

Under the margin condition, the regret is maximized when in each round each context has equal probability to be selected by the Greedy algorithm. The equal probability implies that in expectation $T_1 = T_2 = \ldots = T_k = \frac{T}{k}$. *Also the regret is upper bounded as follows,*

$$\textit{Reg}(T) \leq 2\beta t_{\min} + 82\beta\gamma\sqrt{Tk}\log(T) \ . \tag{6.174}$$

Moreover $\beta = \max_{\substack{1\leq i\leq k, 1\leq t\leq T\\ a\in R^p:\|a\|=1}} \langle x_i^t, a\rangle \leq (1 + c_1\sigma(\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)}))$ *with probability atleast* $1-2\delta$ *and with* $\kappa_\omega = 1$, $\sigma = \frac{c}{\sqrt{p}}$ *with probability atleast* $1-\delta\exp(-\eta_1 w^2(A)) - 6\delta$

$$\textit{Reg}(T) \leq O\left(\sqrt{p}\cdot\gamma\cdot\beta\cdot\log(T)\cdot\sqrt{Tk}\right) \tag{6.175}$$

We first derive bounds on the parameter $t_{\min}$ in Lemma 27. The multi-parameter setting requires a warm start of T_0 rounds, where T_0 is computed as,

$$T_0 = \frac{4k\gamma^2 r^2\beta^2}{\sigma^4} \ . \tag{6.176}$$

This is required for the margin condition of Lemma 28 to be satisfied with high probability. To see that, when $T_0 = \frac{k\gamma^2 r^2\beta^2}{\sigma^4}$, $T_{i,e_i} \geq \frac{4\gamma^2 r^2\beta^2}{\sigma^4}$ for all contexts $1 \leq i \leq k$ and all episodes

$e_i, 1 \leq i \leq k$. Then for any context combination i, j for $r = c_3\sigma\sqrt{\log(Tk)}$, we have the following,

$$\begin{aligned}
\langle \mu_i^t + g_i^t, \Delta_i^{(e_i)} \rangle - \langle \mu_j^t + g_j^t, \Delta_j^{(e_j)} \rangle &= \langle x_i^t, \Delta_i^{(e_i)} \rangle - \langle x_j^t, \Delta_j^{(e_j)} \rangle \\
&\leq \left| \langle x_i^t, \Delta_i^{(e_i)} \rangle \right| + \left| \langle x_j^t, \Delta_j^{(e_j)} \rangle \right| \\
&\leq \beta \|\Delta_i^{(e_i)}\|_2 + \beta \|\Delta_j^{(e_j)}\|_2 \\
&\leq \frac{\sigma^2}{r} , \qquad (6.177)
\end{aligned}$$

where in the last line we use that when T_{i,e_i} with high probability $\|\Delta_i^{e_i}\|_2, \|\Delta_j^{e_j}\|_2 \leq \frac{\sigma^2}{2\beta r}$. Let $i = \underset{1 \leq l \leq k}{\text{argmax}} \langle x_l^t, \theta_l^* \rangle = \langle \mu_l^t + g_l^t, \theta_l^* \rangle$ be the optimal context in round t. In the margin condition, we also assume that $\langle g_i^t, \theta_i^* \rangle \leq c_3\sigma\sqrt{\log(Tk)}$. We show that over T rounds the assumption is not satisfied only for a constant number of rounds. First note that for any context l, $\langle g_l^t, \theta_l^* \rangle$ is a $N(0, \sigma)$ Gaussian random variable. Therefore using Gaussian random variable tail bounds, we get,

$$P\left(\left| \langle g_l^t, \theta_l^* \rangle \right| \geq c_3\sigma\sqrt{\log(Tk)} \right) \leq \exp\left(-c_4 \log(Tk)\right) . \qquad (6.178)$$

Now there are a total of Tk realizations of $\langle g_l^t, \theta_l^* \rangle$ with $1 \leq l \leq k, 1 \leq t \leq T$. Consider the binomial random variable $\nu \sim \text{Binomial}(Tk, \exp\left(-c_4 \log(Tk)\right))$. Now $E[\nu] = Tk \ \exp\left(-c_4 \log(Tk)\right) = \exp\left(-c_4 \log(Tk) + \log(Tk)\right) \leq 1$ where we assume that constants c_3, c_4 are chosen such that the expectation is less than 1. Therefore by a tail bound for binomials,

$$P\left(\nu \geq 1 + \sqrt{\frac{1}{2}\log(1/\delta)} \right) \leq \delta . \qquad (6.179)$$

Therefore combining (6.176) and (6.179) the margin condition is satisfied with probability atleast $1 - \delta \exp(-\eta_1 w^2(A)) - 3\delta$ when,

$$\begin{aligned}
t_{\min} &\geq T_0 + 1 + \sqrt{\frac{1}{2}\log(1/\delta)} \\
&\geq \frac{4k\gamma^2 r^2 \beta^2}{\sigma^4} + 1 + \sqrt{\frac{1}{2}\log(1/\delta)} . \qquad (6.180)
\end{aligned}$$

Now to compute the regret, let $i = \underset{1 \leq l \leq k}{\text{argmax}} \langle x_l^t, \theta_l^* \rangle = \langle \mu_l^t + g_l^t, \theta_l^* \rangle$ be the actual optimal context in round t and $j' = \underset{1 \leq l \leq k; l \neq i}{\text{argmax}} \langle x_l^t, \theta_l^* + \Delta_l^{e_l} \rangle = \langle \mu_l^t + g_l^t, \theta_l^* + \Delta_l^{e_l} \rangle$ be the maximum estimated context rewards other than context i. Now according to (6.179), according to the margin condition except for $1+\sqrt{\frac{1}{2}\log(1/\delta)}$ rounds with high probability we have $\langle g_i^t, \theta_i^* \rangle \geq r$ for some $r \leq c_3\sigma\sqrt{\log(Tk)}$. Now for context i to be be selected over context j' we have the following condition,

$$\begin{aligned}\langle g_i^t, \theta_i^* \rangle &\geq r + \langle \mu_{j'}^t + g_{j'}^t, \Delta_{j'}^{e_{j'}} - \langle \mu_i^t + g_i^t, \Delta_i^{e_i^t} \rangle \\ &\geq r + \frac{\sigma^2}{r} \,, \end{aligned} \tag{6.181}$$

where the second inequality is from equation (6.177). Now from Lemma 29 we have established the following condition,

$$P\left(\langle g_i^t, \theta_i^* \rangle \geq r + \frac{\sigma^2}{r} \,\middle|\, \langle g_i^t, \theta^* \rangle \geq r \right) \geq \frac{1}{20} \,, \tag{6.182}$$

that is, context i is the estimated optimal context in 1 out of 20 times when context i is actually the optimal context. Now let T_{i,e_i}^* be the number of times context i is actually optimal in episode e_i. Then the number of times context i is estimated to be optimal is a binomial random variable: Binomial$(T_{i,e_i}^*, 1/20)$. Therefore applying Chernoff bounds for the binomial random variable $\left(T_{i,e_i}^*, \frac{1}{20}\right)$,

$$\begin{aligned} &P\left[T_{i,e_i} \leq \frac{T_{i,e_i}^*}{20} - \frac{T_{i,e_i}^*}{40}\right] \leq \exp\left(-\frac{T_{i,e_i}^*}{160}\right) \\ \Rightarrow &P\left[T_{i,e_i}^* \geq 40T_{i,e_i}\right] \leq \exp\left(-\frac{T_{i,e_i}}{4}\right) \end{aligned} \tag{6.183}$$

This is for any context i and episode e_i. Now taking a union bound over all contexts $1 \leq i \leq k$ and episodes $1 \leq e_i \leq \log T$ and using $T_{i,e_i} \geq c_9(w(A) + \sqrt{\log\log T} + \sqrt{\log k} + \sqrt{\log(1/\delta)})^2 \log^2(Tk)$ we get,

$$P[T_{i,e_i}^* \geq 40T_{i,e_i}] \leq \exp\left(-\frac{c_9^2(w^2(A) + \log(1/\delta))\log^2(Tk)}{4}\right) \leq \delta \,. \tag{6.184}$$

With this result the regret can be upper bounded as follows with probability atleast $1 -$

$\delta \exp(-\eta_1 w^2(A)) - 4\delta$

$$\begin{aligned}
\text{Regret}(x^t, i^1, \ldots, x^T, i^T) &\leq 2\beta t_{\min} + \sum_{i=1}^{k} \sum_{e_i=1}^{e_{i,\max}} \left(\sum_{t=1}^{T_{i,e_i}} \beta \|\theta_i^* - \hat{\theta}_i^{e_i}\|_2 + \sum_{t=1}^{T^*_{i,e_i}} \beta \|\theta_i^* - \hat{\theta}_i^{e_i}\|_2 \right) \\
&\leq 2\beta t_{\min} + \sum_{i=1}^{k} \sum_{e_i=1}^{e_{i,\max}} \left(\sum_{t=1}^{T_{i,e_i}} 41\beta \|\theta_i^* - \hat{\theta}_i^{e_i}\|_2 \right) \\
&\leq 2\beta t_{\min} + \sum_{i=1}^{k} \sum_{e_i=1}^{e_{i,\max}} \left(\sum_{t=1}^{T_{i,e_i}} 41\beta \frac{\gamma}{\sqrt{T_{i,e_i-1}}} \right) \\
&\leq 2\beta t_{\min} + \sum_{i=1}^{k} \sum_{e_i=1}^{e_{i,\max}} 82\beta\gamma\sqrt{T_{i,e_i}} \\
&\leq 2\beta t_{\min} + \sum_{i=1}^{k} 82\beta\gamma\sqrt{T_i}\log T_i \\
&\leq 2\beta t_{\min} + 82\beta\gamma\sqrt{Tk}\log T \ , \qquad (6.185)
\end{aligned}$$

where in the second inequality we have used the result (6.184), in the fourth inequality we have used $T_{i,e_i} = 2T_{i,e_i-1}$, in the fifth inequality we have used $e_{i,\max} \leq \log T_i$ and in the last inequality we have used $T_i = T/k$ gives the maximum regret and $\log T_i \leq \log T$.

Substituting the value of γ assumed earlier and noting $\beta = \max_{\substack{1\leq i\leq k, 1\leq t\leq T \\ a\in R^p : \|a\|=1}} \langle x_i^t, a\rangle \leq \max_{\substack{1\leq i\leq k, 1\leq t\leq T \\ a\in R^p : \|a\|=1}} \big(\langle \mu_i^t, a\rangle + \langle g_i^t, a\rangle\big) \leq (1 + c_1\sigma(\sqrt{\log(Tk)} + \sqrt{\log(1/\delta)}))$ with probability atleast $1 - 2\delta$ following from Lemma 31 proves the stated result. ■

Chapter 7

Conclusions

In this thesis, motivated by practical applications we studied four high-dimensional regression problems when data is non (sub)-Gaussian and/or independent. Our statistical analysis improves our understanding of the behavior of high-dimensional regression models under non-standard data assumptions. We also propose new algorithms, based on the analysis results, which have provably better performance compared to prior work.

In Chapter 3, we obtain parameter estimation error rates for high-dimensional regression when the design matrix and noise are sub-exponential. The results follow from the unified analysis frameworks for sub-Gaussian data established in prior literature [11] but using advanced probability theory tools like generic chaining to extend results to the sub-exponential data setting. We obtain sharp bounds on RE sample complexity for a large class of atomic norms. For the ℓ_1 norm, the results are sharper than in previous literature [3, 85]. We also give a precise characterization for the regularization parameter for parameter estimation. The upper bounds for the estimation error rates for sub-exponential data are worse by a multiplicate $\sqrt{\log p}$ factor compared to when data is sub-Gaussian.

In Chapter 4, we presented a general framework for the analysis of non-asymptotic error and structured recovery for norm regularized quantile regression for any atomic norm. Our results are based on extending the general analysis framework outlined in [11, 105] using insights from the geometry of the problem. In particular we introduce the Number of InterPolated Samples (NIPS) as critical for determining the sample complexity for consistent recovery. We prove that once the number of samples crosses the NIPS threshold, we start recovering the true parameter. This phase transition phenomena for norm regularized quantile regression problems has not been discussed in prior literature. We also prove that NIPS is of the order of square of the

Gaussian width of the error set for many atomic norms - which is the same order as that for regularized least squares regression and match results from previous work for the ℓ_1 norm [16].

In Chapter 5, we considered the problem of parameter estimation in mean and quantile SIM when the covariates are elliptically symmetric. We establish a common framework for mean SIM under which all existing estimators are compared. We also design a new estimator, borrowing ideas from the sliced inverse regression algorithm, which unlike existing estimators, is sample efficient, robust to heavy tails/outliers and can handle non-monotonic functions. We prove that the quantile regression estimator assuming a linear model can consistently estimate the parameter for quantile SIM. We provide non-asymptotic estimation error bounds for both mean and quantile single index models under general atomic norm constraints. The theoretical results are validated by experiments on synthetic data.

In Chapter 6, we analyzed the structured linear contextual bandit problem under the smoothed analysis framework. Our analysis significantly improves on the bounds obtained in [77]. While previous work have found it difficult to extend exploration strategies to the structured setting with simultaneously exploiting the structure in the parameter, our analysis shows that a simple greedy algorithm achieves sublinear regret under the smoothed bandits framework.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems. In *Conference on Learning Theory (COLT)*, 2011.

[2] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[3] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

[4] S. Agrawal and N. Goyal. Near-Optimal Regret Bounds for Thompson Sampling. *Journal of the ACM*, 64, 2017.

[5] P. Alquier, V. Cottet, and G. Lecue. Estimation Bounds and Sharp Oracle Inequalities of Regularized Procedures with Lipschitz Loss Functions. *CoRR arXiv:1702.01402*, 2017.

[6] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Inform. Inference*, 3(3):224–294, 2014.

[7] A. Argyriou, R. Foygel, and N. Srebro. Sparse Prediction with the \$k\$-Support Norm. In *Advances in Neural Information Processing Systems (NIPS)*, Apr. 2012.

[8] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex Optimization with Sparsity-Inducing Norms. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

[9] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, 2001.

[10] N. M. Ball and R. J. Brunner. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, 19(7):1049–1106, 2010.

[11] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with Norm Regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[12] A. Banerjee, Q. Gu, V. Sivakumar, and Z. S. Wu. Random quadratic forms with dependence: Applications to restricted isometry and beyond. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[13] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley and Sons, 1978.

[14] H. Bastani, M. Bayati, and K. Khosravi. Mostly exploration-free algorithms for contextual bandits. *CoRR arXiv:1704.09011*, 2018.

[15] E. M. L. Beale, M. G. Kendall, and D. W. Mann. The Identification of a Particular Nonlinear Time Series System. *Biometrika*, 54(3/4):357–366, 1967.

[16] A. Belloni and V. Chernozhukov. l1-Penalized Quantile Regression in High-Dimensional Sparse Models. *The Annals of Statistics*, 39(1):82–130, 2011.

[17] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[18] A. Bietti, A. Agarwal, and J. Langford. Practical evaluation and optimization of contextual bandit algorithms. *CoRR arXiv:1802.04064*, 2018.

[19] M. Bogdan, E. v. d. Berg, W. Su, and C. Emmanuel. Statistical Estimation and Testing via the Sorted L1 Norm. *CoRR arXiv:1310.1969*, 2013.

[20] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[21] P. T. Boufounos and R. G. Baranuik. 1-Bit Compressive Sensing. In *42nd Annual Conference on Information Sciences and Systems*, 2008.

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[23] D. R. Brillinger. The Identification of a Particular Nonlinear Time Series System. *Biometrika*, 64:509–515, 1977.

[24] D. R. Brillinger. A generalized linear model with Gaussian regressor variables. In *Selected Works of David Brillinger*, pages 589–606, 2012.

[25] L. Brown. *Fundamentals of Statistical Exponential Families*. Institue of Mathematical Statistics, 1986.

[26] S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*, volume 5. NOW, 2012.

[27] S. Cambanis, S. Huang, and P. Embrechts. On the Theory of Elliptically Contoured Distributions. *Journal of Multivariate Analysis*, 11:368–385, 1981.

[28] E. J. Candès and B. Recht. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[29] E. J. Candes, J. Romberg, and T. Tao. Robust Uncertainty Principles : Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[30] E. J. Candes and T. Tao. Decoding by Linear Programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[31] E. J. Candes and T. Tao. The Dantzig selector : statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[32] G. C. Cawley and N. L. Talbot. Gene Selection in Cancer Classification Using Sparse Logistic Regression with Bayesian Regularization. *Bioinformatics*, 22(19):2348–2355, 2006.

[33] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[34] S. Chatterjee. *High Dimensional Statistical Models: Applications to Climate*. PhD thesis, University of Minnesota, 2015.

[35] S. Chatterjee, S. Chen, and A. Banerjee. Generalized Dantzig Selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[36] S. Chatterjee, S. Liess, and A. Banerjee. Understanding Dominant Factors for Precipitation over Great Lakes Region. In *International workshop on Climate Informatics*, 2015.

[37] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly. Sparse Group Lasso: Consistency and Climate Applications. In *SIAM International Conference on Data Mining (SDM)*, 2012.

[38] S. Chen and A. Banerjee. Structured Estimation with Atomic Norms: General Bounds and Applications. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[39] S. Chen and A. Banerjee. Structured Matrix Recovery via the Generalized Dantzig Selector. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[40] S. Chen and A. Banerjee. Robust Structured Estimation with Single-Index Models. In *International Conference on Machine Learning (ICML)*, 2017.

[41] S. Chen and A. Banerjee. Sparse Linear Isotonic Models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[42] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1996.

[43] W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[44] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic Linear Optimization Under Bandit Feedback. In *Conference on Learning Theory (COLT)*, 2008.

[45] D. Das. *Bayesian Sparse Regression with Application to Data-Driven Understanding of Climate*. PhD thesis, Temple University, 2015.

[46] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. D. Langleben. Classifying Spatial Patterns of Brain Activity with Machine Learning Methods: Application to Lie Detection. *Neuroimage*, 28(3):663–668, 2005.

[47] S. Diaz, J. Kattge, J. H. Cornelissen, W. I. J., S. Lavorel, S. Dray, B. Reu, M. Kleyer, C. Wirth, I. C. Prentice, and et al. The Global Spectrum of Plant Form and Function. *Nature*, 529(7585):167–171, 2004.

[48] S. Dirksen, G. Lecué, and H. Rauhut. On the Gap Between Restricted Isometry Properties and Sparse Recovery Conditions. *IEEE Transactions on Information Theory*, 64(8):5478–5487, 2018.

[49] S. Dirksen and S. Mendelson. Robust One-Bit Compressed Sensing with Partial Circulant Matrices. *CoRR arXiv:1812.06719*, 2018.

[50] N. Draper and H. Smith. *Applied Regression Analysis*. Wiley, 1966.

[51] N. Duan and K.-C. Li. Slicing Regression: A Link-Free Regression Method. *The Annals of Statistics*, 19:509–530, 1991.

[52] M. Efroymson. Stepwise Regression - a Backward and Forward Look. *Eastern Regional Meetings of the Institute of Mathematical Statistics*, 1966.

[53] H. R. F. Krahmer, S. Mendelson. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

[54] J. Fan, Y. Fan, and E. Barut. Adaptive Robust Variable Selection. *The Annals of Statistics*, 42(4):324–351, 2014.

[55] J. Fan, Q. Li, and Y. Wang. Robust Estimation of High-Dimensional Mean Regression. *CoRR arXiv:1410.2150*, 2014.

[56] J. Fan, Q. Li, and Y. Wang. Estimation of High Dimensional Mean Regression in the Absence of Symmetry and Light Tail Assumptions. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 79(1):247–265, 2017.

[57] K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman & Hall, 1990.

[58] F. Fazayeli, A. Banerjee, J. Kattge, F. Schrodt, and P. Reich. Uncertainty Quantification Matrix Completion Using Bayesian Hierarchical Matrix Factorization. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2014.

[59] L. Goldstein, S. Minsker, and X. Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *CoRR arXiv:1609.01025*, 2015.

[60] Y. Gordon. Some Inequalitites for Gaussian Processes and Applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[61] Q. Gu, J. Trzasko, and A. Banerjee. Scalable algorithms for locally low-rank matrix modeling. *Knowledge and Information Systems (KAIS)*, 2018.

[62] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

[63] R. v. Handel. Probability in High Dimensions. Technical report, Princeton University, 2014.

[64] W. Hardle. *Applied Nonparametric Regression*. Econometric Society Monographs. Cambridge University Press, 1992.

[65] W. Hardle, M. Muller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer-Verlag, Berlin, 2004.

[66] R. R. Hocking and R. N. Leslie. Selection of the Best Subset in Regression Analysis. *Technometrics*, 9(4):531–540, 1967.

[67] J. L. Horowitz. *Semiparametric and Nonparametric Methods in Econometrics*. Springer, 2009.

[68] D. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning (ICML)*, 2014.

[69] P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Statistics*, 53(1):73–101, 1964.

[70] H. Ichimura. Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. *Journal of Econometrics*, 58:71–120, 1993.

[71] L. Jacob, O. Obozinski, and J. P. Vert. Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML)*, 2009.

[72] H. Jarvis, W. U. Bajwa, G. Raz, and R. Nowak. Toeplitz Compressed Sensing Matrices with Applications to Sparse Channel Estimation. *IEEE Transactions on Information Theory*, 56(11):5862–5875, 2010.

[73] A. Javanmard and H. Javadi. Dynamic Pricing in High Dimensions. *Journal of Machine Learning Research (JMLR)*, 2018.

[74] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal Methods for Sparse Hierarchical Dictionary Learning. In *International Conference on Machine Learning (ICML)*, 2010.

[75] J. Jia and K. Rohe. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9:1150–1172, 2015.

[76] B. Kai, R. Li, and H. Zou. New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models. *The Annals of Statistics*, 39:305–332, 2011.

[77] S. Kannan, J. Morgenstern, A. Roth, B. Waggoner, and Z. S. Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *CoRR arXiv:1801.04323*, 2018.

[78] K. Kato. Group Lasso for High Dimensional Sparse Quantile Regression Models. *CoRR arXiv:1103.1458*, 2011.

[79] D. Kelker. Distribution Theory of Spherical Distributions and a Location-Scale Parameter Generalization. *Sankhya A*, 32:419–430, 1970.

[80] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[81] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *CoRR arXiv:1312.3580*, 2013.

[82] J. Langford, L. Li, and T. Zhang. Sparse Online Learning via Truncated Gradient. In *Advances in Neural Information Processing Systems (NIPS)*, number 1, 2008.

[83] J. Langford and T. Zhang. The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[84] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *CoRR arXiv:1401.2188*, 2014.

[85] G. Lecué and S. Mendelson. Sparse Recovery Under Weak Moment Assumptions. *Journal of the European Mathematical Society*, 19:881–904, 2017.

[86] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Berlin, 1991.

[87] S. Lemm, B. Blankertz, T. Dickhaus, and K. R. Muller. Introduction to Machine Learning for Brain Imaging. *Neuroimage*, 56(2):387–399, 2011.

[88] K.-c. Li. Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association (JASA)*, 86(414):1–14, 1991.

[89] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *International World Wide Web Conference (WWW)*, 2010.

[90] X. Li, T. Zhao, X. Yuan, and H. Liu. The flare package for high dimensional linear regression and precision matrix estimation in r. *Journal of Machine Learning Research (JMLR)*, 16(1):553–557, 2015.

[91] Y. J. Li and J. Zhu. L_1-norm Quantile Regression. *Journal of Computational and Graphical Statistics*, 17:163–185, 2008.

[92] Q. Lin, X. Li, D. Huang, and J. S. Liu. On the Optimality of Sliced Inverse Regression in High Dimensions. *CoRR arXiv:1701.06009*, 2017.

[93] Q. Lin, Z. Zhao, and J. S. Liu. On Consistency and Sparsity for Sliced Inverse Regression in High Dimensions. *CoRR arXiv:1507.03895*, 2015.

[94] J. Liu and J. Ye. Moreau-Yosida Regularization for Grouped Tree Structure Learning. In *Advances in Neural Information Processing Systems*, 2010.

[95] X. Liu, P. Cao, A. R. Goncalves, D. Zhao, and A. Banerjee. Modeling alzheimers disease progression with fused laplacian sparse group lasso. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2018.

[96] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online Dictionary Learning for Sparse Coding. In *International Conference on Machine Learning (ICML)*, 2009.

[97] Y. Mansour, A. Slivkins, and Z. S. Wu. Competing bandits: Learning under competition. In *ITCS*, 2018.

[98] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 1989.

[99] N. Meinshausen and P. Bühlmann. High-dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[100] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[101] I. Melnyk and A. Banerjee. Estimating Structured Vector Autoregressive Model. In *International Conference on Machine Learning (ICML)*, 2016.

[102] S. Mendelson. Learning without concentration. *Journal of the ACM, To appear*, 2015.

[103] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subGaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17:1248–1282, 2007.

[104] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

[105] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A Unified Framework for High-Dimensional Analysis of \$M\$-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538–557, 2012.

[106] M. Neykov, Q. Lin, and J. Liu. Signed Support Recovery for Single Index Models in High-Dimensions. *CoRR arXiv:1511.02270*, 2015.

[107] G. Nowak, T. Hastie, J. R. Pollack, and R. Tibshirani. A Fused Lasso Latent Feature Model for Analyzing Multi-Sample aCGH Data. *Biostatistics*, 12(4):776–791, 2011.

[108] R. I. Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *CoRR arXiv:1312.2903*, 2013.

[109] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp Time-Data Tradeoffs for Linear Inverse Problems. *IEEE Transactions on Information Theory*, 64(6):4129–4158, 2018.

[110] Y. Plan and R. Vershynin. Robust 1-bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach. *IEEE Transactions on Information Theory*, 59(1):482–494, feb 2013.

[111] Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

[112] M. Raghavan, A. Slivkins, J. W. Vaughan, and Z. S. Wu. The externalities of exploration and how data diversity helps exploitation. In *Conference on Learning Theory (COLT)*, pages 1724–1738, 2018.

[113] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising Model Selection Using 1 -regularized Logistic Regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[114] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, 2010.

[115] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transaction on Information Theory*, 59(6):3434–3447, 2013.

[116] F. Santosa and W. W. Symes. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

[117] S. K. Shevade and S. S. Keerthi. A Simple and Efficient Algorithm for Gene Selection Using Sparse Logistic Regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[118] V. Sivakumar and A. Banerjee. High-Dimensional Structured Quantile Regression. In *International Conference on Machine Learning (ICML)*, 2017.

[119] V. Sivakumar, A. Banerjee, and P. Ravikumar. Beyond Sub-Gaussian Measurements: High-Dimensional Structured Estimation with Sub-Exponential Designs. In *Advances in Neural Information Processing Systems*, 2015.

[120] J. L. Starck, D. L. Donoho, and E. J. Candes. Astronomical Image Representation by the Curvlet Transform. *Astronomy & Astrophysics*, 398(2):785–800, 2003.

[121] M. Talagrand. *The Generic Chaining*. Springer Berlin, 2005.

[122] M. Talagrand. *Upper and Lower Bounds of Stochastic Processes*. Springer, 2014.

[123] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal Royal Statistical Society*, 58(1):267–288, 1996.

[124] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory - A Renaissance*. 2015.

[125] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

[126] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Verlag, 1996.

[127] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, Cambridge, 2012.

[128] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[129] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[130] M. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press (To appear), 2019.

[131] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using L1 -constrained quadratic programmming (Lasso). *IEEE Transaction on Information Theory*, 55(5):2183–2201, 2009.

[132] H. Wang and A. Banerjee. Online Alternating Direction Method. In *International Conference on Machine Learning (ICML)*, 2012.

[133] H. Wang, G. Li, and C. Tsai. Regression Coefficient and Autoregressive Order Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 69(1):63–78, 2007.

[134] L. Wang, Y. Wu, and R. Li. Quantile Regression for Analyzing Heterogeneity in Ultra-high Dimension. *Journal of the American Statistical Association*, 107:214–222, 2012.

[135] L. Wasserman. *All of Nonparametric Statistics*. Springer, New York, 2006.

[136] G. S. Watson. Smooth regression analysis. *Sankhya A*, 26:359–372, 1964.

[137] S. Weisberg. *Applied Linear Regression*. Series in Probability and Statistics. Wiley, 2013.

[138] T. Z. Wu, K. Yu, and Y. Yu. Single-Index Quantile Regression. *Journal of Multivariate Analysis*, 101(7):1607–1621, 2010.

[139] Y. C. Wu and Y. F. Liu. Variable Selection in Quantile Regression. *Statistica Sinica*, 19:801–817, 2009.

[140] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical Models via Generalized Linear Models. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[141] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. On Poisson Graphical Models. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[142] M. S. Yeung, J. Tegner, and J. J. Collins. Reverse Engineering Gene Networks Using Singular Value Decomposition and Robust Regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–6168, 2002.

[143] M. Yuan and Y. Lin. Model Selection and Estimation in Regression With Grouped Variables. *Journal of the Royal Statistical Society*, 68(1):49–67, 2006.

[144] P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research (JMLR)*, 7:2541–2563, 2006.

[145] L. Zhu, M. Huang, and R. Li. Semiparametric quantile regression with high-dimensional covariates. *Statistica Sinica*, 22(4):1379–1401, 2012.

[146] H. Zou and M. Yuan. Composite Quantile Regression and the Oracle Model Selection Theory. *The Annals of Statistics*, 36:1108–1126, 2008.