The Pennsylvania State University
The Graduate School
College of Engineering

HIGH DIMENSIONAL STATISTICAL LEARNING AND DECISION MAKING

A Dissertation in
Industrial Engineering
by
Xue Wang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2019

The dissertation of Xue Wang was reviewed and approved* by the following:

Tao Yao
Associate Professor of Industrial Engineering
Dissertation Advisor, Chair of Committee

Uday V. Shanbhag
Gary and Sheila Chaired Professor of Industrial Engineering

Ethan Xingyuan Fang
Assistant Professor of Industrial Engineering and Professor of Statistics

Runze Li
Verne M. Willaman Professor of Statistics

Janis Terpenny
Peter and Angela Dal Pezzo Department Head of IME and Professor

*Signatures are on file in the Graduate School.

Abstract

This dissertation concerns three problems in learning and decision-making with high-dimensional information. The formulations of these problems are featured with limit sample size requirements and/or efficient computation schemes.

We first study a regularized version of the sample average approximation (SAA). The theory on the traditional SAA scheme for stochastic programming (SP) dictates that the number of samples should be polynomial in the number of problem dimensions in order to ensure proper optimization accuracy. We study a modification to the SAA in the scenario where the global minimizer is either sparse or can be approximated by a sparse solution. By making use of a regularization penalty referred to as the folded concave penalty (FCP), we show that, if an FCP-regularized SAA formulation is solved locally, then the required number of samples can be significantly reduced in approximating the global solution of a convex SP: the sample size is only required to be poly-logarithmic in the number of dimensions. The efficacy of the FCP regularizer for nonconvex SPs is also discussed. As an immediate implication of our result, a flexible class of folded concave penalized sparse M-estimators in high-dimensional statistical learning may yield a sound performance even when the problem dimension cannot be upper-bounded by any polynomial function of the sample size.

In the second problem, we consider the linear constrained nonconvex programming problem. A broad class of learning problems can be formulated as the nonconvex optimization with linear constraints. It is believed that the second order optimal solution yields better out of sample performance since it can avoid part of the saddle points. The classic algorithms require matrix inversion to ensure the second order optimality for constrained optimization problems, which is computationally intensive when high-dimensional issue presents. We propose a novel accelerated interior-point gradient method (AIP-GM). A unique feature of the proposed AIP-GM is the total absence of the need for matrix inversion. As a consequence, the per-iteration cost is significantly lower than the canonical second-order methods. For general smooth non-convex objective function, we show

the new algorithm gives $\tilde{O}(\epsilon^{-7/4})$ iteration complexity dependences in perturbation ϵ on optimal condition.

For the third problem, we study a minimax concave penalized multi-armed bandit algorithm under generalized linear model (G-MCP-Bandit) for a decision-maker facing high-dimensional data in online learning and decision-making process. We demonstrate that the G-MCP-Bandit algorithm asymptotically achieves the optimal cumulative regret in the sample size dimension T , $O(\log T)$, and further attains a tight bound in the covariate dimension d, $O(\log d)$. In addition, we develop a linear approximation method, the 2-step weighted Lasso procedure, to identify the MCP estimator for the G-MCP-Bandit algorithm under non-iid samples. Under this procedure, the MCP estimator matches the oracle estimator with high probability and converges to the true parameters with the optimal convergence rate. Finally, through experiments based on synthetic data and two real datasets (warfarin dosing dataset and Tencent search advertising dataset), we show that the G-MCP-Bandit algorithm outperforms other benchmark algorithms, especially when there is a high level of data sparsity or the decision set is large.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

In the internet era, one of the most significant features in learning and decision-making problems is the involvement of high dimensional data. The high dimensional data may include more information, and it enables better modeling ability potentially. However, the high-dimensionality issue also poses challenges in both computation cost and statistical efficiency. Using classic approaches, we may need to solve ultra large scale problems with massive samples, which then yields very complicated models. In many real-world applications, the sample collection is expensive (e.g., healthcare) and/or the computation efficiency requirement is intense (e.g., ads recommendation). Therefore, it is necessary to study the settings with the limit sample size and/or computation cost. In this dissertation, we consider three particular problems that cover the topics in sample size requirement and efficient computation algorithm for learning and decision-making problems. More specifically, the following problems are discussed:

1. A regularized sample average approximation scheme for high-dimensional stochastic programming
2. An accelerated interior point gradient method for large scale linear constrained nonconvex programming
3. A contextual bandit algorithm for online learning and decision making with high-dimensional features

The rest of this chapter aims to provide some general backgrounds.

1.1 Statistical Learning and Sample Average Approximation

Statistical learning refers to the tools for understanding data [41]. Broadly speaking, those tools can be classified as supervised, unsupervised or semi-supervised. In this dissertation, we focus on the supervised statistical learning model, which involve building a statistical learning model for estimating the response based on input variables. The problem of this nature can be found in various domains, such as operation management, stock price prediction and personalized medicine.

Let $\{\boldsymbol{x}_j, y_j\},\ j = 1, 2, ..., n$ be the sample set whose samples are randomly drawn from a population with density function $f(\boldsymbol{x}, y, \boldsymbol{\beta})$, where $\boldsymbol{x} \in \mathbb{R}^{p\times 1}$, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ and $y \in \mathbb{R}$. We denote $\mathcal{L}(\boldsymbol{\beta})$ as the negative logarithm likelihood function:

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{j=1}^{n} \log\left[f(\boldsymbol{x}_j, y_j, \boldsymbol{\beta})\right]. \tag{1.1.1}$$

The $\boldsymbol{\beta}$ with smaller objective function value $\mathcal{L}(\boldsymbol{\beta})$ will have higher likelihood. The minimizer of $\mathcal{L}(\boldsymbol{\beta})$ is also referred as the maximum likelihood estimator (MLE) $\boldsymbol{\beta}_{MLE}$:

$$\boldsymbol{\beta}_{MLE} = \arg\min \mathcal{L}(\boldsymbol{\beta}). \tag{1.1.2}$$

In this dissertation, we mainly concentrate on two MLE models: linear least square regression and logistic regression. We also want to point out that our results in chapter 2 to 4 can also work on more general settings.
Linear least square regression is defined as:

$$\boldsymbol{\beta}_{least} = \arg\min \sum_{j=1}^{n} (\boldsymbol{x}_j^T\boldsymbol{\beta} - y_j)^2. \tag{1.1.3}$$

If we denote the residual of $\boldsymbol{x}_j^T\boldsymbol{\beta} - y_j$ as ϵ_j, one may show that the linear least squared estimator $\boldsymbol{\beta}_{least}$ is the estimator with smallest sum squared of squared residual (e.g., $\sum_j \epsilon_j^2$). As one of the most popular statistical learning model, the applications of linear least squared regression can be found in many areas (e.g., geodesy [64] and finance [67]).

Different from linear least squared regression, logistic regression is designed for the case that the response is binary. We have $y_j \in \{0, 1\}$, $\boldsymbol{x}_j \in \mathbb{R}^{p\times 1}$ and we want to fit an approximated model to use $\boldsymbol{x}_j$ to predict y_j:

$$\hat{y}_j = \frac{1}{e^{-\boldsymbol{x}_j^T\boldsymbol{\beta}} + 1}, \tag{1.1.4}$$

where $\hat{y}_j$ is the estimator of binary response y_j. When $\boldsymbol{x}_j^T\boldsymbol{\beta} \to +\infty$, we will have $\hat{y}_j \to 0$ and $\hat{y}_j \to 1$ if $\boldsymbol{x}_j^T\boldsymbol{\beta} \to -\infty$. The log-loss function of logistic regression is defined as:

$$\boldsymbol{\beta}_{logistic} = \arg\min \sum_{j=1}^{n} \left[(y_j - 1)\boldsymbol{x}_j^T\boldsymbol{\beta} + \log(e^{-\boldsymbol{x}_j^T\boldsymbol{\beta}} + 1)\right]. \tag{1.1.5}$$

Many problems in operations research and management science involve logistic regression. In dynamic pricing and assortment problem, the manager usually can only observe the binary purchase response and a common approach is to use logistic regression to fit a demand model and make decision based on it. More details can be found in [81]. In healthcare research area, we could also use logistic regression to identify the biomarkers [47].

From the view of stochastic programming (SP), we can also treat those problems as the instances of the sample average approximation (SAA). Denote by W a random vector with probability distribution $\mathbb{P}$ and support $\mathcal{W} \subseteq \mathbb{R}^q$. Define by $f(\,\cdot\,,\,\cdot\,) : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$ a deterministic mapping, where $\mathcal{X} \subseteq \mathbb{R}^p_+$ for some integer $p > 0$ is a compact and convex feasible region. Let $\mathbb{E}[f(\mathbf{x}, W)] = \int_{\mathcal{W}} f(\mathbf{x}, w)\mathbb{P}(dw)$. Assume that, for every $\mathbf{x} \in \mathcal{X}$, the function $f(\mathbf{x},\,\cdot\,)$ is measurable and integrable on $\mathcal{W}$. Then, the SP formulation of consideration is given as:

$$\min_{\mathbf{x}\in\mathcal{X}}\{F(\mathbf{x}) := \mathbb{E}[f(\mathbf{x}, W)]\}. \tag{1.1.6}$$

When distribution $\mathbb{P}$ is unknown and only finite sample can be collected, people may use the sample average approximation (SAA) instead. The optimization problem of SAA can be formulated as follow:

$$\min_{\mathbf{x}\in\mathcal{X}}\{F_n(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}, W_j)]\}. \tag{1.1.7}$$

When sample are iid draw from $\mathbb{P}$, we will have $\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}, W_j)$ converge to $\mathbb{E}[f(\mathbf{x}, W)]$. Thus if we collect enough data, it is expected the solution of SAA will be very close to the true solution of the original SP problem. Although the philosophy of SAA is not exactly the same to MLE, they both consider making the best use of the finite sample to infer the population information. It worths to think about solving the statistical learning problem from the view of SAA.

1.2 Sparse Inducing Penalties

In modern data science, the high-dimensional problem ($p \gg n$) becomes more and more important. It is well known that most traditional statistical procedures may fail to work when the dimension of parameters is much greater than the sample size (e.g., overparameteric setting). A typical example is the genetic data analysis. In neuroblastoma data set ([65]), we have gene expression profiles with 10707 genes from 251 patients. One common technique to tackle the overparameteric problem is to assume sparsity. Although a huge number of variables are collected, we believe that only a small portion of them are relevant to the problem. Screening out the nuisance variables is the key to improve model interpretability and prediction power. From the statistical view, an ideal model would be directly penalizing on the number of non-zero parameters:

$$\min \ \mathcal{L}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_0, \tag{1.2.8}$$

where $\mathcal{L}(\boldsymbol{\beta})$ is the MLE function in (1.1.2), λ is the tuning parameter to control the sparse level and $\|\boldsymbol{\beta}\|_0$ is the cardinality of $\boldsymbol{\beta}$. Problem (1.2.8) involves minimizing a non-convex non-lipschitz function. The complexity to find a global optimal solution can be NP-hard [16]. From the view of computation, its computation cost can be intense when the parameter dimension p is very large.

As an alternative method to L_0 penalty, the LASSO introduced in [83] is a popular tool for the high-dimensional learning problem. The LASSO is formulated as follow:

$$\min \ \mathcal{L}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{1.2.9}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p} |\beta^i|$ and $\boldsymbol{\beta} = (\beta^1, ..., \beta^p)$. Since (1.2.9) is a convex problem,

its global optimal solution can be efficiently computed by many standard convex optimization packages (e.g. cvx [38]). However, theoretical studies reveal that the LASSO requires a critical irrepresentable condition [97] to guarantee statistical performance and the solution could be biased. To solve this defect, folded concave penalties (FCP) were proposed. Two famous instances are smoothly clipped absolute deviation (SCAD) penalty [29] and minimax concave penalty (MCP) [94]. The motivation of these penalties is to connect the LASSO with l_0 penalty. SCAD uses spline function that begins with LASSO at around 0 and turns to constant function value when the input becomes large enough. The function form of SCAD penalty, P_{SCAD} is defined as:

$$P'_{SCAD}(x) = \lambda \mathbb{1}_{x \leq \lambda} + \frac{(a\lambda - \beta)_+}{a-1} \mathbb{1}(x > \lambda), \tag{1.2.10}$$

where $P'_{SCAD}(x)$ is the first order derivative of $P_{SCAD}(x)$, $\mathbb{1}$ is the indicator function, $(\cdot)_+ = \max\{\cdot, 0\}$, λ is the tuning parameter to control the sparsity level and $a > 1$. In [29], the authors suggests $a = 3.7$. If the spline function begins with lasso only at $x = 0$ and then goes towards to the l_0 penalty, SCAD penalty becomes MCP. We may define the penalty function of MCP as follow:

$$P'_{MCP}(x) = \left(\lambda - \frac{t}{a}\right)_+ . \tag{1.2.11}$$

Unlike the LASSO, these two penalty functions do not require the irrepresentable condition [97] to reach the variable selection and correct the system bias of LASSO method. But the price to pay is its computational intractability due to the non-convexity. Various local algorithms [29, 30, 33, 47] have been proposed to lessen computational burden. Recent progress made by [49] is to introduce the modern mixed integer program for obtaining global optimality with higher computational cost.

1.3 Linear Constrained Nonconvex Programming Problem

The linear constrained nonconvex programming problem refers to problem in the following structure:

$$\begin{aligned} \min_{x} \quad & F(x) \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0, \end{aligned} \tag{1.3.12}$$

where $A \in \mathbb{R}^{m\times n}$ and $F : \mathbb{R}^n_+ \to \mathbb{R}$ is a continuous function on $\mathbb{R}^n_+$ and smooth on $\mathbb{R}^n_{++}$. Many real world problem can be formulated into this formulation or their solutions can be approximately got from it (e.g., [39, 47, 48, 49, 51]). In machine learning and statistical learning area, many interesting problems can be formulated as:

$$\min \ \mathcal{L}(x) + P(x), \tag{1.3.13}$$

where $\mathcal{L}$ is loss function and P is regularization function. When P is absent, (1.3.12) reduces the classic machine/statistical learning problem, such as, regressions, SVM, neural network and so on. To overcome the overfitting or induce the sparsity solution, usually we need to add the extra regularization function which are commonly chosen as l_q norm ($q \in (0, 2]$), SCAD [29] and MCP [94]. In this case, the problem we want to minimize can be non-smooth, non-differentiable and/or non-convex. In various instances of (1.3.13), we can transform them into (1.3.12) using variable substitution technique [21, 43, 45]. Moreover, many problems in management science field are naturally represented in (1.3.12), such as, portfolio selection [22, 54], risk management [70], network optimization[6]. In game theory, it is also known that the solution to Nash equilibrium is one to one corresponding to the solution of linear complementarity problem, which can be solved by (1.3.12)[53, 100].

In general, (1.3.12) has better modeling power than the simple unconstrained problem. However, the linear constraints may introduce extra difficulty on algorithm design. Many algorithms are proposed in the last two decades, such as the alternating direction method of multipliers (ADMM) [12], stochastic gradient

descent (SGD) and their variants are proposed. Particularly, [29, 94] argue that the non-convex objective function of (1.3.12) may ensure better statistical performance.

1.4 Online Learning and Decision Making Models

Online decision making models contain two parts, the learning part and the decision making part. To motivate the necessity of online decision making models, we will start with the classic decision making model, which can be formulated as:

$$\min_{\boldsymbol{z}} \quad D(\boldsymbol{z}, \boldsymbol{\beta}) \tag{1.4.14}$$

$$\text{s.t.} \quad g_v(\boldsymbol{z}, \boldsymbol{\beta}) \quad v = 1, 2, ..., \tag{1.4.15}$$

where D is the loss function or negative utility function, g_v are the constraints, $\boldsymbol{z}$ is the vector of the decision variables and $\boldsymbol{\beta}$ is the vector of model parameters. In classic decision making models, we usually assume $\boldsymbol{\beta}$ is known or can be well estimated from enough existing data. Those models have several drawbacks. First, in real-world applications, we only collect the customers/users/patients data instead of $\boldsymbol{\beta}$, and the amount and the quality of the data may not allow us to get a good estimation of $\boldsymbol{\beta}$. Second, after making new decisions, we will be able to collect more data. The classic models don't allow us to use new data to improve decision making. Those drawbacks can be addressed by incorporating with an online learning module:

$$\boldsymbol{\beta}^t = \arg\min \ \mathcal{L}_t(\boldsymbol{\beta}) + \sum_{i=1}^{p} P(|\beta^i|), \tag{1.4.16}$$

where $\mathcal{L}_t$ is the statistical loss function with the sample up to time t and P is the sparse inducing penalty (e.g., LASSO, SCAD, and MCP). $\boldsymbol{\beta}^t$ changes with time and we expect $\boldsymbol{\beta}^t$ will converge to the true $\boldsymbol{\beta}$ with sample size increasing. Based on evolving $\boldsymbol{\beta}^t$, the quality of decision making will also get better and better. In this dissertation, we will focus on a special case of online decision making model: multi-armed bandit model with the generalized linear structure.

Let us consider a sequential arrival process $t \in \{1, 2, ..., T\}$. At each time step t, a single user, prescribed by a vector of user covariates, $\boldsymbol{x}_t \in \mathbb{R}^{1\times d}$, arrives. All covariate vector $\{\boldsymbol{x}_t\}_{t\geq 0}$ are observable to a decision-maker and are i.i.d. distributed according to an unknown distribution. The decision-maker has access to a decision

set $\mathcal{K} = \{1, 2, ..., K\}$, and the reward for decision $i \in \mathcal{K}$ on a user with a covariates vector $\boldsymbol{x}$ is defined as:

$$\mathbb{E}[R_i(\boldsymbol{x})] = f_i(\boldsymbol{x}^T \boldsymbol{\beta}_i^{true}), \tag{1.4.17}$$

where $f_i(\cdot)$ is the utility function with decision i and $\boldsymbol{\beta}_i^{true} \in \mathbb{R}^{1\times d}$ is the unknown coefficient vector for decision $i \in \mathcal{K}$. The decision maker needs to make the decision under the environment with uncertainty due to the absence of true decision parameter vector $\boldsymbol{\beta}_i^{true}$. We denote the decision-maker's policy as $\pi = \{\pi_t\}_{t\geq 0}$, where $\pi_t \in \mathcal{K}$ is the decision prescribed by policy π at time t. To benchmark the performance of policy π, we introduce an oracle policy $\pi^* = \{\pi_t^*\}_{t\geq 0}$ under which the decision-maker knows the true values of the covariates vector $\boldsymbol{\beta}_i^{true}$ for all $i \in \mathcal{K}$ and chooses the best decision to maximize its expected reward:

$$\pi_t^* \doteq \arg\max_i \mathbb{E}[R_i(\boldsymbol{x}_t)]. \tag{1.4.18}$$

Obviously, the decision-maker's reward is upper-bounded by the oracle policy. We then define the expected regret at time t for the observed user covariates $\boldsymbol{x}_t$ under policy π as:

$$r_t \doteq \mathbb{E}\left[\max_i R_i(\boldsymbol{x}_t) - R_{\pi_t}(\boldsymbol{x}_t).\right] \tag{1.4.19}$$

It is the expected reward difference between the optimal oracle policy π^* and the decision-maker's policy π at time t. Our goal is to explore the policy π that minimizes the cumulative regret up to time T, $R_T \doteq \sum_{i=1}^{T} r_t$. Many real-world problems can be solved by the linear multi-armed bandit model, such as online news recommendation, personal medicine, and adaptive clinic trial. More interesting details can be found in [9, 37].

1.5 Potential Contributions of the Dissertation Research

When facing high-dimensionality data, classic statistical learning approaches may fail to work. The sparse inducing penalty is very necessary to ensure the good quality solution. Convex sparse inducing penalized models (e.g., LASSO) can be efficiently calculated by convex solver, but the solution quality may have an

extra bias. Non-convex sparse inducing penalized (e.g., SCAD and MCP) models potentially have better statistical performance, but from the view of computation, it could be challenging. Globally solving the non-convex penalized models is NP-hard. The alternative way is to consider the computation efficient local algorithm with desirable statistical performance. We also explore online learning and decision making problem with high-dimensional data.

The rest of the dissertation is organized as follows. In Chapter 2 we study a modification to the SAA by incorporating the FCP regularization. This modification targets the high-dimensional SP problems with sparsity. We show that when the solution is sparse or can be approximated by a sparse solution, the regularization can significantly reduce the required number of samples in some high-dimensional SP applications: Compared to the conventional SAA approach that requires the sample size to grow polynomially in the number of dimensions, the RSAA stipulates number of samples that is only poly-logarithmic in the dimensionality. In Chapter 3, we discuss a fast algorithm for non-convex optimization problem with linear constraints. We propose an accelerated interior point gradient method and prove a better convergence rate than the classic $O(1/\epsilon^2)$ result. In Chapter 4, we consider the minimax concave penalized multi-armed bandit algorithm for the online decision making problem with high-dimensional covariates. We prove our approach can match the optimal regret bound theoretically. Numerical tests on both simulated and real-world data validate it. In Chapter 5, we summarize the dissertation studies and present future research directions.

We would like to alert the reader that notations in this dissertation are defined locally within a chapter, and do not apply to other chapters unless declared.

Chapter 2
Sample Average Approximation with Sparsity-Inducing Penalty for High-Dimensional Stochastic Programming

2.1 Introduction

We are interested in solving stochastic programming (SP) when the problem dimension is high but the global solution is approximately sparse. Denote by W a random vector with probability distribution $\mathbb{P}$ and support $\mathcal{W} \subseteq \mathbb{R}^q$ for some $q > 0$. Define by $f(\cdot, \cdot) : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$ a deterministic mapping, where $\mathcal{X} \subseteq \mathbb{R}^p_+$ for some integer $p > 0$ is a compact and convex feasible region. Let $\mathbb{E}[f(\mathbf{x}, W)] = \int_{\mathcal{W}} f(\mathbf{x}, w)\mathbb{P}(dw)$. Assume that, for every $\mathbf{x} \in \mathcal{X}$, the function $f(\mathbf{x}, \cdot)$ is measurable and integrable on $\mathcal{W}$. Then, the SP formulation of consideration is given as:

$$\min_{\mathbf{x}\in\mathcal{X}} \{F(\mathbf{x}) := \mathbb{E}[f(\mathbf{x}, W)]\}. \tag{2.1.1}$$

Throughout the chapter, we assume that $\mathcal{X}$ is defined only by coordinate-wise constraints, that is, $\mathcal{X} := \{\mathbf{x} = (x_i) : x_i \in X_i,\ i = 1, ..., p\}$ for some $X_i \subseteq \mathbb{R}_+$ for all $i = 1, ..., p$. Notice that the non-negativity constraints are not restrictive, in that we may always represent a negative variable by the difference of two non-negative variables.

In addition, we will restrict our discussions to the cases where the solution to the original SP, denoted $\mathbf{x}^{\min} \in \arg\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$, can be well approximated by a sparse solution. More precisely, we assume that there exists $\hat{\mathbf{x}}^{\min}$ that satisfies

$$F(\hat{\mathbf{x}}^{\min}) - F(\mathbf{x}^{\min}) \leq \hat{\varepsilon} \tag{2.1.2}$$

for some $\hat{\varepsilon} \geq 0$. We denote that $\mathcal{S} := \{i : \hat{x}_i^{\min} > 0\}$ and $\mathcal{S}^c := \{i : \hat{x}_i^{\min} = 0\}$. Here $\mathcal{S}$ can be understood as the index set for the most contributing dimensions with $|\mathcal{S}|$ assumed small and satisfying $|\mathcal{S}| << p$ and $|\mathcal{S}| < n$. In the special case when $\hat{\varepsilon} = 0$, we know that $\hat{\mathbf{x}}^{\min}$ is an exact solution to (2.1.1).

Under the above setting, one of the most commonly used techniques to solve the SP, the sample average approximation (SAA), is undesirably restrictive on the sample size in some scenarios. The SAA approximates the objective function of (2.1.1) by

$$F_n(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}, W^j), \tag{2.1.3}$$

where $\{W^j : 1, \ldots, n\}$ is a sequence of independently and identically distributed (i.i.d.) random samples of W. Denote that $\mathbf{x}^{SAA} \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_n(\mathbf{x})$. Much literature has discussed the efficacy of $\mathbf{x}^{SAA}$ in approximating $\mathbf{x}^{\min}$ (see [60, 77]). It has been shown in the celebrated work by Shapiro and co-authors [44, 77, 77] that to ensure the optimization accuracy, the required number of samples should be larger than the number of dimensions and should grow polynomially with the increase of dimensionality. In specific, to ensure:

$$\mathbf{P}[F(\mathbf{x}^{SAA}) - F(\mathbf{x}^{\min}) \leq \epsilon] \geq 1 - \alpha, \tag{2.1.4}$$

for any $\epsilon \in (0, 1]$ and $\alpha \in (0, 1]$, the sample size n should satisfy

$$n \gtrsim \frac{p}{\epsilon^2}\ln\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}, \tag{2.1.5}$$

where $x \gtrsim y$ for any $x, y \in \mathbb{R}$ means $x \geq \tilde{c}y$, for some constant $\tilde{c} > 0$ that are independent of α, ϵ, p, and $|\mathcal{S}|$, but may depend polynomially on some other problem quantities. Consider (2.1.5) in a problem with perhaps hundreds of thousands of dimensions, which is not rare in actual applications of SP. The SAA then likely

requires more than millions or even tens of millions of samples for the SAA to perform properly. The overhead in generating these samples, before conducting any optimization-related computation, may have already become prohibitive. Especially considering the case where the most contributing dimensions are in tens or hundreds, such a sample size requirement seems unreasonably demanding.[1]

Seeking to address the above issue, this work studies a modification to (2.1.3) by adding a regularization term to encourage sparsity. This term is in the form of a folded concave penalty (FCP) as first introduced by [29] and [94] to some statistical learning problems. We refer to this modification the regularized SAA (RSAA), which is formulated as:

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{ F_{n,\lambda}(\mathbf{x}) := F_n(\mathbf{x}) + \sum_{i=1}^{p} P_\lambda(x_i) \right\}, \tag{2.1.6}$$

where P_λ with parameters $a > 0$ and $\lambda > 0$ is a special form of FCP called the minimax concave penalty (MCP)[94]:

$$P_\lambda(\tau) := \int_0^\tau \frac{(a\lambda - t)_+}{a}\, dt = \begin{cases} \lambda\tau - \frac{\tau^2}{2a} & \text{if } 0 \le \tau \le a\lambda; \\ \frac{1}{2}a\lambda^2 & \text{if } \tau > a\lambda. \end{cases} \tag{2.1.7}$$

We show in this chapter that the RSAA allows the dimension to be (much) more than the sample size. In specific, when $\hat{\varepsilon} = 0$, to achieve the same optimization quality in (2.1.4), the sample size requirement for the global minimizer to RSAA is

$$n \gtrsim \frac{|\mathcal{S}|}{\epsilon^3}\left(\ln\frac{p}{\epsilon}\right)^{1.5} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}, \tag{2.1.8}$$

under no assumption of convexity. Compared to (2.1.5), the required sample size of RSAA only depends polynomially on $|\mathcal{S}|$ and $\ln p$, instead of p. Although, as a tradeoff, the dependency on ϵ becomes worse after regularization, we believe that such a tradeoff can be well compensated by the efficiency in handling high dimensionality at least for some applications.

Perhaps more importantly, we further consider stationary points that satisfy the significant subspace second-order necessary condition (S^3ONC) [49], which is

[1] This is because, if only we would know which dimensions are nonzero, we may equivalently reduce the problem to one that has only tens or hundreds of dimensions. Then, according to (2.1.5), the required sample size would likely be only in thousands.

weaker than the second-order KKT condition. When $\hat{\varepsilon} = 0$, we show that, if an S^{3}ONC solution is achieved by a(n arbitrary) descent local algorithm starting at an all-zero vector, then the sample size is required to be

$$n \gtrsim \frac{|\mathcal{S}|^{2.5}}{\epsilon^4} \left(\ln \frac{p}{\epsilon} \right)^2 + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}, \tag{2.1.9}$$

if $f(\,\cdot\,, W)$ is convex for almost every $W \in \mathcal{W}$. Furthermore, assume in addition that F is differentiable and strongly convex. Then a smaller sample size is allowed, that is:

$$n \gtrsim \frac{|\mathcal{S}|^{1.5}}{\epsilon^3} \left(\ln \frac{p}{\epsilon} \right)^{1.5} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}. \tag{2.1.10}$$

Both bounds are worse than (2.1.8) in terms of $|\mathcal{S}|$ and/or ϵ, but present similar levels of efficacy in addressing high dimensionality as in (2.1.8). Meanwhile, the computational overhead in solving for an S^{3}ONC solution is largely reduced compared to that in solving for a global solution.

Furthermore, it is worthwhile to mention a special case to demonstrate RSAA's efficacy. Assume again that $f(\,\cdot\,, W)$ is convex for almost every $W \in \mathcal{W}$, function F is differentiable and strongly convex, and $\hat{\varepsilon} = 0$. If all of the most contributing dimensions have a reasonably large magnitude that differentiates them from zero, that is, the value of $\min_{i \in \mathcal{S}} |x_i^{\min}|$ is above a certain threshold dependent only on $|\mathcal{S}|$ and the modulus of strong convexity, then the required sample size becomes as small as:

$$n \gtrsim \frac{|\mathcal{S}|}{\epsilon^2} \ln \frac{p}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha} \tag{2.1.11}$$

for an S^{3}ONC solution. In contrast, under the same set of assumptions, the best known bound on the performance of traditional SAA is still (2.1.5), this means that, at least for some scenarios, the proposed RSAA may achieve a non-trivial improvement to SAA in handling high dimensionality without any compromise in terms of dependencies on $|\mathcal{S}|$, ϵ, and α. A summary of comparisons between RSAA and SAA is provided in Table 2.1 given $\hat{\varepsilon} = 0$.

When the exact global solution to the SP is not sparse but can be approximated by a sparse solution, i.e., $\hat{\varepsilon} > 0$, it turns out that the sample size should grow polynomially in $\hat{\varepsilon}$ and that there can also be a residual suboptimality gap linear in

Table 2.1: A summary of sample size requirement to guarantee optimization quality of (2.1.4) when $\hat{\varepsilon} = 0$ as defined in (2.1.2). The "Global" column indicates whether the approximation formulation being solved globally (✓) or locally (×) is one of the conditions for the bounds on "n" of the same row; the "$f(\cdot, W)$ convex" and the "$\min_{i\in\mathcal{S}} \hat{x}_i^{\min} \geq$ threshold" columns indicate whether (✓) or not (×) Function $f(\cdot, W)$ being convex for a.e. $W \in \mathcal{W}$ and $\min_{i\in\mathcal{S}} \hat{x}_i^{\min}$ being above a certain threshold are conditions for the corresponding bounds on "n", respectively.

	$n \gtrsim$	Global	$f(\cdot, W)$ convex	F strongly convex & differen -tiable	$\min_{i\in\mathcal{S}} \hat{x}_i^{\min} \geq$ threshold
SAA	$\frac{p}{\epsilon^2}\ln\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}$	✓	×	×	×
RSAA	$\frac{\lvert\mathcal{S}\rvert}{\epsilon^3}\left(\ln\frac{p}{\epsilon}\right)^{1.5} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}$	✓	×	×	×
	$\frac{\lvert\mathcal{S}\rvert^{2.5}}{\epsilon^4}\left(\ln\frac{p}{\epsilon}\right)^{2} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}$	×	✓	×	×
	$\frac{\lvert\mathcal{S}\rvert^{1.5}}{\epsilon^3}\left(\ln\frac{p}{\epsilon}\right)^{1.5} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}$	×	✓	✓	×
	$\frac{\lvert\mathcal{S}\rvert}{\epsilon^2}\ln\frac{p}{\epsilon} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}$	×	✓	✓	✓

$\hat{\varepsilon}$. However, the poly-logarithmic dependency of sample size requirement on the dimensionality is maintained.

Since second-order KKT condition implies S^{3}ONC, all numerical algorithms that ensure the second-order KKT condition (e.g., [10, 19, 63, 92, 93]) guarantee S^{3}ONC. Some of these algorithms such as the interior point methods in [10] are fully polynomial-time approximation schemes (FPTAS). Meanwhile, as we will illustrate later, computing the global minimizer may also be possible via a mixed integer programming reformulation.

Regularizing the SP solution schemes with a sparsity-inducing penalty for an important class of SP formulations has been discussed by some literature, such as [3], which focuses on the computational complexity when a stochastic optimization algorithm incorporates an ℓ_1-norm penalty. To our knowledge, no theoretical analysis has been established to qualify the performance of the sparsity-inducing penalties in terms of approximating the true SP problem by the sample average approximation.

Our results may also have implications to the understanding of a flexible class

of high-dimensional sparse learning problems for M-estimation with the FCP. In fact, the SAA (2.1.3) can be considered as a formulation of an M-estimator with f representing a statistical loss function, and the SP problem (2.1.1) is the corresponding population version of the learning problem with F measuring the generalization error. Such a correspondence is also noted by [60]. Following this correspondence, the RSAA (2.1.6) is then the formulation of the sparse learning problem that incorporates the FCP as a regularizer. Our findings imply that high-dimensional M-estimation is possible through the regularization of the FCP, even if the problem dimension cannot be bounded by any polynomial function of the sample size.

While most existing literature on high-dimensional learning such as [11, 15, 29, 49, 52, 61, 87, 89, 94, 95, 96] either focuses on linear regression models or relies on additional conditions such as the (restricted) strong convexity, our analyses do not rely on those assumptions and may apply to a more general M-estimation problem. We would also like to comment that much literature has been devoted to studying an alternative regularizer, the ℓ_1-norm regularizer, or a.k.a., the Lasso. For many reported simulated experiments, numerical comparisons between Lasso and FCP have been reported by [29, 33, 49, 49, 87, 89] in supportive of relative outperformance of the latter. Some theoretical explanations of such outperformance are also provided by [29, 33, 49] in some special cases of high-dimensional learning.

The rest of this chapter is organized as following: Section 2.2 presents our assumptions and the necessary optimality conditions. Section 2.3 summarizes our major results. Proofs for those results are presented in Section 2.4. Section 2.5 discusses different approaches in solving for a desired local/global solution. Section 2.6 presents some preliminary numerical results. Finally, Section 2.7 concludes the work. Throughout the chapter we will denote by $\|\cdot\|$, $|\cdot|$, and $\|\cdot\|_{\mathbf{p}}$ $(1 \leq \mathbf{p} \leq \infty)$ for a vector the ℓ_2, ℓ_1, and $\ell_{\mathbf{p}}$ norm, while $|\cdot|$ for a finite set denotes the cardinality of the set. For any scalars x and y, we denote by $x \bigvee y$ (and by $x \bigwedge y$) the larger (smaller, resp.) number between the two. We will also use "a.s." as an abbreviation for "almost surely", and "a.e." for "almost every".

2.2 Settings and Necessary Conditions

2.2.1 Assumptions

Our analysis relies on the following assumptions.

Assumption A.

A.1 For any $\mathbf{x} \in \mathcal{X}$, the following inequality holds

$$\mathbb{E}[\exp\left(t\left[f(\mathbf{x},\, W) - F(\mathbf{x})\right]\right)] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right), \qquad \forall t \in \mathbb{R},$$

for some $\sigma > 0$.

A.2 There exists a measurable and deterministic function $L : \mathcal{W} \to \mathbb{R}$ such that

$$\mathbb{E}[\exp\left(t\left[L(W) - L_\mu\right]\right)] \leq \exp\left(\frac{\sigma_L^2 t^2}{2}\right), \quad \forall t \in \mathbb{R},$$

for some $\sigma_L > 0$ and $L_\mu := \mathbb{E}[L(W)] \geq 1$ and that

$$\sup_{\mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}} \{|f(\mathbf{x}_1,\, W) - f(\mathbf{x}_2,\, W)| - L(W)\|\mathbf{x}_1 - \mathbf{x}_2\|\} \leq 0, \quad a.e.\ W \in \mathcal{W}.$$

A.3 For almost every $W \in \mathcal{W}$, function $f(\mathbf{x},\, W)$ is twice differentiable in $\mathbf{x}$ and satisfies

$$\frac{\partial^2 f(\mathbf{x},\, W)}{(\partial x_i)^2} \leq L_{\mathcal{H}}, \quad \forall i \in \{1, ..., p\},\ \mathbf{x} = (x_i) \in \mathcal{X}$$

for some $L_{\mathcal{H}} > 0$.

A.4 Assume that $\mathcal{X}$ is defined by coordinate-wise constraints with $\mathcal{X} := \{\mathbf{x} = (x_i) : x_i \in X_i,\, i = 1, ..., p\}$ for some $X_i \subseteq \mathbb{R}_+$ for all $i = 1, ..., p$, and that there exist two hypercubes $\mathbb{H}(0, R) := \{\mathbf{x} \in \mathbb{R}_+^p : \mathbf{x} \leq R\}$, for some $R \geq 1$, and $\mathbb{H}(0, 1) := \{\mathbf{x} \in \mathbb{R}_+^p : \mathbf{x} \leq 1\}$ such that $\mathbb{H}(0, 1) \subseteq \mathcal{X} \subseteq \mathbb{H}(0, R)$.

A.5 Function $f(\,\cdot\,, W)$ is convex for almost every $W \in \mathcal{W}$.

We will also make stipulations on the choices of the penalty parameters a and λ.

Condition B. Let the penalty parameters (a, λ) of the MCP as in (2.1.7) satisfy that $a < L_{\mathcal{H}}^{-1}$, $a \leq 1$ and $\lambda > 0$.

Assumption A.1 and A.2 are essentially subgaussian. The same set of assumptions are standard for sample complexity analyses of the conventional SAA as in [77]. Meanwhile, A.3 and A.5 are verifiable regularities of the objective function. More specifically, Assumption A.3 essentially assumes that the largest eigenvalue of the Hessian matrix of the SAA formulation is bounded from above almost surely and Assumption A.5 requires that the SAA formulation is almost surely convex. Assumption A.4 requires that the constraints are component-wise rectangle constraints. In addition, it is also required that the feasible region contain an inner hypercube and is compact. For some of our theoretical results (as in Theorem 2.3.5), Assumption A.5 is not required. Condition B is non-restrictive, since the parameters a and λ are user-specified.

Under Assumption A.2, there exists another measurable and deterministic function, denoted by $L_{|\mathcal{S}|} : \mathcal{W} \to \mathbb{R}$, and a constant, denoted by $L_{\mu,s} : 1 \leq L_{\mu,s} \leq L_\mu$, such that

$$\mathbb{E}[\exp\left(t\left[L_{|\mathcal{S}|}(W) - L_{\mu,s}\right]\right)] \leq \exp\left(\frac{\sigma_L^2 t^2}{2}\right), \tag{2.2.12}$$

for all $t \in \mathbb{R}$, and that $\sup_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \cap \{\mathbf{x}: x_i = 0, j \in \mathcal{S}^c\}} \{|f(\mathbf{x}_1, W) - f(\mathbf{x}_2, W)| - L_{|\mathcal{S}|}(W)\|\mathbf{x}_1 - \mathbf{x}_2\|\} \leq 0$ for almost every $W \in \mathcal{W}$. In some cases, such as when F_n is quadratic, $L_{\mu,s}$ may be nontrivially smaller than L_μ especially if p is large.

2.2.2 Necessary Conditions for Local Optimality

We focus on local solutions to (2.1.6) that satisfy some necessary conditions for local minimality. Telling from (2.1.7), $P_\lambda(t)$ is twice differentiable in t for all $t \in [0, a\lambda)$. In the meantime, $F_n(\mathbf{x})$ is almost surely twice differentiable under Assumption A.3 for any $\mathbf{x} \in \mathcal{X}$. We consider the following necessary conditions:

First-order necessary condition (FONC): The solution $\mathbf{x}^* \in \mathcal{X}$ satisfies that

$$\langle \nabla F_n(\mathbf{x}^*) + (P'_\lambda(x_i^*) : 1 \leq i \leq p), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{2.2.13}$$

Significant subspace second-order necessary condition (S^3 ONC): The solution $\mathbf{x}^* :=$

$(x_i^* : 1 \leq i \leq p) \in \mathcal{X}$ satisfies FONC. Furthermore, for all $i \in \{i : x_i^* \in (0, \min\{1, a\lambda\})\}$, it holds that $\frac{\partial^2[F_n(\mathbf{x})+\sum_{i=1}^p P_\lambda(x_i)]}{(\partial x_i)^2}\Big|_{\mathbf{x}=\mathbf{x}^*} \geq 0$.

The S^{3}ONC is derived from the observation that a local minimal solution to the original problem must be a local minimizer in the subspace that considers only a single nonzero variable (See also [23, 49]). One may easily check that any second-order KKT point satisfies the S^{3}ONC.

2.3 Major Results

Our major results concern two propositions and four theorems. Propositions 2.3.1 and 2.3.2 provide sample size estimates for all S^{3}ONC solutions within the set $\{\mathbf{x} : F(\mathbf{x}) - F(\hat{\mathbf{x}}^{\min}) \leq \Gamma\}$ for some prescribed $\Gamma \geq 0$. Those bounds vary with different regularities on f or F. Then Theorems 2.3.5, 2.3.7, and 2.3.8 discuss some special S^{3}ONC solutions: the global solutions or the local solutions generated with some naive initialization. Finally, Theorem 2.3.9 presents the special case where the RSAA improves over the conventional SAA nearly without any compromise.

2.3.1 Sample Size Estimation for All S^{3}ONC Solutions

We will use the following short-hand notation:

$$N^*(c_1) := \frac{\sigma^2}{\epsilon^2} \ln \frac{c_1}{\alpha} + \frac{\sigma^2|\mathcal{S}|}{\epsilon^2} \ln \frac{c_1 R L_\mu p}{\epsilon} + \sigma_L^2 \cdot \ln \frac{c_1 p}{\alpha}, \tag{2.3.14}$$

where $c_1 > 0$.

Proposition 2.3.1. *Suppose that Assumptions A.1-A.3, and Condition B hold. Let $|\mathcal{S}| \geq 1$, $4p^2 \geq n$, $\lambda = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$ for arbitrary $\delta : 0 < \delta < 1/2$ and $\rho : 0 \leq \rho \leq 1/2$. Consider an S^{3}ONC solution $\mathbf{x}^*$ to (2.1.6) that satisfies $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ almost surely for some $\Gamma \geq 0$. For any $\alpha : 0 < \alpha \leq 1$, $\epsilon : 0 < \epsilon \leq 1$ and $\hat{\varepsilon} \geq 0$:*

1. *if it holds that, for some problem-independent constant $c_2 > 0$,*

$$n \geq N_1 \bigvee c_2 \cdot N^*(c_2) \tag{2.3.15}$$

where $N_1 := \sigma^2 \left(\frac{1}{\epsilon}\right)^{\frac{1}{2\delta}} |\mathcal{S}|^{\frac{1-2\rho}{2\delta}} \bigvee \sigma^2 |\mathcal{S}|^{\frac{2\rho}{1-2\delta}} \left(c_2 \frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{c_2 R L_\mu p}{\min\{\epsilon, \sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}}$, then $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon} + \Gamma$ with probability lower bounded by $1 - \alpha$;

2. *if Assumption A.5 is satisfied and it holds that, for some problem-independent constant* $c_2 > 0$,

$$n \geq N_2 \bigvee c_2 \cdot N^*(c_2), \tag{2.3.16}$$

where $N_2 := \sigma^2 \cdot |\mathcal{S}|^{\frac{1-\rho}{\delta}} \left(\frac{R}{\epsilon}\right)^{\frac{1}{\delta}} \bigvee \sigma^2 |\mathcal{S}|^{\frac{2\rho}{1-2\delta}} \cdot \left(c_2 \frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{c_2 R L_\mu p}{\min\{\epsilon, \sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}}$, *then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon}$ *with probability lower bounded by* $1 - \alpha$.

Proof. The proof is postponed till Section 2.4.2.3. □

We assume in the following proposition that F is differentiable and strongly convex with constant $\mathcal{U}_{\mathcal{H}}$ such that, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$F(\mathbf{x}_1) - F(\mathbf{x}_2) \geq \langle \nabla F(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\mathcal{U}_{\mathcal{H}}}{2} \|\mathbf{x}_1 - \mathbf{x}_2\|^2, \tag{2.3.17}$$

for some $U_{\mathcal{H}} > 0$, where $\nabla F(\mathbf{x}_2)$ is a gradient of F at $\mathbf{x}_2$. Due to the increased regularity, we may have a different sample size requirement.

Proposition 2.3.2. *Consider an* S^3*ONC solution* $\mathbf{x}^*$ *to* (2.1.6) *that satisfies* $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ *almost surely for some* $\Gamma \geq 0$. *Suppose that Assumption A and Condition B hold. Let* $4p^2 \geq n$, $|\mathcal{S}| \geq 1$ *and* $\lambda = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$ *for arbitrary* $\delta : 0 < \delta < 1/2$ *and* $\rho : 0 \leq \rho \leq 1/2$. *Assume, in addition, that* F *is differentiable and strongly convex to satisfy* (2.3.17). *For any* $\alpha : 0 < \alpha \leq 1$, $\epsilon : 0 < \epsilon \leq 1$, *and* $\hat{\varepsilon} \geq 0$, *if it holds that, for some problem-independent constant* $c_3 > 0$,

$$n \geq c_3 \cdot N^*(c_3) \bigvee N_3 \tag{2.3.18}$$

where $N_3 := \frac{\sigma^2 |\mathcal{S}|^{\frac{1-2\rho}{2\delta}}}{\mathcal{U}_{\mathcal{H}}^{\frac{1}{2\delta}}} \left[\left(\frac{c_3}{\epsilon}\right)^{\frac{1}{2\delta}} + \left(\frac{c_3\hat{\varepsilon}}{\epsilon^2}\right)^{\frac{1}{2\delta}}\right] \bigvee \frac{\sigma^2}{|\mathcal{S}|^{\frac{2\rho}{2\delta-1}}} \left(c_3 \frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{c_3 R L_\mu p}{\min\{\epsilon, \sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}}$, *then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3(\epsilon + \hat{\varepsilon})$ *with probability lower bounded by* $1 - \alpha$.

Proof. The proof is postponed till Section 2.4.2.4. □

Remark 2.3.3. *The assumption of* $4p^2 \geq n$ *can be easily relaxed but is imposed for notational simplification in our derivations. Meanwhile, it is possible that* (2.3.17) *is satisfied but* $F_n(\cdot) = \frac{1}{n}\sum_{i=1}^{j} f(\cdot, W^j)$ *is not strongly convex. For an example,*

we may consider the case of linear regression, which is often solved with the SAA in the form of the least squares problem. When $n < p$, the least squares problem may not be strongly convex, but the population version of the linear regression problem (which is the corresponding SP problem) usually have a strongly convex objective.

Remark 2.3.4. *Consider the global minimizer, denoted $\mathbf{x}^{SAA}$, to the conventional SAA formulation in* (2.1.3) *within the feasible region $\mathcal{X}$. In [77], it is shown (after some immediate conversion of notations from Theorem 5.18 therein) that to achieve an optimization accuracy of $F(\mathbf{x}^{SAA}) - F(\mathbf{x}^{\min}) \leq \epsilon$ with lower-bounded probability $1 - \alpha$, the stipulated sample size follows*

$$n \geq \frac{c_a \sigma^2}{\epsilon^2} \left[p \ln \frac{c_a L_\mu R}{\epsilon} + \ln \frac{c_a}{\alpha} \right] \bigvee \sigma_L^2 \cdot \ln \frac{c_a}{\alpha} =: N_{SAA}. \tag{2.3.19}$$

for some constants $c_a > 0$. In contrast, Propositions 2.3.1 and 2.3.2 indicate that, in nonconvex, convex, and strongly convex cases, RSAA requires the sample sizes to be at least $N_1 \bigvee c_2 N^*(c_2)$ in (2.3.15)*, $N_2 \bigvee c_2 N^*(c_2)$ in* (2.3.16)*, or $N_3 \bigvee c_3 N^*(c_3)$ in* (2.3.18)*, respectively. For all the three cases, it is easily verifiable that N^* is always dominantly better than N_{SAA} in terms of dependency, while as a tradeoff, N_1, N_2, and N_3 may become more sensitive to the reduction in ϵ than the conventional SAA. A detailed comparison will be made in the next subsection.*

2.3.2 Sample Size Estimates for Some Special S^{3}ONC Solutions

We consider, in Theorem 2.3.5, the performance of a global minimal solution $\mathbf{x}^*$, in the sense that $F_{n,\lambda}(\mathbf{x}^*) = \inf_{\mathbf{x} \in \mathcal{X}} F_{n,\lambda}(\mathbf{x})$ almost surely. Then in Theorems 2.3.7, 2.3.8, and 2.3.9, we study the S^{3}ONC solutions with a better objective value than an all-zero vector, denoted by $\mathbf{0}$. In particular, Theorem 2.3.9 identifies the best performing case for RSAA.

Recalling the definition of N^* in (2.3.14), we have the following results on the global solution.

Theorem 2.3.5. *Suppose that Assumptions A.1-A.3, and Condition B hold. Let $4p^2 \geq n$, $|\mathcal{S}| \geq 1$, and $\lambda = \frac{\sigma^{1/3}}{n^{1/6} |\mathcal{S}|^{1/4}}$. Consider a global solution $\mathbf{x}^*$ to* (2.1.6)*. For*

any $\alpha : 0 < \alpha \leq 1$, $\epsilon : 0 < \epsilon \leq 1$, *and* $\hat{\varepsilon} \geq 0$, *if*

$$n \geq \frac{c_4 \sigma^2 |\mathcal{S}|}{\epsilon^3} \left[1 + \frac{(1+\hat{\varepsilon})^{\frac{3}{2}}}{a^3} \left(\ln \frac{c_4 R L_\mu p}{\min\{\epsilon,\, \sigma^{1/3}\}} \right)^{\frac{3}{2}} \right] \bigvee c_4 \cdot N^*(c_4), \tag{2.3.20}$$

is satisfied for some problem-independent constant $c_4 > 0$, *then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon}$ *with probability lower bounded by* $1 - \alpha$.

Proof. Since the global solution is also a local minimal solution, $\mathbf{x}^*$ also satisfies the S^{3}ONC almost surely. In addition, since $F(\mathbf{x}^*) \leq F(\mathbf{x}^{\min}) \leq F(\hat{\mathbf{x}}^{\min})$, we may invoke Part 1 of Proposition 2.3.1 with $\Gamma = 0$, $\delta = \frac{1}{6}$, and $\rho = \frac{1}{3}$ to obtain the desired results. □

Remark 2.3.6. *Theorem 2.3.5 stipulates the minimal assumptions on* F_n, *but, as a tradeoff, it requires the global optimization of* (2.1.6). *Computing* (2.1.6) *globally is challenging, because the MCP is nonconvex. [35] showed that* (2.1.6) *in some special cases is strongly NP-hard. This motivates us to further consider a class of solutions that only satisfy certain necessary conditions for local minimality.*

Theorem 2.3.7. *Suppose that Assumption A and Condition B hold. Let* $4p^2 \geq n$, $|\mathcal{S}| \geq 1$, *and* $\lambda = \frac{\sigma^{1/2}}{n^{1/4}|\mathcal{S}|^{3/8}}$. *Consider an* S^3*ONC solution* $\mathbf{x}^*$ *to* (2.1.6) *that satisfies* $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\mathbf{0})$ *almost surely. For any* $\alpha : 0 < \alpha \leq \frac{1}{2}$, $\epsilon : 0 < \epsilon \leq 1$ *and* $\hat{\varepsilon} \geq 0$, *if*

$$n \geq \frac{c_5 \sigma^2 |\mathcal{S}|^{\frac{5}{2}}}{\epsilon^4} \left[R^4 + \frac{(1 + L_{\mu,s} R + \hat{\varepsilon})^2}{a^4} \left(\ln \frac{c_5 R L_\mu p}{\min\{\epsilon,\, \sigma^{1/2}\}} \right)^2 \right] \bigvee c_5 N^*(c_5) \tag{2.3.21}$$

is satisfied for some problem-independent constants $c_5 > 0$, *then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon}$ *with probability lower bounded by* $1 - 2\alpha$.

Proof. The proof is postponed till Section 2.4.2.5. □

Theorem 2.3.8. *Suppose that Assumption A and Condition B hold. Let* $4p^2 \geq n$, $|\mathcal{S}| \geq 1$, *and* $\lambda = \frac{\sigma^{1/3}}{n^{1/6}|\mathcal{S}|^{1/4}}$. *Also assume that* F *is differentiable and strongly convex as in* (2.3.17). *Consider an* S^3*ONC solution* $\mathbf{x}^*$ *to* (2.1.6) *that satisfies* $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\mathbf{0})$ *almost surely. For any* $\alpha : 0 < \alpha \leq \frac{1}{2}$, $\epsilon : 0 < \epsilon \leq 1$, *and* $\hat{\varepsilon} \geq 0$, *if*

$$n \geq \frac{c_6\sigma^2|\mathcal{S}|^{\frac{3}{2}}}{\epsilon^3}\left[\frac{1}{\mathcal{U}_{\mathcal{H}}^3}+\frac{\hat{\varepsilon}^3}{\mathcal{U}_{\mathcal{H}}^3\epsilon^3}+\frac{(1+L_{\mu,s}R+\hat{\varepsilon})^{\frac{3}{2}}}{a^3}\left(\ln\frac{c_6RL_\mu p}{\min\{\epsilon,\,\sigma^{1/3}\}}\right)^{\frac{3}{2}}\right] \bigvee c_6N^*(c_6), \quad (2.3.22)$$

is satisfied for some problem-independent constant $c_6 > 0$, then $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3(\epsilon + \hat{\varepsilon})$ with probability lower bounded by $1 - 2\alpha$.

Proof. The proof is postponed till Section 2.4.2.6. □

Theorem 2.3.9. *Consider an S^3ONC solution $\mathbf{x}^*$ to (2.1.6). Suppose that the same set of assumptions hold as in Theorem 2.3.8. Let $\lambda = \frac{1}{|\mathcal{S}|^{1/4}}$. Assume additionally $\hat{\varepsilon} = 0$ and $\min_{i\in\mathcal{S}}|\hat{x}_i^{\min}| > \frac{|\mathcal{S}|^{1/4}+\sqrt{|\mathcal{S}|^{1/2}+2\mathcal{U}_{\mathcal{H}}}}{\mathcal{U}_{\mathcal{H}}}$, where $\hat{x}_i^{\min}$ is the i-th dimension of $\hat{\mathbf{x}}^{\min}$. For any $\alpha:\, 0 < \alpha \leq \frac{1}{2}$ and $\epsilon:\, 0 < \epsilon \leq 1$, if*

$$n \geq \frac{c_7\sigma^2|\mathcal{S}|}{\epsilon^2}\left(\frac{1+L_{\mu,s}R}{a^2}\ln\frac{c_6RL_\mu p}{\epsilon}\right)\bigvee c_7N^*(c_7), \quad (2.3.23)$$

for some problem-independent constant $c_7 > 0$, then $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq \epsilon$ with probability lower bounded by $1 - 2\alpha$.

Proof. The proof is postponed till Section 2.4.2.7. □

Remark 2.3.10. *We notice that the choices of λ are different among the above theorems. At the minimum, the above theorems ensure the existence of proper λ's that ensure the sound performance of the RSAA in all the scenarios discussed above. In practice, λ can also be determined by a simple cross-validation procedure, which is a commonly adopted scheme in penalized statistical learning to tune the parameter of the sparsity-inducing penalties.*

Remark 2.3.11. *We would like to compare the sample size requirement of the RSAA as presented in the results above with that of the conventional SAA.*

- *We see that N_{SAA} as in (2.3.19) depends polynomially in the problem dimension p. In contrast, Theorems 2.3.5, 2.3.7, 2.3.8, and 2.3.9 reveal that the global solutions and some computable local solutions to RSAA require the sample size to be polynomial in $\ln p$ and $|\mathcal{S}|$. We regard it as a demonstration of the RSAA's capability in handling high dimensionality, as now exponentially increased p can be compensated by polynomially increasing n.*

- *As a tradeoff to the potential advantage mentioned above, the RSAA's performance has a worse dependency on ϵ than the conventional SAA in general. More specifically, N_{SAA} increases at a rate of $O(\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon})$. In contrast, RSAA follows a rate of $O(\frac{1}{\epsilon^3}\cdot(\ln\frac{1}{\epsilon})^{3/2})$ if minimized globally (under Assumptions A.1-A.3), or $O(\frac{1}{\epsilon^4}\cdot(\ln\frac{1}{\epsilon})^2)$ if solved locally with a naive initialization (additionally under Assumption A.5). Furthermore, under some assumption of differentiability and strong convexity, if $\hat{\varepsilon}\leq O(1)\cdot\epsilon$ for some problem-independent constant $O(1)$, then a local solution with a naive initialization retains the rate of $O(\frac{1}{\epsilon^3}\cdot(\ln\frac{1}{\epsilon})^{3/2})$, which is the same as the global minimizer. We think that compromising the dependency on ϵ to achieve a non-trivial reduction in the dependency on p can be worthwhile in many high dimensional SP applications, where p can be redundantly very large but the suboptimality gap ϵ is not required to be very small.*

- *Theorem 2.3.9 identifies a case where RSAA non-trivially reduces the dependency on p while the growth of the required sample size maintains at the same rate as the conventional SAA in terms of ϵ.*

- *The RSAA's dependencies on σ and σ_L are almost the same as those of the SAA. Meanwhile, RSAA becomes dependent on some other quantities that originally do not influence the SAA's performance: a, $|\mathcal{S}|$, and $\mathcal{U}_{\mathcal{H}}$. Moreover, in some cases, the RSAA may be more sensitive to the increase in the Lipschitz-like constant $L_{\mu,s}$ as defined in* (2.2.12) *and the radius of the feasible region, R. Nonetheless, those dependencies all maintain to be polynomial.*

Remark 2.3.12. *By allowing $\hat{\varepsilon}\geq 0$, our results apply to the cases where the exact solution to the SP is dense, but can be approximated by a sparse solution. We can see that, when $\hat{\varepsilon}>0$, RSAA will require more samples and may incur a residual suboptimality gap no greater than $O(1)\cdot\hat{\varepsilon}$.*

Remark 2.3.13. *Our results may also have potentially important implications to high-dimensional M-estimation. One may consider the following correspondence between our setting and the setting for a high-dimensional learning problem: (i) Eq.* (2.1.3) *can be thought of as an in-sample statistical loss function; (ii) the (global/local) solution to RSAA formulation* (2.1.6) *can be considered as a folded*

concave penalized sparse estimator; (iii) the SP formulation (2.1.1) *can be considered as the population version of the (unpenalized) learning problem (a.k.a., expected risk or generalization error); and (iv) The suboptimality gap* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min})$ *is then a performance measure*[2] *of the estimator* $\mathbf{x}^*$*. The above conversion is also noted by [60]. Under this conversion, we can easily tell from Theorems 2.3.5, 2.3.7, and 2.3.8 that a global solution or an* S^3*ONC solution initialized at an all-zero vector can achieve a reasonable upper bound on the* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min})$ *even in the undesirable scenarios where the dimension* p *cannot be upper bounded by any polynomial of* n*. The same setting has been discussed by [32] for the linear regression model, by [33] for several M-estimation models, and by [52, 89] under restricted strong convexity (RSC, which is some variation of strong convexity in certain subset of the feasible region). In contrast, our results may be applicable to a wider class of M-estimators without the RSC assumption. In particular, if we consider the estimator that globally minimizes the RSAA, nonconvexity in the statistical loss function is also allowed.*

Remark 2.3.14. *We would also like to remark that the sparsity of an* S^3*ONC solution is dependent on* λ *and* Γ*. The correlations between those quantities and the sparsity level are in fact characterized by Lemma 2.4.4 in the subsequent section. Although the formula seem nontrivial, we think that the general trend is clear; that is, larger* λ*, and smaller* Γ *may result in fewer nonzeros in the* S^3*ONC solution. Our numerical experiments in Section 2.6 also show that the number of nonzero dimensions can be well constrained at an* S^3*ONC solution.*

2.4 Technical Proofs

We will first present a set of preliminary results in Subsection 2.4.1 and then provide the proofs for the claimed results in Subsection 2.4.2. A sketch of proof is provided in Subsection 2.4.2.1.

[2] $F(\mathbf{x}^*) - F(\mathbf{x}^{\min})$ is also referred to as the "excess risk" in a learning problem. See for example [8].

2.4.1 Some Preliminary Results

In this subsection, we present a couple of observations that are useful to our proofs. Firstly, we observe that MCP as in (2.1.7) has the following properties:

(i) $P_\lambda(t)$ is non-decreasing and concave in $t \in \Re_+$ with $P_\lambda(0) = 0$ and $P_\lambda(t) > 0$ if $t > 0$;

(ii) $P_\lambda(t)$ is differentiable for all $t \in \Re_+$ and twice differentiable for any $t \in [0,\, a\lambda) \cup (a\lambda,\, \infty)$;

(iii) The first derivative $P'_\lambda(t) = 0$ for any $t \geq a\lambda$;

(iv) $0 \leq P'_\lambda(t) \leq \lambda$ and $0 \leq P_\lambda(t) \leq P_\lambda(a\lambda) = \frac{a\lambda^2}{2}$ for any $t \geq 0$;

(v) The second derivative $P''_\lambda(t) = -\frac{1}{a}$ for any $t \in [0,\, a\lambda)$ and $P''_\lambda(t) = 0$ for any $t > a\lambda$.

Secondly, consider an S^{3}ONC solution $\mathbf{x}^* \in \mathcal{X}$ under Assumption A.5. Recall that S^{3}ONC implies FONC. Then, from the definition of FONC in Eq. (2.2.13) and Assumption A.5, we know that, if $\mathbf{x}^*$ satisfies the FONC, then it holds that

$$F_n(\mathbf{x}^*) + \sum_{i=1}^{p} P'_\lambda(x_i^*)x_i^* \leq F_n(\mathbf{x}) + \sum_{i=1}^{p} P'_\lambda(x_i^*)x_i,\ \forall \mathbf{x} = (x_i) \in \mathcal{X},\quad a.s.,\quad (2.4.24)$$

which immediately yields that

$$F_n(\mathbf{x}^*) + \sum_{i=1}^{p} P'_\lambda(x_i^*)x_i^* \leq F_n(\hat{\mathbf{x}}^{\min}) + \sum_{i=1}^{p} P'_\lambda(x_i^*)\hat{x}_i^{\min},\quad a.s.$$

Together with (a) $\hat{x}_i^{\min} = 0$ for all $i \in \mathcal{S}^c$, (b) $\mathbf{x}^* \geq 0$, and (c) Property (iv) of P_λ, it is then straightforward to obtain:

$$\begin{aligned} F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) &\leq \sum_{i=1}^{p} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) \\ &\leq \sum_{i\in\mathcal{S}} P'_\lambda(x_i^*)|\hat{x}_i^{\min} - x_i^*| + \sum_{i\in\mathcal{S}^c} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) \\ &\overset{(a)}{=} \sum_{i\in\mathcal{S}} P'_\lambda(x_i^*)|\hat{x}_i^{\min} - x_i^*| + \sum_{i\in\mathcal{S}^c} P'_\lambda(x_i^*)\cdot(-x_i^*) \\ &\overset{(b),(c)}{\leq} \lambda \sum_{i\in\mathcal{S}} |\hat{x}_i^{\min} - x_i^*|,\quad a.s. \end{aligned} \quad (2.4.25)$$

Similarly, with (a) $\hat{x}_i^{\min} = 0$ for all $i \in \mathcal{S}^c$, (b) $\mathbf{x}^* \geq 0$, and (c) Property (iv) of P_λ, again,

$$\begin{aligned} F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) \leq & \sum_{i=1}^{p} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) \\ \leq & \sum_{i\in\mathcal{S}} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) + \sum_{i\in\mathcal{S}^c} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) \\ \overset{(a)}{=} & \sum_{i\in\mathcal{S}} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) + \sum_{i\in\mathcal{S}^c} P'_\lambda(x_i^*) \cdot (-x_i^*) \\ \overset{(b)}{\leq} & \sum_{i\in\mathcal{S}} P'_\lambda(x_i^*)(\hat{x}_i^{\min}) \overset{(c)}{\leq} \lambda \sum_{i\in\mathcal{S}} |\hat{x}_i^{\min}|, \quad a.s. \end{aligned} \tag{2.4.26}$$

Thirdly, consider an S^{3}ONC solution $\mathbf{x}^* \in \mathcal{X}$ again. One has that

$$x_i^* \notin (0,\, \min\{a\lambda,\, 1\}) \text{ for any } i = \{1, ..., p\}, \text{ almost surely.} \tag{2.4.27}$$

To see this, suppose that for an arbitrary dimension $i \in \{1, ..., p\}$, it holds that $x_i^* \in (0,\, \min\{a\lambda,\, 1\})$. Since $\frac{\partial^2 F_n(\mathbf{x})}{(\partial x_i)^2} \leq L_{\mathcal{H}}$ for all $\mathbf{x} \in \mathcal{X}$ almost surely as an immediate result of Assumption A.3, combined with $a < L_{\mathcal{H}}^{-1}$ under Condition B and Property (v) of P_λ, we have that $\frac{\partial^2 F_{n,\lambda}(\mathbf{x})}{(\partial x_i)^2}\Big|_{\mathbf{x}=\mathbf{x}^*} = \left[\frac{\partial^2 F_n(\mathbf{x})}{(\partial x_i)^2} - \frac{1}{a}\right]_{\mathbf{x}=\mathbf{x}^*} < 0$, almost surely. The satisfaction of this inequality contradicts with the S^{3}ONC, that is, for all $i = 1, ..., p$,

$$\mathbb{P}\left[\left\{\frac{\partial^2 F_n(\mathbf{x}^*)}{(\partial x_i)^2} \leq L_{\mathcal{H}}\right\} \cap \{\mathbf{x}^* \text{ satisifies S}^3\text{ONC}\} \cap \{x_i^* \in (0,\, \min\{a\lambda,\, 1\})\}\right] = 0.$$

Notice that

$$\begin{aligned} & \mathbb{P}\left[\left(\left\{\frac{\partial^2 F_n(\mathbf{x}^*)}{(\partial x_i)^2} \leq L_{\mathcal{H}}\right\} \cap \{\mathbf{x}^* \text{ satisfies S}^3\text{ONC}\}\right) \cup \{x_i^* \in (0,\, \min\{a\lambda,\, 1\})\}\right] \\ = & \mathbb{P}\left[\{x_i^* \in (0,\, \min\{a\lambda,\, 1\})\}\right] + \mathbb{P}\left[\left\{\frac{\partial^2 F_n(\mathbf{x}^*)}{(\partial x_i)^2} \leq L_{\mathcal{H}}\right\} \cap \{\mathbf{x}^* \text{ satisfies S}^3\text{ONC}\}\right] \\ & - \mathbb{P}\left[\left\{\frac{\partial^2 F_n(\mathbf{x}^*)}{(\partial x_i)^2} \leq L_{\mathcal{H}}\right\} \cap \{\mathbf{x}^* \text{ satisfies S}^3\text{ONC}\} \cap \{x_i^* \in (0,\, \min\{a\lambda,\, 1\})\}\right] \end{aligned}$$

which means that

$$1 = \mathbb{P}\left[\{x_i^* \in (0,\, \min\{a\lambda,\, 1\})\}\right] + 1 - 0, \quad \forall i = 1, ..., p$$

$$\Longrightarrow \mathbb{P}\left[\{x_i^* \in (0,\, \min\{a\lambda,\, 1\})\}\right] = 0, \quad \forall i = 1, ..., p$$
$$\Longrightarrow \mathbb{P}\left[\{x_i^* \notin (0,\, \min\{a\lambda,\, 1\}),\, \forall i = 1, ..., p\}\right] = 1$$

Combined with Properties (i) and (iii) of P_λ, it further implies that

$$\begin{aligned} P_\lambda(a\lambda)\|\mathbf{x}^*\|_0 \geq \sum_{i=1}^{p} P_\lambda(x_i^*) \geq\, & P_\lambda(\min\{a\lambda,\, 1\})\|\mathbf{x}^*\|_0 \\ = \left(\lambda \min\{a\lambda,\, 1\} - \frac{\min\{a^2\lambda^2,\, 1\}}{2a}\right)\|\mathbf{x}^*\|_0, \quad & a.s. \end{aligned} \tag{2.4.28}$$

Fourthly, the following two useful lemmas are some quick results from Assumption A.2 and are taken from [77] after some slight changes.

Lemma 2.4.1. *(a). Under Assumption A.2, for any $t > 0$,*

$$\sup_{\mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}} \left\{ \left| \sum_{j=1}^{n} f(\mathbf{x}_1, W^j)/n - \sum_{j=1}^{n} f(\mathbf{x}_2, W^j)/n \right| -\, (L_\mu + t)\|\mathbf{x}_1 - \mathbf{x}_2\| \right\} \leq 0,$$

with probability at least $1 - 2\exp\left(-\frac{nt^2}{2\sigma_L^2}\right)$.

(b). Under Assumption A.2, for any $t > 0$,

$$\sup_{\substack{\mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X} \cap \\ \{\mathbf{x}:\, x_i = 0,\, i \in \mathcal{S}^c\}}} \left\{ \left| \sum_{j=1}^{n} f(\mathbf{x}_1, W^j)/n - \sum_{j=1}^{n} f(\mathbf{x}_2, W^j)/n \right| -\, (L_{\mu,s} + t)\|\mathbf{x}_1 - \mathbf{x}_2\| \right\} \leq 0,$$

with probability at least $1 - 2\exp\left(-\frac{nt^2}{2\sigma_L^2}\right)$.

Proof. **To show (a):** Firstly, by Assumption A.2, one has $\sup_{\mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}}\{|f(\mathbf{x}_1, W^j) - f(\mathbf{x}_2, W^j)| -\, L(W^j)\|\mathbf{x}_1 - \mathbf{x}_2\|\} \leq 0$ for all $j = 1, ..., n$ almost surely. Combining the inequalities for all $j = 1, ..., n$, we obtain

$$\sup_{\mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}} \left\{ \sum_{j=1}^{n} |f(\mathbf{x}_1, W^j) - f(\mathbf{x}_2, W^j)| - \sum_{j=1}^{n} L(W^j)\|\mathbf{x}_1 - \mathbf{x}_2\| \right\} \leq\, 0, \quad a.s.$$

By triangular inequality and dividing both sides by n, we have

$$\sup_{\mathbf{x}_1,\, \mathbf{x}_2 \in \mathcal{X}} \left\{ \left| \sum_{j=1}^{n} f(\mathbf{x}_1, W^j)/n - \sum_{j=1}^{n} f(\mathbf{x}_2, W^j)/n \right| \right.$$

$$- \sum_{j=1}^{n} n^{-1} L(W^j) \|\mathbf{x}_1 - \mathbf{x}_2\| \Bigg\} \leq 0 \quad a.s.$$

By the second part of Assumption A.2, we can invoke the well-known large deviation theorem on subgaussian i.i.d. random variables and obtain

$$\mathbb{P}\left[\left| n^{-1} \sum_{j=1}^{n} L(W^j) - L_\mu \right| \geq t \right] \leq 2 \exp\left(-\frac{nt^2}{2\sigma_L^2}\right) \tag{2.4.29}$$

for any $t > 0$. Combining the above,

$$\sup_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}} \left\{ \left| \sum_{j=1}^{n} f(\mathbf{x}_1, W^j)/n - \sum_{j=1}^{n} f(\mathbf{x}_2, W^j)/n \right| - (L_\mu + t)\|\mathbf{x}_1 - \mathbf{x}_2\| \right\} \leq 0,$$

with probability at least $1 - 2\exp\left(-\frac{nt^2}{2\sigma_L^2}\right)$, as claimed.

To show (b): Under Assumption A.2, it obtains that (2.2.12) holds. Then, the same argument to prove Part (a) immediately leads to the desired result in Part (b). □

Lemma 2.4.2. *(a). Under Assumption A.2, for any fixed* $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$*, it holds that* $|F(\mathbf{x}_1) - F(\mathbf{x}_2)| \leq L_\mu \|\mathbf{x}_1 - \mathbf{x}_2\|$.

(b). Under Assumption A.2, for any fixed $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \cap \{\mathbf{x} : x_i = 0, i \in \mathcal{S}^c\}$*, it holds that* $|F(\mathbf{x}_1) - F(\mathbf{x}_2)| \leq L_{\mu,s} \|\mathbf{x}_1 - \mathbf{x}_2\|$.

Proof. **To show (a):** By Assumption A.2, we have,

$$\begin{aligned} L_\mu \|\mathbf{x}_1 - \mathbf{x}_2\| = \mathbb{E}[L(W) \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|] &\geq \mathbb{E}[|f(\mathbf{x}_1, W) - f(\mathbf{x}_2, W)|] \\ &\geq |\mathbb{E}[f(\mathbf{x}_1, W)] - \mathbb{E}[f(\mathbf{x}_2, W)]| = |F(\mathbf{x}_1) - F(\mathbf{x}_2)|, \end{aligned}$$

which is immediately the claimed result.

To show (b): Under Assumption A.2, Inequality (2.2.12) holds. Then, with the same argument to prove Part (a), we immediately obtain the desired result in Part (b). □

2.4.2 Proof of Major Results

This section presents the proofs for our claimed theoretical results. We first present a sketch of the proof in Subsection 2.4.2.1. Then, two useful lemmas that serve as the pillar of our analysis are presented in Subsection 2.4.2.2. The proofs for the aforementioned propositions and theorems as our major results are provided subsequently in Subsections from 2.4.2.3 to 2.4.2.7.

2.4.2.1 Sketch of Proof

Our proof is organized as following:

Step 1: In Lemma 2.4.3, we show how well the objective function of the SP problem F can be approximated by the objective function of the SAA problem F_n at a feasible solution that satisfies the sparsity assumption in addition to the standard assumptions for the SAA (Assumptions A.1 and A.2). More specifically, we derive a bound on the probability for the point-wise difference between $F(\mathbf{x})$ and $F_n(\mathbf{x})$ to be contained within a prescribed level $\epsilon > 0$ when $\|\mathbf{x}\|_0 \leq \tilde{p}$ for any $\tilde{p} : 1 \leq \tilde{p} \leq p$. It turns out that, if sparsity holds (i.e., if $\tilde{p}$ is small), the approximation quality is less sensitive to the problem dimension p compared to the conventional SAA by [77, 77, 78].

Step 2: To exploit the results from Step 1, Lemma 2.4.4 then shows that, once Assumption A.3 holds (i.e., the diagonals of the Hessian matrix of the SAA formulation is bounded from the above), we can guarantee that any S^3ONC solution is sparse. Furthermore, the number of nonzeros can be controlled by tuning the penalty parameters a and λ. As a result, through properly choosing the values for a and λ, we ensure that $\tilde{p}$ can indeed be a small number at the S^3ONC solution. Lemma 2.4.4 also explicates the number of nonzeros at an S^3ONC solution as a function in parameterization of a, λ, and the global suboptimality of that S^3ONC solution.

Step 3: Combining results from Steps 1 and 2, we may obtain the claimed results for Propositions 2.3.1 and 2.3.2 in Subsection 2.4.2.3 by choosing the proper pair of parameters (a, λ). The bounds derived in both propositions are in parameterization of the suboptimality gap Γ in solving the RSAA. Note that Proposition 2.3.2 makes use of additional inequalities from strong convexity and thus provides a sharper bound than Proposition 2.3.1.

Step 4: Employing bounds on the approximation quality from Propositions 2.3.1 and 2.3.2, which are in parameterization of Γ, we then consider the S^{3}ONC solutions where Γ can be explicated. In particular, we focus on two cases. (i) We first consider the global solutions where $\Gamma = 0$. By employing the propositions shown in Step 3, we can immediately derive Theorem 2.3.5 by properly choosing a and λ. (ii) Under Assumption A.5 (i.e., the unpenalized SAA formulation is convex) we then look at those solutions that have a better objective value than an all-zero solution. This immediately leads to all our results in Theorems 2.3.7-2.3.9.

2.4.2.2 Two Pillar Lemmas

This section provide two pillar lemmas that lay the foundation of our analyses and constitutes Step 1 of our proof sketch in Subsection 2.4.2.1.

Lemma 2.4.3. *Suppose that Assumptions A.1 and A.2 hold. For any scalar $t > 0$ and any integer $\tilde{p} : p \geq \tilde{p} > 0$, the following inequality holds:*

$$\mathbf{P}\left[\sup_{\mathbf{x}\in\mathcal{X}:\,\|\mathbf{x}\|_0\leq\tilde{p}}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\, W^j) - F(\mathbf{x})\right| \leq\, t\right] \geq\, 1 - 2\left\lceil\left(\frac{12\sqrt{\tilde{p}}RL_\mu}{t}\right)^{\tilde{p}}\binom{p}{\tilde{p}}\right\rceil \cdot \exp\left(-\frac{nt^2}{8\sigma^2}\right) - 2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right).$$

Proof. We can divide the feasible region $\mathcal{X}$ by a net of finitely many grids $V(t) := \{\mathbf{x}^k,\, k = 1, 2, ...\} \subseteq \mathcal{X}$, such that for any $\mathbf{x} \in \mathcal{X} \cap \{\mathbf{x} : \|\mathbf{x}\|_0 \leq \tilde{p}\}$, there always exists an $\mathbf{x}^k \in V(t)$ that satisfies $\|\mathbf{x}^k - \mathbf{x}\| \leq \frac{t}{6L_\mu}$. Since $\mathcal{X} \subseteq \mathbb{H}(0, R)$, it is easily verifiable that one can always find such a net of grids if $|V(t)| = \left\lceil(\frac{12\sqrt{\tilde{p}}RL_\mu}{t})^{\tilde{p}}\binom{p}{\tilde{p}}\right\rceil$. Corresponding to every grid $\mathbf{x}^k$, there is a subset of the feasible region $\mathcal{X}_k := \left\{\mathbf{x} \in \mathcal{X} : \|\mathbf{x} - \mathbf{x}^k\| \leq \frac{t}{6L_\mu}\right\}$. As per our construction, we know that $\mathcal{X} \cap \{\mathbf{x} : \|\mathbf{x}\|_0 \leq \tilde{p}\} = \left(\cup_{\mathbf{x}^k\in V(t)}\mathcal{X}_k\right) \cap \{\mathbf{x} : \|\mathbf{x}\|_0 \leq \tilde{p}\}$. Therefore, it holds surely that

$$\sup_{\mathbf{x}\in\mathcal{X}\cap\{\mathbf{x}:\,\|\mathbf{x}\|_0\leq\tilde{p}\}}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\, W^j) - F(\mathbf{x})\right| \leq \max_{k=1,\ldots,|V(t)|}\sup_{\mathbf{x}\in\mathcal{X}_k}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\, W^j) - F(\mathbf{x})\right| \quad (2.4.30)$$

Now, consider the following events:

$$\mathcal{E}_1(t) := \left\{ \max_{\mathbf{y}\in V(t)} \left| \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{y},\, W^j) - F(\mathbf{y}) \right| \leq\, t/2 \right\}$$

$$\mathcal{E}_2 := \left\{ \sup_{\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}} \left| \sum_{j=1}^{n} f(\mathbf{x}_1, W^j)/n - \sum_{j=1}^{n} f(\mathbf{x}_2, W^j)/n \right| -\, 2L_\mu \|\mathbf{x}_1 - \mathbf{x}_2\| \leq 0 \right\}$$

$$\begin{aligned}\mathcal{E}_3(k) := \Bigg\{ \sup_{\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}_k} \Bigg| & \sum_{j=1}^{n} f(\mathbf{x}_1, W^j)/n \\ & - \sum_{j=1}^{n} f(\mathbf{x}_2, W^j)/n \Bigg| -\, 2L_\mu \|\mathbf{x}_1 - \mathbf{x}_2\| \leq 0 \Bigg\}, \quad k = 1, ..., |V(t)|.\end{aligned}$$

It is easily verifiable that $\mathcal{E}_2 \subseteq \mathcal{E}_3(k)$ for any $k = 1, ..., |V(t)|$. Conditioning on $\mathcal{E}_2$, we have that for any $k = 1, ..., |V(t)|$:

$$\begin{aligned}& \sup_{\mathbf{x}\in\mathcal{X}_k} \left| \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\, W^j) - F(\mathbf{x}) \right| \\ \leq & \sup_{\mathbf{x}\in\mathcal{X}_k} \left| \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\, W^j) - \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\, W^j) \right| + \left| F(\mathbf{x}) - F(\mathbf{x}^k) \right| \\ & + \left| \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\, W^j) - F(\mathbf{x}^k) \right|\end{aligned}$$

$$\begin{aligned}& \overset{\mathcal{E}_2\subseteq\mathcal{E}_3(k)}{\leq} \sup_{\mathbf{x}\in\mathcal{X}_k} 2L_\mu \left\|\mathbf{x} - \mathbf{x}^k\right\| + \left| F(\mathbf{x}) - F(\mathbf{x}^k) \right| + \left| \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\, W^j) - F(\mathbf{x}^k) \right| \\ & \overset{\text{Lemma 2.4.2}}{\leq} \sup_{\mathbf{x}\in\mathcal{X}_k} 2L_\mu \left\|\mathbf{x} - \mathbf{x}^k\right\| + L_\mu \left\|\mathbf{x} - \mathbf{x}^k\right\| + \left| \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\, W^j) - F(\mathbf{x}^k) \right| \\ & = \frac{t}{2} + \left| \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\, W^j) - F(\mathbf{x}^k) \right|, \qquad a.s.\end{aligned}$$

Therefore, conditioning on the simultaneous occurrence of both $\mathcal{E}_1(t)$ and $\mathcal{E}_2$, we

have

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}\cap\{\mathbf{x}:\,\|\mathbf{x}\|_0\le\tilde{p}\}}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{x},\,W^j)-F(\mathbf{x})\right| \\
\le&\max_{k=1,\ldots,|V(t)|}\sup_{\mathbf{x}\in\mathcal{X}_k}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{x},\,W^j)-F(\mathbf{x})\right| \\
\le&\frac{t}{2}+\max_{k=1,\ldots,|V(t)|}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{x}^k,\,W^j)-F(\mathbf{x}^k)\right|\le\frac{t}{2}+\frac{t}{2}=t,\quad a.s.
\end{aligned}$$

Now it suffices to bound the probability for $\mathcal{E}_1(t)$ and $\mathcal{E}_2$.

(i). To consider $\mathcal{E}_1(t)$, we know by union bound that

$$\begin{aligned}
&\mathbf{P}\left[\max_{\mathbf{y}\in V(t)}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{y},\,W^j)-F(\mathbf{y})\right|>\frac{t}{2}\right] \\
&\qquad\qquad\le\sum_{\mathbf{y}\in V(t)}\mathbf{P}\left[\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{y},\,W^j)-F(\mathbf{y})\right|>\frac{t}{2}\right]
\end{aligned}$$

Due to Assumption A.1, we may invoke the large deviation theorem on sub-gaussian i.i.d. random variables to obtain that, for any $t>0$, it holds that $\mathbf{P}\left[\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{y},\,W^j)-F(\mathbf{y})\right|\ge t\right]\le 2\exp\left(-\frac{nt^2}{2\sigma^2}\right)$ for any $\mathbf{y}\in V(t)$. Therefore, we may continue as

$$\begin{aligned}
&\mathbf{P}[\mathcal{E}_1(t)]=\mathbf{P}\left[\max_{\mathbf{y}\in V(t)}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{y},\,W^j)-F(\mathbf{y})\right|\le t/2\right] \\
\ge&1-2|V(t)|\cdot\exp\left(-\frac{nt^2}{8\sigma^2}\right)\ge 1-2\left[\left(\frac{12\sqrt{\tilde{p}}RL_\mu}{t}\right)^{\tilde{p}}\binom{p}{\tilde{p}}\right]\exp\left(-\frac{nt^2}{8\sigma^2}\right) \quad (2.4.31)
\end{aligned}$$

(ii). To consider $\mathcal{E}_2$, we invoke Lemma 2.4.1 (in which we let $t:=L_\mu$ only within that lemma), we know that

$$\mathbf{P}\left[\mathcal{E}_2\right]\ge 1-2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \quad (2.4.32)$$

Now, invoking both the De Morgan's Law and the union bound to combine all the above, we obtain the desired result. □

Lemma 2.4.4. *Suppose that Assumptions A.1-A.3 and Condition B hold. Let* $\hat{\varepsilon} \geq 0$ *and* $\mathbf{x}^* \in \mathcal{X}$ *be an* S^3*ONC solution. For any integer* $\tilde{p} : \tilde{p} \geq |\mathcal{S}|$ *and any scalars* $t > 0$, $\hat{\varepsilon} \geq 0$, *and* $\Gamma \geq 0$, *if* $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ *almost surely,* $\frac{nt^2}{8\sigma^2} \geq \ln\left(\frac{12pRL_\mu}{t}\right)$, $a\lambda \leq 1$ *and*

$$P_\lambda(a\lambda) > \frac{\Gamma + 2t\sqrt{\tilde{p}+1} + \hat{\varepsilon}}{\tilde{p} - |\mathcal{S}| + 1}, \tag{2.4.33}$$

then $\|\mathbf{x}^*\|_0 \leq \tilde{p}$ *with probability at least*

$$\begin{aligned}\mathbf{P}^*(t,\tilde{p}) := \; & 1 - 2\exp\left(-\frac{(\tilde{p}+1)nt^2}{8\sigma^2}\right) \cdot \frac{1}{1-\exp\left(-\frac{nt^2}{8\sigma^2}\right)} \\ & - 2p\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) - 2\exp\left(-(\tilde{p}+1)\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right) \\ & \cdot \frac{1}{1-\exp\left(-\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)}\end{aligned} \tag{2.4.34}$$

Proof. If $\tilde{p} > p$, then $\|\mathbf{x}^*\|_0 \leq p < \tilde{p}$ with probability 1, while $\mathbf{P}^*(t,\tilde{p}) \leq 1$ for any $t > 0$ and $\tilde{p} \geq |\mathcal{S}|$. Thus the desired result holds if $\tilde{p} > p$. The rest of the proof then considers only the case where $\tilde{p} \leq p$.

For arbitrary integers $\tilde{p} : p \geq \tilde{p} \geq |\mathcal{S}|$ and $k : 1 \leq k \leq p - \tilde{p}$, consider the events

$$\mathcal{E}_a(\tilde{p}+k) := \{\|\mathbf{x}^*\|_0 = \tilde{p}+k\}; \quad \mathcal{E}_b := \{F_n(\hat{\mathbf{x}}^{\min}) - F_n(\mathbf{x}^*) \leq 2t\sqrt{\tilde{p}+k} + \hat{\varepsilon}\}$$

and

$$\mathcal{E}_c := \left\{\sup_{\mathbf{x}\in\mathcal{X}:\, \|\mathbf{x}\|_0 \leq \tilde{p}+k} |F_n(\mathbf{x}) - F(\mathbf{x})| \leq t\sqrt{\tilde{p}+k}\right\}.$$

Firstly, we want to show that $\mathbf{P}[\mathcal{E}_a(\tilde{p}+k) \cap \mathcal{E}_b] = 0$. To this end, consider another two events

$$\mathbb{A} := \{\forall i : x_i^* \notin (0,\, a\lambda)\}$$

$$\mathbb{B} := \left\{F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma\right\}.$$

If we recall Property (iv) of P_λ and the assumption that $a\lambda \leq 1$, it holds that

$$\left.\begin{aligned} &\forall i:\ x_i^* \notin (0,\, a\lambda) \Longrightarrow \textstyle\sum_{i=1}^{p} P_\lambda(x_i^*) = \ \|\mathbf{x}^*\|_0 P_\lambda(a\lambda) \\ &F_{n,\lambda}(\mathbf{x}^*) \leq\ F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma \end{aligned}\right\} \tag{2.4.35}$$

$$\Longrightarrow F_n(\mathbf{x}^*) + \|\mathbf{x}^*\|_0 P_\lambda(a\lambda) \leq\ F_n(\hat{\mathbf{x}}^{\min}) + |\mathcal{S}| P_\lambda(a\lambda) + \Gamma \tag{2.4.36}$$

Meanwhile,

$$\left.\begin{aligned} &(2.4.36) \\ &\|\mathbf{x}^*\|_0 = \ \tilde{p} + k \\ &F_n(\hat{\mathbf{x}}^{\min}) - F_n(\mathbf{x}^*) \leq\ 2t\sqrt{\tilde{p}+k} + \hat{\varepsilon} \end{aligned}\right\} \tag{2.4.37}$$

$$\Longrightarrow (\tilde{p} + k - |\mathcal{S}|) P_\lambda(a\lambda) \leq\ 2t\sqrt{\tilde{p}+k} + \hat{\varepsilon} + \Gamma \tag{2.4.38}$$

However, (2.4.38) contradicts with the assumed inequality (2.4.33), that is, the event $\{(2.4.38)\}$ is a sub-event of the complement of the event $\{(2.4.33)\}$. Further noticing that $\{(2.4.33)\}$ holds surely as per our assumption, therefore, $\{(2.4.38)\} = \emptyset$. Combining this with the observations that (2.4.35) $\Rightarrow$ (2.4.36), and (2.4.37) $\Rightarrow$ (2.4.38) as well as the definitions of $\mathbb{A}$, $\mathbb{B}$, $\mathcal{E}_a(\tilde{p}+k)$ and $\mathcal{E}_b$, we know that $\mathbb{A} \cap \mathbb{B} \cap \mathcal{E}_a(\tilde{p}+k) \cap \mathcal{E}_b = \emptyset$. Since $\mathbf{P}(\mathbb{A} \cap \mathbb{B}) = 1$ by assumption and by (2.4.28) with $a\lambda \leq 1$, it therefore obtains that

$$\begin{aligned} 1 =& \mathbf{P}[(\mathbb{A} \cap \mathbb{B}) \cup (\mathcal{E}_a(\tilde{p}+k) \cap \mathcal{E}_b)] \\ =& \mathbf{P}[\mathbb{A} \cap \mathbb{B}] + \mathbf{P}[\mathcal{E}_a(\tilde{p}+k) \cap \mathcal{E}_b] - \mathbf{P}[\mathbb{A} \cap \mathbb{B} \cap \mathcal{E}_a(\tilde{p}+k) \cap \mathcal{E}_b] \\ =& 1 + \mathbf{P}[\mathcal{E}_a(\tilde{p}+k) \cap \mathcal{E}_b] + 0 \\ \Longrightarrow\ & \mathbf{P}[\mathcal{E}_a(\tilde{p}+k) \cap \mathcal{E}_b] = 0. \end{aligned} \tag{2.4.39}$$

Secondly, we want to show that $\mathbf{P}[\bar{\mathcal{E}}_c] \geq\ \mathbf{P}[\mathcal{E}_c(\tilde{p}+k)]$, where $\bar{\mathcal{E}}_c$ is the complement of $\mathcal{E}_c$. To this end, consider one more event $\mathbb{C} := \{F(\mathbf{x}^{\min}) \leq F(\mathbf{x}^*)\}$, which satisfies that $\mathbf{P}[\mathbb{C}] = 1$ by the definition of $\mathbf{x}^{\min}$. We observe that, since $\|\hat{\mathbf{x}}^{\min}\|_0 = |\mathcal{S}|$,

$$\left.\begin{aligned} &\sup_{\mathbf{x} \in \mathcal{X}:\, \|\mathbf{x}\|_0 \leq \tilde{p}+k} |F_n(\mathbf{x}) - F(\mathbf{x})| \leq\ t\sqrt{\tilde{p}+k} \\ &F(\mathbf{x}^{\min}) \leq F(\mathbf{x}^*) \\ &\|\mathbf{x}^*\|_0 = \ \tilde{p} + k \end{aligned}\right\}$$

$$\implies \begin{cases} -F_n(\mathbf{x}^*) \leq -F(\mathbf{x}^*) + t\sqrt{\tilde{p}+k} \\ F_n(\hat{\mathbf{x}}^{\min}) \leq F(\hat{\mathbf{x}}^{\min}) + t\sqrt{\tilde{p}+k} \leq F(\mathbf{x}^{\min}) + t\sqrt{\tilde{p}+k} + \hat{\varepsilon} \\ \|\mathbf{x}^*\|_0 = \tilde{p}+k \\ F(\mathbf{x}^{\min}) \leq F(\mathbf{x}^*) \end{cases}$$

which immediately leads to the simultaneous satisfaction of both $F_n(\hat{\mathbf{x}}^{\min}) - F_n(\mathbf{x}^*) \leq 2t\sqrt{\tilde{p}+k}+\hat{\varepsilon}$ and $\|\mathbf{x}^*\|_0 = \tilde{p}+k$. Therefore, $\mathbb{C}\cap\mathcal{E}_c\cap\mathcal{E}_a(\tilde{p}+k) \subseteq \mathcal{E}_b\cap\mathcal{E}_a(\tilde{p}+k)$ and thus $\mathbf{P}[\mathbb{C}\cap\mathcal{E}_c\cap\mathcal{E}_a(\tilde{p}+k)] \leq \mathbf{P}[\mathcal{E}_b\cap\mathcal{E}_a(\tilde{p}+k)]$. Since we have shown above that $\mathbf{P}[\mathcal{E}_b\cap\mathcal{E}_a(\tilde{p}+k)] = 0$, we know that $\mathbf{P}[\mathbb{C}\cap\mathcal{E}_c\cap\mathcal{E}_a(\tilde{p}+k)] = 0$. Further recall that we have also known that $\mathbf{P}(\mathbb{C}) = 1$. Therefore, by both the De Morgan's Law and the union bound, under the assumption of (2.4.33),

$$0 \geq 1 - \mathbf{P}[\bar{\mathcal{E}}_a(\tilde{p}+k)] - \mathbf{P}[\bar{\mathcal{E}}_c] - (1-\mathbf{P}(\mathbb{C})) \Longrightarrow \mathbf{P}[\bar{\mathcal{E}}_c] \geq \mathbf{P}[\mathcal{E}_a(\tilde{p}+k)], \quad (2.4.40)$$

where $\bar{\mathcal{E}}_a(\tilde{p}+k)$ and $\bar{\mathcal{E}}_c$ are complements of $\mathcal{E}_a(\tilde{p}+k)$ and $\mathcal{E}_c$.

Lastly, using the upper bound on $\mathbf{P}[\bar{\mathcal{E}}_c]$ provided by Lemma 2.4.3, we obtain

$$\begin{aligned} &\mathbf{P}[\mathcal{E}_a(\tilde{p}+k)] \\ &\leq 2\left[\left(\frac{12RL_\mu\sqrt{\tilde{p}+k}}{t\sqrt{\tilde{p}+k}}\right)^{\tilde{p}+k}\binom{p}{\tilde{p}+k}\right]\cdot\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}\right) + 2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\ &\leq 2\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2} + (\tilde{p}+k)\ln\left(\frac{12RL_\mu}{t}\right) + (\tilde{p}+k)\cdot\ln p\right) \\ &\quad + 2\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}\right) + 2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \qquad (2.4.41) \\ &= 2\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2} + (\tilde{p}+k)\ln\left(\frac{12pRL_\mu}{t}\right)\right) \\ &\quad + 2\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}\right) + 2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right). \qquad (2.4.42) \end{aligned}$$

To get (2.4.41) we make use of the facts that $\binom{p}{\tilde{p}+k} \leq p^{\tilde{p}+k}$ and that $\lceil x \rceil \leq x+1$ for any $x \geq 0$.

Notice that if $\|\mathbf{x}^*\|_0 > \tilde{p}$, it must hold that $\|\mathbf{x}^*\|_0 \in \{\tilde{p}+1, ..., p\}$ and that by

the union bound:

$$\mathbf{P}[\{\|\mathbf{x}^*\|_0 \in \{\tilde{p}+1,\, ...,\, p\}\}] \leq \sum_{k=1}^{p-\tilde{p}} \mathbf{P}[\{\|\mathbf{x}^*\|_0 = \tilde{p}+k\}]. \tag{2.4.43}$$

We therefore can find an upper bound to $\mathbf{P}[\{\|\mathbf{x}^*\|_0 \in \{\tilde{p}+1,\, ...,\, p\}\}]$ by invoking (2.4.42). That upper bound writes as

$$\begin{aligned}
&\mathbf{P}[\{\|\mathbf{x}^*\|_0 \in \{\tilde{p}+1,\, ...,\, p\}\}] \leq \sum_{k=1}^{p-\tilde{p}} [\mathcal{E}_a(\tilde{p}+k)] \\
\leq &\sum_{k=1}^{p-\tilde{p}} 2\exp\left(-\frac{(\tilde{p}+k)nt^2}{8\sigma^2} + (\tilde{p}+k)\ln\left(\frac{12pRL_\mu}{t}\right)\right) \\
&+ 2\sum_{k=1}^{p-\tilde{p}} \exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}\right) + 2(p-\tilde{p})\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
= &\, 2\exp\left(-(\tilde{p}+1)\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right) \\
&\cdot \frac{1-\exp\left(-(p-\tilde{p})\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)}{1-\exp\left(-\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)} + 2(p-\tilde{p})\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
&+ 2\exp\left(-\frac{(\tilde{p}+1)nt^2}{8\sigma^2}\right) \cdot \frac{1-\exp\left(-\frac{(p-\tilde{p})nt^2}{8\sigma^2}\right)}{1-\exp\left(-\frac{nt^2}{8\sigma^2}\right)} \qquad (2.4.44) \\
\leq &1 - \mathbf{P}^*(t,\tilde{p}), \qquad (2.4.45)
\end{aligned}$$

where to achieve (2.4.44) we invoke the sum of a geometric series and to obtain (2.4.45) we make use of the assumptions that $\frac{nt^2}{8\sigma^2} \geq \ln\left(\frac{12pRL_\mu}{t}\right)$ and $\tilde{p} \leq p$. The desired result then follows immediately. □

2.4.2.3 Proof of Proposition 2.3.1

For an arbitrary $\epsilon:\ 0 < \epsilon \leq 1$, denote that

$$\mathcal{E}_A := \left\{|F(\mathbf{x}^*) - F_n(\mathbf{x}^*)| \leq \frac{\epsilon}{2}\right\};\quad \mathcal{E}_B := \left\{\left|F(\hat{\mathbf{x}}^{\min}) - F_n(\hat{\mathbf{x}}^{\min})\right| \leq \frac{\epsilon}{2}\right\}. \tag{2.4.46}$$

We examine the two parts of the proposition:

(i). For Part 1, according to (2.3.15), $0 < \epsilon \leq 1$, and $|\mathcal{S}| \geq 1$, as well as $a \leq 1$, we obtain $n \geq N_1 \geq \sigma^2 = (\frac{\sigma^{2\delta}}{1})^{\frac{1}{\delta}} \geq (\frac{a\sigma^{2\delta}}{1})^{\frac{1}{\delta}}$. Combined with $0 \leq \rho \leq \frac{1}{2}$, we

know that $a\lambda = \frac{a\sigma^{2\delta}}{n^{\delta}|\mathcal{S}|^{\rho}} \leq 1$. Conditioning on the event $\mathcal{E}_A \cap \mathcal{E}_B$, under the assumption that $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ almost surely, it holds almost surely that

$$\begin{aligned} F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) - \epsilon - \hat{\varepsilon} \leq F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^{\min}) - \epsilon &\leq F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) \\ &\leq |\mathcal{S}| \cdot P_\lambda(a\lambda) + \Gamma = |\mathcal{S}| \cdot \frac{a\lambda^2}{2} + \Gamma \\ &= \frac{a\sigma^{4\delta}}{2n^{2\delta}} |\mathcal{S}|^{1-2\rho} + \Gamma. \end{aligned} \tag{2.4.47}$$

Since $a \leq 1$, if $n \geq N_1 \geq \sigma^2 \left(\frac{1}{\epsilon}\right)^{\frac{1}{2\delta}} |\mathcal{S}|^{\frac{1-2\rho}{2\delta}} \geq \sigma^2 \left(\frac{a}{\epsilon}\right)^{\frac{1}{2\delta}} |\mathcal{S}|^{\frac{1-2\rho}{2\delta}}$, then (2.4.47) implies that $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon} + \Gamma$. Therefore, to show the first part of the proposition, it suffices to prove that there exists a problem-independent constant $c_2 > 0$ such that, if $n \geq N_1 \bigvee c_2 N^*(c_2)$ as in (2.3.15), then the event $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least $1 - \alpha$, which will be shown soon afterwards.

(ii). For Part 2, according to (2.3.16), $0 < \epsilon \leq 1$, and $R \geq 1$, combined with $|\mathcal{S}| \geq 1$ and $0 \leq \rho \leq \frac{1}{2}$, we know that $n \geq N_2 \geq \sigma^2 \geq \left(\frac{a\sigma^{2\delta}}{1}\right)^{\frac{1}{\delta}} \Longrightarrow a\lambda = \frac{a\sigma^{2\delta}}{n^{\delta}|\mathcal{S}|^{\rho}} \leq 1$. Conditioning on the event $\mathcal{E}_A \cap \mathcal{E}_B$, under Assumption A.5, we obtain from (2.4.26) that

$$\begin{aligned} F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) - \epsilon - \hat{\varepsilon} \leq F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^{\min}) - \epsilon &\leq F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) \\ &\leq \lambda |\hat{\mathbf{x}}^{\min}| = \frac{\sigma^{2\delta}}{n^{\delta}} |\mathcal{S}|^{1-\rho} R \end{aligned} \tag{2.4.48}$$

Hence, if $n \geq N_2 \geq \left(\frac{|\mathcal{S}|^{1-\rho} R \sigma^{2\delta}}{\epsilon}\right)^{\frac{1}{\delta}}$, then (2.4.48) implies that $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon}$. Therefore, to show the second part of the proposition, it also suffices to show that there exists a problem-independent constant $c_2 > 0$ such that, if $n \geq N_2 \bigvee c_2 N^*(c_2)$ as in (2.3.16), then the event $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least $1 - \alpha$.

The following provides probability lower bound for the occurrence of $\mathcal{E}_A \cap \mathcal{E}_B$. Such a bound applies to both (i) and (ii) above.

We have shown above that $a\lambda \leq 1$ for both (i) and (ii), and we also have let Assumptions A.1-A.3 and Condition B hold. Under the assumption that

$F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ almost surely, we may invoke Lemma 2.4.4, where we assume for now that

$$\frac{nt^2}{8\sigma^2} \geq \ln\left(\frac{12pRL_\mu}{t}\right) \tag{2.4.49}$$

which will be shown soon afterwards. It then follows that, for any integer $\tilde{p} \geq |\mathcal{S}|$ such that $\tilde{p} > |\mathcal{S}| + \frac{2t\sqrt{\tilde{p}+1}+\Gamma+\hat{\varepsilon}}{P_\lambda(a\lambda)} - 1 \Longleftrightarrow \sqrt{\tilde{p}+1} > \frac{t}{P_\lambda(a\lambda)} + \sqrt{\frac{t^2}{[P_\lambda(a\lambda)]^2} + |\mathcal{S}| + \frac{\Gamma+\hat{\varepsilon}}{P_\lambda(a\lambda)}}$, it holds that $\|\mathbf{x}^*\|_0 \leq \tilde{p}$ with probability at least $\mathbf{P}^*(t,\tilde{p})$ as defined in (2.4.34). Further notice that, since $\|\hat{\mathbf{x}}^{\min}\|_0 = |\mathcal{S}|$, for any $\tilde{p} \geq |\mathcal{S}|$ it holds that $\mathcal{E}_A \cap \mathcal{E}_B \supseteq \left\{\sup_{\mathbf{x}\in\mathcal{X}:\, \|\mathbf{x}\|_0 \leq \tilde{p} \bigwedge p} \left|\frac{1}{n}\sum_{j=1}^n f(\mathbf{x}, W^j) - F(\mathbf{x})\right| \leq \epsilon/2\right\} \cap \{\|\mathbf{x}^*\|_0 \leq \tilde{p}\}$. Hence we may combine Lemma 2.4.3 (in which we let $t = \frac{\epsilon}{2}$ and rescale $\tilde{p}$ only within that lemma into $p \bigwedge \tilde{p}$), and Lemma 2.4.4 (in which we let $\tilde{p} = \left\lfloor \frac{4t^2}{[P_\lambda(a\lambda)]^2} + 4|\mathcal{S}| + \frac{4(\Gamma+\hat{\varepsilon})}{P_\lambda(a\lambda)} \right\rfloor$ here and we will also let $t = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$ soon afterwards) through both the De Morgan's Law and the union bound to obtain that $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least

$$\begin{aligned}
\mathcal{P}^* :=\ & \left[\mathbf{P}^*\left(t,\tilde{p}\right) - 2\left\lceil\left(\frac{24\sqrt{\tilde{p}}RL_\mu}{\epsilon}\right)^{(p\bigwedge\tilde{p})}\binom{p}{p\bigwedge\tilde{p}}\right\rceil\right. \\
&\left.\cdot\exp\left(-\frac{n\epsilon^2}{32\sigma^2}\right) - 2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right)\right]_{\tilde{p}=\left\lfloor \frac{16t^2}{a^2\lambda^4}+\frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2}+4|\mathcal{S}|\right\rfloor} \\
\geq\ & 1 - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2}\right) - 2(p+1)\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
& - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} + \left[p\bigwedge\left\lfloor\frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}|\right\rfloor\right]\ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right) \\
& - \frac{2\exp\left(-\left[\left\lfloor\frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}|\right\rfloor + 1\right]\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)}{1 - \exp\left(-\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)} \\
& - 2\exp\left(-\frac{\left(\left\lfloor\frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}|\right\rfloor + 1\right)nt^2}{8\sigma^2}\right)\cdot\frac{1}{1-\exp\left(-\frac{nt^2}{8\sigma^2}\right)},
\end{aligned} \tag{2.4.50}$$

where we may plug in $t = \frac{\sigma^{2\delta}}{n^\delta|\mathcal{S}|^\rho}$ in the next.

Now we want to show the satisfaction of (2.4.49). Observe that, with $t = \lambda =$

$\frac{\sigma^{2\delta}}{n^{\delta}|\mathcal{S}|^{\rho}}$, $\delta < \frac{1}{2}$, $\rho \leq \frac{1}{2}$, $4p^2 \geq n$ and $p \geq |\mathcal{S}| \geq 1$, we know that

$$\begin{aligned}
&\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right) \\
&= \frac{n^{1-2\delta}}{8\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} - \ln\left(\frac{12n^{\delta}|\mathcal{S}|^{\rho}pRL_\mu}{\sigma^{2\delta}}\right) \\
&\geq \frac{n^{1-2\delta}}{8\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} - \ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}}\right) \\
&= \frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} + \frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} - \ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}}\right) \qquad (2.4.51)
\end{aligned}$$

Observe that, if $n \geq \left[12\sigma^{2-4\delta}|\mathcal{S}|^{2\rho} \bigvee 16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}\ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}}\right)\right]^{1/(1-2\delta)}$, then $\frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} \geq \ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}}\right) \bigvee \frac{12}{16}$. Therefore, we know that (2.4.51) $\geq \frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} \geq \frac{12}{16} \geq \ln 2$. This inequality implies (2.4.49).

The above provides a lower bound on the probability for the event of interest. The rest of the proof seeks to simplify this bound. We have shown above that (2.4.51) $\geq \frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} \geq \ln 2$. This inequality implies both $\exp(-\frac{nt^2}{8\sigma^2}) \leq 1/2$ and

$$\exp\left(-\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right) \leq \frac{1}{2}.$$

Further observing $\frac{t^2}{\lambda^4} = \frac{n^{2\delta}|\mathcal{S}|^{2\rho}}{\sigma^{4\delta}}$, we may combine the above with (2.4.50) to obtain

$$\begin{aligned}
\mathcal{P}^* \geq\ & 1 - 2\exp\left(-\frac{16t^2}{a^2\lambda^4}\cdot\left[\frac{nt^2}{16\sigma^2} + \frac{nt^2}{16\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right) \\
& \cdot \frac{1}{1 - \exp\left(-\left[\frac{nt^2}{16\sigma^2} + \frac{nt^2}{16\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)} \\
& - 2\exp\left(-\frac{\left(\left\lfloor\frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}|\right\rfloor + 1\right)nt^2}{8\sigma^2}\right)\cdot\frac{1}{1-\exp\left(-\frac{nt^2}{8\sigma^2}\right)} \\
& - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} + \left\lfloor\frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}|\right\rfloor\ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right) \\
& - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2}\right) - 2(p+1)\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
\geq\ & 1 - 2\exp\left(-\frac{16t^2}{a^2\lambda^4}\cdot\frac{nt^2}{16\sigma^2}\right)\cdot\frac{1}{1-\frac{1}{2}} - 2\exp\left(-\frac{16t^2}{a^2\lambda^4}\cdot\frac{nt^2}{8\sigma^2}\right)\cdot\frac{1}{1-\frac{1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&- 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} + \left\lfloor \frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}| \right\rfloor \ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right) \\
&- 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2}\right) - 2(p+1)\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
\geq & 1 - 8\exp\left(-\frac{n}{a^2\sigma^2}\right) - 2(p+1)\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2}\right) \\
&- 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} + \left\lfloor \frac{16n^{2\delta}|\mathcal{S}|^{2\rho}}{a^2\sigma^{4\delta}} + \frac{8(\Gamma+\hat{\varepsilon})|\mathcal{S}|^{2\rho}n^{2\delta}}{a\sigma^{4\delta}} + 4|\mathcal{S}| \right\rfloor \ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right)
\end{aligned}
$$

Combined with the above, it is easily verifiable that, if n is large enough to satisfy both $n \geq \sigma^2 \left[12|\mathcal{S}|^{2\rho} \bigvee 16|\mathcal{S}|^{2\rho} \ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}}\right)\right]^{1/(1-2\delta)}$ and

$$
\begin{aligned}
n \geq a^2\sigma^2 \ln\frac{32}{\alpha} + \frac{2\sigma_L^2}{L_\mu^2}\ln\left(\frac{8(p+1)}{\alpha}\right) + \frac{\sigma^2}{\epsilon^2}\left(64\cdot\ln\frac{8}{\alpha} + 256\cdot|\mathcal{S}|\ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right) \\
\bigvee \sigma^2\left[\frac{64}{\epsilon^2}\left(\frac{16|\mathcal{S}|^{2\rho}}{a^2} + \frac{8(\Gamma+\hat{\varepsilon})|\mathcal{S}|^{2\rho}}{a}\right)\ln\frac{24RL_\mu p^{5/2}}{\epsilon}\right]^{\frac{1}{1-2\delta}},
\end{aligned}
$$

then $\mathcal{P}^* \geq 1-\alpha$. Therefore, recalling that $a \leq 1$, $L_\mu \geq 1$, $p \geq |\mathcal{S}| \geq 1$ and $\epsilon \leq 1$, there exists a problem-independent constant $c_2 > 0$ such that the above stipulation of n is satisfied if

$$
n \;\geq\; c_2^{\frac{1}{1-2\delta}}\sigma^2|\mathcal{S}|^{\frac{2\rho}{1-2\delta}} \;\cdot\; \left(\frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2}\ln\frac{24RL_\mu p}{\min\{\epsilon,\sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}} \bigvee c_2 \;\cdot\; N^*(c_2). \quad (2.4.52)
$$

Combining the above with (i) Eq. (2.4.47) and (ii) Eq. (2.4.48) yields the desired results for part 1 and part 2 of the proposition, respectively. □

2.4.2.4 Proof of Proposition 2.3.2

For an arbitrary $\epsilon:\ 0 < \epsilon \leq 1$, let us consider the events that

$$
\mathcal{E}_A := \left\{|F(\mathbf{x}^*) - F_n(\mathbf{x}^*)| \leq \frac{\epsilon}{2}\right\};\ \ \mathcal{E}_B := \left\{\left|F(\hat{\mathbf{x}}^{\min}) - F_n(\hat{\mathbf{x}}^{\min})\right| \leq \frac{\epsilon}{2}\right\} \quad (2.4.53)
$$

Conditioning on the event $\mathcal{E}_A \cap \mathcal{E}_B$, under Assumption A.5, we obtain from (2.4.25) that, almost surely,

$$
F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) - \epsilon - \hat{\varepsilon} \leq F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^{\min}) - \epsilon
$$

$$
\begin{aligned}
&\leq F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) \leq \lambda \sum_{i\in\mathcal{S}} |\hat{x}_i^{\min} - x_i^*| \\
&= \lambda\sqrt{|\mathcal{S}|}\sqrt{\sum_{i\in\mathcal{S}} \|\hat{x}_i^{\min} - x_i^*\|^2}
\end{aligned} \tag{2.4.54}
$$

Further invoking (2.3.17), which immediately leads to $F(\mathbf{x}) - F(\mathbf{x}^{\min}) \geq \frac{\mathcal{U}_{\mathcal{H}}}{2}\|\mathbf{x} - \mathbf{x}^{\min}\|^2$ for all $\mathbf{x} \in \mathcal{X}$, we may continue the above as, almost surely (conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$),

$$
\begin{aligned}
&F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) - \epsilon - \hat{\varepsilon} \\
\leq &\lambda\sqrt{|\mathcal{S}|}\sqrt{\sum_{i\in\mathcal{S}} \|\hat{x}_i^{\min} - x_i^*\|^2} \leq \lambda\sqrt{|\mathcal{S}|}\cdot\|\mathbf{x}^* - \hat{\mathbf{x}}^{\min}\| \\
\leq &\lambda\sqrt{|\mathcal{S}|}\cdot\|\mathbf{x}^* - \mathbf{x}^{\min}\| + \lambda\sqrt{|\mathcal{S}|}\cdot\|\hat{\mathbf{x}}^{\min} - \mathbf{x}^{\min}\| \\
\leq &\lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2}{\mathcal{U}_{\mathcal{H}}}(F(\mathbf{x}^*) - F(\mathbf{x}^{\min}))} + \lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2}{\mathcal{U}_{\mathcal{H}}}(F(\hat{\mathbf{x}}^{\min}) - F(\mathbf{x}^{\min}))} \\
\leq &\lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2}{\mathcal{U}_{\mathcal{H}}}(F(\mathbf{x}^*) - F(\mathbf{x}^{\min}))} + \lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}}}.
\end{aligned}
$$

Solving the inequality for $\sqrt{F(\mathbf{x}^*) - F(\mathbf{x}^{\min})}$, we have, almost surely (conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$),

$$
\sqrt{F(\mathbf{x}^*) - F(\mathbf{x}^{\min})} \leq \frac{\lambda\sqrt{\frac{2|\mathcal{S}|}{\mathcal{U}_{\mathcal{H}}}} + \sqrt{\frac{2\lambda^2|\mathcal{S}|}{\mathcal{U}_{\mathcal{H}}} + 4(\hat{\varepsilon} + \epsilon) + 4\lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}}}}}{2} \tag{2.4.55}
$$

Therefore, combined with $\lambda = \frac{\sigma^{2\delta}}{n^{\delta}|\mathcal{S}|^{\rho}}$, we know that

$$
F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq \left(\sqrt{\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}}n^{2\delta}}} + \sqrt{\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}}n^{2\delta}} + \sqrt{\frac{2\sigma^{4\delta}\hat{\varepsilon}|\mathcal{S}|^{1-2\rho}}{n^{2\delta}\mathcal{U}_{\mathcal{H}}}} + (\hat{\varepsilon} + \epsilon)}\right)^2
$$

almost surely (conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$).

Notice that if $n \geq \sigma^2\left(\frac{8|\mathcal{S}|^{1-2\rho}}{\mathcal{U}_{\mathcal{H}}\epsilon}\right)^{\frac{1}{2\delta}} \bigvee \sigma^2\left(\frac{8\hat{\varepsilon}|\mathcal{S}|^{1-2\rho}}{\mathcal{U}_{\mathcal{H}}\epsilon^2}\right)^{\frac{1}{2\delta}} \bigvee \sigma^2$, then the following three inequalities hold: (a) $a\lambda = a\frac{\sigma^{2\delta}}{n^{\delta}|\mathcal{S}|^{\rho}} \leq 1$; (b) $\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}}n^{2\delta}} \leq \frac{\epsilon}{16}$; and (c) $\sqrt{\frac{2\sigma^{4\delta}\hat{\varepsilon}|\mathcal{S}|^{1-2\rho}}{n^{2\delta}\mathcal{U}_{\mathcal{H}}}} \leq \frac{\varepsilon}{2}$.

Thus,

$$\begin{aligned}
&\left(\sqrt{\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}}n^{2\delta}}}+\sqrt{\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}}n^{2\delta}}+\sqrt{\frac{2\sigma^{4\delta}\hat{\varepsilon}|\mathcal{S}|^{1-2\rho}}{n^{2\delta}\mathcal{U}_{\mathcal{H}}}}+(\hat{\varepsilon}+\epsilon)}\right)^2\\
\leq&\left(\frac{\sqrt{\epsilon}}{4}+\sqrt{\frac{25\epsilon}{16}+\hat{\varepsilon}}\right)^2=\frac{26\epsilon}{16}+\hat{\varepsilon}+\sqrt{\frac{25\epsilon^2}{64}+\frac{\epsilon\hat{\varepsilon}}{4}}\\
\leq&\frac{26\epsilon}{16}+\hat{\varepsilon}+\sqrt{\frac{25\epsilon^2}{64}+\frac{\epsilon\hat{\varepsilon}}{4}+\frac{\hat{\varepsilon}^2}{25}}\\
=&\left(\frac{26}{16}+\frac{5}{8}\right)\epsilon+\left(1+\frac{1}{5}\right)\hat{\varepsilon}=\frac{9}{4}\epsilon+\frac{6}{5}\hat{\varepsilon}
\end{aligned}\tag{2.4.56}$$

Hence, if $n\geq\sigma^2|\mathcal{S}|^{\frac{1-2\rho}{2\delta}}\left[\left(\frac{8}{\mathcal{U}_{\mathcal{H}}\epsilon}\right)^{\frac{1}{2\delta}}+\left(\frac{8\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}\epsilon^2}\right)^{\frac{1}{2\delta}}\right]\bigvee\sigma^2$, then (2.4.55) implies that $F(\mathbf{x}^*)-F(\mathbf{x}^{\min})\leq 3\epsilon+3\hat{\varepsilon}$ almost surely (conditioning on $\mathcal{E}_A\cap\mathcal{E}_B$). Therefore, to achieve the desired result of the proposition, it suffices to show that, if n additionally satisfies

$$n\geq c_3^{\frac{1}{1-2\delta}}\cdot\sigma^2|\mathcal{S}|^{\frac{2\rho}{1-2\delta}}\cdot\left(\frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2}\ln\frac{24RL_\mu p}{\min\{\epsilon,\,\sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}}\bigvee c_3\cdot N^*(c_3)$$

for some universal constant $c_3>0$, then the event $\mathcal{E}_A\cap\mathcal{E}_B$ occurs with probability at least $1-\alpha$, which can be shown by the same argument as in the proof for Proposition 2.3.1 as in Section 2.4.2.3 in showing (2.4.52). Further noticing that we can let $c_3\geq 2$ to further satisfy that $c_3N^*(c_3)\geq\frac{2\sigma^2}{\epsilon^2}\ln\frac{2}{\alpha}\geq\sigma^2$ (since $\alpha\leq 1$ and $\epsilon\leq 1$), we then have the desired result. □

2.4.2.5 Proof of Theorem 2.3.7

We first want to show that, if $\lambda=\frac{\sigma^{2\delta}}{n^\delta|\mathcal{S}|^\rho}$, then $F_{n,\lambda}(\mathbf{0})-F_{n,\lambda}(\hat{\mathbf{x}}^{\min})\leq 2L_\mu R\sqrt{|\mathcal{S}|}$ with a lower bounded probability. To this end, we observe that $\|\mathbf{0}-\hat{\mathbf{x}}^{\min}\|=\|\hat{\mathbf{x}}^{\min}\|\leq R\sqrt{|\mathcal{S}|}$. This combined with Lemma 2.4.1 (where we let $t=L_{\mu,s}$ in that lemma) in Section 2.4.1, we know that

$$|F_n(\mathbf{0})-F_n(\hat{\mathbf{x}}^{\min})|\leq 2L_{\mu,s}R\sqrt{|\mathcal{S}|},\tag{2.4.57}$$

with probability at least $1 - 2\exp(-\frac{n(L_{\mu,s})^2}{2\sigma_L^2})$. Furthermore, since $F_n(\mathbf{0}) = F_{n,\lambda}(\mathbf{0})$ and $F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) = F_n(\hat{\mathbf{x}}^{\min}) + \sum_{i=1}^p P_\lambda(\hat{x}_i^{\min})$, we have that

$$\begin{aligned} F_{n,\lambda}(\mathbf{0}) - F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) &= F_n(\mathbf{0}) - F_n(\hat{\mathbf{x}}^{\min}) - \sum_{i=1}^p P_\lambda(\hat{x}_i^{\min}) \\ &\leq F_n(\mathbf{0}) - F_n(\hat{\mathbf{x}}^{\min}) \leq 2L_{\mu,s}R\sqrt{|\mathcal{S}|} \end{aligned} \tag{2.4.58}$$

with a lower bounded probability $1 - 2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2})$.

Then, we may invoke both the De Morgan's Law and the union bound to combine the above with Part 2 of Proposition 2.3.1, where we let $\delta = \frac{1}{4}$ and $\Gamma = 2L_{\mu,s}R\sqrt{|\mathcal{S}|}$. As a result, there exists a problem-independent constant $\tilde{c}_5 > 0$ such that, if

$$\begin{aligned} n \geq\ & \sigma^2 \cdot |\mathcal{S}|^{4-4\rho} \left(\frac{R}{\epsilon}\right)^4 \bigvee \tilde{c}_5 \cdot N^*(c_5) \\ & \bigvee \tilde{c}_5 \cdot \sigma^2 |\mathcal{S}|^{4\rho} \cdot \left(\frac{1 + 2L_{\mu,s}R\sqrt{|\mathcal{S}|} + \hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{\tilde{c}_5 R L_\mu p}{\min\{\epsilon,\, \sigma^{1/2}\}} \right)^2 \end{aligned} \tag{2.4.59}$$

then $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon}$ with probability lower bounded by $1 - \alpha - 2\exp(-\frac{n(L_{\mu,s})^2}{2\sigma_L^2})$. Recall again that $a \leq 1$. Then, inequality (2.4.59) holds with $2\exp(-\frac{n(L_{\mu,s})^2}{2\sigma_L^2}) \leq \alpha$, if $\rho = 3/8$ and if n is large enough to satisfy both of the following inequalities

$$n \geq 2\sigma_L^2 \cdot \ln \frac{2}{\alpha} \geq \frac{2\sigma_L^2}{L_{\mu,s}^2} \ln \frac{2}{\alpha} \tag{2.4.60}$$

where the last inequality is due to $L_{\mu,s} \geq 1$.

$$\begin{aligned} n \geq\ & \sigma^2 \cdot |\mathcal{S}|^{5/2} \left(\frac{R}{\epsilon}\right)^4 \bigvee \tilde{c}_5 \cdot \sigma^2 |\mathcal{S}|^{5/2} \cdot \left(\frac{1 + 2L_{\mu,s}R + \hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{\tilde{c}_5 R L_\mu p}{\min\{\epsilon, \sigma^{1/2}\}} \right)^2 \\ & \bigvee \tilde{c}_5 \cdot N^*(c_5). \end{aligned} \tag{2.4.61}$$

The above immediately leads to the desired result by observing that $\tilde{c}_5 N^*(\tilde{c}_5) \geq 2\sigma_L^2 \cdot \ln \frac{2}{\alpha}$ if $\tilde{c}_5 \geq 2$. □

2.4.2.6 Proof of Theorem 2.3.8

Following the same argument as in the proof for Theorem 2.3.7, we have $F_{n,\lambda}(\mathbf{0}) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + 2L_{\mu,s}R\sqrt{|\mathcal{S}|}$ with lower-bounded probability $1-2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2})$. We may invoke both the De Morgan's Law and the union bound to combine the above with Proposition 2.3.2, where we let $\delta = \frac{1}{6}$, $\rho = 1/4$ and $\Gamma = 2L_{\mu,s}R\sqrt{|\mathcal{S}|}$. As a result, $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3(\epsilon + \hat{\varepsilon})$ with probability lower bounded by $1-\alpha-2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2})$, for n satisfying

$$n \geq \tilde{c}_6|\mathcal{S}|^{3/2}\sigma^2\left[\left(\frac{1}{\mathcal{U}_{\mathcal{H}}\epsilon}\right)^3 + \left(\frac{\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}\epsilon^2}\right)^3\right] \bigvee \tilde{c}_6 N^*(c_6)$$
$$\bigvee \tilde{c}_6\sigma^2|\mathcal{S}|^{3/4} \cdot \left(\frac{1 + 2L_{\mu,s}R\sqrt{|\mathcal{S}|} + \hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{\tilde{c}_6 R L_\mu p}{\min\{\epsilon, \sigma^{1/3}\}}\right)^{\frac{3}{2}} \quad (2.4.62)$$

Therefore, since $a \leq 1$ and $L_{\mu,s} \geq 1$, if one stipulates both

$$n \geq 2\sigma_L^2 \cdot \ln\frac{2}{\alpha} \geq \frac{2\sigma_L^2}{L_{\mu,s}^2} \ln\frac{2}{\alpha} \Longrightarrow 2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2}) \leq \alpha$$

and, for some problem-independent $\tilde{c}_6 > 0$,

$$n \geq \tilde{c}_6\sigma^2|\mathcal{S}|^{3/2}\left[\left(\frac{1}{\mathcal{U}_{\mathcal{H}}\epsilon}\right)^3 + \left(\frac{\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}\epsilon^2}\right)^3\right]$$
$$\bigvee \tilde{c}_6\sigma^2|\mathcal{S}|^{3/2} \cdot \left(\frac{1 + 2L_{\mu,s}R + \hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{\tilde{c}_6 R L_\mu p}{\min\{\epsilon, \sigma^{1/3}\}}\right)^{\frac{3}{2}} \bigvee \tilde{c}_6 N^*(\tilde{c}_6),$$

we know that $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3(\epsilon + \hat{\varepsilon})$ with probability lower bounded by $1 - 2\alpha$. This immediately leads to the desired result by further noticing that $\tilde{c}_6 N^*(\tilde{c}_6) \geq 2\sigma_L^2 \cdot \ln\frac{2}{\alpha}$ if $\tilde{c}_6 \geq 2$. □

2.4.2.7 Proof of Theorem 2.3.9

Consider again

$$\mathcal{E}_A := \left\{|F(\mathbf{x}^*) - F_n(\mathbf{x}^*)| \leq \frac{\epsilon}{2}\right\}; \quad \text{and} \quad \mathcal{E}_B := \left\{\left|F(\hat{\mathbf{x}}^{\min}) - F_n(\hat{\mathbf{x}}^{\min})\right| \leq \frac{\epsilon}{2}\right\}.$$

Following the same steps as in the proof for Proposition 2.3.2, it obtains that (2.4.55) holds almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$. When $\hat{\varepsilon} = 0$ and $\lambda = \frac{\sigma^{2\delta}}{n^{\delta}|\mathcal{S}|^{\rho}}$ with $\rho = \frac{1}{4}$ and $\delta = 0$, (2.4.55) immediately yields:

$$F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) = F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^{\min}) \leq \left(\sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_{\mathcal{H}}}} + \sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_{\mathcal{H}}} + \epsilon} \right)^2$$

almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$. Since it is assumed that F is differentiable and strongly convex as in (2.3.17) with constant $\mathcal{U}_{\mathcal{H}}$, we know that $F(\mathbf{x}) - F(\mathbf{x}^{\min}) \geq \frac{\mathcal{U}_{\mathcal{H}}}{2}\|\mathbf{x} - \mathbf{x}^{\min}\|^2$ for all $\mathbf{x} \in \mathcal{X}$ and that $\hat{\mathbf{x}}^{\min} = \mathbf{x}^{\min}$ (because we have let $\hat{\varepsilon} = 0$). Therefore,

$$\begin{aligned} \frac{\mathcal{U}_{\mathcal{H}}}{2}\|\mathbf{x}^* - \hat{\mathbf{x}}^{\min}\|^2 &\leq \left(\sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_{\mathcal{H}}}} + \sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_{\mathcal{H}}} + \epsilon} \right)^2 \\ &\overset{0<\epsilon\leq 1}{\leq} \left(\sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_{\mathcal{H}}}} + \sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_{\mathcal{H}}} + 1} \right)^2 \\ \Longrightarrow \min_{i\in\mathcal{S}} \hat{x}_i^{\min} - \min_{i\in\mathcal{S}} x_i^* \leq \|\mathbf{x}^* - \hat{\mathbf{x}}^{\min}\| &\leq \sqrt{\frac{2}{\mathcal{U}_{\mathcal{H}}}} \cdot \left(\sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_{\mathcal{H}}}} + \sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_{\mathcal{H}}} + 1} \right) \end{aligned}$$

almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$, where we have made use of the assumption that $\mathbf{x}^*$, $\hat{\mathbf{x}}^{\min} \in \mathcal{X} \subseteq \Re_+^p$. Therefore, if

$$\min_{i\in\mathcal{S}} \hat{x}_i^{\min} > \frac{|\mathcal{S}|^{1/4} + \sqrt{|\mathcal{S}|^{1/2} + 2\mathcal{U}_{\mathcal{H}}}}{\mathcal{U}_{\mathcal{H}}},$$

it holds that $\min_{i\in\mathcal{S}} x_i^* > 0$ almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$. Further invoking (2.4.27) with $a\lambda = \frac{a}{|\mathcal{S}|^{1/4}} \leq 1$, we know that $\min_{i\in\mathcal{S}} x_i^* \geq a\lambda$, and thus $P'_\lambda(x_i^*) = 0$ for all $i \in \mathcal{S}$ and $P'_\lambda(x_i^*) \geq 0$ for all $i = 1, ..., p$ due to Properties (iii) and (iv) of MCP in Section 2.4.1. If we recall (2.4.25) and the fact that $\hat{x}_i^{\min} = 0$ for all $i \notin \mathcal{S}$, conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$,

$$\begin{aligned} F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) &\leq \sum_{i=1}^{p} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) \leq \sum_{i=1}^{p} P'_\lambda(x_i^*)\hat{x}_i^{\min} \\ &= \sum_{i\in\mathcal{S}} P'_\lambda(x_i^*)\hat{x}_i^{\min} + \sum_{i\notin\mathcal{S}} P'_\lambda(x_i^*)\hat{x}_i^{\min} = 0, \quad a.s. \end{aligned}$$

The above inequality yields that $F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^{\min}) \leq \epsilon$ almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$.

Now, to achieve the desired result of the theorem, it suffices to show that, if n satisfies

$$n \geq c_7 \cdot \sigma^2 |\mathcal{S}| \cdot \left(\frac{1 + L_{\mu,s} R}{a^2 \epsilon^2} \ln \frac{24 R L_\mu p}{\epsilon} \right) \bigvee c_7 \cdot N^*(c_7) \tag{2.4.63}$$

for some universal constant $c_7 > 0$, then the event $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least $1 - 2\alpha$. To this end, notice that $\hat{\varepsilon} = 0$. We may use the same argument as in the proof for Proposition 2.3.1 in Section 2.4.2.3 in showing (2.4.52) and obtain that $\mathbf{P}[\mathcal{E}_A \cap \mathcal{E}_B] \geq 1 - \alpha$ if

$$n \geq \hat{c}_7 \cdot \sigma^2 |\mathcal{S}|^{1/2} \cdot \left(\frac{1 + \Gamma}{a^2 \epsilon^2} \ln \frac{24 R L_\mu p}{\epsilon} \right) \bigvee \hat{c}_7 \cdot N^*(\hat{c}_7) \tag{2.4.64}$$

for some universal constant $\hat{c}_7 > 0$.

Recall the assumption that $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\mathbf{0})$ almost surely. Since $F_{n,\lambda}(\mathbf{0}) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + 2L_{\mu,s} R \sqrt{|\mathcal{S}|}$ with lower-bounded probability $1 - 2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2})$ (to see this, we can repeat the steps in showing (2.4.57) in Subsection 2.4.2.5), we may let $\Gamma = 2L_{\mu,s} R \sqrt{|\mathcal{S}|}$. It is then easily verifiable from (2.4.64) that there exists such a problem-independent constant $c_7 > 0$ such that, if (2.4.63) holds, then $2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2}) \leq \alpha$ and the desired result holds. □

2.5 Some Discussions on Solution Schemes for RSAA

This section discusses two classes of solution techniques to ensure the desired S^{3}ONC solutions: local schemes (in Subsection 2.5.1) and a global technique (in Subsection 2.5.2).

2.5.1 Local Optimization for RSAA

The S^{3}ONC is weaker than the second-order KKT condition. Therefore, any algorithm that guarantees the second-order KKT condition can satisfy the stipulations made by Part 2 of Proposition 2.3.1 and those by Proposition 2.3.2. Furthermore, among those algorithms, any descent algorithm that guarantees the second-order

KKT condition can ensure the conditions as in Theorems 2.3.7 to 2.3.9, if initialized with an all-zero solution.

Algorithms that ensure the second-order KKT condition have been discussed by much literature. For instance, [10, 19, 63, 92, 93] provide algorithms with different convergence and complexity results. In particular, one of these algorithms, the interior point algorithm (IPA) presented by [10], is a descent, and fully polynomial-time approximation scheme (FPTAS) for a local solution that satisfies the desired second-order necessary condition, when $\mathcal{X}$ consists of a set of box constraints. In the special case where (2.1.6) is a quadratic program, [93] proposes a potential reduction (PR) algorithm and shows its convergence to a second-order KKT solution.

To facilitate the solution schemes we may reformulate the objective function into a twice continuously differentiable function. Specifically, according to [49], we have the following equivalence

$$P_\lambda(x) = \min_{\eta\in[0,a\lambda]} \frac{1}{2a}\eta^2 - \frac{1}{a}\eta x + \lambda x,$$

for which the optimizer admits a closed form:

$$\eta^{\min}(x) := \begin{cases} x & \text{if } 0 \le x \le a\lambda; \\ a\lambda & \text{if } x > a\lambda. \end{cases} \tag{2.5.65}$$

Therefore, we have the equivalence between the original regularized problem $\min_{\mathbf{x}\in\mathcal{X}} F_n(\mathbf{x}) + \sum_{i=1}^p P_\lambda(x_i)$ and an optimization problem with additional dummy variables:

$$\min_{\mathbf{x}\in\mathcal{X},\eta=(\eta_i)\in[0,a\lambda]^p} G_n(\mathbf{x}) := F_n(\mathbf{x}) + \sum_{i=1}^p \left(\frac{1}{2a}\eta_i^2 - \frac{1}{a}\eta_i x_i + \lambda x_i\right) \tag{2.5.66}$$

where η is the vector of dummy variables. Notice that Problem (2.5.66) is convex in η.

One can show that the second-order KKT condition to the reformulated program (2.5.66) implies the S^3ONC of (2.1.6). To see this, observe that, at a second-order KKT point $(\mathbf{x}^*, \eta^*)$ the first-order KKT condition also holds. Due to the convexity of (2.5.66) in η, it holds that $\eta^* = \eta^{\min}(x^*)$. Also by the definition of the second-

order KKT condition, we know that

$$d^\top \begin{bmatrix} \nabla^2 F_n(x^*) & -\frac{1}{a}I \\ -\frac{1}{a}I & \frac{1}{a}I \end{bmatrix} d \geq 0, \text{ for all } d \text{ in the critical set.} \tag{2.5.67}$$

To check if S^{3}ONC is satisfied, we only need to consider the case where $x_i \in (0, \min\{1, a\lambda\})$. According to (2.5.65), it holds that $\eta_i^* \in (0, \min\{1, a\lambda\})$. As an immediate result, (2.5.67) implies that the submatrix $\begin{bmatrix} \frac{\partial^2 F_n(x^*)}{\partial x_i^2} & -1/a \\ -1/a & 1/a \end{bmatrix}$ is positive semi-definite. Invoking Schur complement condition, it obtains that $0 \leq \frac{\partial^2 F_n(x^*)}{\partial x_i^2} - \frac{1}{a} = \frac{\partial^2 [F_n(\mathbf{x}) + \sum_{i=1}^p P_\lambda(x_i)]}{(\partial x_i)^2}\Big|_{\mathbf{x}=\mathbf{x}^*}$, where the last identity is immediate from the definition of P_λ for $x_i \in (0, \min\{1, a\lambda\})$. By its definition, the S^{3}ONC holds.

The reformulated problem (2.5.66) then satisfies all the assumptions for some existing FPTASs that guarantee a second-order KKT point, such as the interior point method by [10].

2.5.2 Global Optimization for RSAA

The global minimizer is a local minimizer, and, thus, also satisfies the S^{3}ONC. To compute this solution, the RSAA formulation can be equivalently formulated as a mixed integer program. Let Assumption A.3 hold and $a\lambda \leq 1$. This inequality is not restrictive as a and λ are user-specified parameters for P_λ. Then, as per (2.4.27), one can immediately rewrite the RSAA formulation into the following

$$\begin{aligned} \min \;\; & F_n(\mathbf{x}) + P_\lambda(a\lambda) \cdot \left(\mathbf{1}^\top \mathbf{z}_1 + \mathbf{1}^\top \mathbf{z}_2\right) \\ s.t. \;\; & \mathbf{x} \geq a\lambda \cdot \mathbf{z}_2 - \mathcal{M}\mathbf{z}_1; \quad \mathbf{x} \leq \mathcal{M} \cdot \mathbf{z}_2 \\ & -\mathbf{x} \geq a\lambda \cdot \mathbf{z}_1 - \mathcal{M}\mathbf{z}_2; \quad \mathbf{x} \geq -\mathcal{M} \cdot \mathbf{z}_1 \\ & \mathbf{x} \in \mathcal{X}; \quad \mathbf{z}_1, \mathbf{z}_2 \in \{0, 1\}^p. \end{aligned}$$

where $\mathcal{M}$ is a big-M and can be any scaler greater than $R + a\lambda$ in our case and where $P_\lambda(a\lambda) = \frac{a\lambda^2}{2}$. In particular, if Assumption A.5 holds, F_n is convex almost surely and the above formulation falls into the category of mixed integer convex programming, which admits numerical solvers to ensure global optimality. [49]

presents MILP reformulations when F_n is a quadratic but not necessarily convex function.

2.6 Preliminary Numerical Results

This section presents a preliminary set of numerical experiments following similar setups with [33**?**]. Specifically, we consider the following SP problem

$$\min\{\mathbb{E}[(\boldsymbol{\varrho}\mathbf{x} - \beta)^2] : \mathbf{x} \in [0,\, 5]^p\}, \tag{2.6.68}$$

where the relationship between $\boldsymbol{\varrho}$ and β is governed by $\beta = \boldsymbol{\varrho}\mathbf{x}^{\min} + \omega$ with $\mathbf{x}^{\min} = [3;\, 1.5;\, 0;\, 0;\, 2;\, \mathbf{0}_{p-5}]$. Let the ω be a standard normally distributed random variable; that is $\omega \sim \mathcal{N}(0,\, 1)$. Also assume that $\boldsymbol{\varrho} \sim \mathcal{N}_p(0, \Sigma)$, which is a p-variate normally distributed random variable with covariance matrix defined by $\Sigma = (\varsigma_{ij}) \in \Re^{p\times p}$ and $\varsigma_{ij} = 0.5^{|i-j|}$. It is easily verifiable that the optimal solution to the SP problem in (2.6.68) is $\mathbf{x}^{\min}$.

We compare the following approaches to solving (2.6.68) in problems with different choices of sample sizes and dimensions:

SAA: A global minimal solution to SAA in (2.1.3) computed using Mosek.

RSAA-local: An S^{3}ONC solution to RSAA in (2.1.6) generated by the PR algorithm as discussed in Section 2.5.1. The PR is initialized with an (approximate) all-zero solution. Our theories in Section 2.3 have predicted that such a local solution can approximate (2.6.68) globally.

RSAA-global: A global solution to RSAA in (2.1.6) solved with Mosek through the reformulation given in Section 2.5.2.

All experiments are conducted in Matlab on a computer with 2.2 GHz Intel Core i7 processor and 16GB memory. Mosek is invoked via Matlab to generate solutions for SAA and RSAA-global. For both RSAA-local and RSAA-global, the parameters for FCP are fixed as $\lambda = 0.5$ and $a = 0.9$. We would also like to remark that, since the PR algorithm requires the starting point to be an interior point, we approximate the all-zero solution by $10^{-4} \cdot \mathbf{1}$ for the PR's initialization.

For every $(n,\, p)$ combination, we replicate each solution scheme five times with independently generated samples for each repetition. We report the average,

maximal, and minimal suboptimality gaps as measured by $F(\cdot) - F(\mathbf{x}^{\min})$ in Tables 2.2 and 2.5. In Tables 2.2, we fix the number of samples $n = 100$ and gradually increase p from 10 to 1500. From this table, we can observe a clear trend that the solution quality of SAA deteriorates dramatically. In contrast, the suboptimality gaps are well contained by the proposed RSAA, even if the RSAA is only solved locally (as shown in the "RSAA-local" column). When $p = 1400$, RSAA-global is noticeably better than RSAA-local, as the former has a smaller maximal suboptimality gap than the latter. Nonetheless, the two different types of solutions yield almost the same quality in approximating (2.6.68). Note that our theories in fact provide a sharper performance bound for RSAA-global than RSAA-local. Therefore, the closely similar numerical performance between RSAA-global and RSAA-local is an indication that our bounds for RSAA-local may not be tight enough for at least the special case in the numerical experiments.

Figure 2.1 shows the dependence between the suboptimality gap and p. Particularly, in Figure 2.1.(a), the suboptimality gaps of SAA increase faster than linearly in p. In contrast, the suboptimality gaps for both RSAA-local and RSAA-global increase very slowly when p grows, as shown in both Figures 2.1.(a) and 1.(b).

Table 2.2 reports the computational time of the three different approaches. We notice that SAA is the most efficient among the three. RSAA-local incurs a noticeable increase in the computational efforts than SAA. Nonetheless, considering the substantial improvement generated by the RSAA-local in solution quality, we argue that the additional amount of computational cost is reasonable. RSAA-global is significantly slower than RSAA-local, even though the two have almost the same solution quality in our experiments.

Table 2.3 shows the sparsity of the solutions generated by the three different schemes. We can see from this table that SAA generates dense solutions in all the test instances, while both RSAA-local and RSAA-global can maintain sparsity in the output solutions.

We further compare the three approaches in problems that have different sample sizes n and a fixed number of dimensions $p = 100$. The comparison is presented in Table 2.4 and Figure 2.2. By comparison, we see that the solution quality of both RSAA-local and RSAA-global increase rapidly with the growth of n. Their rates are significantly faster than SAA.

In summary, our numerical results verify our theoretical predictions that the

Table 2.2: Comparison in solution quality measured by the suboptimality gaps for problems with different numbers of dimensions p and a fixed sample size $n = 100$.

p	SAA			RSAA-local			RSAA-global		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
10	0.13	0.22	4.79	0.04	0.11	0.00	0.04	0.11	0.00
30	0.466	0.617	0.31	0.04	0.06	0.02	0.04	0.06	0.02
50	1.05	1.25	0.76	0.05	0.09	0.00	0.05	0.09	0.00
70	2.42	4.09	1.55	0.03	0.05	0.01	0.03	0.05	0.01
90	11.8	17.4	8.91	0.04	0.06	0.02	0.04	0.06	0.02
200	366.56	488.31	279.27	0.02	0.06	0.01	0.02	0.06	0.01
300	1.25e3	1.57e3	1.04e3	0.02	0.04	0.00	0.02	0.04	0.00
400	2.48e3	2.74e3	2.18e3	0.03	0.07	0.01	0.03	0.07	0.01
500	3.40e3	3.75e3	3.00e3	0.03	0.06	0.00	0.03	0.06	0.00
600	4.89e3	5.18e3	4.35e3	0.02	0.04	0.01	0.02	0.04	0.01
700	6.21e3	6.41e3	5.75e3	0.02	0.04	0.00	0.02	0.04	0.00
800	7.96e3	8.54e3	7.34e3	0.02	0.03	0.01	0.02	0.03	0.01
900	9.92e3	1.06e4	9.44e3	0.04	0.10	0.01	0.04	0.10	0.01
1000	1.17e4	1.31e4	1.04e4	0.03	0.08	0.01	0.03	0.08	0.01
1100	1.32e4	1.43e4	1.19e4	0.03	0.08	0.01	0.03	0.08	0.01
1200	1.51e4	1.58e4	1.35e4	0.04	0.09	0.01	0.04	0.09	0.01
1300	1.73e4	1.85e4	1.59e4	0.01	0.03	0.00	0.01	0.03	0.00
1400	1.88e4	1.97e4	1.81e4	0.07	0.15	0.03	0.07	0.14	0.03
1500	2.18e4	2.34e4	2.10e4	0.03	0.08	0.01	0.03	0.08	0.01

RSAA is particularly effective when n is much smaller than than p. In such a case, RSAA may significantly improve solution quality over SAA.

2.7 Conclusion

We propose the RSAA, a modification to the SAA by incorporating a regularization scheme called the FCP. This modification targets the high-dimensional SP problems with sparsity. We show that when the solution is sparse or can be approximated by a sparse solution, the regularization can significantly reduce the required number of samples in some high-dimensional SP applications: Compared to the conventional SAA approach that requires the sample size to grow polynomially in the number of dimensions, the RSAA stipulates number of samples that is only poly-logarithmic in the dimensionality.

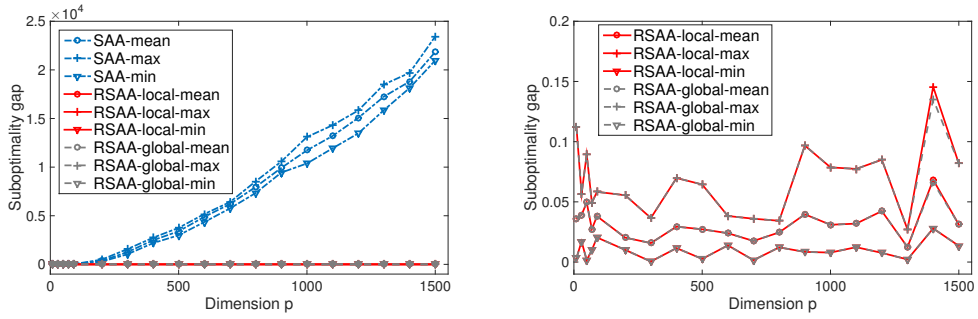


Figure 2.1: Comparison of suboptimality gaps of solutions generated by SAA, local optimization of RSAA, and global optimization of RSAA when $n = 100$ and p increases from 10 to 1500. "SAA-mean", "SAA-max", and "SAA-min" are the average, maximal, and minimal suboptimality gaps of SAA out of the five replications, "RSAA-local-mean", "RSAA-local-max", and "RSAA-local-min" are the average, maximal, and minimal suboptimality gaps of RSAA-local, "RSAA-global-mean", "RSAA-global-max", and "RSAA-global-min" are the average, maximal, and minimal suboptimality gaps of RSAA-global.

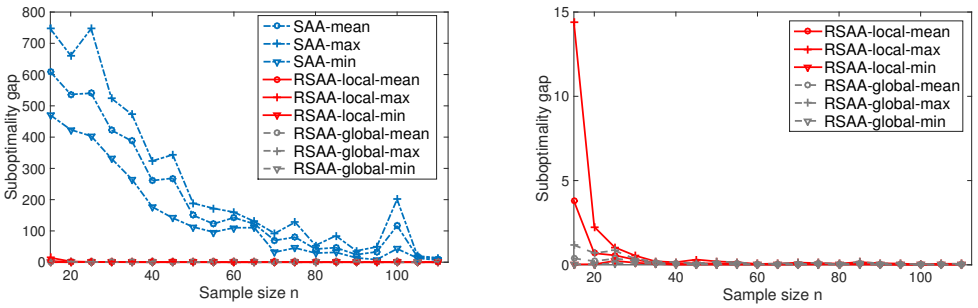


Figure 2.2: Comparison of suboptimality gaps of solutions generated by SAA, local optimization of RSAA, and global optimization of RSAA when $p = 100$ and n increases from 15 to 110. "SAA-mean", "SAA-max", and "SAA-min" are the average, maximal, and minimal suboptimality gaps of SAA out of the five replications, "RSAA-local-mean", "RSAA-local-max", and "RSAA-local-min" are the average, maximal, and minimal suboptimality gaps of RSAA-local, "RSAA-global-mean", "RSAA-global-max", and "RSAA-global-min" are the average, maximal, and minimal suboptimality gaps of RSAA-global.

Table 2.3: The numbers of nonzeros in the solutions generated by SAA, RSAA-local, and RSAA-global, when $n = 100$.

p	SAA			RSAA-local			RSAA-global		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
10	10	10	10	3	3	3	3	3	3
30	30	30	30	3	3	3	3	3	3
50	50	50	50	3	3	3	3	3	3
70	70	70	70	3	3	3	3	3	3
90	90	90	90	3	3	3	3	3	3
200	200	200	200	3	3	3	3	3	3
300	300	300	300	3	3	3	3	3	3
400	400	400	400	3	3	3	3	3	3
500	500	500	500	3	3	3	3	3	3
600	600	600	600	3	3	3	3	3	3
700	700	700	700	3	3	3	3	3	3
800	800	800	800	3	3	3	3	3	3
900	900	900	900	3	3	3	3	3	3
1000	1000	1000	1000	3	3	3	3	3	3
1100	1100	1100	1100	3	3	3	3	3	3
1200	1200	1200	1200	3	3	3	3	3	3
1300	1300	1300	1300	3	3	3	3	3	3
1400	1400	1400	1400	3.8	6	3	3	3	3
1500	1500	1500	1500	3	3	3	3	3	3

Table 2.4: Comparison of the average computational time out of the five replications for problems with different dimensionality p and fixed sample size $n = 100$.

p	SAA (s)	RSAA-local (s)	RSAA-global (s)	p	SAA (s)	RSAA-local (s)	RSAA-global (s)
10	3.19	1.71	9.77	700	3.42	20.92	241.68
30	3.21	4.08	13.22	800	3.38	34.13	1220.89
50	3.20	3.86	17.31	900	3.42	40.34	1425.75
70	3.17	4.46	30.28	1000	3.42	34.59	2693.44
90	3.13	8.55	27.31	1100	3.38	33.50	4014.09
200	3.06	19.03	7.21	1200	3.66	37.62	3686.88
300	3.13	15.82	45.60	1300	3.89	39.30	11658.30
400	3.35	14.02	157.64	1400	3.38	54.65	16927.54
500	3.33	19.34	134.08	1500	3.37	63.68	13463.53
600	3.40	20.92	240.10				

Table 2.5: Comparison in solution quality measured by the suboptimality gaps for problems with different sample sizes n and a fixed number of dimensions $p = 100$.

n	SAA			RSAA-local			RSAA-global		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
15	608.79	746.59	470.60	3.80	14.39	0.03	0.38	1.17	0.03
20	536.58	660.87	423.64	0.69	2.25	0.03	0.22	0.70	0.03
25	540.28	746.75	403.39	0.57	1.04	0.23	0.37	0.85	0.09
30	422.14	523.62	331.26	0.31	0.55	0.13	0.26	0.35	0.13
35	387.38	472.50	265.12	0.12	0.21	0.06	0.12	0.21	0.06
40	261.00	323.83	176.91	0.09	0.15	0.01	0.09	0.15	0.01
45	268.50	343.60	141.38	0.10	0.31	0.01	0.05	0.08	0.01
50	149.85	188.51	112.81	0.08	0.20	0.02	0.08	0.20	0.02
55	122.59	172.12	96.07	0.06	0.15	0.01	0.06	0.15	0.01
60	142.53	159.97	110.20	0.03	0.05	0.02	0.03	0.05	0.02
65	122.31	130.33	110.29	0.04	0.07	0.01	0.04	0.07	0.01
70	69.64	92.05	32.02	0.05	0.13	0.01	0.05	0.13	0.01
75	80.03	127.81	45.62	0.07	0.11	0.02	0.07	0.11	0.02
80	42.01	53.67	29.14	0.04	0.07	0.02	0.04	0.07	0.02
85	46.52	84.56	31.37	0.07	0.16	0.02	0.07	0.16	0.02
90	24.21	36.26	14.04	0.03	0.09	0.01	0.03	0.09	0.01
95	32.96	48.93	8.22	0.03	0.07	0.00	0.03	0.07	0.00
100	116.52	201.05	42.98	0.02	0.03	0.01	0.02	0.03	0.01
105	17.20	19.94	13.04	0.03	0.06	0.01	0.03	0.06	0.01
110	10.48	13.88	6.41	0.02	0.06	0.01	0.02	0.06	0.01

Although the incorporation of FCP renders the RSAA formulation nonconvex, we argue that any S^3ONC solution achieved by a decent algorithm starting at the all-zero vector is good enough to ensure the optimization performance of the local solution. The S^3ONC is a necessary condition (for local minimality) weaker than the second-order KKT condition. Numerical algorithms to ensure the second-order KKT condition are known from the literature. Furthermore, under some conditions on the feasible region, the S^3ONC solutions admit an FPTAS. We also discuss a mixed integer convex reformulation to the RSAA formulation that allows for exact, though exponential-time in the worst case, computation of the global solution. Our preliminary numerical experiments have verified our theoretical predictions.

Chapter 3
Fast Algorithm for Non-convex Optimization

3.1 Introduction

In this chapter we are interested in the problem with the following structure:

$$\begin{aligned} \min \quad & F(x) \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0, \end{aligned} \tag{3.1.1}$$

where $A \in \mathbb{R}^{m\times n}$ and $F : \mathbb{R}^n_+ \to \mathbb{R}$ is a continuous function on $\mathbb{R}^n_+$ and smooth on $\mathbb{R}^n_{++}$. In the constrained non-convex optimization literature, one of the major research directions is the first order methods , such as, the gradient project method [71], the alternating direction method of multipliers (ADMM) approach [12, 42, 88] and the first order interior method [10, 34, 39]. They can only ensure the first order optimal condition and the iteration complexity is $O(1/\epsilon^2)$. Furthermore, the first order interior point method can even handle the problem without Lipschitz derivatives (e.g. l_q penalized problem with $0 < q < 1$). Beside the first order methods, the second order methods (e.g., the cubic regularization method and the second order interior point method) are also discussed in many literature. The cubic regularization method (e.g. [4, 18, 63]) are designed for the smooth optimization, which iteration complexity is $\tilde{O}(\epsilon^{-3/2})$, where $\tilde{O}(\cdot)$ means the complexity is $O(\cdot)$ complexity up to some $\log(1/\epsilon)$ terms. The second order interior point method

[10, 39] also has $\tilde{O}(\epsilon^{-3/2})$ complexity and can also address the non-Lipschitz problem. However, those second order methods require matrix inversion at each iteration, which can be very time consuming for large scale optimization. Recently, based on the power method and accelerated gradient decent, [4, 18] discuss the approximated the second order algorithms which have the $\tilde{O}(\epsilon^{-7/4})$ iteration complexity and don't require inversing hessian matrix. In [4], the algorithm is based on the cubic regularization method for the unconstrained problem. They leverage the accelerated gradient descent (AGD) for matrix inversion and Lanczos method to solve the sub-problem efficiently. The method in [18] contains two parts: first use the Lanczos method to approximate the negative curvature direction and then use the AGD to refine the direction. Those literature give us an idea that the approximated second order method can be even faster than vanilla first order methods (e.g. gradient descent). But they can only be applied to the unconstrained, smooth problem with Lipschitz hessian matrix. Even the simple $l1$ penalized regression fails to satisfy the above requirements.

One natural direction to address non-smoothness and/or non-Lipschitz hessian matrix limitations is to add extra linear constraints, and directly handle the constrained problem. However, the linear constraints may introduce extra difficulty on algorithm design and analysis. The analysis of [4, 18] depends on the good convergence rate of AGD. If we directly immigrate their methods to the constrained case, the original good complexity result of AGD fails and we may not have $\tilde{O}(\epsilon^{-7/4})$ result. In order to utilize the good complexity of AGD, we need to use special techniques to "unconstrain" the constrained problem. Here we borrow the spirit of the interior point method to conduct such unconstraining job.

Our contributions are as follows:

1. We construct an accelerated interior point gradient method (AIP-GM) for linear constrained non-convex programming optimization. The complexity is bounded by $\tilde{O}(\epsilon^{-7/4})$, which will lead to an ϵ-KKT solution or an ϵ-global minimizer with high probability. Our complexity result is better than the classic first order method $\mathrm{O}(\epsilon^{-2})$ (e.g.[42]).

2. Our method can also guarantee the second order necessary condition at the limit point. When algorithm terminates, the reduced Hessian H_{reduce} will satisfy $H_{reduce} \succeq O(-\sqrt{\epsilon}I)$. Furthermore, when the objective function has

Table 3.1: Runtime comparison for non-convex optimization

	iteration	hessian free	constraint	second order guarantee	Lipschitz Gradient	Lipschitz Hessian
Gradient Projection	$O(\epsilon^{-2})$	✓	✓	×	✓	✓
ADMM[42]	$O(\epsilon^{-2})$	✓	✓	×	✓	✓
First order Interior point[39]	$O(\epsilon^{-2})$	✓	✓	×	×	×
Cubic regularization[20, 63]	$O(\epsilon^{-3/2})$	×	✓	✓	✓	✓
Second order Interior point[39]	$O(\epsilon^{-3/2})$	×	✓	✓	×	×
Accelerated Method[18]	$\tilde{O}(\epsilon^{-7/4})$	✓	×	✓	✓	✓
Linear Cubic regularization[4]	$\tilde{O}(\epsilon^{-7/4})$	✓	×	✓	✓	✓
AIP-GM (**Theorem 3.2.11**)	$\tilde{O}(\epsilon^{-7/4})$	✓	✓	✓	×	×

the strictly saddle property[36], the second order necessary condition can further ensure that we reach an approximated local minimizer.

3. Our method can handle the Non-Lipschitz objective function (e.g. l_q minimization problem with $0 < q < 1$). We only require weaker conditions than classic Lipschitz gradient and Lipschitz hessian.

4. Our method only needs the gradient calculation and matrix-vector multiplication. Thus our method is Hessian-free and suitable for large scale optimization.

The rest of this chapter is structured as follows. Section 3.2 introduces the preliminaries and the main theorem. In section 3.3, we present and explain the sub-routines to replace the matrix inversion in the traditional second order interior point method. The detail of proof will be found in section 3.4.

3.2 Preliminaries and Main theorem

We use $\|\cdot\|$ to denote the Euclidean norm of a vector and the spectral norm of a matrix. $\|\|_\infty$ denotes the element of a vector which has the largest absolute value. For a symmetric matrix A, we denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ as the maximum and minimum eigenvalue of A respectively. We denote $A \succeq B$ as $A - B$ is positive semidefinite. We also introduce the following definitions.

Definition 3.2.1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is L_1-scaled Lipschitz if $\|X\nabla f(x + Xd_1) - X\nabla f(x+Xd_2)\| \leq L_1\|d_1 - d_2\|$ for all d_1, d_2 such that $x+Xd_1$ and $x+Xd_2$ are in the strictly interior of the feasible region. $\|d_1\| \leq r, \|d_2\| \leq r$ for some $r < 1$. $X = diag(x)$.*

Definition 3.2.2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is L_2-scaled Lipschitz Hessian if $\|X(\nabla^2 f(x+Xd_1) - \nabla^2 f(x+Xd_2))X\| \leq L_2\|d_1 - d_2\|$ for all d_1, d_2 such that $x+Xd_1$ and $x + Xd_2$ are in the strictly interior of the feasible region. $\|d_1\| \leq r, \|d_2\| \leq r$ for some $r < 1$. $X = diag(x)$.*

The **Definition 3.2.1** and **3.2.2** is a scaled version of the traditional definitions for smoothness. It generalizes the traditional definition of Lipschitz and Lipschitz Hessian. Denote $G(d) = F(x + Xd) \Rightarrow \nabla G(d) = X\nabla F(x + Xd)$, $\nabla^2 G(d) =$

$X\nabla^2 F(x+Xd)X$. If $F(x)$ is l_1-Lipschitz and/or l_2-Lipschitz hessian, $G(d)$ will also be $l_1\|X\|_\infty$-Lipschitz and/or $l_2\|X\|_\infty^2$-Lipschitz hessian. If we further assume $\|X\|_\infty < R$, we can conclude that l_1-Lipschitz and/or l_2-Lipschitz hessian implies Rl_1-scaled Lipschitz and/or $R^2 l_2$- scaled Lipschitz hessian. Some non-Lipschitz functions can still be scaled-Lipschitz. For example, let $F(x) = x^q,\ 0 \le x \le R,\ 0 < q < 1 \Rightarrow \nabla F(x) = qx^{q-1}$ and we will have $F(x)$ is not Lipschitz when x is close to 0. $X\nabla F(x+Xd) = xq(x+xd)^{q-1} = qx^q(1+e)^{q-1} \Rightarrow \|X\nabla F(x+Xd_1) - X\nabla F(x+Xd_2)\| = qx^q\|(1+d_1)^{q-1} - (1+d_2)^{q-1}\| \le qx^q\|(q-1)(1+d)^{q-2}(d_1-d_2)\| \le q(q-1)R^q(1-r)^{q-2}\|d_1-d_2\|$. It leads that $F(x)$ is a $q(q-1)R^q(1-r)^{q-2}$-scaled Lipchitz function. As we will take the non-Lipschitz objective function into consideration, throughout this project we will only assume scaled Lipschitz and scaled Lipschitz hessian hold without further mentioned.

Assumption A. *The feasible region of* (3.1.1) *is bounded, non-empty and has strictly interior. Furthermore, we assume there is a* $0 < R < +\infty$ *such that* $\|x\|_\infty < R$ *for all feasible point. The objective function* $F(x)$ *in* (3.1.1) *is* L_1*-scaled Lipschitz and* L_2*-scaled Lipschitz Hessian.*

Definition 3.2.3. *A function* $f : \mathbb{R}^n \to \mathbb{R}$ *is* μ_1*-strongly convex if* $\frac{\mu_1}{2}\|x-y\| \le f(y) - f(x) - \nabla f(x)^T(y-x)$ *for all* x, y.

The **Definition 3.2.3** are standard in the literature to characterize the convex level of the function.

Definition 3.2.4. *Given* $\epsilon > 0$, $x \in \mathbb{R}^n$ *is an* ϵ_1*-KKT point for* (3.1.1) *if there exists* $\lambda \in \mathbb{R}^m$ *such that:*

1. $Ax = b, x \ge 0$
2. $\nabla f(x) + A^T\lambda \ge -\epsilon$
3. $\|X(\nabla f(x) + A^T\lambda)\| \le \epsilon$, $X = diag(x)$

Definition 3.2.5. *Given* $\epsilon_1, \epsilon_2 > 0$, *we say* $x \in \mathbb{R}^n$ *satisfies* ϵ_2, ϵ_2*-KKT2 condition if:*

1. x *is a* ϵ_1 *KKT point.*
2. $d^T(X\nabla^2 f(x)X + \epsilon_2 I)d \ge 0$ *for all* d *such that* $AXd = 0$.

The **Definition 3.2.4** and **3.2.5** describe the first and second order optimality condition that we will use in this project. Notice that they are not the same as the traditional ϵ-KKT point's definition. The reason why we need them is that the classic definition of ϵ-KKT point may not suitable to handle the objective function with the non-Lipschitz gradient. The new definitions work the non-Lipschitz gradient case and are also sufficient to imply the classic ϵ-KKT point's definitions when the objective function is Lipschitz. Reader can refer the Definition 2 and Proposition 2 in [39] for more details.

When measuring the complexity of the total running time, we may need to count the computation costs for many different type s of operations, such as evaluation of the $\nabla f(x)$ and matrix vector multiplication. To simplify our analysis, we make the following assumption:

Assumption B. *1. The following operations' computation costs are bounded by $O(T_1)$:*

(a) evaluate $\nabla f(x)$ at any x.

(b) evaluate $\nabla^2 f(x)v$ at any x and v

(c) up to $n \times n$ matrix and n dimension vector multiplication.

2. The vector vector multiplication up to n dimension takes up to $O(T_2)$ computation cost.

3. Other vector vector operations (addition and subtraction) take $O(0)$ computation cost.

Assumption B is a mild assumption. For L_2- scaled Lipschitz hessian $f(x)$, (b) can be approximated evaluated with Hessian-free technique, which costs 2 gradient evolutions and 4 vector-vector multiplications. Thus the total cost of (b) is bounded by $O(2T_1 + 4T_2)$. And as vector-vector multiplication is a special case of matrix-vector multiplication, $O(2T_1 + 4T_2) < O(6T_1) = O(T_1)$. The details of Hessian-free technique will be discussed later. Since the computation cost for vector multiplication usually dominates the vector addition and subtraction's cost, here we won't count the addition and subtraction's computation cost.

3.2.1 Technique Lemmas

With the basic definition and assumptions in place, we now introduce the convergence results for two accelerated methods, Nestervo' accelerated gradient descent method [62] for strongly convex function and Lanczos method [46] for the approximated minimum eigenvector.

Lemma 3.2.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mu > 0$-strongly convex function and L_1-Lipschitz. Let $\Delta_f = f(x_0) - f(x^*)$, where x^* is the optimal solution. For any $k \geq 1 + \sqrt{\frac{L_1}{\mu}} \log(\frac{4L_1^2 \Delta_f}{\mu\epsilon})$:*

$$f(x_k) - f(x^*) \leq \frac{\epsilon}{2L_1}$$

It tells that in order to achieve an $\epsilon/2L_1$-approximated minimizer, the iteration dependence on μ and ϵ are $k = O(\sqrt{\frac{1}{\mu}} \log(\frac{1}{\mu\epsilon}))$. Compared the non-accelerated gradient decent method (e.g [13]) for μ-strongly convex function ($k = O(\frac{1}{\mu} \log(\frac{1}{\mu\epsilon}))$), the accelerated gradient decent (AGD) has better dependence μ. Next show the classic result of Lanczos method

Lemma 3.2.7. *Let $H \in \mathbb{R}^{n \times n}$ by a symmetric and $H \succeq 0$. Lanczos method will return unit vectorv in $k = O(\sqrt{\frac{1}{\epsilon}} \log(\frac{n}{\delta}))$ step, such that*

$$v^T H v \geq \lambda_{\max}(H)(1 - \epsilon)$$

With probability $1 - \delta$. Where $\lambda_{\max}(H)$ denotes the leading enginvalue of H.

The above Lemma shows that Lanczos method is efficient on approximated the leading (maximum) eigenvector. Supposing we know that the maximum eigenvalue is upper bounded by some $L_1 \geq 0$, it is easy to verify that $L_1 I - H \succeq (L_1 - \lambda_{\max}(H))I \succeq 0$ for any H. The leading eigenvalue for $L_1 I - H$ is $L_1 - \lambda_{\min}(H)$, where $\lambda_{\min}(H)$ is the minimum eigenvalue of H. Thus if we apply Lanczos for $L_1 I - H$, we will be able to find a v such that $v^T(L_1 I - H)v \geq (L_1 - \lambda_{\min})(1 - \epsilon)$. Rearrange and we will have $v^T H v \leq L_1\epsilon + \lambda_{\min}(H)(1 - \epsilon)$. Hence v is an $L_1\epsilon$-approximated minimum eigenvector of H. We may also use the Lanczos method to approximate the minimum eigenvector. We summarize it in the following corollary.

Corollary 3.2.8. *Let $H \in \mathbb{R}^{n \times n}$ by a symmetric and $L_1 I \succeq H \succeq -L_1 I$. Applying Lanczos method to $L_1 I - H$ will return unit vectorv in $k = O(\sqrt{\frac{2L_1}{\epsilon}} \log(\frac{n}{\delta}))$ step,*

such that

$$v^T H v \leq \lambda_{\min}(H) + \epsilon$$

With probability $1 - \delta$. *Where* $\lambda_{\min}(H)$ *denotes the minimum eigenvalue of* H.

We substitute the ϵ in **Lemma 3.2.7** by $\epsilon/(2L_1)$. As we apply the Lanczos method to $L_1 I - H$, $v^T H v \leq L_1\epsilon/(2L_1) + \lambda_{\min}(H)(1 - \epsilon/(2L_1)) \leq \lambda_{\min}(H) + \frac{L_1 - \lambda_{\min}(H)}{2L_1}\epsilon \leq \lambda_{\min}(H) + \epsilon$.

3.2.2 Main Results

The main theorem is shown as follows. Our method can be described in two levels. The upper level (**Algorithm 1**) construct the subproblem of second order interior point method and lower level (**Algorithm 2**) solve it approximately. We first show our result on the upper-level algorithm.

Theorem 3.2.9. *Suppose* **Assumption A** *holds. Let* F_0 *be a lower bound on the optimal objective function value,* x_0 *is an approximated analytic center* x_c *of the feasible region and* $\sum_{i=1}^n \log(x_i^0) - \sum_{i=1}^n \log(x_i^c) \leq \frac{\rho}{3}(F(x^0) - F_0)$. *Set* $\eta_1, \eta_3 = \frac{1}{60\sqrt{\rho}}$, $\eta_2 = \frac{L_1}{369\sqrt{\rho^3 L_2^2(2+\Lambda_{\max})}}$, $\Lambda_{\max} = 2(F(x^0) - F_0)L_1^2/\epsilon + L_2\sqrt{L_1/\epsilon} + \sqrt{\epsilon/L_1}$, $\rho = \max\{\frac{2+\frac{1}{L_2}}{\epsilon}, 1.1\}$, $\beta = \min\{\frac{11}{12}, \sqrt{\frac{1}{\rho L_2^2}}\}$ *and* $L_2 > 1/2$. *Then at least one of the following events will happen:*

1. *Algorithm 1 will generate a point satisfies* $\epsilon, \sqrt{\epsilon}-KKT2$ *condition*
2. *Algorithm 1 will generate a* ϵ*-minimizer, i.e.* $F(x^t) - F(x^*) \leq \epsilon$, *where* x^* *is a global minimizer*

before

$$t = O\left(\frac{8L_2(F(x_0) - F_0)\left(2 + \frac{2}{L_2}\right)^{3/2}}{\epsilon^{3/2}}\right)$$

Furthermore, if the objective function is strictly saddle and $Ax = b$ *absence, at least one of the following events will happens:*

1. *Algorithm 1 will generate a* ϵ*-approximated local minimizer.*
2. *Algorithm 1 will generate a* ϵ*- approximated global minimizer*

Algorithm 1 Approximated second order interior

Input $f(x)$; $A,\ b$; x^0, F_0, δ, $\rho$and η_i $i = 1, 2, 3$

Let $\phi(x) = \rho(F(x) - F_0) + \sum_{i=1}^{n} \log(x_i)$, where $\rho > 0$

For t = 1,2,... until convergence

Use lanczos method to get an $\eta_1/2$-minimum eigenvalue α of $X\nabla^2\phi(x^{t-1})X$

Set $\mu_{\min} = \alpha$

Set $\mu_{\max} = 2\frac{F(x^0)-F_0}{\beta^2} + L_2\beta + \frac{1}{\rho\beta}$

$d = \text{fast\_trust\_region}(\phi, A, x^{t-1}, \beta, \mu_{\min}, \mu_{\max}, \eta_i,\ i = 1, 2, 3)$

$x^{t+1} = x^t + Xd$

EndFor

Return x^t.

Theorem 3.2.9 implies that with the sub-routine fast_trust_region instead of exact solution, the second order interior point will also converge within $O(\epsilon^{-3/2})$ iterations. And in the following theorem, we will show that this sub-routine is very cheap:

Theorem 3.2.10. *Suppose settings in* **Theorem 3.2.9** *hold. The computation cost of Algorithms 2 is upper bounded by:*

$$O\left(n^2 T_2 + \log\left(\frac{\mu_{\max} - \mu_{\min}}{\eta_1}\right) \max\left\{\sqrt{\frac{L_1}{\eta_1}} \log\left(\frac{L_1}{\eta_1\eta_2}\right), \sqrt{\frac{2L_1}{\eta_3}} \log\left(\frac{n^2}{\delta^2}\right)\right\} T_1\right) \tag{3.2.2}$$

And its dependence on ϵ is $O(\frac{1}{\epsilon^{1/4}} \log(1/\epsilon)^2)$.

Algorithm 2 fast_trust_region

Input $\phi(x)$; A; $\tilde{x}$, β, $\mu_{\max}$, $\mu_{\min}$, δ, ϵ and $\eta_i,\ i = 1, 2, 3$

Generate N which contains the orthonormal basis spanning the null space of AX, where $X = \text{diag}(\tilde{x})$ and set $\mu_{\min,0} = \mu_{\min}$.

While $\mu_{\max} - \mu_{\min} \geq \eta_1$

If $\mu > 3\eta_1 + \mu_{\min,0}$

Let $\mu = \frac{1}{2}(\mu_{\max} + \mu_{\min})$

Solve $\bar{d}$ from min $\frac{1}{2}\bar{d}(NX\nabla^2\phi(\tilde{x})XN + \mu I)\bar{d} + \nabla\phi(\tilde{x})XN\bar{d}$ via AGD with tolerance η_2

Else

Use lanczos method to get an η_3-minimum eigenvector v of $NX\nabla^2\phi(\tilde{x})XN$ with probability $1 - \delta$

Set $\bar{d} = -\beta * v * \text{sign}(v^T\nabla\phi(\tilde{x})XN)$

EndIf

If $\|\bar{d}\| < \beta$

$\mu_{\max} = \mu$

Else

$\mu_{\min} = \mu$

EndIf

EndWhile

Return $d = N\bar{d}$

Combining the results in **Theorem 3.2.9** and **Theorem 3.2.10** together, we are ready to show our main results:

Theorem 3.2.11. *Suppose* **Assumption A** *and* **B** *and settings in* **Theorem 3.2.9** *hold. The total computation cost for* **Algorithm 1** *is upper bounded by:*

$$O\left(\frac{L_2(F(x^0) - F_0)}{\epsilon^{3/2}}\left(n^2T_2 + \sqrt{\frac{L_1}{\sqrt{\epsilon}}}\log\left(\frac{\Lambda_{\max} - L_1}{\sqrt{\epsilon}}\right)\right.\right.$$
$$\left.\left.\max\left\{\log\left(\frac{L_1L_2^{3/2}}{\epsilon^2}\right), \log\left(\frac{n^2t^2}{\delta^2}\right)\right\}\right)T_1\right) \tag{3.2.3}$$

Where $\Lambda_{\max}$ *and* t *is same as in* **Theorem 2.4**.

If we ignored the logarithm part, the complexity is bounded by

$$\tilde{O}\left(\frac{L_2\sqrt{L_1}(F(x^0)-F_0)}{\epsilon^{3/2}}\left(\frac{n^2}{\sqrt{L_1}}T_2+\frac{1}{\epsilon^{1/4}}T_1\right)\right) \tag{3.2.4}$$

For the problem only with non-negative constraints, box constraints or low rank constraints, the complexity bound can be further improved to:

$$\tilde{O}\left(\frac{L_2\sqrt{L_1}(F(x^0)-F_0)}{\epsilon^{3/2}}\left(\frac{1}{\epsilon^{1/4}}T_1\right)\right)=\tilde{O}\left(\frac{L_2\sqrt{L_1}(F(x^0)-F_0)}{\epsilon^{7/4}}T_1\right) \tag{3.2.5}$$

3.3 Interpretation on AIP-GM

Here we consider the problem (3.1.1). Per **Assumption A**, we can build an local upper bound for $F(x)$ at some feasible point x:

$$F(x+Xd)\leq F(x)+\nabla F(x)Xd+\frac{1}{2}d^TX\nabla^2F(x^-)Xd+\frac{L_2}{6}\|d\|_2^3 \tag{3.3.6}$$

Where $\|d\|\leq r$ and $X=\mathrm{diag}(x)$. It can also be expressed by:

$$F(x+Xd)-F(x)\leq \nabla F(x)Xd+\frac{1}{2}d^TX\nabla^2F(x^-)Xd+\frac{L_2}{6}\|d\|_2^3 \tag{3.3.7}$$

An direct way is to minimize the right hand side iteratively:

$$\begin{aligned}\min\quad & \nabla F(x)Xd+\frac{1}{2}d^TX\nabla^2F(x)Xd+\frac{L_2}{6}\|d\|_2^3\\ \text{s.t.}\quad & AXd=0\\ & \|d\|\leq r\end{aligned} \tag{3.3.8}$$

It means that we need to optimization a cubic function over some convex set. In general, it is hard to solve globally. Even if the problem has a special structure and can ensure the global minimizer, we still may not guarantee iteratively solving (3.3.8) will generate a sequence that converges to the desired solution. Instead of solving (3.3.8), in the remaining part of this section, we will show that a properly designed second order interior can overcome those issues.

To utilized the interior point method, the first step is to relax the non-negative

constraints and build the potential function $\phi(x)$:

$$\phi(x) = \rho(F(x) - F_0) - \sum_{i=1}^{p} \log(x_i) \tag{3.3.9}$$

Where $\rho > 0$ and F_0 is a lower bound of optimal objective function value for (3.1.1). Denote $\Delta\phi(x)$ as $\phi(x) - \phi(x)$ and we will have:

$$\begin{aligned}
\Delta\phi(x) &= \phi(x + Xd) - \phi(x) \\
&= \rho(F(x + Xd) - F_0) - \sum_{i=1}^{p} \log(x_i + x_i d_i) - (\rho(F(x) - F_0) - \sum_{i=1}^{p} \log(x_i)) \\
&= \overbrace{\rho(F(x + Xd) - F(x))}^{①} + \overbrace{\sum_{i=1}^{p} \log\left(\frac{x_i}{x_i + x_i d_i}\right)}^{②}
\end{aligned}$$

For the term ①, we will have:

$$\begin{aligned}
① = \rho(F(x + Xd) - F(x)) &\leq \rho\left(F(x) + \nabla F(x)Xd + \frac{1}{2}d^T X\nabla^2 F(x^-)Xd + \frac{L_2}{6}\|d\|^3 - F(x)\right) \\
&= \rho\left(\nabla F(x)Xd + \frac{1}{2}d^T X\nabla^2 F(x^-)Xd + \frac{L_2}{6}\|d\|^3\right)
\end{aligned}$$

For the term ②, if $\|d\| \leq \beta < 1, X = \text{diag}(x)$, then via Lemma 1 in [91]:

$$② = \sum_{i=1}^{p} \log(\frac{x_i}{x_i + x_i d_i}) \leq -e^T d + \frac{\beta^2}{2(1-\beta)}$$

Where e is the one vector. Therefore we will have:

$$\Delta\phi(x) \leq \rho\left(\nabla F(x)Xd + \frac{1}{2}d^T X\nabla^2 F(x^-)Xd - \frac{1}{\rho}e^T d\right) + \frac{\rho L_2}{6}\|d\|_2^3 + \frac{\beta^2}{2(1-\beta)} \tag{3.3.10}$$

To achieve a reduction for potential function $\phi(x)$, we minimize a quadratic function subject to an ellipsoid constraint.

$$\begin{aligned}
\min \quad & q(d) = \left(\nabla F(x)X - \frac{1}{\rho}e^T\right)d + \frac{1}{2}d^T x\nabla^2 F(x)Xd \\
\text{s.t.} \quad & AXd = 0
\end{aligned} \tag{3.3.11}$$

$$\|d\|^2 \leq \beta^2$$

If we denote N as the matrix contains the orthonormal basis span the null space of AX, the above problem will reduce to:

$$\begin{aligned} \min_{v} \quad & \frac{1}{2} v^T (NX\nabla^2 f(x) XN)v + \nabla f(x) XNv \\ \text{s.t} \quad & \|v\| \leq \beta \end{aligned} \tag{3.3.12}$$

We will use the **Algorithm 2** to search for a proper approximated solution to (3.3.12). **Algorithm 2** contains two stages : binary search the dual variable μ and approximate v. At the optimal solution pair μ^* and v^* of (3.3.12), we will have

$$(NX\nabla f(\tilde{x})XN + \mu^* I)v^* + \nabla f(\tilde{x})XN = 0 \tag{3.3.13}$$

$$(NX\nabla f(\tilde{x})XN + \mu^* I) \succeq 0, \quad \mu^* \geq \{0, -\lambda_{\min}(NX\nabla f(\tilde{x})XN + \mu^* I)\} \tag{3.3.14}$$

As $(NX\nabla f(\tilde{x})XN + \mu^* I) \succeq 0$, we can say that v^* is the optimal solution to the following convex problem:

$$\min \quad \frac{1}{2} v^T (NX\nabla f(\tilde{x})XN + \mu^* I)v + \nabla f(\tilde{x})XNv \tag{3.3.15}$$

Supposing that we have an approximated $\tilde{\mu}$, v^e will satisfy:

$$v^e = \arg\min \quad \frac{1}{2} v^T (NX\nabla f(\tilde{x})XN + \tilde{\mu} I)v + \nabla f(\tilde{x})XNv \tag{3.3.16}$$

Since (3.4.30) allow a close form solution $v^e = -(NX\nabla f(\tilde{x})XN + \tilde{\mu} I)^{-1} \nabla f(\tilde{x})XN$. It is not hard to show that $ll(\tilde{\mu}) = \|v^e\|^2$ is monotonic decreasing in μ. As $ll(\mu^*) = \beta^2$, $ll(\tilde{\mu}) < \beta^2$ for $\tilde{\mu} > \mu^*$. Therefore by binary search on μ, we will eventually come to an $\tilde{\mu}$ such that $\max\{\mu^*, \eta_1 - \lambda_{\min}(N^T X\nabla^2 f(x)NX)\} < \mu \leq \mu^* + \eta_1$ from AGD step in **Algorithm 2**. It ensures that the AGD will only work on at least η_1-strongly convex function. When μ^* is very close to $-\lambda_{\min}(N^T X\nabla^2 f(x)NX)$ (the hard case in trust region problem), we will switch to the lanczos method, which is not depends on the degree of convexity. With proper accurate levels, the approximated scheme **Algorithm 2** will also work as good as we exactly solve the sub-problem and we summarize it in the following lemma.

Lemma 3.3.1. *Suppose μ^* corresponding to optimal solution v^* of* (3.3.12)*. If*

we set $\eta_1, \eta_3 = \frac{1}{60\sqrt{\rho}}$, $\eta_2 = \frac{L_1}{360\rho^2 L_2^2(2+\Lambda_{\max})}$. *If* $\frac{1}{2}\mu^*\|v^*\|^2 \geq \frac{2}{3\sqrt{\rho^3 L_2^2}}$, **Algorithm 2** *will return a* $d = Nv^*$ *such that* $q(d) \leq -\frac{5}{12\sqrt{\rho^3 L_2^2}}$. *Otherwise,* d *will lead to an approximated KKT2 solution in the next step.*

This lemma tells us that with a suitable accuracy the approximated solution from **Algorithm 2** will remain sufficient descent by $\frac{5}{12\sqrt{\rho^3 L_2^2}}$ amount before reaching an approximated KKT2 point. The total number that we call **Algorithm 2** will be $O(\rho^{3/2})$. The outer loop of **Algorithm 2** is a binary search with fixed upper and lower bounds. Therefore the iteration of is $\tilde{O}(1)$. For inner loop, we will either use AGD or Lanczos method. For AGD case, the problem we consider will always be at least $O(\rho^{-1/2})-$strongly convex and the complexity would be $\tilde{O}(\rho^{1/4})$ due to **Lemma 3.2.6**. For the Lanczos case, as $\eta_3 = O(\rho^{-1/2})$, via **Corollary 3.2.8** the complexity will be $\tilde{O}(\rho^{1/4}\log(1/\delta^2))$ with probability $1-\delta$. If we merge them up and choose $\delta = O(\delta_0/\rho^{3/2})$, we will have the total complexity would be $\tilde{O}(\rho^{3/2}\cdot 1\cdot(\rho^{1/4}+\rho^{1/4}\log(1/\delta^2))) = \tilde{O}(\rho^{7/4})$ with probability $1-O(\delta_0)$. From **Theorem 3.2.9**, we have $\rho = \max\{\frac{2+1/L_2}{\epsilon}, 1.1\}$. Hence the total complexity for **Algorithm 1** would be $\tilde{O}(\epsilon^{-7/4})$.

3.4 Technical Proofs

3.4.1 Hessian Free Technique

Our method can be implemented hessian-free. In both AGD and Lanczos method we only require gradient calculation and the matrix vector multiplying, i.e. $NX\nabla^2 F(x)XN$ multiplying some vector d. As we do not require F to be L_2-Lipschitz hessian but only L_2-scaled Lipschitz hessian. The hessian free techniques introduced in [4, 18] can't not be directly applied.

As we have L_2-scaled Lipschitz hessian:

$$\|\nabla F(x+Xd_1)X - \nabla F(x+Xd_2)X - X\nabla^2 F(x+Xd_2)X(d_1-d_2)\| \leq \frac{1}{2}L_2\|d_1-d_2\|^2 \tag{3.4.17}$$

We want to approximate $N^T X\nabla^2 F(\tilde{x})XN\bar{d}$ by only evaluating the $\nabla F(x)$. Set

$x = \tilde{x}$, $d_2 = 0$, $d_1 = h\bar{d}$ in (3.4.17) with $0 < h < 1$:

$$\begin{aligned} &\|\nabla F(\tilde{x} + hXd_1)X - \nabla F(\tilde{x})X - X\nabla^2 F(x)Xhd_1\| \leq \frac{h^2}{2} L_2 \|d_1\|^2 \\ &\frac{\|\nabla F(\tilde{x} + hXd_1)X - \nabla F(\tilde{x})X - X\nabla^2 F(x)Xhd_1\|}{h} \leq \frac{h}{2} L_2 \|d_1\|^2 \end{aligned} \tag{3.4.18}$$

Thus we could approximated $N^T X\nabla^2 F(\tilde{x})XN\bar{d}$ in three step:

1. $p_1 = N\bar{d}$
2. $p_2 = \frac{\nabla F(\tilde{x}+hXp_1)X - \nabla F(\tilde{x})X}{h}$, which small enough h.
3. $p_3 = N^T p_2$

As when h is small enough, we will have p_2 close enough to $X\nabla^2 F(x)XN\bar{d}$ and finally p_3 will well approximate $N^T X\nabla^2 F(\tilde{x})XN\bar{d}$. In this procedure, we do not need to compute and store the hessian direct, which can significantly reduce the time and space cost. It make our method be able to handle the large scale optimization problem.

3.4.2 Proof of Theorem 3.2.9

Proof. To achieve a reduction for potential function $\phi(x)$, we minimize a quadratic function subject to linear and ellipsoid constraints.

$$\begin{aligned} \min \quad & q(d) = \left(\nabla F(x)X - \frac{1}{\rho}e^T\right) d + \frac{1}{2} d^T X\nabla^2 F(x)Xd \\ \text{s.t.} \quad & AXd = 0 \\ & \|d\|^2 \leq \beta^2 \end{aligned} \tag{3.4.19}$$

Let

$$Q = X\nabla^2 F(x)X, \quad c = \nabla F(x^-)X - \frac{1}{\rho}e \quad \text{and} \quad \bar{A} = AX, \bar{d} = Xd \tag{3.4.20}$$

Then the above problem becomes:

$$\min \quad q_Q(\bar{d}) = \frac{1}{2}\bar{d}^T Q\bar{d} + c\bar{d}$$

$$\begin{aligned} \text{s.t.} \quad & \bar{A}\bar{d} = 0 \\ & \|\bar{d}\|_2^2 \leq \beta^2 \end{aligned} \tag{3.4.21}$$

The optimal condition is:

$$\begin{aligned} (Q + \mu I)\bar{d} + c + \bar{A}^T\lambda = 0, \quad \mu \geq \max\{0, -\text{eig}_{\min}(H)\} \\ \|\bar{d}\|_2 \leq \beta, \quad \bar{A}d = 0, \quad \mu(\|\bar{d}\|_2 - \beta) = 0 \end{aligned} \tag{3.4.22}$$

$H = N^T Q N, g = N^T c$, where N is an orthonormal basis spanning the null space of $\bar{A}$. We can further reduce the above problem to:

$$\begin{aligned} \min \quad & q_H(v) = \frac{1}{2} v H v + g^T v \\ \text{s.t.} \quad & \|v\|_2^2 \leq \beta^2 \end{aligned} \tag{3.4.23}$$

The sufficient and necessary conditions are

$$(H + \mu I)v = -g, \quad \mu(\|v\| - \beta) = 0, \quad \mu \geq \max\{0, -\text{eig}_{\min}(H)\}, \|v\| \leq \beta \tag{3.4.24}$$

Based on the optimality condition, we will be able to build an upper bound on the objective function and summarized as follows:

Lemma 3.4.1. (3.4.22) *and* (3.4.24) *share the same optimal dual variable* μ^**, which is associated with* $\|\bar{d}\| \leq \beta$ *and* $\|v\| \leq \beta$*. Let* $\bar{d}^*$ *and* v^* *be the optimal solution to* (3.4.22) *and* (3.4.24)*. Then we will have:*

$$\begin{aligned} q(d^*) = q_Q(\bar{d}^*) = \frac{1}{2}(\bar{d}^*)^T Q \bar{d}^* + c^T \bar{d}^* = qH(v^*) \\ = \frac{1}{2}(v^*)^T H v^* + g^T v^* \leq -\frac{\mu^*}{2}\|\bar{d}^*\|^2 = -\frac{\mu^*}{2}\|v^*\|^2 \end{aligned} \tag{3.4.25}$$

Furthermore, for any feasible v*, we will have* $q_Q(\bar{d}) = q_H(v)$ *if* $\bar{d} = Nv$*.*

From above Lemma, we know that it is equivalent to solve (3.4.22) and (3.4.22) and the optimal objective is upper bounded to be non-positive. When this upper bound is small enough, we will able to make sure that $\Delta\phi(x)$ is also negative. Plug

$\frac{1}{2}(\bar{d}^*)^T Q\bar{d}^* + c^T\bar{d}^* le - \frac{\mu^*}{2}\|\bar{d}^*\|^2$ back to (3.3.10):

$$\begin{aligned}\Delta\phi(x) &\leq -\frac{\rho}{2}\|\bar{d}\|_2^2\mu + \frac{\rho L_2}{6}\|d\|_2^3 + \frac{\beta^2}{2(1-\beta)} \\ &\overset{①}{\leq} -\frac{\rho}{2}\|\bar{d}\|_2^2\mu + \left(\frac{\rho L_2\beta^3}{6} + \frac{\beta^2}{2(1-\beta)}\right) \\ &= -\frac{\rho}{2}\|\bar{d}\|_2^2\mu + \left(\frac{\rho L_2}{6}\beta + \frac{1}{2(1-\beta)}\right)\beta^2 \end{aligned} \tag{3.4.26}$$

① is because $\|d\| \leq \beta$. Here we pick

$$\beta = \min\{\frac{11}{12}, \sqrt{\frac{1}{\rho L_2^2}}\}, \quad \rho > 1$$

Put β back into (3.4.26) and we will have

$$\begin{aligned}\Delta\phi(x) &\overset{①}{\leq} -\frac{\rho}{2}\|\bar{d}\|_2^2\mu + \left(\frac{\rho L_2\beta + 1}{6}\right)\beta^2 \\ &= -\frac{\rho}{2}\|\bar{d}\|_2^2\mu + \left(\frac{\sqrt{\rho}+1}{6}\right)\frac{1}{\rho L_2^2} \\ &\overset{②}{\leq} \rho\left(-\frac{1}{2}\|\bar{d}\|_2^2\mu + \frac{1}{3\sqrt{\rho^3}L_2^2}\right)\end{aligned}$$

① is from $\beta \leq \frac{11}{12} \Rightarrow \frac{1}{2(1-\beta)} \leq \frac{1}{6}$ and ② can be get from $\rho > 1 \Rightarrow \sqrt{\rho} \leq \rho \Rightarrow (\sqrt{\rho}+1)/6\rho \leq 1/3\sqrt{\rho}$. If we want the $\Delta\phi(x)$ is sufficient decent, we may impose the condition that

$$\frac{1}{2}\|\bar{d}\|_2^2\mu \geq \frac{2}{3\sqrt{\rho^3}L_2^2} \tag{3.4.27}$$

That is:

$$\Delta\phi(x) \leq \rho\left(-\frac{1}{2}\|\bar{d}\|^2\mu + \frac{1}{3\sqrt{\rho^3}L_2^2}\right) \leq -\rho\frac{1}{3\sqrt{\rho^3}L_2^2} = -\frac{1}{3\sqrt{\rho}L_2^2}$$

If we have $\frac{1}{2}\|\bar{d}\|_2^2\mu \geq \frac{2}{3\sqrt{\rho^3}L_2^2}$ for all step, then we will come to a $\frac{1}{3\sqrt{\rho}L_2^2}$ approximated global minimizer. The next step is to show that if we pick ρ to be large enough, we will have the x is a ϵ_1, ϵ_2-KKT2 solution when the $\Delta\phi(x)$ fails to be sufficient decent.

Lemma 3.4.2. *Set* $\rho = \frac{2+1/L_2}{\epsilon}$, $\beta = \min\{\frac{11}{12}, \frac{1}{\sqrt{\rho L_2^2}}\}$. *If* $L_2 \geq 2$, *we will reach a* $\epsilon, \sqrt{\frac{\epsilon}{2+1/L_2}}$*-KKT2 solution when* $\frac{1}{2}\|\bar{d}\|_2^2\mu \geq \frac{1}{\sqrt{\rho^3 L_2^2}}$ *fails.*

The above Lemma shows that by carefully choosing ρ, we can ensure that the when algorithm fail to find a sufficient decent, we will already reach an approximated KKT2 point.

The next step is to show the iteration complexity. As during the procedure, we always have $\Delta\phi(x) \leq -\frac{1}{3\sqrt{\rho}L_2^2}$. It will lead to two facts. First, $\phi(x^t)$ is upper bounded by $\phi(0)$; Second at each iteration $\phi(x)$ will sufficiently decrease by $\frac{1}{3\sqrt{\rho}L_2^2}$ if we have exact solution to the trust region sub-problem.

Here supposing we can not solve the trust region sub-problem exactly but only can derive an approximated solution. We need to figure out suitable accuracy for the approximated solution to make sure **Algorithm 1** converge to the correct solution. To achieve this goal, two things need to happen:

1. If $\frac{1}{2}\mu^*\|\bar{d}^*\|_2^2 \leq \frac{2}{3\sqrt{\rho^3 L_2^2}}$, we must be able to ensure the next step converge to an approximated KKT2 point.

2. If $\frac{1}{2}\mu^*\|\bar{d}^*\|_2^2 \geq \frac{2}{3\sqrt{\rho^3 L_2^2}}$, we must be able to show that the approximated solution will also lead to a sufficient decent, i.e. $\Delta\phi(x^k) = O(\frac{1}{\sqrt{\rho}})$

Where μ^* and $\bar{d}^*$ is the optimal solution to (3.4.21). Per **Lemma 3.4.2**, we know that if $\frac{1}{2}\mu^*\|\bar{d}^*\|_2^2 \leq \frac{1}{\sqrt{\rho^3 L_2^2}}$, we will terminate with an approximated KKT2 point. Therefore the approximated solution must satisfy $q_Q(\bar{d}) \leq q_Q(\bar{d}^*) + \frac{1}{3\sqrt{\rho^3 L_2^2}} \Rightarrow q_Q(\bar{d}) \leq -\frac{2}{3\sqrt{\rho^3 L_2^2}}$. Here $\bar{d}$ is the approximated solution. If $\frac{1}{2}\mu^*\|\bar{d}^*\|_2^2 \geq \frac{2}{3\sqrt{\rho^3 L_2^2}}$ and the approximated solution satisfies:

$$q_H(v) = \frac{1}{2}\bar{d}H\bar{d} + g\bar{d} \leq -\frac{5}{12\sqrt{\rho^3 L_2^2}}$$
$$\Rightarrow \Delta\phi(x^k) \leq \rho\left(q_H(v) + \frac{1}{3\sqrt{\rho^3 L_2^2}}\right) \leq -\frac{1}{12\sqrt{\rho}L_2^2}$$

If we can solve for an approximated solution with $q_Q(\bar{d}) \leq -\frac{5}{12\sqrt{\rho^3 L_2^2}}$, we will also have sufficient descent of $\Delta\phi(x^{k+1}) = O(\frac{1}{\sqrt{\rho}})$.

In summary, for case 1, we will focus on solving for an approximated solution with $q_Q(\bar{d}) - q_Q(\bar{d}^*) \leq \frac{1}{3\sqrt{\rho^3 L_2^2}}$ or $\frac{1}{2}\tilde{\mu}\|\bar{d}\|^2 \leq \frac{2}{3\sqrt{\rho^3 L_2^2}}$.For case 2, we require $q_H(v) \leq$

$-\frac{5}{12\sqrt{\rho^3 L_2^2}}$ and we will be able to be able to remain sufficient decent before reaching an approximated KKT2 solution.

In **Algorithm 2**, we approximately solve the trust region sub-problem (3.3.12) via a two stage framework: binary search the dual variable μ and approximate $\bar{d}$. Let first consider the scenario that we can have exact solution $\bar{d}^e$ for a given μ.

At the optimal solution pair μ^* and v^*, we will have

$$(N^T X \nabla f(\tilde{x}) X N + \mu^* I) v^* + \nabla f(\tilde{x}) X N = 0 \tag{3.4.28}$$

As $(N X \nabla f(\tilde{x}) X N + \mu^* I) \succeq 0$, we can say that v^* is the optimal solution to the following convex problem:

$$\min \quad \frac{1}{2} v^T (N^T X \nabla f(\tilde{x}) X N + \mu^* I) v + \nabla f(\tilde{x}) X N v \tag{3.4.29}$$

For a given $\tilde{\mu}$, v^e will satisfy:

$$v^e = \arg\min \quad \frac{1}{2} v^T (N^T X \nabla f(\tilde{x}) X N + \tilde{\mu} I) v + \nabla f(\tilde{x}) X N v \tag{3.4.30}$$

Since (3.4.30) allow a close form solution $v^e = -(N X \nabla f(\tilde{x}) X N + \tilde{\mu} I)^{-1} \nabla f(\tilde{x}) X N$. It is not hard to show that $ll(\tilde{\mu}) = \|V^e\|^2$ is monotonic decreasing in μ. As $ll(\mu^*) = \beta^2$, $ll(\tilde{\mu}) < \beta^2$, if $\tilde{\mu} > \mu^*$. Therefore by binary search on μ, we will eventually come to an $\tilde{\mu}$ such that $\min\{\mu^*, \eta_1\} < \mu \leq \mu^* + \eta_1$ from **Algorithm 2**. If we switch to the lanczos method branch in **Algorithm 2**, we denote $\bar{d}^e = \bar{d}$.

We characterize necessary searching accuracy on μ in the following Lemma.

Lemma 3.4.3. *Suppose for a given μ in* **algorithm 2** *and we can solve d from AGD step with infinite precision:*

1. $\mu_{\max} \leq \frac{2(F(x^0)-F_0)L_1^2}{\epsilon} + \frac{L_2\sqrt{L_1}}{\sqrt{\epsilon}} + \sqrt{\frac{\epsilon}{L_1}}$, $\mu_{\min} \geq 0$
2. *If $\mu^* \leq \frac{1}{6\sqrt{\rho}}$, we will have an approximated KKT2 solution with $\tilde{\mu} = \frac{1}{6\sqrt{\rho}}$.*
3. *If $\mu^* > \frac{1}{6\sqrt{\rho}}$ and $\mu^* \leq \tilde{\mu} - \mu^* \leq \frac{1}{60\sqrt{\rho}}$, we will have $q_H(v^e) \leq -\frac{1}{2\sqrt{\rho^3 L_2}}$ or sufficiently determine an approximated KKT2 solution.*

From **Lemma 3.4.3**, we show that the smallest $\tilde{\mu}$ we need to consider is $\frac{1}{6\sqrt{\rho}}$. And we can set the minimum difference between $\mu_{\max}$ and $\mu_{\min}$ being $\frac{1}{60\sqrt{\rho}}$ to make sure $\tilde{\mu} - \mu^* \leq \frac{1}{60\sqrt{\rho}}$. It is because that $\mu_{\min} \leq \mu^*$ and $\tilde{\mu} = \mu_{\max}$. Therefore we could

set $\eta_1, \eta_3 = \frac{1}{60\sqrt{\rho}}$ and $\eta_2 = 0$ in **Algorithm 2** to reach the results in **Lemma 3.4.3**. However, we only want to find an approximated solution v to (3.4.30) instead of exactly solving it. Incorporating with the statements in **Lemma 3.2.6**, we will have at each iteration of **Algorithm 1**, we will either find a feasible direction which will lead to objective function decreasing by $\frac{5}{12\sqrt{\rho^3 L_2}}$ or declare an approximated KKT2 point in the next step. It is equivalent to claim that the difference potential function will decrease at least by

$$\Delta\phi(x^t) \leq \rho\left(q(d) + \frac{1}{3\sqrt{\rho^3 L_2}}\right) \leq -\frac{1}{12\sqrt{\rho} L_2}$$

before converging to an approximated KKT2 point. As we will require to initialize with an approximated analytic center, which is the maximizer of:

$$\begin{aligned} \max \quad & \sum_{i=1}^{p} \log(x_i) \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{3.4.31}$$

An approximated analytic center leads to a feasible solution such that $\sum_{i=1}^{p} \log(x_0) \geq \sum_{i=1}^{p} \log(x_i^*) - \epsilon_0$. If we initial with such approximated analytic center, we will have:

$$\sum_{i=1}^{p} \log(x_i^t) \leq \sum_{i=1}^{p} \log(x_i^0) + \epsilon_0 \tag{3.4.32}$$

From the definition for $\phi(x)$, we will have:

$$\phi(x) \geq \rho(F(x^*) - F_0) - \sum_{i-1}^{p} \log(x_i^0) - \epsilon_0 := \phi_{floor} \tag{3.4.33}$$

Therefore the number of iteration would be bounded by:

$$t = \left\lfloor \frac{\phi(0) - \phi_{floor}}{\frac{1}{12\sqrt{\rho} L_2}} \right\rfloor$$

$$= \left\lfloor \frac{\rho(F(x_0) - F_0) - \sum_{i=1}^{p} \log(x_i^0) - (\rho(F(x^*) - F_0) - \sum_{i=1}^{p} \log(x_i^*))}{\frac{1}{12\sqrt{\rho}L_2}} \right\rfloor$$
$$= \left\lfloor \frac{\rho(F(x_0) - F(x^*)) - \rho(\sum_{i=1}^{p} \log(x_i^0) - \sum_{i=1}^{p} \log(x_i^*))}{\frac{1}{12\sqrt{\rho}L_2}} \right\rfloor$$

Since we assume $\sum_{i=1}^{p} \log(x_i^0) - \sum_{i=1}^{p} \log(x_i^*)) \leq \frac{1}{3}(F(x^*) - F_0)$:

$$\begin{aligned} t &= \left\lfloor \frac{\rho(F(x_0) - F(x^*)) - \rho(\sum_{i=1}^{p} \log(x_i^0) - \sum_{i=1}^{p} \log(x_i^*))}{\frac{1}{12\sqrt{\rho}L_2}} \right\rfloor \\ &\leq \left\lfloor \frac{\frac{2}{3}\rho(F(x_0) - F(x^*))}{\frac{1}{12\sqrt{\rho}L_2}} \right\rfloor \\ &= \left\lfloor 8L_2\sqrt{\rho^3}\left(F(0) - F(x^*)\right) \right\rfloor \end{aligned}$$

As we set $\rho = \max\{\frac{2+2/L_2}{\epsilon}, 1.1\}$, for small enough ϵ, we will have:

$$\begin{aligned} t &\leq \left\lfloor 8L_2\sqrt{\rho^3}\left(F(0) - F(x^*)\right) \right\rfloor \\ &\leq \left\lfloor \frac{8L_2\left(F(0) - F(x^*)\right)\left(2 + 2/L_2\right)^{3/2}}{\epsilon^{3/2}} \right\rfloor \\ &= O\left(\frac{8L_2\left(F(0) - F(x^*)\right)\left(2 + 2/L_2\right)^{3/2}}{\epsilon^{3/2}} \right) \end{aligned}$$

If the objective function is strictly saddle and $Ax = b$ is absent, an approximated second order necessary solution will also be an approximated local minimizer [36]. The last two statements in **Theorem 3.2.9** follow. □

3.4.3 Proof of Theorem 3.2.10

Proof. From **Algorithm 2**, we know that the outer loop is a binary search with upper bound $\mu_{\max}$ and lower bound $\mu_{\min}$. Thus it will terminate in $\log(\frac{\mu_{\max} - \mu_{\min}}{\eta_1})$ iterations. In each iteration, we will either meet a AGD step or Lanczos step.

For AGD step, we will always have $\mu \geq \mu_{\min,0} + 3\eta_1/2$. Since we set $\mu_{\min,0} > -\lambda_{\min}(N^T\nabla^2 F(x)N) - \eta_1/2$, μ will always greater than $\lambda_{\min}(N^T X \nabla^2 F(x) X N) + \eta_1$.

From the optimality condition (15):

$$(NX\nabla^2 F(\tilde{x})XN + \mu I) \succeq (NX\nabla^2 F(\tilde{x})XN + (-\lambda_{\min}(NX\nabla^2 F(\tilde{x})XN) + \eta_1)I) \succeq \eta_1 I \tag{3.4.34}$$

Therefore for AGD step, we will always work on an at least η_1- strongly convex function. Via the **Lemma 3.2.6**, the complexity for a single AGD step will be $O(\sqrt{\frac{L_1}{\eta_1}}\log(\frac{L_1}{\eta_1\eta_2}))$. If we face the Lanczos step, the complexity will be $O(\sqrt{\frac{2L_1}{\eta_3}}\log(\frac{n^2}{\delta^2}))$ via bf Corollary 2.3. Combine those observations, we will have for the loop part, the total number of iteration would be upper bounded by:

$$O\left(\log\left(\frac{\mu_{\max}-\mu_{\min}}{\eta_1}\right)\max\{\sqrt{\frac{L_1}{\eta_1}}\log\left(\frac{L_1}{\eta_1\eta_2}\right), \sqrt{\frac{2L_1}{\eta_3}}\log\left(\frac{n^2}{\delta^2}\right)\}\right) \tag{3.4.35}$$

Before the beginning of the loop, we will need to calculate N matrix, which contains the orthonormal basis spanning AX and it can be done by QR factorization on AX, which requires $O(n^2)$ times matrix vector multiplication. In cooperate with **Assumption B**, we will have the total computation cost for **Algorithm 2** will be:

$$O\left(n^2T_2 + \log\left(\frac{\mu_{\max}-\mu_{\min}}{\eta_1}\right)\max\{\sqrt{\frac{L_1}{\eta_1}}\log\left(\frac{L_1}{\eta_1\eta_2}\right), \sqrt{\frac{2L_1}{\eta_3}}\log\left(\frac{n^2}{\delta^2}\right)T_1\}\right) \tag{3.4.36}$$

From **Algorithm 1**, we know that $\mu_{\max}$ is upper bounded by $O(\epsilon^{-1})$ and $\mu_{\min}$ is lower bounded by 0 globally. η_1, η_3 is on the order of $O(\rho^{-1/2}) = O(\epsilon^{1/2})$ and $\eta_2 = O(\epsilon^{5/2})$. Plug them into (3.4.36) we will have the computation cost is bounded by

$$O\left(n^2T_2 + \sqrt{\frac{L_1}{\sqrt{\epsilon}}}\log\left(\frac{\Lambda_{\max}-L_1}{\sqrt{\epsilon}}\right)\max\left\{\log\left(\frac{L_1L_2^{3/2}}{\epsilon^2}\right), \log\left(\frac{n^2t^2}{\delta^2}\right)\right\}T_1\right) \tag{3.4.37}$$

Its dependence on ϵ is $O(\epsilon^{-1/4}\log(1/\epsilon)^2)$. □

3.4.4 Proof of Theorem 3.2.11

Proof. Follow the last statement in **Theorem 3.2.10**, we will have the computation cost for **Algorithm 2** is bounded by:

$$O\left(n^2T_2 + \sqrt{\frac{L_1}{\sqrt{\epsilon}}}\log\left(\frac{\Lambda_{\max}-L_1}{\sqrt{\epsilon}}\right)\max\left\{\log\left(\frac{L_1L_2^{3/2}}{\epsilon^2}\right), \log\left(\frac{n^2t^2}{\delta^2}\right)\right\}T_1\right) \tag{3.4.38}$$

Combine the iteration upper bound in **Theorem 3.2.9**, (4) and (5) follow immediately. To show (6), we will need to introduce special techniques for find N when A has special structure. Without exploring the special structure of AX, the computation cost of projecting a vector on to the $AXd = 0$ is almost the same as OR factorization on AX. For general AX case, our method could be as efficient as than any algorithm involving the step-wise projection on to $AXd = 0$. When AX has special structure (e.g. low-rank, box constraints), the projection operation can work very efficiently and can reach $O(nT_2)$. In this scenario, we design the special procedure to derive N which computation cost is also $O(nT_2)$. We summarize them in the following Lemma.

Lemma 3.4.4. *1. If we only have box constraints, i.e. $A = (I\ I)$, N has closed form solution*

$$N = \begin{pmatrix} -x_{n/2+1}/\sqrt{x_{n/2+1}^2 + x_1^2} & & & \\ & -x_{n/2+2}/\sqrt{x_{n/2+2}/^2 + x_2^2} & & \\ & & \cdots & \\ & & & -x_n/\sqrt{x_n^2 + x_{n/2}^2} \\ x_1/\sqrt{x_{n/2+1}^2 + x_1^2} & & & \\ & x_2/\sqrt{x_{n/2+2}/^2 + x_2^2} & & \\ & & \cdots & \\ & & & x_{n/2}/\sqrt{x_n^2 + x_{n/2}^2} \end{pmatrix}$$

2. If A is row independent, $rank(A) = s \ll n$ and A is element-wise non-zero.

There exist a upper s bands matrix M such that

$$M = \begin{pmatrix} m_{11} & 0 & \dots & \dots & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \dots & 0 \\ m_{s1} & \dots & m_{ss} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & m_{n,n-2s} & \dots & m_{n,n-2} \end{pmatrix}$$

M contains the basis span the null space of A. N can be calculated in $O(2snT_2) \approx O(sT_1)$ with M.

Based on the above Lemma, we can see that if A has some special structures, we do not need to reply on RQ factorization to find N but can only require $O(nT_2)$ computation cost, which almost match the cost like special case in projection operation. We can replace the computation cost for N in **Theorem 3.2.10** ($O(n^2T_2)$) by $O(T_1)$ and the (6) follows.

□

3.4.5 Proof of Lemma 3.4.1

Proof. Let

$$p = Q\bar{d}^* + c - \bar{A}^T\lambda$$

Then from the optimality condition (3.4.22):

$$\mu^* = \frac{\|p\|}{\|\bar{d}^*\|}, \quad \bar{d}^* = -\frac{\|\bar{d}^*\|p}{\|p\|} \tag{3.4.39}$$

And

$$\begin{aligned} \frac{1}{2}(\bar{d}^*)^T Q\bar{d}^* + c^T\bar{d}^* &= (\bar{d}^*)^T(Q\bar{d} + c) - \frac{1}{2}(\bar{d}^*)^T Q\bar{d}^* \\ &\overset{①}{=} (\bar{d}^*)^T(Q\bar{d}^* + c - \bar{A}\lambda) - \frac{1}{2}(\bar{d}^*)^T Q\bar{d}^* \\ &\overset{②}{=} (\bar{d}^*)^T p - \frac{1}{2}(\bar{d}^*)^T Q\bar{d}^* \end{aligned}$$

$$
\begin{aligned}
&\overset{③}{=} -\|\bar{d}^*\|\|p\| - \frac{1}{2}(\bar{d}^*)^T Q\bar{d}^* \\
&\overset{④}{=} -\|\bar{d}^*\|_2^2\mu^* - \frac{1}{2}(\bar{d}^*)^T Q(\bar{d}^*) \\
&\overset{⑤}{=} -\|\bar{d}^*\|_2^2\mu^* - \frac{1}{2}v^T N^T QNv \\
&\leq -\|\bar{d}^*\|_2^2\mu^* - \frac{1}{2}\lambda_{\min}(N^T QN)\|v\|^2 \\
&\overset{⑥}{=} -\|\bar{d}^*\|_2^2\mu^* + \frac{\lambda_{\min}(H)}{2}\|(\bar{d}^*)^T N^T N\bar{d}^*\|_2^2 \\
&\overset{⑦}{\leq} -\frac{\mu^*}{2}\|\bar{d}^*\|_2^2
\end{aligned} \tag{3.4.40}
$$

The above equations provide an upper bound of the function value of (3.4.21). ① is because we have $\bar{A}\bar{d}^* = 0$ in (3.4.22). In ② we replace p by $Q\bar{d} + c - \bar{A}^T\lambda$. ③ and ④ are from (3.4.39). As N is the orthonormal basis of the null space of A, we will have $Nv = \bar{d}$. ⑤ and ⑥ follow. And the ⑦ is because $\mu \leq -\lambda_{\min}(H)$ in (3.4.24).

Furthermore, as $Nv = \bar{d}$, we will have $\|v\|^2 = v^T v = \bar{d}^T N^T N\bar{d} = \|\bar{d}\|^2$ and

$$
\frac{1}{2}v^T Hv + gv = \frac{1}{2}\bar{d}^T N^T NQN^T N\bar{d} + cN^T N\bar{d} = \frac{1}{2}\bar{d}^T Q\bar{d} + c\bar{d} \tag{3.4.41}
$$

And it is easy to verify that $v^* = N^T\bar{d}^*$, μ^* also satisfy the optimal condition (3.4.24). And we will have

$$
q_H(v^*) = q_Q(\bar{d}^*) \leq -\frac{\mu^*}{2}\|\bar{d}^*\|^2 = -\frac{\mu^*}{2}\|v^*\|^2
$$

The last statement is because $X^{-1}d = \bar{d} \Rightarrow q(d) = q_q(\bar{d})$. □

3.4.6 Proof of Lemma 3.4.2

Proof. Here we consider two cases: **case 1.** If $\|\bar{d}\| < \beta$, then we will have $\mu = 0$. From the optimal condition for (3.4.21):

$$
\begin{aligned}
Q\bar{d} - \bar{A}\lambda + c &= 0 \\
\bar{A}\bar{d} = 0, \quad N^T QN &\succeq 0
\end{aligned}
$$

It leads to

$$X\nabla^2 F(x^-)X\bar{d} - AX\lambda + X\nabla F(x^-) = \frac{1}{\rho}e \geq 0 \tag{3.4.42}$$

case 2. If $\|\bar{d}\| = \beta$:

$$\frac{1}{\sqrt{\rho^3}L_2^2} > \frac{1}{2}\|\bar{d}\|_2^2\mu = \frac{1}{2}\mu\beta^2 \Rightarrow \mu \leq \frac{2}{\sqrt{\rho^3}\beta^2 L_2^2}$$

From the optimal condition:

$$\begin{aligned}(Q+\mu I)\bar{d} - \bar{A}\lambda + c = 0\\ \bar{A}\bar{d} = 0, \|\bar{d}\| = \beta\\ N^T(Q+\mu I)N \succeq 0\end{aligned}$$

We will have:

$$\begin{aligned}X\nabla^2 F(x^-)X\bar{d} - AX\lambda + X\nabla F(x^-) &= -\mu\bar{d} + \frac{1}{\rho}e\\ &\leq \mu\|\bar{d}\|e + \frac{1}{\rho}e\\ &\leq \left(\frac{2}{\sqrt{\rho^3}\beta L_2^2} + \frac{1}{\rho}\right)e\\ &\overset{\Phi}{\leq} \left(\frac{2}{L_2} + 1\right)\frac{1}{\rho}e\end{aligned}$$

Φ is due to $\beta \leq \frac{1}{\sqrt{\rho L_2^2}}$. And similarly:

$$\begin{aligned}X\nabla^2 F(x^-)X\bar{d} - AX\lambda + X\nabla F(x^-) &= -\mu\bar{d} + \frac{1}{\rho}e\\ &\geq -\mu\|\bar{d}\|e + \frac{1}{\rho}e\\ &\geq \left(-\frac{2}{\sqrt{\rho^3}\beta^2 L_2^2} + \frac{1}{\rho}\right)e\\ &\geq \left(-\frac{2}{L_2} + 1\right)\frac{1}{\rho}e\end{aligned}$$

If we assume $L_2 \geq 2 \Rightarrow \frac{2}{L_2} \leq 1$:

$$0 \leq \left(-\frac{2}{L_2}+1\right)\frac{1}{\rho}e \leq X\nabla^2 F(x^-)X\bar{d} - AX\lambda + X\nabla F(x^-) \leq \left(-\frac{2}{L_2}+1\right)\frac{1}{\rho}e \leq \frac{2}{\rho}e \tag{3.4.43}$$

Thus for both cases, we will have:

$$\begin{aligned}
0 \leq X\nabla^2 F(x^-)X\bar{d} - AX\lambda + X\nabla F(x^-) &\leq \frac{2}{\rho}e \\
0 \leq X\nabla^2 F(x^-)X\bar{d} - AX\lambda - X\left(\nabla F(x) - \nabla F(x^-)\right) + X\nabla F(x) &\leq \frac{2}{\rho}e \\
0 \leq X\nabla^2 F(x^-)X\bar{d} - AX\lambda - X\overbrace{\nabla^2 F(\tau x + (1-\tau)x^-)(x-x^-)}^{\Phi} + X\nabla F(x) &\leq \frac{2}{\rho}e \\
0 \leq X\nabla^2 F(x^-)X\bar{d} - AX\lambda - X\nabla^2 F(\tau x + (1-\tau)x^-)X\bar{d} + X\nabla F(x) &\leq \frac{2}{\rho}e \\
0 \leq X(\nabla^2 F(x^-) - F(\tau x + (1-\tau)x^-))X\bar{d} - AX\lambda + X\nabla F(x) &\leq \frac{2}{\rho}e
\end{aligned} \tag{3.4.44}$$

Φ is from Taylor expansion, where $\tau \in (0,1)$. Since we have lipschitz hessian:

$$\|\nabla^2 F(x^-) - F(\tau x + (1-\tau)x^-)\| \leq L_2\tau\|x - x^-\| \leq L_2\tau R\beta \tag{3.4.45}$$

It will lead to:

$$\begin{aligned}
\frac{2}{\rho}e &\geq X(\nabla^2 F(x^-) - F(\tau x + (1-\tau)x^-))X\bar{d} - AX\lambda + X\nabla F(x) \\
&\geq -L_2\tau\beta^2 e - AX\lambda + X\nabla F(x) \\
\Rightarrow -AX\lambda + X\nabla F(x) &\leq e\left(\frac{2}{\rho} + L_2\tau\beta^2\right) \\
\Rightarrow -AX\lambda + X\nabla F(x) &\leq e\left(\frac{2}{\rho} + L_2\tau\beta^2\right)
\end{aligned}$$

Similarly:

$$\begin{aligned}
0 &\leq X(\nabla^2 F(x^-) - F(\tau x + (1-\tau)x^-))X\bar{d} - AX\lambda + X\nabla F(x) \\
&\leq R^3 L_2\tau\beta^2 e - AX\lambda + X\nabla F(x)
\end{aligned}$$

$$\Rightarrow -AX\lambda + X\nabla F(x) \geq e\left(\frac{2}{\rho} - L_2\tau\beta^2\right) \tag{3.4.46}$$

As we have $\beta \leq \sqrt{\frac{1}{\rho L_2^2}}$:

$$|L_2\tau\beta^2| \leq L_2\frac{1}{\rho L_2^2} = \frac{1}{\rho L_2} \tag{3.4.47}$$

Therefore:

$$\begin{aligned} e\left(\frac{2}{\rho} - L_2\tau\beta^2\right) \leq -AXy + X\nabla F(x) \leq e\left(\frac{2}{\rho} + L_2\tau\beta^2\right) \\ \frac{e}{\rho}\left(2 - \frac{1}{L_2}\right) \leq -AXy + X\nabla F(x) \leq \frac{e}{\rho}\left(2 + \frac{1}{L_2}\right) \end{aligned}$$

If we assume that $L_2 \geq 2$ and pick $\rho = \frac{2+\frac{1}{L_2}}{\epsilon}$:

$$\begin{aligned} \frac{e}{\rho}\left(2 - \frac{1}{L_2}\right) &\geq 0 \\ \frac{1}{\rho}\left(2 + \frac{1}{L_2}\right) &= \epsilon \\ \Rightarrow 0 \leq -AXy + X\nabla F(x) &\leq \epsilon e \end{aligned} \tag{3.4.48}$$

Since we always have $AX\bar{d} = 0$ and x^- is a strictly interior point, $x = X\bar{d} + x^-$ will also be a feasible interior point, which means $Ax = b$ is satisfied. Till now, we show that the solution x is an first order ϵ-KKT point. The next step is to prove it also satisfied the second order necessary condition. Here we involve the linear constraints, the second order necessary will become the reduced hessian to be positive definite instead of the original hessian. The reduced hessian is defined as:

$$H_{reduce} = N^T\nabla^2 F(x)N \tag{3.4.49}$$

If we reconsider the **case 1** and **case 2**, we will have

$$\begin{aligned} N^T X^{-1}(Q + \mu I)X^{-1}N \succeq 0 \\ N(\nabla^2 F(x) + \mu I)N \succeq 0 \end{aligned}$$

$$(H_{reduce} + \mu I) \succeq 0$$

As we will stop at $\frac{1}{2}\|\bar{d}\|_2^2\mu \leq \frac{2}{3\sqrt{\rho^3 L_2^2}} \leq \frac{1}{\sqrt{\rho^3 L_2^2}}$. If $\mu > 0$, we will have

$$\begin{aligned} \|\bar{d}\| &= \beta \\ \Rightarrow \mu &\leq \frac{1}{\sqrt{\rho^3 L_2^2 \beta^2}} \\ \mu &\leq \frac{1}{\sqrt{\rho}} = \sqrt{\frac{\epsilon}{2 + \frac{1}{L_2}}} \end{aligned}$$

Thus we must have:

$$H_{reduce} \succeq -\sqrt{\frac{\epsilon}{2 + \frac{1}{L_2}}} \tag{3.4.50}$$

Now the second order necessary condition is satisfied. □

3.4.7 Proof of Lemma 3.4.3

Proof. For part 1, from the optimality condition (3.4.24), we know that $\mu^* \geq 0$, therefore we only need to search $x_{\min} \geq 0$. For $\mu_{\max}$, combine **Lemma 3.4.2** and (3.4.20):

$$\frac{1}{2}\mu^* \|\bar{d}^*\|_2^2 \leq -\frac{1}{2}(\bar{d}^*)^T X \nabla^2 F(x^t) X \bar{d}^* - (\nabla F(x^t) X - \frac{1}{\rho})\bar{d}^*$$

As F is L_2-scaled Lipschitz hessian:

$$\|F(x^t + X\bar{d}^*) - F(x^t) - \nabla F(x^t) X \bar{d}^* - \frac{1}{2}(\bar{d}^*)^T X \nabla^2 F(x^t) X \bar{d}^*\| \leq \frac{1}{6} L_2 \|\bar{d}^*\|^3 \tag{3.4.51}$$

Thus

$$\frac{1}{2}(\bar{d}^*)^T X \nabla^2 F(x^t) X \bar{d}^* \leq F(x^t + X\bar{d}^*) - F(x^t) - \nabla F(x^t) X \bar{d}^* + \frac{1}{6} L_2 \|\bar{d}^*\|^3$$

And

$$\begin{aligned}\frac{1}{2}\mu^*\|\bar{d}^*\| \leq & -\frac{1}{2}(\bar{d}^*)^T X\nabla^2 F(x^t)X\bar{d}^* - (\nabla F(x^t)X - \frac{1}{\rho})\bar{d}^* \\ \leq & F(x^t + X\bar{d}^*) - F(x^t) + L_2\|\bar{d}^*\|^3 + \frac{1}{\rho}\bar{d}^* \\ \leq & F(x^0) - F_0 + L_2\|\bar{d}^*\|_2^3 + \frac{1}{\rho}\|\bar{d}^*\|\end{aligned}$$

As $\|\bar{d}^*\| \leq \beta$

$$\frac{1}{2}\mu^*\|\bar{d}^*\| \leq F(x^0) - F_0 + L_2\beta^3 + \frac{1}{\rho}\beta$$

If $\|\bar{d}^*\| < \beta$, we will have $\mu^* = 0$. And supposing $\mu^* > 0$, we must have $\|\bar{d}^*\| = \beta > 0$ and we upper bounded the non-zero μ^* by

$$\mu^* \leq 2\frac{F(x^0) - F_0}{\beta^2} + L_3R^3\beta + \frac{1}{\rho\beta}$$

Thus our search space on μ is upper bound by $2\frac{F(x^0)-F_0}{\beta^2} + L_3R^3\beta + \frac{1}{\rho\beta}$ and we only need to set $\mu_{\max} = 2\frac{F(x^0)-F_0}{\beta^2} + L_3R^3\beta + \frac{1}{\rho\beta}$.
For part 2, as $\tilde{\mu} \leq \frac{1}{6\sqrt{\rho}}$:

$$\frac{1}{2}\tilde{\mu}\|\bar{d}\|^2 \leq \frac{1}{2}\frac{1}{6\sqrt{\rho}}\beta^2 = \frac{1}{12\sqrt{\rho^3}L_2^2} < \frac{1}{\sqrt{\rho^3}L_2^2} \tag{3.4.52}$$

Via **Lemma 3.4.2**, we know it $\tilde{\mu}$ will lead to an approximated KKT2 solution.
For part 3, we consider two cases $\mu^* + \lambda_{\min}(H) \leq 10\eta_1$ and $\mu^* + \lambda_{\min}(H) \geq 10\eta_1$.

For the first case, we will switch to lanczos method. Since F is L_2-scaled Lipschitz, we will have $L_1 I \succeq H \succeq -L_1 I$. Per **Corollary 2.3**, we will be able to solve for a unit vector $\tilde{v}$ such that:

$$\tilde{v}^T H\tilde{v} \leq \lambda_{\min}(H) + \eta_3 \tag{3.4.53}$$

And we choose the $v = -\tilde{v}\beta\text{sign}(g^T v)$:

$$\frac{1}{2}v^T Hv + g^T v \leq \frac{1}{2}v^T Hv \leq \frac{1}{2}\beta^2(\lambda_{\min}(H) + \eta_3) \leq \frac{-\mu^* + 10\eta_1 + \eta_3}{2}\beta^2 \tag{3.4.54}$$

If $\frac{1}{2}\mu^*\|\bar{d}^*\|^2 \geq \frac{2}{3\sqrt{\rho^3 L_2^3}}$, we will have

$$\frac{1}{2}v^T Hv + g^T v \leq -\frac{2}{3\sqrt{\rho^3 L_2}} + \frac{1}{24\sqrt{\rho^3 L_2^2}} + \frac{\eta_3}{2}\beta^2 \leq -\frac{1}{2\sqrt{\rho^3 L_2^2}}$$

It means that v will lead to a sufficient decent.
On the other hand, if $\frac{1}{2}\mu^*\|\bar{d}^*\|^2 < \frac{2}{3\sqrt{\rho^3 L_2^3}}$, we will have $\frac{1}{2}\tilde{\mu}\|v\|^2 \leq \frac{1}{2}(\mu^* + \frac{1}{12\sqrt{\rho}} + \eta_3)\beta^2 \leq \frac{1}{\sqrt{\rho^3 L_2^2}}$. Via **Lemma 3.4.2** we can conclude that the new solution will also be an KKT2 solution.

For the second case, we first consider the difference between $\|\bar{d}^*\|^2$ and $\|\bar{d}\|^2$:

$$\begin{aligned}
&\|\bar{d}^*\|_2^2 - \|\tilde{d}\|_2^2 \\
=&\|\bar{d}^*\|_2^2 - \|(H+\tilde{\mu}I)^{-1}g\|_2^2 \\
=&\|\bar{d}^*\|_2^2 - \|(H+\tilde{\mu}I)^{-1}(H+\mu^*I)(H+\mu^*I)^{-1}g\|_2^2 \\
=&\|\bar{d}^*\|_2^2 - \|(H+\tilde{\mu}I)^{-1}(H+\mu^*I)\bar{d}^*\|_2^2 \\
=&\bar{d}^*(I-(H+\mu^*I)(H+\tilde{\mu}I)^{-2}(H+\mu^*I))\bar{d}^*
\end{aligned}$$

The main idea of above proof utilize the fact that $(H+\mu^*)\bar{d}^* = (H+\tilde{\mu})\bar{d} = -g$ and $(H+\mu^*) \succ 0$. Next we want to leverage the condition $\mu^* < \tilde{\mu} < \mu^* + \eta_1$:

$$\begin{aligned}
&\bar{d}^*(I-(H+\mu^*I)(H+\tilde{\mu}I)^{-2}(H+\mu^*I))\bar{d}^* \\
=&\bar{d}^*(I-(H+\tilde{\mu}I+\mu^*-\tilde{\mu})(H+\tilde{\mu}I)^{-2}(H+\tilde{\mu}I+\mu^*-\tilde{\mu}))\bar{d}^* \\
=&\bar{d}^*(I-I-(\mu^*-\tilde{\mu})^2(H+\tilde{\mu}I)^{-2}-2(\mu^*-\tilde{\mu})(H+\tilde{\mu}I)^{-1})\bar{d}^* \\
=&\bar{d}^*(-(\mu^*-\tilde{\mu})^2(H+\tilde{\mu}I)^{-2}-2(\mu^*-\tilde{\mu})(H+\tilde{\mu}I)^{-1})\bar{d}^*
\end{aligned}$$

Via holder inequality, we will have:

$$\begin{aligned}
&\bar{d}^*(-(\mu^*-\tilde{\mu})^2(H+\tilde{\mu}I)^{-2}-2(\mu^*-\tilde{\mu})(H+\tilde{\mu}I)^{-1})\bar{d}^* \\
\leq&\|\bar{d}^*\|_2^2\left(\|-(\mu^*-\tilde{\mu})^2(H+\tilde{\mu}I)^{-2}-2(\mu^*-\tilde{\mu})(H+\tilde{\mu}I)^{-1}\|\right)
\end{aligned}$$

As $(H+\tilde{\mu}I) \succeq \tilde{\mu} + \lambda_{\min}(H) > 0$:

$$\|-(\mu^*-\tilde{\mu})^2(H+\tilde{\mu}I)^{-2}-2(\mu^*-\tilde{\mu})(H+\tilde{\mu}I)^{-1}\|$$

$$
\begin{aligned}
&\leq \left(-\frac{(\mu^* - \tilde{\mu})^2}{(\tilde{\mu} + \lambda_{\min}(H))^2} - 2\frac{\mu^* - \tilde{\mu}}{\tilde{\mu} + \lambda_{\min}(H)} \right) \\
&= \left(-\frac{(\mu^* - \tilde{\mu})^2 + 2(\mu^* - \tilde{\mu})(\tilde{\mu} - \mu^* + \mu^* + \lambda_{\min}(H))}{(\tilde{\mu} + \lambda_{\min}(H))^2} \right) \\
&= \left(\frac{(\mu^* - \tilde{\mu})^2 + 2(\tilde{\mu} - \mu^*)(\mu^* + \lambda_{\min}(H))}{(\tilde{\mu} + \lambda_{\min}(H))^2} \right) \\
&= \left(1 - \frac{(\mu^* + \lambda_{\min}(H))^2}{(\tilde{\mu} + \lambda_{\min}(H))^2} \right)
\end{aligned}
$$

As $\mu^* + \lambda_{\min}(H) \geq 10\eta_1$ and $\tilde{\mu} \leq \mu^* + \eta_1$

$$
\left(1 - \frac{(\mu^* + \lambda_{\min}(H))^2}{(\tilde{\mu} + \lambda_{\min}(H))^2} \right) \leq \left(1 - \frac{(\mu^* + \lambda_{\min}(H))^2}{(\mu^* + \eta_1 + \lambda_{\min}(H))^2} \right) \leq \left(1 - \frac{1}{(1 + \frac{\eta_1}{\mu^* + \lambda_{\min}(H)})^2} \right) \leq \left(1 - \frac{1}{(1 + \frac{1}{1}} \right)
$$

Therefore $\|\bar{d}^*\|^2 - \|\bar{d}\| \leq \frac{1}{5}\|\bar{d}^*\|$. The next job is to show that if $\|\bar{d}^*\|$ is not far away from $\|\tilde{d}\|$, the function value $q(\bar{d}^*)$ will also not be far away from $q(\tilde{d})$:

$$
\begin{aligned}
&\frac{1}{2}dHd + gd - \frac{1}{2}d^*Hd^* - gd^* \\
\overset{①}{=} &-\frac{1}{2}\tilde{d}(\tilde{\mu}\tilde{d} + g) + g\tilde{d} + \frac{1}{2}\bar{d}^*(\mu^*\bar{d}^* + g) - g\bar{d}^* \\
= &-\frac{1}{2}\tilde{\mu}\|d\|_2^2 + \frac{1}{2}\mu^*\|d^*\|_2^2 + \frac{1}{2}gd - \frac{1}{2}gd^* \\
= &-\frac{1}{2}\tilde{\mu}\|\tilde{d}\|_2^2 + \frac{1}{2}\mu^*\|\bar{d}^*\|_2^2 + \frac{1}{2}(g\tilde{d} + \tilde{d}H\bar{d}^* - \tilde{d}H\bar{d}^*) - \frac{1}{2}g\bar{d}^* \\
= &-\frac{1}{2}\tilde{\mu}\|\tilde{d}\|_2^2 + \frac{1}{2}\mu^*\|\bar{d}^*\|_2^2 + \frac{1}{2}(g + H\bar{d}^*)\tilde{d} - \frac{1}{2}(g + H\tilde{d})\bar{d}^* \\
= &-\frac{1}{2}\tilde{\mu}\|\tilde{d}\|_2^2 + \frac{1}{2}\mu^*\|\bar{d}^*\|_2^2 - \frac{1}{2}\mu^*\bar{d}^*\tilde{d} + \frac{1}{2}\tilde{\mu}\tilde{d}\bar{d}^* \\
= &\frac{1}{2}\mu^*(\|\bar{d}^*\|_2^2 - \|\tilde{d}\|_2^2) + \frac{1}{2}(\tilde{\mu} - \mu^*)(-\|\tilde{d}\|_2^2 + \tilde{d}^*\bar{d}) \\
\overset{②}{\leq} &\frac{1}{10}\mu^*\beta^2 + \frac{1}{2}\frac{1}{60\sqrt{\rho}}(\beta^2 + \beta^2) \\
= &\frac{1}{10}\mu^*\|\bar{d}^*\|^2 + \frac{1}{60\sqrt{\rho}}\beta^2
\end{aligned}
$$

① uses $(H + \tilde{\mu}I)\tilde{d} = -g$ and $(H + \mu^*I)\bar{d}^* = -g$. ② uses $\|\bar{d}^*\|^2 - \|\bar{d}\| \leq \frac{1}{5}\|\bar{d}^*\|$ and

$\tilde{\mu} - \mu^* \leq \eta_1 = \frac{1}{60\sqrt{\rho}}$. If $\frac{1}{2}\mu^*\|\bar{d}^*\|^2 \geq \frac{2}{3\sqrt{\rho^3 L_2^2}}$:

$$\frac{1}{2}dHd + gd \leq \frac{1}{2}d^*Hd^* + gd^* + \frac{1}{10}\mu^*\|\bar{d}^*\|^2 + \frac{1}{60\sqrt{\rho}}\beta^2$$

As $\frac{1}{2}d^*Hd^* - gd^* \leq -\frac{1}{2}\mu^*\|\bar{d}^*\|^2$,

$$\begin{aligned}\frac{1}{2}dHd + gd &\leq -\frac{2}{5}\mu^*\|\bar{d}^*\|^2 + \frac{1}{60\sqrt{\rho}}\beta^2 \\ &\leq -\frac{8}{15\sqrt{\rho^3 L_2^2}} + \frac{1}{60\sqrt{\rho^3 L_2^2}} \\ &\leq -\frac{31}{60\sqrt{\rho^3 L_2^2}} \leq -\frac{1}{2\sqrt{\rho^3 L_2^2}}\end{aligned}$$

If $\frac{1}{2}\mu^*\|\bar{d}^*\|^2 < \frac{2}{3\sqrt{\rho^3 L_2^2}}$, we will have:

$$\frac{1}{2}dHd + gd - \frac{1}{2}d^*Hd^* - gd^* \leq \frac{1}{10}\mu^*\|\bar{d}^*\|^2 + \frac{1}{60\sqrt{\rho}}\beta^2 \leq \frac{1}{12\sqrt{\rho^3 L_2^2}}$$

In conclusion, if we combine the results for case 1 and case 2 together, we will have

$$\begin{aligned}q(\bar{d}) = \frac{1}{2}v^THv + g^Tv &\leq \max\{-\frac{1}{2\sqrt{\rho^3 L_2^2}}, -\frac{1}{2\sqrt{\rho^3 L_2^2}}, \frac{1}{2}(v^*)^THv^* + g^Tv^* + \frac{1}{12\sqrt{\rho^3 L_2^3}}\} \\ &\leq \max\{-\frac{1}{2\sqrt{\rho^3 L_2^2}}, q(\bar{d}^*) + \frac{1}{12\sqrt{\rho^3 L_2^3}}\}\end{aligned}$$

□

3.4.8 Proof of Lemma 3.3.1

Proof. If we are using the lanczos method, we will have $v = v^e$ by definition. Now we only need to consider the AGD method branch. The problem we solve in AGD branch is:

$$\min \quad \frac{1}{2}v^T(NX\nabla f(\tilde{x})XN + \tilde{\mu}I)v + \nabla f(\tilde{x})XNv$$

Denote the above function as $F_{\tilde{\mu}}(v)$, we will have $F_{\tilde{\mu}}(v)$ is at least η_1-strongly convex:

$$\frac{\eta_1}{2}\|v - v^e\|^2 \leq F_{\tilde{\mu}}(v) - F_{\tilde{\mu}}(v^e) - \nabla F_{\tilde{\mu}}(v^e)(v - v^e)$$

As $\nabla F_{\tilde{\mu}}(v^e) = 0$, we will have:

$$\frac{\eta_1}{2}\|v - v^e\| \leq F_{\tilde{\mu}}(v) - F_{\tilde{\mu}}(v^e) \Rightarrow \|v^e\|^2 - \|v\|^2 \leq \frac{F_{\tilde{\mu}}(v) - F_{\tilde{\mu}}(v^e)}{\eta_1} \leq \frac{\eta_2}{L_1\eta_1}$$

Where the last inequality use the fact that $F_{\tilde{\mu}}(v) - F_{\tilde{\mu}}(v^e) \leq \frac{\eta_2}{L_1}$ if we adopt the AGD method. Since $F_{\tilde{\mu}}(v) = q(v) + \frac{1}{2}\tilde{\mu}\|v\|^2$, combine it with $F_{\tilde{\mu}}(v) - F_{\tilde{\mu}}(v^e) \leq \frac{\eta_2}{L_1}$:

$$q(v) - q(v^e) \leq \frac{\eta_2}{L_1} + \frac{1}{2}\tilde{\mu}(\|v^e\|^2 - \|v\|^2) \leq \frac{\eta_2(2+\tilde{\mu})}{2L_1\eta_1} \leq \eta_2\frac{2+\Lambda_{\max}}{2L_1\eta_1}$$

The last inequality is because $\tilde{\mu} < \mu_{\max} = \Lambda_{\max}$. If $\frac{1}{2}\mu^*\|v^*\|^2 \geq \frac{2}{3\sqrt{\rho^3}L_2}$, we will require $q_H(v) - q_H(v^e) \leq \frac{1}{12\sqrt{\rho^3}L_2^2}$ therefore we could set

$$\eta_2 \leq \frac{L_1\eta_1}{6\sqrt{\rho^3}L_2^2(2+\Lambda_{\max})} = \frac{L_1}{360\rho^2L_2^2(2+\Lambda_{\max})} \tag{3.4.55}$$

to fulfill this requirement.
If $\frac{1}{2}\mu^*\|\bar{d}^*\|^2 < \frac{2}{3\sqrt{\rho^3}L_2}$, as we know $q_H(V) > q_H(V^e)$. Therefore if $q_H(v^e)$ can determine an approximated KKT2 solution, $q_H(v)$ will also remain insufficient descent and stop at an approximated KKT2 solution. Combine with $q(d) = q_H(v)$, the final result follows.

□

3.4.9 Proof of Lemma 3.4.4

Proof. We first consider the case 1, where $A = (I\ I)$. Denote the diag matrix X as $\begin{pmatrix} x & \\ & s \end{pmatrix}$ where $x = \text{diag}(x_1, ..., x_{n/2})$ and $s = \text{diag}(x_{n/2+1}, ..., x_n)$ and we will have:

$$AX = \begin{pmatrix} I & I \end{pmatrix}\begin{pmatrix} x & \\ & s \end{pmatrix} = \begin{pmatrix} x & s \end{pmatrix}. \tag{3.4.56}$$

We can find one the null space is

$$N = \begin{pmatrix} -s \\ x \end{pmatrix} = \begin{pmatrix} -s_1 & & & \\ & -s_2 & & \\ & & \cdots & \\ & & & -s_n \\ x_1 & & & \\ & x_2 & & \\ & & \cdots & \\ & & & x_n \end{pmatrix}.$$

Since N is column orthogonal, we only need to normalized it into:

$$\tilde{N} = \begin{pmatrix} -s_1/\sqrt{s_1^2+x_1^2} & & & \\ & -s_2/\sqrt{s_2^2+x_2^2} & & \\ & & \cdots & \\ & & & -s_n/\sqrt{s_n^2+x_n^2} \\ x_1/\sqrt{s_1^2+x_1^2} & & & \\ & x_2/\sqrt{s_2^2+x_2^2} & & \\ & & \cdots & \\ & & & x_n/\sqrt{s_n^2+x_n^2} \end{pmatrix}. \tag{3.4.57}$$

The result related to the box constraints in **Lemma 3.4.4** follows. Next we focus on the case 2. By solve the following linear programming problem:

$$\min 0 \quad \text{s.t.} Ad = 0$$

we will be able to seperate A in to basis part B and non-basis part C. One can verify $\tilde{N} = \begin{pmatrix} -B^{-1}C \\ I \end{pmatrix}$ contains basis that span the null space of A. ($A\tilde{N} = (B\ C)\tilde{N} = -C + C = 0$ and $\text{rank}(\tilde{N}) = n - \text{rank}(A)$). N has a special structure. Except the first s rows, the remaining part is an identity matrix. Therefore with column operations we will be able to have a M matrix described in the **Lemma 3.4.4**. N can be transfer into M. Note that M only need to be calculate once and it has sparse structure. The computation cost and store cost can be very small.

Based on M, we only need to calculate the orthonormal matrix N. As $AM = 0$,

we will have $AXX^{-1}M = 0$, which means $X^{-1}M$ is a matrix that contains the basis spanning the null space of AX. Since X^{-1} is diagonal, sparse structure of $X^{-1}M$ is the same as M. If we use gram schmidt orthogonalization procedure to find N from $X^{-1}M$, the computation cost for find every orthonormal vector would be bounded by $O(sT_2)$. It is because the the $i-th$ column is naturally orthogonal to the $(i+s)-th$ column and the column after that. Hence to computation such N could only require $O(snT_2)$ cost. As $O(T_2)$ is the time for vector-vector multiplication and $O(T_1)$ is the upper bound for matrix-vector multiplication. We will have $O(snT_2) = O(sT_1)$. □

3.5 Conclusion

We discuss an accelerated interior point gradient method for nonconvex programming with linear constraints. We integrate the accelerated gradient descent method with Lanczos method and show that the worst case complexity of gradient based algorithm will be approximately upper bounded by $\tilde{O}(\epsilon^{-7/4})$ with high probability. Our method doesn't involve matrix inversion calculation that can be very time consuming for large scale optimization. Compared with classic first-order methods, our method breaks the $O(\epsilon^{-2})$ barriers.

Chapter 4
MCP Multi-Armed Bandit Model with High-Dimensional Covariates

4.1 Introduction

Individual-level data have become increasingly accessible in the Internet era, and decision-makers have accelerated data accumulation with extraordinary speed in a variety of industries, including health care, retail, advertising, etc. The growing availability of user-specific data, such as demographics, geographics, medical records, and searching/browsing history, provides decision-makers with unprecedented opportunities to tailor decisions to individual users. For example, doctors can personalize treatments for patients based on their medical history, clinical tests, and biomarkers; search engines can offer personalized advertisements for users based on their queries, demographics, and geographics. These user-specific data are often collected sequentially over time, during which decision-makers adaptively learn to predict the expected rewards based on users' responses to each available decision as a function of the user-specific data (i.e., the user's covariates) and optimally adjust decisions to maximize their rewards – an *online* learning and decision-making process.

This online learning and decision-making process requires a thoughtful balance between exploration and exploitation. Consider a decision-maker who selects decisions for incoming users and obtains rewards based on users' responses to these

decisions. To maximize his expected rewards, the decision-maker first needs an accurate predictive model for users' responses, which is typically uncertain at the beginning but can be partially learned through collecting samples of users' responses. On the one hand, the decision-maker could select a decision that yields the "highest", based on his best knowledge so far, expected reward (i.e., exploitation). Yet, this decision can be suboptimal, as the selection is based on the rough prediction of users' responses due to limited samples. Even worse, the decision-maker could incorrectly estimate the expected reward of the true optimal decision to be low and never have a chance to correct such a mistake (as the decision-maker will not select the true optimal decision due to the current low reward prediction, he will not generate additional samples to be able to learn and correct his incorrect estimation). On the other hand, the decision-maker can improve his predictive ability and learn users' responses by collecting more response samples, which often are obtained through random clinical trials and/or user experiments and are typically costly (i.e., exploration). The exploration and exploitation dilemma has been extensively studied in the multi-armed bandit model ([69]), but the growing dimensionality and availability of data have added another layer of complexity to the bandit model.

In practice, individual-level data are typically presented in a high-dimensional fashion, which poses significant computational and statistical challenges in the online learning and decision-making process. Traditional statistical methods, such as Ordinary Least Squares (OLS), require a large number of samples (e.g., the sample size must be larger than the covariate dimension) to be deemed computationally feasible. Under high-dimensional settings, learning the accurate predictive models requires a substantial amount of samples, which are obtained, if possible, through costly trials or experiments. Take the search advertising industry for example. Search advertising occurs when an Internet user searches certain keyword(s) (i.e., a query) in an online search engine and then the search engine displays both search results, in response to the user's query, and some sponsored ads, in response to the query and user-specific information. In order to select the ad that maximizes its revenue, the search engine must have accurate estimations on users' clicking probabilities in response to the displayed ads – Click-Through Rate (CTR).

However, the search engine's ability to accurately predict CTR is often crippled by the high-dimensional search advertising data coupled with limited samples. Counting more than three quarters of a million distinct words and their combinations

([66]), there are nearly infinite possible queries the user can submit to the search engine. For example, from 2003 to 2012, Google answered 450 billion unique queries, and it has estimated that 16% to 20% of queries submitted every day have never been used before ([58]). Hence, to accurately estimate a single ad's CTR to these queries, the search engine requires billions, if not trillions, of samples. The craving for samples will be further intensified if the search engine practices personalized advertising by taking users' individual information (such as demographics and geographics) into consideration. However, the available samples for the search engine to learn and predict CTR are greatly limited. Consider a 45 days new marketing campaign promoting a sales event or merchandise, during which time an average ad is expected to reach approximately one third of a million users ([79, 90]). Among these users, a very small portion can be selected to perform costly experiments to learn CTR, and that number is much smaller comparing to the size of queries and individual data.

In this work, we propose a new algorithm, the G-MCP-Bandit algorithm, for online learning and decision-making processes in high-dimensional settings. Our algorithm follows the ideas of the bandit model and develops a ϵ-decay random sampling method to balance the exploration-and-exploitation trade-off. We allow the decision-maker's reward function to follow the generalized linear model ([55]), which is a large class of models including the linear model, the logistic model, the Poisson regression model, etc., and we adopt the Minimax Concave Penalized (MCP) method ([94]) to improve the parameter estimations and predict the expected rewards in high-dimensional settings.

In the high-dimensional statistics literature, MCP is developed to explore and recover the latent sparse data structure for high-dimensional data. Compared to traditional statistical methods (e.g., OLS), MCP uses significantly fewer data samples and delivers better performance in high-dimensional settings ([94]). Although it is statistically favorable to adopt MCP, solving the MCP estimator (an NP-complete problem) could be computationally challenging. We propose a linear approximation method, the 2-step weighted Lasso procedure (2sWL), under the bandit setting as an efficient approach to tackle this challenge. We show that the MCP estimator solved by the 2sWL procedure matches the oracle estimator with high probability and converges to the true parameter with the optimal convergence rate. Since the bandit model mixes the exploitation and exploration phases, sam-

ples generated under the exploitation phase may be non-iid. Therefore, we adopt a matrix perturbation technique to derive new oracle inequalities for the MCP estimator under non-iid samples. To the best of our knowledge, this work is the first one that applies MCP to handle non-iid samples.

We theoretically demonstrate that the G-MCP-Bandit algorithm can significantly improve the cumulative regret bound in high-dimensional settings comparing to existing bandit algorithms. In particular, we benchmark the G-MCP-Bandit algorithm to an oracle policy, in which all parameter vectors are common knowledge, and adopt the expected cumulative regret (i.e., the difference in rewards achieved by the oracle policy and the G-MCP-Bandit algorithm) as the performance measure. We show that the cumulative regret of the G-MCP-Bandit algorithm over T users (i.e., a sample size of T) is at most $O(\log T)$, which is the optimal/lowest theoretical bound for all possible algorithms ([37]). Further, we show that the G-MCP-Bandit algorithm also attains a tight bound in the covariate dimension d, $O(\log d)$. We believe that our work is the first one in high-dimensional settings that attains the logarithmic dependence on both the sample size dimension and the covariate dimension, which are of particular importance in high-dimensional data with limited samples and suggest that the G-MCP-Bandit algorithm can bring substantial regret reduction comparing to existing bandit algorithms.

Through two synthetic-data-based experiments, we benchmark the G-MCP-Bandit algorithm's performance to other state-of-the-art bandit algorithms designed both in low-dimensional settings, OLS-Bandit by [37] and OFUL by [1], and in high-dimensional settings, Lasso-Bandit by [9]. We find that the G-MCP-Bandit algorithm performs favorably in both experiments. In particular, when the sample size is not extremely small[1], the G-MCP-Bandit algorithm appears to be able to accurately learn the parameter estimations with limited samples and therefore have the lowest cumulative regret. Furthermore, we observe that the benefits of the G-MCP-Bandit algorithm over other benchmark algorithms seems to increase with the data's sparsity level and the size of the decision set.

Finally, we evaluate the G-MCP-Bandit algorithm's performance through two real-data-based experiments, warfarin dosing data and Tencent search advertising data, where the technical assumptions specified for the theoretical analysis of the G-

[1]When the sample size is extremely small, the decision-maker has little information to learn. Therefore, all algorithms perform equally poorly.

MCP-Bandit algorithm's expected cumulative regret may not hold. We observe that the G-MCP-Bandit algorithm continues to perform favorably in both experiments. In particular, in the warfarin dosing experiment (formulated as a 3-armed bandit problem with 93 covariates), the G-MCP-Bandit algorithm needs the fewest patient samples (i.e., merely 50 patients) to provide better dosing decisions than actual physicians. Similarly, in the Tencent search advertising experiment (formulated as a 3-armed bandit problem with hundreds of thousands of covariates), the G-MCP-Bandit algorithm, after observing 140 users, can consistently generate better average revenue than other benchmark algorithms under the linear model. Further, we observe that the choice of the underlying reward model can significantly influence the G-MCP-Bandit algorithm's performance. In particular, under the logistic model, which is a special case of the generalized linear model, the G-MCP-Bandit algorithm merely needs 20 users to outperform other benchmark algorithms. This observation suggests that understanding the context of the underlying managerial problem and identifying the appropriate model for the G-MCP-Bandit algorithm can be critical and bring the decision-maker substantial revenue improvement.

4.2 Literature Review

This research is closely related to the exploration-exploitation trade off in the multi-armed bandit literature. [68, 80] follow the non-parametric approach and consider that the arm reward can be any smooth non-parametric function. Under this approach, the expected cumulative regret has an exponential dependence on the covariate dimension d, which is undesirable under high-dimensional settings where d can be extremely large. Such exponential dependence can be improved by following the parametric approach. [7] proposes the UCB algorithm for a linear bandit model, where the arm reward can be approximated by a linear combinations of covariates. Since [7], other UCB-type algorithms(e.g., [2, 25, 26, 73]) and Bayesian-type algorithms (e.g., [5, 74]) have been proposed and shown to improve on the expected cumulative regret. Yet, allowing the adversary and without regulating the sample generating process, the statistical performance of the parameter vector estimation in the learning process may suffer. As a result, the expected cumulative regret bound typically has a sublinear dependence on the sample size dimension T (e.g., $O(\sqrt{T})$) and a polynomial dependence on the covariate dimension d. However,

in high-dimensional settings, where the covariate dimension and the sample size dimension can be exceedingly large, these algorithms can perform poorly.

By introducing a forced sampling approach to the linear bandit model, [37] ensure that enough samples generated in their algorithm possess desired iid property and show that their proposed OLS-Bandit algorithm can achieve $O(\log T)$ dependence on the sample size dimension T in low-dimensional settings. Following a similar approach, [9] propose the Lasso-Bandit algorithm, which attains a poly-logarithmic dependence on the sample size dimension $O(\log^2 T)$ and the covariate dimension $O(\log^2 d)$ in high-dimensional settings. In this work, we allow the reward function to follow the generalized linear model, which contains a wide family of models that includes the linear bandit model. We propose a ϵ-decay random sampling method and show that our proposed G-MCP-Bandit algorithm continues to achieve the optimal cumulative regret bound on the sample size dimension $O(\log T)$ and attain a tight bound in the covariate dimension $O(\log d)$ in high-dimensional settings. We believe that our work is the first one that attains the logarithmic dependence on both the sample size dimension and the covariate dimension in high-dimensional settings.

Our research is also connected to the statistical learning literature. In high-dimensional statistics, Lasso type methods ([83]) have become the golden standard for high-dimensional learning ([56, 57, 86, 95]). Yet, Lasso-type regularizations may lead to estimation bias, and strong conditions are needed for analyzing its theoretical performance guarantee ([28]). Recently, [94] proposes MCP, a non-convex penalty method, which entails better statistical properties, such as the unbiasedness and a strong oracle property for high-dimensional sparse estimation, and requires weaker conditions than Lasso ([33, 56, 101]). Although it is statistically favorable to adopt MCP, solving the MCP estimator (an NP-complete problem) could be computationally challenging ([49, Liu et al.]). Various approximation methods have been developed in the literature. For example, [29] use the local quadratic approximation, [31, 33, 98, 101] adopt the local linear approximation, [94] choose the path following algorithm, and [Liu et al.] propose the second-order approximation. Our proposed solution procedure (the 2sWL procedure) is analogous to the local linear approximation and guarantees that the solution has desirable statistical properties for theoretical analysis and can be efficiently solved. In the literature, the theoretical analysis of MCP's statistical properties relies on the

assumption that all samples are iid, which is hardly the case under bandit models. This work also contribute to the statistical learning literature by deriving new oracle inequalities for MCP under non-iid samples.

4.3 Model Settings

Consider a sequential arrival process $t \in \{1, 2, ..., T\}$. At each time step t, a single user (e.g., consumer or patient), described by a high-dimensional feature covariate vector $\boldsymbol{x}_t \in \mathbb{R}^{1\times d}$, arrives. The covariate vector combines all available (but not necessarily valuable for the decision-maker to base his decision on) user-specific data, such as demographics, geographics, browsing/shopping history, and medical records. Upon arrival, users' covariate vectors $\{\boldsymbol{x}_t\}_{t\geq 0}$ become observable to the decision-maker and are iid distributed according to an unknown distribution $\mathcal{P}_x$.

Based on the user's covariate vector $\boldsymbol{x}$, the decision-maker will select a decision from a decision set $\mathcal{K} = \{1, 2, ..., K\}$ to maximize his expected reward. The user will respond to the chosen decision $k \in \mathcal{K}$, and such response will generate a reward for the decision-maker. Take the search advertising for example. The search engine can recommend one of K different ads to the user; the user can respond to the recommended ad by clicking, which generates revenue for the search engine. We denote this reward under the chosen decision k as R_k, which follows a distribution $\mathbb{P}(R_k|\boldsymbol{x}^T\boldsymbol{\beta}_k^{true})$, where $\boldsymbol{x}$ is the user's covariate vector and $\boldsymbol{\beta}_k^{true}$ is the unknown parameter vector corresponding to decision k.

We present the reward function in terms of the generalized linear model ([55]), which is a large class of models including the linear model, the logistic model, the Poisson regression model, etc. For example, if we assume that R_k is a σ-gaussian random variable with mean $\boldsymbol{x}^T\boldsymbol{\beta}_k^{true}$, then we can define the density function of the distribution $\mathbb{P}(R_k|\boldsymbol{x}^T\boldsymbol{\beta}_k^{true})$ as $g(R_k = r|\boldsymbol{x}^T\boldsymbol{\beta}_k^{true}) = (1/\sqrt{2\pi\sigma^2})\exp(-\frac{(r-\boldsymbol{x}^T\beta_k^{true})^2}{2\sigma^2})$, which is the standard setting for the classic linear multi-armed bandit model where the reward takes a linear form: $R_k(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}_k^{true} + \epsilon$ ([5, 7]). The cumulative regret performance of the linear bandit algorithms has been extensively studied by [25] and [37], among others, under low-dimensional settings and by [9] under high-dimensional settings. The generalized linear model adopted in this work facilitates us to go beyond the classic linear bandit model, as the reward may take a nonlinear form in practice. For instance, the search engine collects revenue only when a user has

clicked the recommended ad; otherwise, the search engine earns nothing – a logistic model by nature. By specifying R_k as a binary random variable (e.g., $R_k \in \{0, 1\}$), we can define the mass function of the distribution $\mathbb{P}(R_k|\boldsymbol{x}^T\boldsymbol{\beta}_k^{true})$ as $g(R_k = 1|\boldsymbol{x}^T\boldsymbol{\beta}_k^{true}) = 1/(1+\exp(-\boldsymbol{x}^T\boldsymbol{\beta}_k^{true}))$ and $g(R_k = 0|\boldsymbol{x}^T\boldsymbol{\beta}_k^{true}) = \exp(-\boldsymbol{x}^T\boldsymbol{\beta}_k^{true})/(1+\exp(-\boldsymbol{x}^T\boldsymbol{\beta}_k^{true}))$, which is a logistic bandit model with the binary reward ([27, 75, 76]).

The parameter vector $\boldsymbol{\beta}_k^{true}$ is high-dimensional with latent sparse structure, and we denote $\mathcal{S}_k = \{j : \beta_{k,j}^{true} \neq 0\}$ as the index set for significant covariates, which have non-zero coefficient parameters and therefore are important for the decision-maker to predict the user's response. This index set is also unknown to the decision-maker. We define the number of significant covariates as $|\mathcal{S}_k|$, which is typically much smaller than the dimension of the covariate vector.

The decision-maker's objective is to maximize his expected cumulative reward. Denote the decision-maker's current policy as $\pi = \{\pi_t\}_{t\geq 0}$, where $\pi_t \in \mathcal{K}$ is the decision prescribed by policy π at time t. To benchmark the performance of policy π, we first introduce an *oracle policy* $\pi^* = \{\pi_t^*\}_{t\geq 0}$ under which the decision-maker knows the true parameter vector values $\boldsymbol{\beta}_k^{true}$ for all $k \in \mathcal{K}$ and chooses the best decision to maximize his expected reward:

$$\pi_t^* = \arg\max_{k\in\mathcal{K}} \left\{\mathbb{E}[R_k|\boldsymbol{x}_t, \boldsymbol{\beta}_k^{true}]\right\} = \arg\max_{k\in\mathcal{K}} \left\{\int_{-\infty}^{+\infty} r_k dG(r_k|\boldsymbol{x}_t^T\boldsymbol{\beta}_k^{true})\right\},$$

where $G(r_k|\boldsymbol{x}_t^T\boldsymbol{\beta}_k^{true})$ is the cumulative distribution function for R_k. Note that in practice, the parameter vector $\boldsymbol{\beta}_k^{true}$ is unknown to the decision-maker, and therefore the construction and definition of the oracle policy directly imply that the decision-maker's reward under policy π is upper-bounded by that of the oracle policy. We therefore define the decision-maker's expected cumulative regret up to time T under the policy π as follows:

$$R^C(T) = \sum_{t=1}^{T} \mathbb{E}[R_t^{\pi_t^*} - R_t^{\pi_t}],$$

which is the expected reward difference between the optimal policy π^* and the decision-maker's alternative policy π. To maximize his expected cumulative reward, the decision-maker is equivalent to explore for the policy π that minimizes the cumulative regret up to time T.

Before presenting the proposed G-MCP-Bandit algorithm, we will first state five technical assumptions necessary for the theoretical analysis of the decision-maker's expected cumulative regret. The first three assumptions are adopted directly from the multi-armed bandit literature, and the last two assumptions from the high-dimensional statistics literature.

A. 1 (Parameter set) There exist positive constants $x_{\max}$, s, $R_{\max}$, $\beta_{\min}$ and b such that for any t and $k \in \mathcal{K}$, we have $\|\boldsymbol{x}_t\|_\infty \leq x_{\max}$, $|\mathcal{S}_k| \leq s$, $|R_k| \leq R_{\max}$, $\beta_{\min} \leq \min_{j \in \mathcal{S}_k, k \in \mathcal{K}} |\beta_{k,j}^{true}|$, $\|\boldsymbol{\beta}_k^{true}\|_1 \leq b$ and all feasible β satisfies $\|\boldsymbol{\beta}\|_1 \leq b$.

The first assumption is a standard assumption in the bandit literature ([73]) and ensures that both the covariate vector $\boldsymbol{x}$ and the coefficient vector $\boldsymbol{\beta}_k$ are upper bounded so that the maximum regret at every time step will also be upper bounded to avoid trivial decisions. Most real world applications, including two real data experiments in §4.6.2 and §4.6.3, satisfy this assumption.

A. 2 (Margin condition) There exists a $C > 0$ such that $\mathbb{P}(0 < |\mathbb{E}[R_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{true}] - \mathbb{E}[R_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{true}]| \leq \gamma) \leq C R_{\max} \gamma$ for $i \neq j$ and $i, j \in \mathcal{K}$.

The second assumption is first introduced in the classification literature by [85]. [37] and [9] adopt this assumption to the linear bandit model, under which the Margin Condition ensures only a fraction of covariates can be drawn near the boundary hyperplane $\boldsymbol{x}^T(\boldsymbol{\beta}_i^{true} - \boldsymbol{\beta}_j^{true}) = 0$ in which rewards for both arms are nearly equal. Clearly, if a large proportion of covariates are drawn from the vicinity of the boundary hyperplane, then for any bandit algorithm, a small estimation error in the decision parameter vectors may lead the decision-maker to choose the suboptimal decision and perform poorly ([9]). Therefore, this margin condition ensures that given a user's covariate vector, decisions can be properly separated from each other and ordered based on their rewards.

A. 3 (Arm optimality) There exists a partition $\mathcal{K}_o$ and $\mathcal{K}_s$ for $\mathcal{K}$. For $k_1 \in \mathcal{K}_s$, we will have $\mathbb{E}[R_{k_1}|\boldsymbol{x}, \boldsymbol{\beta}_{k_1}^{true}] + h < \max_{k \neq k_1} \mathbb{E}[R_k|\boldsymbol{x}, \boldsymbol{\beta}_k^{true}]$ for a positive constant h for every $\boldsymbol{x}$. For $k_2 \in \mathcal{K}_o$, these exists another positive constant p^* such that $\min \mathbb{P}(\boldsymbol{x} \in U_{k_2}) \geq p^*$, where $U_{k_2} \doteq \left\{\boldsymbol{x} | \mathbb{E}[R_{k_2}|\boldsymbol{x}, \boldsymbol{\beta}_{k_2}^{true}] > \max_{k \neq k_2} \mathbb{E}[R_k|\boldsymbol{x}, \boldsymbol{\beta}_k^{true}] + h, k \in \mathcal{K}\right\}$.

The arm optimality condition ([9, 37]) ensures that as the sample size increases, the parameter vectors for optimal decisions can eventually be learned. In particular, this condition separates decisions to an optimal decision subset $\mathcal{K}_o$ and a suboptimal decision subset $\mathcal{K}_s$. Decision i in $\mathcal{K}_o$ is strictly optimal for some users' covariate vectors (denoted by set U_i); otherwise, decision j in $\mathcal{K}_s$ must be strictly suboptimal

for all users' covariate vectors. Therefore, even if there is a small estimation error for decision i in $\mathcal{K}_o$, the decision-maker will be more likely to choose decision i for a user with a covariate vector draw from the set U_i. Accordingly, as sample size T increases, decision-makers can improve their estimations for optimal arms' parameter vectors.

These first three assumptions are directly adopted from the multi-armed bandit literature and have been shown to be satisfied for all discrete distributions with finite support and a very large class of continuous distributions (see [9] for detailed examples and discussions).

A. 4 (Restricted eigenvalue condition) There exists $\kappa > 0$ such that for all feasible $\boldsymbol{\xi}$ satisfying $\|\xi\|_1 \leq b$ and $\boldsymbol{u}$ such that $\|\boldsymbol{u}_{\mathcal{S}_k}^c\|_1 \leq 3\|\boldsymbol{u}_{\mathcal{S}_k}\|_1$, we have $\frac{\kappa}{s}\|\boldsymbol{u}_{\mathcal{S}_k}\|_1^2 \leq \boldsymbol{u}^T\mathbb{E}[\nabla^2\mathcal{L}(\boldsymbol{\xi})]\boldsymbol{u}$, where $\mathcal{L}$ is the log likelihood function, $\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{j=1}^n -\log g(r_j|\boldsymbol{x}_j^T\boldsymbol{\beta})$, and $\{\boldsymbol{x}_j, j = 1, 2, ..., n\}$ are iid random samples with $\boldsymbol{x}_j \in U_k, k \in \mathcal{K}$.

The restricted eigenvalue condition assumption is a standard assumption in high-dimensional statistics and is necessary for the identifiability and consistency of high-dimensional estimators ([31, 33]). This assumption considers the local geometry of the log likelihood function $\mathcal{L}$ with iid samples in U_k. To intuit, note that under low-dimensional settings, the literature ([59]) requires that $\mathcal{L}$ is strongly convex around the true parameter vector $\boldsymbol{\beta}^{true}$ (e.g., the Hessian matrix in OLS estimator is positive-definite and invertible) in order to achieve identifiability of the parameter vector. However, the strong convexity assumption is typically violated in high-dimensional settings, as the sample size can be much smaller than the covariate dimension. Therefore, a weaker condition is adopted: The $\mathcal{L}$ exhibits local strongly convex behavior only in some restricted subspace of u. In high-dimensional linear models, the restricted eigenvalue condition assumption is analogous to the compatibility condition ([9, 14]), restrict strongly convexity condition ([52, 61]), and sparse eigenvalue condition ([31, 96]).

A. 5 (Density function) The negative logarithm of the reward density function $f(r|y) \doteq -\log g(r|y)$ is (i) convex with smooth gradient and hessian in y, and (ii) there exists positive constants σ, σ_2 and σ_3 such that $|f^{'}(r|y)| \leq \sigma$, $f^{''}(r|y) < \sigma_2$ and $|f^{'''}(r|y)| \leq \sigma_3$.

The density function assumption enables us to use the estimated expected reward to statistically infer the true expected reward. Specifically, under this assumption, when the parameter estimator $\boldsymbol{\beta}$ is close enough to the underlying

true parameter vector $\boldsymbol{\beta}^{true}$, the negative logarithm of the reward density function under the estimator $\boldsymbol{\beta}$, $g(\boldsymbol{x}^T\boldsymbol{\beta})$, will converge to that under the true parameter vector $\boldsymbol{\beta}^{true}$, $g(\boldsymbol{x}^T\boldsymbol{\beta}^{true})$. The density function assumption is a fairly weak technical assumption. Many common distributions, such as sub-Gaussian distribution and Bernoulli distribution, satisfy this density function assumption.

4.4 G-MCP-Bandit Algorithm

One of the major challenges for online learning and decision-making problems is discovering the underlying sparse data structure and estimating the parameter vector for high-dimensional data with limited samples. Lasso ([83]) has been proposed as an efficient statistical learning method and adopted in the multi-armed bandit literature ([9]) to hurdle this challenge. However, the Lasso estimator can be biased and performs inadequately, especially when the magnitude of true parameters is not too small ([29]). One way to address this performance issue is to construct new penalty functions that could render unbiased estimators and improve the sparse structure discovery under high-dimensional data with limited samples. In this research, we will adopt the novel MCP method.

4.4.1 Parameter Vector Estimation

For notation convenience, we will omit parameters' subscripts corresponding to the choice of arms, as long as doing so will not cause any misinterpretation. Consider an oracle estimator for an arbitrary arm, $\boldsymbol{\beta}^{oracle}$, which is the parameter estimator when the decision-maker has perfect knowledge of the index set for significant covariates $\mathcal{S}$. In other words, the oracle estimator can be determined by setting $\beta_j = 0$ for $j \in \mathcal{S}^c$ and solving

$$\boldsymbol{\beta}^{oracle}(\boldsymbol{X}, \boldsymbol{r}) \doteq \arg \min_{\substack{\boldsymbol{\beta}_{\mathcal{S}^c}=0 \\ \boldsymbol{\beta}_{\mathcal{S}}}} \left\{ \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} f(r_j | \boldsymbol{x}_j^T \boldsymbol{\beta}) \right\}, \tag{4.4.1}$$

where $\mathcal{A}$ is the available historical data samples and $f(\cdot|\cdot)$ is the negative logarithm of the reward density function defined early. When solving for the oracle estimator, the decision-maker can directly ignore insignificant covariates by forcing their corresponding coefficients to be zero and essentially reduce the high-dimensional

problem to a low-dimensional counterpart. The statistical performance of the oracle estimator is provided in the following lemma.

Lemma 4.4.1. *Let n be the sample size. Under assumption A.1, A.4, and A.5, the following inequality for the oracle estimator holds*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{\frac{8s^2\sigma^2 x_{\max}^2}{\mu_0^2 n}}\right) \geq 1 - \delta_1(n), \tag{4.4.2}$$

where $\delta_1(n) \doteq 2\exp(-\frac{C_h n \mu_0}{2s x_{\max}^2}) + s\exp(-\frac{\mu_0 n}{8s\sigma_2 x_{\max}^2})$, and C_h and μ_0 are positive constants.

Since there are only $|\mathcal{S}|$ significant covariates, which is upper-bounded by s, are free to change in Equation (4.4.1), the optimal statistical performance of the likelihood estimation is commonly recognized as $O(\sqrt{s/n})$ in the literature ([31, 99]), which doesn't include the dependence of the largest eigenvalue in the objective function's Hessian matrix. In Equation (4.4.2), we explicitly include its influence and can directly verify that the largest eigenvalue in the objective function's Hessian matrix is universally upper bounded by $\sigma_2 s x_{\max}^2$ and therefore Equation (4.4.2) reduces to $O(\sqrt{s/n})$ dependence. In other words, the oracle estimator attains the optimal statistical performance.

However, the significant covariates index set $\mathcal{S}$ is typically unknown to the decision-maker in practice, and we will rely on the MCP method to recover this latent sparse structure. To better understand the rationale behind the MCP method, we start with the following weighted Lasso estimator:

$$\boldsymbol{\beta}^W(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) \doteq \arg\min_{\beta} \left\{\frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}} f(r_j|\boldsymbol{x}_j^T\boldsymbol{\beta}) + \sum_{i=1}^{d} w_i|\beta_i|\right\}, \tag{4.4.3}$$

where $\boldsymbol{w} = (w_1, w_2, ..., w_d)$ is a positive weights vector chosen by the decision-maker. Note that when we set $w_i = \lambda$ for all i, $\boldsymbol{\beta}^W(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w})$ reduces to the standard Lasso estimator, which can be biased when the magnitude of true parameters is not too small. To recover the sparse structure and provide an unbiased parameter estimator, an ideal way to select $\{w_i\}$ is to set $w_i = \lambda > 0$ for all $i \in \mathcal{S}^c$ and $w_j = 0$ for all $j \in \mathcal{S}$. By doing so, when the weight λ is large enough, the weighted Lasso estimator converges to the oracle estimator $\boldsymbol{\beta}^{oracle}(\boldsymbol{X}, \boldsymbol{r})$. The benefits of the weighted Lasso method have attracted considerable attention recently, and various

mechanisms have been proposed in the literature aiming to improve the weight selection process ([17, 40, 101]). The MCP method, adopted in our work, reflect such a process.

In particular, we define the following MCP penalty function:

$$P_{\lambda,a}(x) \doteq \int_0^{|x|} \max\left(0, \lambda - \frac{1}{a}|t|\right) dt,$$

where a and λ are positive parameters selected by the decision-maker, and the MCP estimator can be presented as follows:

$$\boldsymbol{\beta}^{MCP}(\boldsymbol{X}, \boldsymbol{r}, \lambda) \doteq \arg\min_{\beta} \mathcal{L}_{\mathcal{A}_k}(\boldsymbol{\beta}) = \arg\min_{\beta} \left\{ \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} f(r_j | \boldsymbol{x}_j^T \boldsymbol{\beta}) + \sum_{i=1}^{d} P_{\lambda,a}(\beta_i) \right\}. \tag{4.4.4}$$

Denote the index set for non-zero coefficients solutions in Equation (4.4.4) as $\mathcal{J} \doteq \{j : \hat{\beta}_j \neq 0\}$. If the absolute value of the MCP estimator in $\mathcal{J}$ is greater than $a\lambda$, then $P_{\lambda,a}(\beta_j)$ become constant parameters for all $j \in \mathcal{J}$. Therefore, we will have $P_{\lambda,a}(\beta_j) = \frac{1}{2}a\lambda^2$ for $j \in \mathcal{J}$ and $P_{\lambda,a}(\beta_j) = 0$ otherwise. In other words, the statistical performance of solving the MCP estimator is equivalent to solving the following problem: $\arg\min_{\boldsymbol{\beta}_{\mathcal{J}^c}=0, \boldsymbol{\beta}_{\mathcal{J}}} \{\frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} f(r_j | \boldsymbol{x}_j^T \boldsymbol{\beta})^2\}$. Hence, if $\mathcal{J} = \mathcal{S}$, then the MCP estimator converges to the oracle estimator.

Solving the MCP estimator can be challenging. [Liu et al.] have shown that it is an NP-complete problem to find the MCP estimator by globally solving Equation (4.4.4). In the next subsection, we propose a local linear approximation method, the 2-step Weighted Lasso (2sWL) procedure, to tackle this challenge, and we demonstrate that the estimator solved by the 2sWL procedure will match the oracle estimator $\boldsymbol{\beta}^{oracle}$ with high probability.

4.4.2 2-Step Weighted Lasso Procedure

The 2sWL procedure consists of two steps. We first solve a standard Lasso problem by setting all positive weights in Equation (4.4.3) to a given parameter λ_0. Then, we use the Lasso estimator obtained in the first step to update the weights vector $\boldsymbol{w}$ by taking the first-order derivatives of the MCP penalty function, and applying this updated weight vector, we re-solve the weighted Lasso problem in Equation

(4.4.3) to obtain the MCP estimator. The procedures of 2sWL at time t can be described as follows:

2-Step Weighted Lasso (2sWL) Procedure:

Require: input parameters a and λ

Step 1: solve a standard Lasso problem

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}^W(\boldsymbol{X}, \boldsymbol{y}, \lambda);$$

Step 2: update $w_j = \begin{cases} P'_{a,\lambda}(|\beta_{1,j}|) & \text{, for } \beta_{1,j} \neq 0 \\ \lambda & \text{, for } \beta_{1,j} = 0 \end{cases}$

and solve a weighted Lasso Problem

$$\hat{\boldsymbol{\beta}}_{2sWL} = \boldsymbol{\beta}^W(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}).$$

As the 2sWL procedure is equivalent to solving the Lasso problem twice, the worst-case computation complexity for 2sWL is on same order as for the standard Lasso problem. In practice, we can initialize the second step procedure with a warm start from the first step of the Lasso solution, which further reduces the computation time.

The following proposition shows that the MCP estimator identified by the 2sWL procedure can recover the oracle estimator with high probability.

Proposition 4.4.2. *Under assumptions A.1, A.4, and A.5, if* $\min\{|\beta_j^{true}|, \beta^{true} \neq 0, j = 1, 2, ..., d\} \geq \left(\frac{96s}{\kappa} + a\right)\lambda$, $a > \frac{96s}{\kappa}$, *the MCP estimator solved under the 2sWL procedure,* $\boldsymbol{\beta}^{MCP}$ *satisfies the following inequality*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{\frac{8s^2\sigma^2 x_{\max}^2}{\mu_0^2 n}}\right) \geq 1 - \delta_1(n) - \delta_2(n, n, \lambda) - \delta_3(n), \quad (4.4.5)$$

where $\delta_2(n, n_1, \lambda) \doteq d \exp\left(-\frac{n\lambda^2}{2x_{\max}^2}\left(\left(\frac{1}{4} - \frac{24ns}{n_1 \kappa a}\right) \min\left\{1, \frac{n\mu_0}{8n_1 s x_{\max}^2}\right\}\right)^2\right)$, $\delta_3(n) \doteq \exp(-C_1 n)$, μ_0 *and* C_1 *are positive constants.*

Comparing to the oracle estimator β^{oracle} in Lemma 4.4.1, the probability bound on the MCP estimator under the 2sWL procedure has two extra terms $\delta_2(n, n, \lambda)$ and $\delta_3(n)$, which depend on the covariate dimension d and the sample size n. Note that as the sample size increases, these two extra terms decrease to 0 at exponential rates. In other words, as the sample size increases, $\boldsymbol{\beta}^{MCP}$ matches the oracle parameters with high probability and converges to the true parameters at the optimal convergence rate.

4.4.3 ϵ-decay Random Sampling Method

As bandit models involve exploitation and exploration, samples generated under exploitation typically are not iid. These non-iid samples pose challenges to the existing MCP literature, which relies on the assumption that samples are iid in establishing the convergence rate and regret bounds (see the proof of Proposition 4.4.2 in §4.4.2).

In this research, to ensure that there are some iid samples generated in the online learning and decision-making process, we propose a ϵ-decay random sampling method, in which the decision-maker draws random samples, with decreasing probability, by randomly selecting decisions from the decision set with equal probability. In particular, the ϵ-decay random sampling method can be described as follows:

ϵ-decay Random Sampling Method: At time t, the decision-maker will draw a random sample, with probability $\min\{1, t_0/t\}$, where t_0 is a pre-determined positive constant. If the seller has decided to draw a random sample at time t, then the decision-maker will randomly select a decision from his decision set with equal probability. Otherwise, the decision-maker will follow a bi-level decision structure, which will be specified later, to determine the optimal decision to maximize his expected reward.

The ϵ-decay random sampling method can balance the exploitation and exploration trade-off by ensuring that the decision-maker does not explore too much to significantly sacrifice his revenue performance (as the number of random samples decays in time) but has sufficient random samples to guarantee the quality of the parameter vector estimation. In particular, we can bound the random sample size in the following proposition.

Proposition 4.4.3. *Let $C_0 \geq 10$, $T > \frac{(t_0+1)^2}{e^2}$, and $t_0 = 2C_0|\mathcal{K}|$. Under the ϵ-decay random sampling method, the random sample size n_k for arm $k \in \mathcal{K}$ up to time T is bounded by*

$$C_0(1 + \log(T+1) - \log(t_0+1)) \leq n_k \leq 3C_0(1 + \log(T) - \log(t_0))$$

with probability at least $1 - 2/(T+1)$.

4.4.4 G-MCP-Bandit Algorithm

After establishing the MCP estimator's statistical property and the ϵ-decay random sampling method, we are ready to present the proposed G-MCP-Bandit algorithm. The execution of the G-MCP-Bandit algorithm can be summarized as follows:

G-MCP-Bandit Algorithm

Require: Input parameters $t_0, h, \lambda_{1,0}, \lambda_{2,0}, a$.

Initialize $\boldsymbol{\beta}_i^{random}(0) = \boldsymbol{\beta}_i^{whole}(0) = \mathbf{0}$, and $\mathcal{R}_{\pi_0} = \mathcal{W}_{\pi_0} = \phi$ for all $i \in \mathcal{K}$.

For $t = 1, 2, ...$ **do**

Observe $\boldsymbol{x}_t$.

Draw a binary random variable $\mathcal{D}_t$, where $\mathcal{D}_t = 1$ with probability $\min\{1, t_0/t\}$.

If $\mathcal{D}_t = 1$

Assign π_t to a random decision $k \in \mathcal{K}$ with probability $\mathbb{P}(\pi_t = k) = 1/|\mathcal{K}|$.

Play decision π_t and observe r_t

Update $\mathcal{R}_{\pi_t} = \mathcal{R}_{\pi_{t-1}} \cup \{\boldsymbol{x}_t, r_t\}$ and $\mathcal{W}_{\pi_t} = \mathcal{W}_{\pi_{t-1}} \cup \{\boldsymbol{x}_t, r_t\}$.

Else

Construct the optimal decision set:

$\Pi_t = \Big\{ i : \mathbb{E}[R_i|\boldsymbol{x}_t, \boldsymbol{\beta}_i^{random}(t-1)] \geq \max_{j \in \mathcal{K}} \mathbb{E}[R_j|\boldsymbol{x}_t, \boldsymbol{\beta}_j^{random}(t-1)] - \frac{1}{2}h,\ i \in \mathcal{K} \Big\}$.

If Π_t is a singleton

Set $\pi_t = \Pi_t$.

Else

Set $\pi_t = \arg\max_{k \in \Pi_t} \mathbb{E}[R_k|\boldsymbol{x}_t, \boldsymbol{\beta}_k^{whole}(t-1)]$.

End If

Play decision π_t, observe r_t, and update $\mathcal{W}_{\pi_t} = \mathcal{W}_{\pi_{t-1}} \cup \{\boldsymbol{x}_t, r_t\}$.

End If

For all $k \in \mathcal{K}$, set $\lambda_1(t) = \lambda_{1,0}\sqrt{1 + \frac{\log d}{\log(t+1)}}$ and $\lambda_2(t) = \lambda_{2,0}\sqrt{\frac{\log(t+1)+\log d}{t+1}}$.

Update parameters $\boldsymbol{\beta}_k^{random}(t)$ via the 2sWL procedure with $(\mathcal{R}_{\pi_t}, \lambda_1(t))$.

Update parameters $\boldsymbol{\beta}_k^{whole}(t)$ via the 2sWL procedure with $(\mathcal{W}_{\pi_t}, \lambda_2(t))$.

End for

Specifically, the decision-maker will start by assigning values for system parameters (t_0, $\mathcal{K}$, $s_{\max}$, and h), which can be optimized through tuning, and initialing two parameter vector estimators ($\boldsymbol{\beta}^{random}$ and $\boldsymbol{\beta}^{whole}$) and two sample datasets ($\mathcal{R}_{\pi_0}$ and $\mathcal{W}_{\pi_0}$, which represent the random sample set and the whole sample set,

respectively). Then, for an incoming user at time t, the decision-maker will draw a random sample with probability $\min\{1, t_0/t\}$. There are two possibilities:

- If the decision-maker decides to draw a random sample, then he will randomly choose a decision k from his decision set $\mathcal{K}$ with equal probability of $1/|\mathcal{K}|$; then, he will implement the chosen decision (i.e., $\pi_t = k$), observe the user's response, and claim the corresponding reward; finally, the decision-maker will include the user's covariate vector and the corresponding reward $\{\boldsymbol{x}_t, r_t\}$ in both sample datasets, $\mathcal{R}_{\pi_t}$ and $\mathcal{W}_{\pi_t}$.

- If the decision-maker decides not to draw a random sample on this incoming user, then he will use the bi-level decision structure to determine his decision. In the upper-level decision-making process, the decision-maker will first construct an optimal decision set Π_t. Specifically, all decisions in the optimal decision set Π_t are estimated, based on the random sample MCP estimator $\boldsymbol{\beta}^{random}$, to yield expected rewards within $h/2$ of the maximum possible reward. If there is only one decision in the optimal decision set Π_t, then the decision-maker will implement this decision as the optimal decision; otherwise, the decision-maker will perform the lower-level decision-making process, in which the decision-maker will estimate, by using the whole sample MCP estimator $\boldsymbol{\beta}^{whole}$, the rewards for all decisions in the optimal decision set Π_t and select the decision that generates the highest expected reward. Then, observing the user's response to the optimal decision and collecting the corresponding reward, the decision-maker will only update the whole sample dataset $\mathcal{W}_{\pi_t}$ by appending the user's covariate vector and the corresponding reward $\{\boldsymbol{x}_t, r_t\}$.

Finally, the decision-maker will reset two parameters, λ_1 and λ_2, and use the 2sWL procedure to update the random sample parameter vector estimator $\boldsymbol{\beta}^{random}$ and the whole sample parameter vector estimator $\boldsymbol{\beta}^{whole}$, based on sample data sets $\mathcal{R}_{\pi_t}$ and $\mathcal{W}_{\pi_t}$, respectively.

The expected cumulative regret upper bound for the G-MCP-Bandit algorithm can be established in the following theorem.

Theorem 4.4.4. *Under assumptions A.1-A.5, let* $t_0 = 2C_0|\mathcal{K}|$, $T \geq T_0$, $\lambda_{1,0} = \frac{\beta_{\min} p^* \kappa}{(2304s + ap^*\kappa)\sqrt{1+\log d}}$, $\lambda_{2,0} = \frac{\sqrt{2}x_{\max}^2}{\frac{1}{4} - \frac{192}{p^*\kappa a}\min\{1, \frac{\mu_0}{p^* s x_{\max}^2}\}}$, *and* $a \geq \frac{2304s}{\kappa p^*}$. *The cumulative*

regret of the G-MCP-Bandit algorithm up to time T is upper bounded:

$$R^C(T) \leq (6R_{\max}|\mathcal{K}|C_0 + 31R_{\max}|\mathcal{K}| + 2e^{4\sigma x_{\max} b} C R_{\max}^3 |\mathcal{K}| x_{\max}^2 C_\beta s^3) \log(T+1) + R_{\max}(T_0 + |\mathcal{K}|) = O(|\mathcal{K}| s^2 (s + \log d) \log T),$$

where T_0, C_0, C_h, μ_0 and C_β are constants independent of T.

Theorem 4.4.4 shows that the expected cumulative regret of the G-MCP-Bandit algorithm over T users is upper-bounded by $O(\log T)$. [37] have shown that under low-dimensional settings, the expected cumulative regret for a linear bandit model is lower-bounded by $O(\log T)$, which is directly applicable to the high-dimensional settings. Further, note that the linear model is a special case of the generalized linear model. Therefore, the expected cumulative regret of the G-MCP-Bandit algorithm is also lower-bounded by $O(\log T)$. In other words, the G-MCP-Bandit algorithm achieves the optimal expected cumulative regret in the sample size dimension. This result comes from the facts that we can ensure $O(\log T)$ random samples at time T via the ϵ-decay random sampling method (Proposition 4.4.3) and that the MCP estimator is able to match the oracle estimator with high probability (Proposition 4.4.2). Further, when compared to the Lasso-Bandit algorithm proposed by [9] for the linear model under high-dimensional settings, the G-MCP-Bandit algorithm reduces the dependence of the expected cumulative regret on the sample size dimension from $O(\log^2 T)$ to $O(\log T)$. As the G-MCP-Bandit algorithm achieves the optimal expected cumulative regret and improves on the cumulative regret performance from existing high-dimensional bandit algorithms in the sample size dimension, we expect that the G-MCP-Bandit algorithm will be able to improve the learning process of the parameter vector estimation with limited samples and perform favorably in the cumulative regret performance even in sample-poor regions.

Theorem 4.4.4 also demonstrates that the cumulative regret of the G-MCP-Bandit algorithm in the high-dimensional covariate vector d is upper-bounded by $O(\log d)$. This bound presents a significant improvement over other classic bandit algorithms ([2, 25, 37]), which yield polynomial dependence on d, and is also a tighter bound than the Lasso-type algorithm (i.e., $O(\log^2 d)$ in [9]). This improvement is of particular importance in high-dimensional settings, in which the covariate dimension can be extremely large, and it suggests that the G-MCP-

Bandit algorithm can bring substantial regret reduction comparing to existing bandit algorithms, which we will illustrate through experiments in §4.6.

4.5 Key Steps of Regret Analysis for the G-MCP-Bandit Algorithm

In this section, we provide the abridged technical proofs for Theorem 4.4.4 – the main theorem in this work. Specifically, we briefly lay out four key steps in establishing the expected cumulative regret upper bound for the G-MCP-Bandit algorithm. In the first step, we highlight the influence of non-iid data, inherited from the multi-armed bandit model, and provide the statistical convergence property for the MCP estimator under partially iid samples. Applying these results to the G-MCP-Bandit algorithm, in the second and third steps, we establish the convergence properties for both the random sample estimator, which is based on samples generated only through the ϵ-decay random sampling method, and the whole sample estimator, which uses all available samples. Finally, in the last step, we establish the total expected cumulative regret by separating the regret up to time T into three segments and providing a bound for each segment. The main structure and sequence of our proving steps described above are first introduced by [9], which presents their expected regret analysis for a linear bandit model (i.e., LASSO-Bandit algorithm) in a similar sequence. We will largely follow their presentation structure, but with different steps, proving techniques, and convergence properties, to illustrate the key steps in analyzing the G-MCP-Bandit algorithm.

4.5.1 General Non-iid Sample Estimator

Note that the restricted eigenvalue condition (A.4 in §4.3) for high-dimensional statistics is typically established for iid samples in the literature. Yet, in this research, we consider the G-MCP-Bandit algorithm, under which only part of the samples are iid, so we first show that the restricted eigenvalue condition continues to hold for partially iid samples (Lemma A.0.6 in E-Companion). Then, we can establish some general results for the MCP estimator under non-iid data.

We denote $\mathcal{W}$ as the whole sample set that contains all users' covariate vectors $\boldsymbol{X}$ and the corresponding rewards $\boldsymbol{r}$ for an arbitrary decision $k \in \mathcal{K}$ up to time

T, and $\boldsymbol{\beta}^{MCP}$ as the MCP estimator for the parameter vector corresponding to decision k. Note that as samples in $\mathcal{W}$ are not iid, standard MCP convergence results ([31, 33]) cannot be directly applied. Recall that we proposed the ϵ-decay random sampling method and that samples generated under this method are iid. Therefore, there exists a subset $\mathcal{A} \subseteq \mathcal{W}$ such that all samples in this subset are iid from the distribution $\mathcal{P}_{\boldsymbol{X}}$. The next step is to show that when the cardinality of $\mathcal{A}$ (i.e., $|\mathcal{A}|$) is large enough, $\boldsymbol{\beta}^{MCP}$ will converge to the true parameters $\boldsymbol{\beta}^{true}$.

Proposition 4.5.1. *Denote the whole sample size as n and the sub-sample set, containing only iid random samples, as $\mathcal{A}$. Under assumptions A.1, A.4, and A.5, if $\beta_{\min} \geq (\frac{96ns}{\kappa|\mathcal{A}|} + a)\lambda$ and $a > \frac{96ns}{\kappa|\mathcal{A}|}$, then for $\zeta \leq \frac{\mu_0|\mathcal{A}|\sqrt{C_2\lambda}}{2n}$, the following inequality hold for the MCP estimator under the 2sWL procedure $\boldsymbol{\beta}^{MCP}$*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_2 \leq \frac{2n\zeta}{|\mathcal{A}|\mu_0}\right) \geq 1 - \delta_2(n, |\mathcal{A}|, \lambda) - \delta_3(|\mathcal{A}|) - \delta_4(n, |\mathcal{A}|, \zeta). \tag{4.5.6}$$

Moreover, if $|\mathcal{A}| \geq \frac{2s^2x_{\max}^2}{\mu_0}$, then we have the following result

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{\frac{8s^2\sigma^2x_{\max}^2 n}{\mu_0^2|\mathcal{A}|^2}}\right) \geq 1 - \delta_1(|\mathcal{A}|) - \delta_2(n, |\mathcal{A}|, \lambda) - \delta_3(|\mathcal{A}|), \tag{4.5.7}$$

where C_2 and μ_0 are positive constants and $\delta_4(n, |\mathcal{A}|, \zeta) \doteq s\exp\left(-\frac{|\mathcal{A}|\mu_0}{8\sigma_2 s x_{\max}^2}\right) + s\exp\left(-\frac{n\zeta^2}{2\sigma^2 x_{\max}^2}\right)$.

Proposition 4.5.1 describes the statistical properties of the non-iid MCP estimators under the 2sWL procedure. First, if we don't require the iid sample size $|\mathcal{A}|$ to be sufficiently large, then the MCP estimator's statistical performance is given by Equation (4.5.6). If we set ζ to be on the order of $O(s/\sqrt{n})$, then $\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|$ is on the order of $O(\sqrt{s^2n/|\mathcal{A}|^2})$, which matches the result of Equation (4.5.7). Meanwhile, however, $\delta_4(n, |\mathcal{A}|, \zeta)$ in Equation (4.5.6) becomes a positive constant asymptotically, which implies that when $|\mathcal{A}|$ is not large enough, the MCP estimator may not warrant good statistical performance. Yet, when we have sufficient iid samples (i.e., $|\mathcal{A}| \geq \frac{2s^2x_{\max}^2}{\mu_0}$), Equation (4.5.7) suggests that the MCP estimator not only guarantees a better statistical convergence ($O(\sqrt{s^2n/|\mathcal{A}|^2})$) but also attains

probability 1 when the whole sample size n and the iid sample size $|\mathcal{A}|$ go to infinity.

Moreover, Proposition 4.5.1 shows the necessity of generating iid random samples in high-dimension bandit settings. Non-iid samples are inevitable in online learning and decision-making process, so ensuring desired asymptotical performance of the parameter vector estimation in high-dimensional settings can only be achieved through generating sufficient number of iid samples, as shown in Proposition 4.5.1. We will show in next two subsections that the size of iid samples generated under the ϵ-decay random sampling method is on the order of $O(\log T)$ and that the size can be further improved to the order of $O(T)$ under the bi-level decision structure in the G-MCP-Bandit algorithm.

4.5.2 Estimator from Random Samples up to Time T

In Proposition 4.5.1, we show that the MCP estimator will converge to the oracle parameter as long as the sample set contains a sufficient number of iid samples. Recall that in our proposed G-MCP-Bandit algorithm, samples generated by the ϵ-decay random sampling method are iid, and the size of these iid samples is on the order of $O(\log(T))$; see Proposition 4.4.3. Combining these observations, we can establish the statistical performance of the MCP estimator under the G-MCP-Bandit algorithm in the following proposition.

Proposition 4.5.2. *Let $t_0 = 2C_0|\mathcal{K}|$, $T \geq \max\{(t_0+1)^2/e^2 - 1, e\}$, $a > 2304s/p^*\kappa$ and $\lambda = C_5\sqrt{1+\log d/\log(T+1)}$. If assumptions A.1, A.3, A.4, and A.5 hold, then the MCP estimator under the G-MCP-Bandit algorithm $\boldsymbol{\beta}^{MCP}$ will satisfy the following inequality*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_1 \leq \min\left\{\frac{1}{\sigma x_{\max}}, \frac{h}{4e\sigma R_{\max} x_{\max}}\right\}\right) \geq 1 - \frac{7}{T+1},$$

where C_0 and C_5 are positive constants.

4.5.3 Estimator from Whole Samples up to Time T

In addition to the iid samples generated by the ϵ-decay random sampling method, other samples can also be iid and used to improve the statistical performance of the MCP estimator. To intuit, recall that in the G-MCP-Bandit algorithm, when

the user is not selected to perform a random sampling, the decision-maker will use the bi-level structure to determine the optimal decision to maximize his expected reward. In the upper-level decision-making process, only iid samples will be used (as $\boldsymbol{\beta}^{random}$ is the MCP estimator based on samples generated only by the ϵ-decay random sampling method) to determine the candidate(s) for the optimal decision set. From Proposition 4.5.2, we know that this random sample MCP estimator will not be far away from its true parameter values. In other words, if we define the event that the random sample MCP estimator at time t is within a given distance from its true parameter as event $\mathcal{E}_6$:

$$\mathcal{E}_6 \doteq \left\{ \|\boldsymbol{\beta}_k^{random}(t) - \boldsymbol{\beta}_k^{true}\|_1 \leq \min\left\{ \frac{1}{\sigma x_{\max}}, \frac{h}{4e\sigma R_{\max} x_{\max}} \right\},\ k \in \mathcal{K} \right\}, \quad (4.5.8)$$

then event $\mathcal{E}_6$ will happen with high probability. Further, conditioning on event $\mathcal{E}_6$, we can directly verify that for any $\boldsymbol{x} \in U_k$, $k \in \mathcal{K}$, the following inequality holds:

$$\mathbb{E}(R_k|\boldsymbol{x}, \boldsymbol{\beta}_k^{random}(t)) \geq \max_{j \neq k} \mathbb{E}(R_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{random}(t)) + \frac{h}{2}. \quad (4.5.9)$$

Therefore, if using Equation (4.5.9) as the selecting criterion, the decision-maker will be able to choose the optimal decision k for any $x \in U_k, k \in \mathcal{K}$ with high probability. Formally, we can bound the total number of times under which event $x \in U_k$ and event $\mathcal{E}_6$ happen simultaneously. In particular, we define $M(i) \doteq \mathbb{E}\left[\sum_{j=1}^{T+1} \mathbb{1}(\boldsymbol{x}_j \in U_k, \mathcal{E}_6, x_j \notin \mathcal{R}_k)|\mathcal{F}_i\right]$ for $i \in \{0, 1, 2, .., T+1\}$, where $\mathcal{F}_i = \{(\boldsymbol{x}_j, r_j) \text{ for } j \leq i\}$ and $\mathcal{R}_k$ is the set containing iid samples generated through the ϵ-decay random sampling method for arm k. Then, $\{M(i)\}$ is a martingale with bounded difference $|M(i) - M(i+1)| \leq 1$ for $i = 0, 1, 2, ..., T$, and we can bound the value of $M(T+1)$ in the following proposition:

Proposition 4.5.3. *If* $T \geq \max\{14, 4C_0|\mathcal{K}|\}$, *then* $\mathbb{P}\left(M(T+1) \leq \frac{p^*(T+1)}{8}\right) \leq \exp\left(-\frac{(p^*)^2 T}{128}\right)$.

Intuitively, Proposition 4.5.3 suggests that with high probability, the actual iid sample size in U_k for decision k will be on the order of $O(T)$ instead of $O(\log T)$. This improvement is the reason why the whole sample MCP estimator $\boldsymbol{\beta}^{whole}$ used in the lower-level decision-making process has a better statistical performance, compared to the random sample MCP estimator $\boldsymbol{\beta}^{random}$ used in the upper-level

decision-making process. Specifically, we can establish the convergence property for the whole sample MCP estimator in the following proposition.

Proposition 4.5.4. *Let* $t_0 = 2C_0|\mathcal{K}|$, $T > T_0$, $\lambda = C_4\sqrt{\frac{\log(T+1)+1+\log d}{T+1}}$, *and* $a > \frac{2304s}{p^*\kappa}$. *If assumptions A.1, A.3, A.4, and A.5 hold, then at time* T *the whole sample MCP estimator under the G-MCP-Bandit algorithm* $\boldsymbol{\beta}^{whole}$ *will satisfy the following inequality:*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{whole}(T) - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{C_\beta \frac{s^2}{T+1}}\right) \geq 1 - \frac{12}{T+1},$$

where C_0, T_0, C_4, *and* C_β *are positive constants.*

4.5.4 Cumulated Regret Up To Time T

Finally, to bound the cumulative regret for the G-MCP-Bandit algorithm, we need to divide the time, up to time T, into three groups and provide a upper bound for each group.

The first group contains all samples before time T_0 and all random samples up to time T. Note that before time T_0 (the explicit expression for T_0 is given in the proof of Theorem 4.4.4 in E-Companion), the decision-maker does not have sufficient samples to accurately estimate covariate parameter vectors. Hence, the reward under the G-MCP-Bandit algorithm will suffer and be sub-optimal compared to that of the oracle case. We can bound the cumulative regret by the worst case performance: $R_{\max}T_0 + R_{\max}|\mathcal{K}|(2 + 6C_0 \log T)$, where the first part of this cumulative regret is for all samples before time T_0 and the second part is for all random samples up to time T.

Next, we will segment the $t > T_0$ case into two groups, depending on whether we can accurately estimate covariate parameter vectors by using only random samples. In particular, the second group includes cases where $t > T_0$ and the random-sample-based estimators are not accurate (i.e., event $\mathcal{E}_6$ doesn't hold). Under those scenarios, inevitably, the decision-maker's decisions will be suboptimal with high probability. However, note that as the size of iid samples increases in t, the probability of event $\mathcal{E}_6$ not occurring decreases. We can bound the cumulative regret for the second group by $7R_{\max}|\mathcal{K}| \log(T + 1)$.

The last group includes scenarios where $t > T_0$ and the random sample estimators

are accurate enough. Benefiting from the improved estimation accuracy (Proposition 4.5.4), we can bound the cumulative regret for the last group by $(24R_{\max}|\mathcal{K}| + 4e^{4\sigma^2 x_{\max} b} C R_{\max}^3 |\mathcal{K}| x_{\max}^2 C_\beta s^3) \log(T)$. Combining the cumulative regret for all three groups, Theorem 4.4.4 directly follows.

4.6 Empirical Experiments

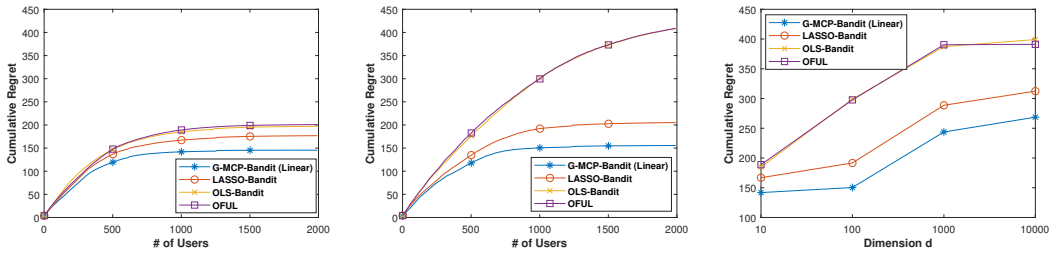
In this section, we will benchmark the G-MCP-Bandit algorithm to OFUL ([1]), OLS-Bandit ([37]), and Lasso-Bandit ([9]). In particular, we seek answers to the following two questions: How does the performance of the G-MCP-Bandit algorithm compare to other bandit algorithms? And how is the performance of the G-MCP-Bandit algorithm influenced by the data availability (T), the data dimensions (s and d), and the size of the decision set (K)?

To this end, we start with two synthetic-data-based experiments in §4.6.1 and conduct two additional experiments based on real datasets, the warfarin dosing patient data in §4.6.2 and the Tencent search advertising data in §4.6.3, respectively. Note that the algorithms and theoretical bounds of OFUL, OLS-Bandit, and Lasso-Bandit are developed under the assumption that the reward function follows the linear model, which is a special case in the G-MCP-Bandit algorithm. Therefore, for fair comparison, we specify the underlying reward function for the G-MCP-Bandit algorithm to follow the same linear model (i.e., the reward under decision k for a user with covariate vector $\boldsymbol{x}$ takes the form of $R_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}_k^{true} + \epsilon$, where ϵ is a σ-gaussian random variable) in all experiments, except the Tencent search advertising data experiment, in which we explore the performance of the G-MCP-Bandit model under both the linear model and the logistic model.

4.6.1 Synthetic Data (Linear Model)

In the first synthetic data experiment, we fix the size of the decision set K and focus on the impacts of the data dimensions, s and d, and the data availability, T, on learning algorithms' cumulative regret performance. In particular, we consider a two-arm bandit setting (i.e., $K = 2$). To simulate different sparsity levels, we vary the covariate dimension $d = \{10, 10^2, 10^3, 10^4\}$ and keep the dimension for significant covariates unchanged at $s = 5$. Therefore, as the covariate dimension

d increases, the data become sparser. The underlying true parameter vectors for covariates are arbitrarily set to be $\boldsymbol{\beta}_1 = (1, 2, 3, 4, 5, 0, 0, ...)$ for the first arm and $\boldsymbol{\beta}_2 = 1.1 \cdot \boldsymbol{\beta}_1$ for the second arm. For each incoming user, we randomly draw her covariate vector from $N(0, I_{d\times d})$ and the error term in the linear model ϵ from $N(0, 1)$. Finally, we use the same parameter λ value in both the Lasso-Bandit algorithm and the G-MCP-Bandit algorithm and select the unique parameter for the G-MCP-Bandit algorithm a at 2. For each algorithm, we perform 100 trials and report the average cumulative regret for OFUL, OLS-Bandit, Lasso-Bandit, and G-MCP-Bandit (under the linear model) in Figure 4.1.



(a) d=10 (b) d=100 (c) d=1000

Figure 4.1: Synthetic study 1: The impact of T and d on the cumulative regret, where $K = 2$ and $s = 5$.

Overall, we observe that the G-MCP-Bandit algorithm significantly outperforms OFUL, OLS-Bandit, and Lasso-Bandit and achieves the lowest cumulative regret. Facing only two decisions/arms, the decision-maker can easily identify the optimal arm, and therefore OFUL and OLS-Bandit, both of which are not specifically designed for high-dimensional settings, perform nearly identically. Lasso-Bandit and G-MCP-Bandit could benefit from their abilities to recover the sparse structure and identify the significant covariates. Therefore, compared to OFUL and OLS-Bandit, Lasso-Bandit and G-MCP-Bandit can improve their parameters estimations, especially under high-dimensional settings, and perform substantially better. Further, the improvement of the cumulative regret performance of G-MCP-Bandit over Lasso-Bandit follows from the facts that the MCP estimator is unbiased and could improve the sparse structure discovery. Next, we will discuss the influence of sample size T and the covariate dimension d on these algorithms' cumulative regret performance.

Figure 4.1(a) and 4.1(b) illustrate the influence of the sample size T on the

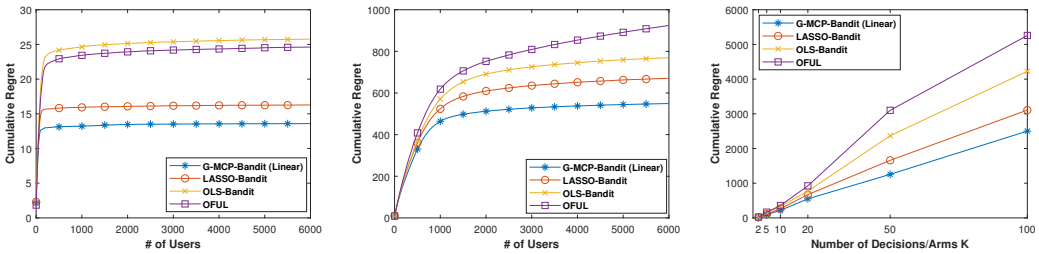
cumulative regret for the cases where $d = 10$ and $d = 100$ (other cases exhibit a similar pattern and are therefore omitted)[2]. As we have proven that G-MCP-Bandit provides the optimal time dependence under both low-dimensional and high-dimensional settings (Theorem 4.4.4), G-MCP-bandit strictly improves on the cumulative regret performance from Lasso-Bandit, especially when T is not too small. Note that facing insufficient samples, all algorithms fail to accurately learn parameter vectors and therefore perform poorly. As the sample size increases, the G-MCP-bandit algorithm is able to, in an expeditious fashion, unveil the underlying sparse data structure, accurately estimate parameter vectors, and outperform all other benchmarks. For example, in Figure 4.1(b), we observe that the regret reduction of G-MCP-Bandit over all other algorithms is larger than 10% when the sample size T is larger than 350. This observation echoes our theoretical findings that the G-MCP-Bandit algorithm attains the optimal regret bound in sample size dimension $O(\log T)$.

We also observe that the benefits of G-MCP-Bandit over other three algorithms appear to increase in the data sparsity level. Figure 4.1(c) presents the influence of the covariate dimension d on the cumulative regret for the case where $T = 1000$. Recall that we fixed the dimension for significant covariates $s = 5$. Therefore, as the covariate dimension d increases, the data become sparser (i.e., d/s increases). As expected, the cumulative regret for all four algorithms increases in the covariate dimension d, but at different rates. On the one hand, both OLS-Bandit and OFUL lack the ability to recover the sparse data structure and are ill suited for high-dimensional problems. On the other hand, Lasso-Bandit and G-MCP-Bandit, which adopt different statistical learning methods for the sparse structure discovery and are designed for high-dimensional problems, have lower cumulative regret that increases in d at a slower rate. Further, we notice that the G-MCP-Bandit algorithm has the least increase in cumulative regret among all four algorithms, which confirms our theoretical finding in Theorem 4.4.4: The G-MCP-Bandit algorithm has a better dependence on the covariate dimension $O(\log d)$ than Lasso-Bandit $O(\log^2 d)$, OFUL, and OLS-Bandit (the last two algorithms have polynomial bounds in d).

In the second synthetic data experiment, we study the influence of the size of decision set by varying $K = \{2, 5, 10, 20, 50, 100\}$ and keeping the data dimensions

[2]In all four experiments where $d \in \{10, 10^2, 10^3, 10^4\}$, we simulated the sample size up to $10,000$ and observe that the G-MCP-Bandit algorithm's cumulative regret seems to be stabilized before $T = 2000$. Therefore, we only plot for the first 2000 samples to avoid duplications.

unchanged ($s = 5$ and $d = 100$). For each decision, we randomly draw the parameter vector for the significant covariates from a uniform distribution, $U(0, 1)$. Finally, we keep other parameters the same as in the first synthetic data experiment. Figure 4.2 plots the average cumulative regret for OFUL, OLS-Bandit, Lasso-Bandit, and G-MCP-Bandit (under the linear model).



(a) K=2 (b) K=20 (c) T=6000

Figure 4.2: Synthetic study 2: The impact of T and K on the cumulative regret, where $d = 100$ and $s = 5$.

We observe that the benefits of adopting G-MCP-Bandit over the other three algorithms increases with the size of the decision set. In particular, as K increases, the cumulative regret gap between G-MCP-Bandit and any other algorithm grows; see Figure 4.2(c). This observation is as expected. To intuit, note that as we add more possible decisions into the decision set, the complexity and difficulty for the decision-maker to select the optimal decision grow for two main reasons. First, the decision-maker will need more samples to identify the significant covariates and estimate the parameter vectors. Second, as the number of decisions increases, the process of comparing the expected rewards among all decisions and selecting the optimal decision becomes more vulnerable to estimation errors. Therefore, we should expect that as the number of arms increases, the amount of samples required for these algorithms to accurately learn the parameter vectors and select the optimal decision will increase as well.

Figure 4.2(a) and Figure 4.2(b) plot the cumulative regret for the case of two arms and twenty arms, respectively. Clearly, the decision-maker needs far more samples before his cumulative regret can be stabilized in the case of twenty arms than in the case of two arms. Therefore, the cumulative regret performance under all algorithms suffers from the increasing size of the decision set. As discussed earlier, the G-MCP-Bandit algorithm attains the optimal bound in the sample

size dimension and is able to learn the sparse data structure and provide accurate unbiased estimators for parameter vectors. Hence, we observe that the benefits of adopting the G-MCP-Bandit algorithm over other algorithms are amplified as the number of arms increases, as illustrated in Figure 4.2(c).

4.6.2 Warfarin Dosing Patient Data (Linear Model)

In the first real-data-based experiment, we considers a health care problem in which physicians determine the optimal personalized warfarin dosage for incoming patients ([24]). Using the same dataset, [9] demonstrate that the Lasso-Bandit algorithm outperforms other existing bandit algorithms, including OFUL-LS ([1]), OFUL-EG ([2]), and OLS-Bandit ([37]). The warfarin dosing patient data contains detailed covariates (the size of covariates used in our experiment is 93) for $5,700$ patients, including demographic, diagnosis, and genetic information that can be used to predict the optimal warfarin dosage.

We apply the G-MCP-Bandit algorithm to the warfarin dosing patient dataset to evaluate its performance in practical decision-making contexts where the technical assumptions specified early in §4.3 may not hold. Following [9], we formulate this problem as a 3-armed bandit with covariates under the linear model.

Figure 4.3 compares the average fraction of optimal/correct dosing decisions under G-MCP-Bandit (under the linear model) to those under OFUL, OLS-Bandit, Lasso-Bandit, actual physicians' decisions, and the oracle policy. We observe that as long as the sample size is not too small (e.g., the number of patients exceeds 40), the G-MCP-Bandit algorithm will outperform physicians' decisions, OLS-Bandit, Lasso-Bandit, and OFUL. However, when there are very limited samples (< 40 patients), the physicians' static decisions (i.e., always recommend medium dose) perform the best, with a stable optimal percentage of 54%. This is because that without sufficient samples, all learning algorithms are unable to accurately learn the parameter vectors for patients' covariates, and consequently they behave suboptimally.

As the sample size increases, all learning algorithms are able to update their estimation of parameter vectors and eventually outperform the physicians' static decisions. Among all learning algorithms, the G-MCP-Bandit algorithm requires the fewest samples (i.e., $T > 40$ for G-MCP-Bandit, $T > 90$ for Lasso-Bandit,

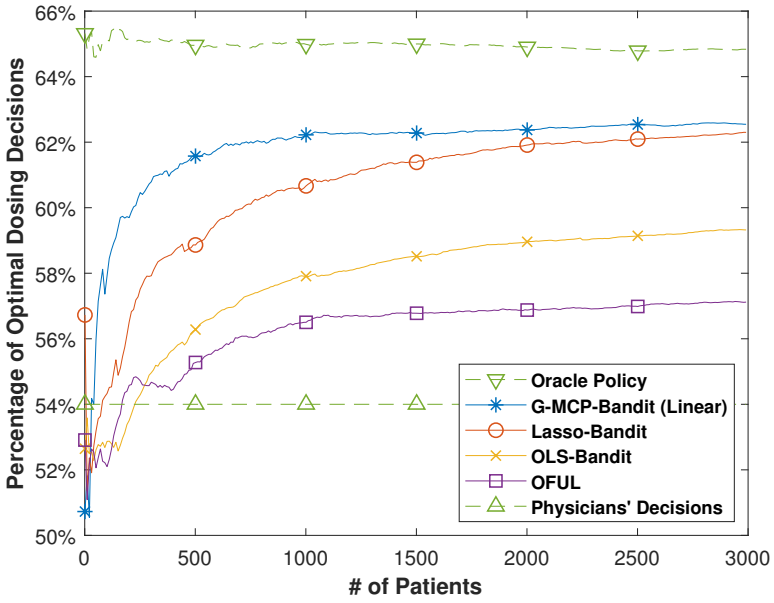


Figure 4.3: Warfarin dosing experiment: The percentage of optimal warfarin dosing decisions.

$T > 180$ for OFUL, $T > 220$ for OLS-Bandit) to provide better dosing decisions than physicians.

4.6.3 Tencent Search Advertising Data (Linear & Logistic Models)

In the last experiment, we scale up the dataset's dimensionality to consider a search advertising problem at Tencent. The Tencent search advertising dataset is collected by Tencent's proprietary search engine, soso.com, and it documents the interaction sessions between users and the search engine ([82]). In the dataset, each session contains a user's demographic information (age and gender), the query issued by the user (combinations of keywords), ads information (title, URL address, and advertiser ID), the user's response (click or not), etc. This dataset is high-dimensional with sparse data structure and contains millions of observations and covariates. To put the size of the dataset into perspective, it contains $149,639,105$ session entries, more than half a million ads, more than one million unique keywords, and more than 26 million unique queries.

For illustration purposes, we focus on a three-ad experiment[3] (with ad IDs 21162526, 3065545, and 3827183). Each of these three ads has an average CTR higher than 2% and more than $100,000$ session entries, which provide reasonably accurate estimation for parameter vectors (see next paragraph for more discussions). In total, there are $849,338$ session entries with $169,744$ unique queries and 8 covariates for users' demographic information. As the search engine receives payment from advertisers only when the user has clicked the sponsored ad, we arbitrarily assume that advertisers will award the search engine \$1, \$5, and \$10 for each clicked ad, respectively.

Figure 4.4 plots the the average revenue performance under OFUL, OLS-Bandit, Lasso-Bandit, a random policy, the oracle policy, and G-MCP-Bandit (under both linear and logistic models). It is worth noting that the "true" oracle policy is impossible to implement, as the true parameter vectors are unknown, or at least have considerable variance even when all session entries in the dataset are used for estimation. Therefore, the oracle policy in the experiment represents the scenario when the search engine has access to all data to estimate these parameter vectors and make ad selection decisions. In addition, we introduce the random policy as another benchmark to simulate the scenario in which the search engine will randomly recommend an ad with equal probability to an incoming user. Finally, note that the CTR prediction is binary in nature (i.e., click or not). We therefore include the G-MCP-Bandit algorithm under the logistic model and compare it to the G-MCP-Bandit algorithm under the linear model to study the influence of the underlying model choice. In the experiment, we simulate incoming users by permuting their covariate vectors randomly. For each algorithm, we perform 100 trials and report the average revenue with 5000 users, which seems to be sufficient for the G-MCP-Bandit algorithm to converge.

We can show that all learning algorithms generate higher average revenue than the random policy for any number of users and that the G-MCP-Bandit algorithm outperforms other algorithms under most scenarios. Specifically, when comparing all algorithms under the same linear model, we observe that the G-MCP-Bandit algorithm (under the linear model) has better average revenue performance than OFUL, OLS-Bandit, and Lasso-Bandit as soon as there are more than 140 users.

[3]We have extended the experiment to include more ads, but we find that doing so will not qualitatively change our observations and insights but considerably increases the computation time. Therefore, we decide to focus on this three-ad experiment in the work.

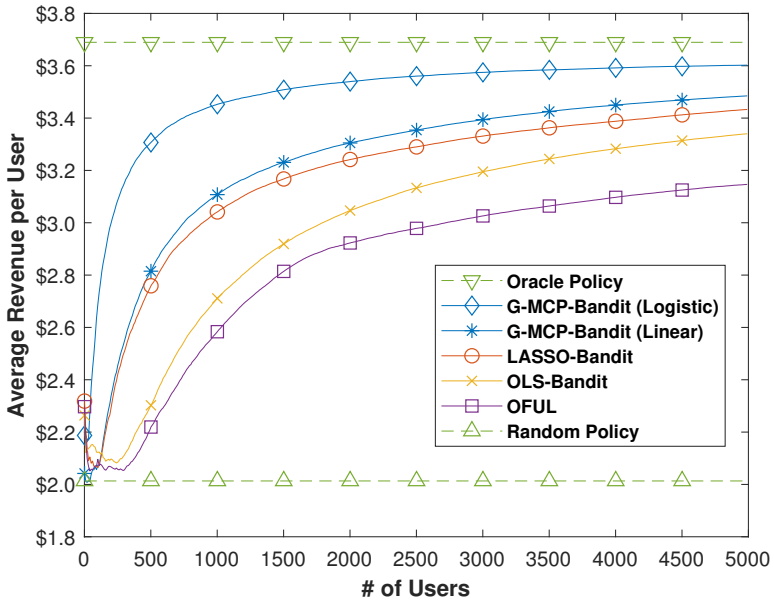


Figure 4.4: Tencent search advertising experiment: The average revenue under different algorithms.

This observation is consistent with that in warfarin dosing experiment in §4.6.2 and suggests that compared to other benchmark algorithms, the G-MCP-Bandit algorithm can improve the parameter vector estimation under high-dimensional data with limited samples and achieve better revenue performance.

Further, we find the choice of underlying models can significantly influence the G-MCP-Bandit algorithm's average revenue performance. Note that the advertisers award the search engine only when users have clicked the recommended ads. Therefore, the search engine's reward function is binary in nature. When comparing the G-MCP-Bandit algorithm under the logistic model to that under the linear model, both of which are special cases of the G-MCP-Bandit algorithm, we observe that the former always dominates the latter for any number of users. In addition, the G-MCP-Bandit algorithm under the logistic model merely needs 20 users to outperform the other three algorithms. This observation suggests that understanding the underlying managerial problem and identifying the appropriate model for the G-MCP-Bandit algorithm can be critical and bring substantial revenue improvement for the decision-maker.

4.7 Technical proofs

To simplify the notations, we denote $\nabla_{\mathcal{A}} F(\boldsymbol{x})$ as the vector with $(\nabla_{\mathcal{A}} F(\boldsymbol{x}))_i = (\nabla F(\boldsymbol{x}))_i,\ i \in \mathcal{A}$, where $(\cdot)_i$ is the i-th element in the vector. Similarly we denote $\nabla^2_{\mathcal{A},\mathcal{B}} F(\boldsymbol{x})$ as the matrix with $(\nabla^2_{\mathcal{A},\mathcal{B}} F(\boldsymbol{x}))_{ij} = (\nabla^2 F(\boldsymbol{x}))_{ij},\ i \in \mathcal{A}, j \in \mathcal{B}$, where $(\cdot)_{ij}$ is the element in i-th column and j-th row. We denote $\lambda_{\min}(\boldsymbol{X})/\lambda_{\max}(\boldsymbol{X})$ as the smallest/largest eigenvalue of matrix $\boldsymbol{X}$.

4.7.1 Proof of Lemma 4.4.1

Proof. Lemma 4.4.1 directly follows Lemma A.0.2 in Appendix: A by setting $|\mathcal{A}| = n$. □

4.7.2 Proof of Proposition 4.4.2

Proof. Proposition 4.4.2 follows Proposition 4.5.1 by setting $|\mathcal{A}| = n$. □

4.7.3 Proof of Proposition 4.4.3

Proof. Under the ϵ-decay random sampling method, the probability of randomly drawing arm k at time t is $\min\{1, t_0/t\}/|\mathcal{K}|$, where $|\mathcal{K}|$ is the number of arms. Hence, at time T, the expected total number of times at which arm k were randomly drawn is

$$\mathbb{E}[n_k] = \frac{1}{|\mathcal{K}|} \sum_{t=1}^{T} \min\left\{1, \frac{t_0}{t}\right\}.$$

When $T > t_0$,

$$\mathbb{E}[n_k] = \frac{1}{|\mathcal{K}|}\left(t_0 + \sum_{t=t_0+1}^{T} \frac{t_0}{t}\right) = \frac{t_0}{|\mathcal{K}|}\left(1 + \sum_{t=t_0+1}^{T} \frac{1}{t}\right). \tag{4.7.10}$$

Since the function $f(t) = 1/t$ is decreasing in t , it can be bounded as follows.

$$\int_t^{t+1} \frac{1}{t} dt < \frac{1}{t} < \int_{t-1}^{t} \frac{1}{t} dt,\ t \geq 2.$$

As $t_0 \geq 1$, for any t from $t_0 + 1$ to T, we have

$$\log(T+1) - \log(t_0+1) < \sum_{t=t_0+1}^{T} \frac{1}{t} < \log(T) - \log(t_0). \tag{4.7.11}$$

Combining (4.7.10) and (4.7.11), we can bound $\mathbb{E}[n_k]$ as follows.

$$\frac{1}{|\mathcal{K}|} t_0(1 + \log(T+1) - \log(t_0+1)) < \mathbb{E}[n_k] < \frac{1}{|\mathcal{K}|} t_0(1 + \log(T) - \log(t_0)). \tag{4.7.12}$$

Since $n_k = \sum_{t=1}^{T} \mathbb{1}\{\text{random sampling for arm } k \text{ at } t\}$, we can view n_k as the summarization of bounded iid random variables. Via Chernoff bound, we can build the connect between n_k and $\mathbb{E}[n_k]$.

$$\mathbb{P}\left(\frac{1}{2}\mathbb{E}[n_k] \leq n_k \leq \frac{3}{2}\mathbb{E}[n_k]\right) > 1 - 2\exp\left(-\frac{1}{10}\mathbb{E}[n_k]\right). \tag{4.7.13}$$

We then relax the $\mathbb{E}[n_k]$ in (4.7.13) with the upper and lower bounds provided in (4.7.12) and the following result is attained.

$$\begin{aligned} &\mathbb{P}\left(\frac{t_0(1 + \log(T+1) - \log(t_0+1))}{2|\mathcal{K}|} \leq n_k \leq \frac{3t_0(1 + \log(T) - \log(t_0))}{2|\mathcal{K}|}\right) \\ &\geq 1 - 2\left(\frac{t_0+1}{e(T+1)}\right)^{\frac{t_0}{10|\mathcal{K}|}}. \end{aligned} \tag{4.7.14}$$

When $t_0 = 2C_0|\mathcal{K}|,\ C_0 \geq 10$, and $T > \frac{(t_0+1)^2}{e^2}$, we can simplify the right-hand size of (4.7.14).

$$1 - 2\left(\frac{t_0+1}{e(T+1)}\right)^{\frac{t_0}{10|\mathcal{K}|}} \geq 1 - 2\left(\frac{e\sqrt{T+1}}{e(T+1)}\right)^{C_0/5} \geq 1 - \frac{2}{T+1}. \tag{4.7.15}$$

□

4.7.4 Proof of Proposition 4.5.1

Proof. In the first step of 2sWL procedure, we are essentially solving the Lasso problem. From Lemma A.0.7, we have $\|\boldsymbol{\beta}^{lasso} - \boldsymbol{\beta}^{true}\|_1 \leq \frac{96ns\lambda}{|\mathcal{A}|\kappa}$ which high

probability. As we assume $\beta_{\min} \geq \left(\frac{96ns}{|\mathcal{A}|\kappa} + a\right)\lambda$ and $\|\boldsymbol{\beta}^{lasso} - \boldsymbol{\beta}^{true}\|_\infty \leq \|\boldsymbol{\beta}^{lasso} - \boldsymbol{\beta}^{true}\|_1$ we have the follow statements hold.

$$|\beta_i^{lasso}| \geq a\lambda,\ i \in \mathcal{S} \text{ and } |\beta_i^{lasso}| \leq \frac{96ns\lambda}{|\mathcal{A}|\kappa},\ i \in \mathcal{S}^c, \tag{4.7.16}$$

where we ignore the subscript in $\mathcal{S}_k$ to simplify the notation. Combining (4.7.16) and $P_\lambda^{'}(|x|) = \max\{0, \lambda - |x|/a\}$, we have the following two results.

$$P_\lambda^{'}(|\beta_i^{lasso}|) = 0 \quad i \in \mathcal{S}, \tag{4.7.17}$$

$$P_\lambda^{'}(|\beta_i^{lasso}|) \geq P_\lambda^{'}\left(\frac{96ns\lambda}{|\mathcal{A}|\kappa}\right) = \left(\lambda - \frac{96ns\lambda}{|\mathcal{A}|\kappa a}\right) \quad i \in \mathcal{S}^c. \tag{4.7.18}$$

Define the event $\mathcal{E}_2$ as follows

$$\mathcal{E}_2 = \left\{\|\nabla_{\mathcal{S}^c}\mathcal{L}(\boldsymbol{\beta}^{oracle})\|_\infty < \lambda - \frac{96ns\lambda}{|\mathcal{A}|\kappa a}\right\}. \tag{4.7.19}$$

From the convexity of $\mathcal{L}(\boldsymbol{\beta})$, we can build a lower bound on the optimal objective function value in the second step of 2sWL.

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}^*) + \sum_j P_\lambda'(|\beta_j^{lasso}|)\cdot|\beta_j^*| \geq\ & \mathcal{L}(\boldsymbol{\beta}^{oracle}) \\ & + \nabla\mathcal{L}(\boldsymbol{\beta}^{oracle})^T(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{true}) + \sum_j P_\lambda'(|\beta_j^{lasso}|)\cdot|\beta_j^*|,\end{aligned} \tag{4.7.20}$$

where $\boldsymbol{\beta}^*$ is the optimal solution of the second step of the 2sWL procedures. From the definition of oracle solution, we have

$$\boldsymbol{\beta}^{oracle} = \arg\min_{\boldsymbol{\beta}_{\mathcal{S}^c}=0} \mathcal{L}(\boldsymbol{\beta}) \Rightarrow 1)\ \nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{oracle}) = 0 \text{ and } 2)\ \boldsymbol{\beta}_{\mathcal{S}^c} = 0. \tag{4.7.21}$$

Combining (4.7.17), (4.7.18), (4.7.20), and (4.7.21), we have

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\beta}^*) + \sum_{j\in\mathcal{S}^c} P_\lambda'(|\beta_j^{lasso}|)\cdot|\beta_j^*| \\ &\geq \mathcal{L}(\boldsymbol{\beta}^{oracle}) + \nabla_{\mathcal{S}^c}\mathcal{L}(\boldsymbol{\beta}^{oracle})^T(\boldsymbol{\beta}^*_{\mathcal{S}^c} - \boldsymbol{\beta}^{oracle}_{\mathcal{S}^c}) + \sum_{j\in\mathcal{S}^c} P_\lambda'(|\beta_j^{lasso}|)\cdot|\beta_j^*|\end{aligned}$$

$$
\begin{aligned}
&= \mathcal{L}(\boldsymbol{\beta}^{oracle}) + \sum_{j\in\mathcal{S}^c} \left(\nabla_j \mathcal{L}(\boldsymbol{\beta}^{oracle})(\beta_j^* - 0) + P_\lambda'(|\beta_j^{lasso}|)\cdot|\beta_j^*|\right) \\
&= \mathcal{L}(\boldsymbol{\beta}^{oracle}) + \sum_{j\in\mathcal{S}^c} P_\lambda'(|\beta_j^{lasso}|)\cdot|\beta_j^{oracle}| \\
&\quad + \sum_{j\in\mathcal{S}^c} \left(\nabla_j \mathcal{L}(\boldsymbol{\beta}^{oracle})\text{sign}(\beta_j^*) + P_\lambda'(|\beta_j^{lasso}|)\right)|\beta_j^*|. \qquad (4.7.22)
\end{aligned}
$$

Using $\mathcal{E}_2$ defined in (4.7.19), (4.7.22) can be simplified as follows.

$$
\mathcal{L}(\boldsymbol{\beta}^*) + \sum_{j\in\mathcal{S}^c} P_\lambda'(|\beta_j^{lasso}|)\cdot|\beta_j^*| \geq \mathcal{L}(\boldsymbol{\beta}^{oracle}) + \sum_{j\in\mathcal{S}^c} P_\lambda'(|\beta_j^{lasso}|)\cdot|\beta_j^{oracle}| + c_0\sum_{j\in\mathcal{S}^c}|\beta_j^*|, \qquad (4.7.23)
$$

where c_0 is a positive constant. Since $\boldsymbol{\beta}^*$ is the optimal solution of the second step in 2sWL, per (4.7.23) we must have $\beta_j^* = 0$ for all $j \in \mathcal{S}^c$. Together with the uniqueness of the solution of (4.4.1), $\boldsymbol{\beta}^{oracle}$ is also the unique optimal solution to the second step in 2sWL, i.e, $\boldsymbol{\beta}^{MCP} = \boldsymbol{\beta}^{oracle}$. Therefore once event $\mathcal{E}_2$ happens, with high probability $\boldsymbol{\beta}^{MCP}$ becomes the oracle solution, which enjoy the optimal statistical performance. We then need to consider the chance that $\mathcal{E}_2$ happens and the result is summarized in Lemma A.0.11. Per Lemma A.0.11, the following $\mathcal{E}_3, \mathcal{E}_4$ and $\mathcal{E}_5$ implies $\mathcal{E}_2$.

$$
\begin{aligned}
\mathcal{E}_3 &= \left\{\|\nabla_{\mathcal{S}^c}\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty \leq \left(1 - \frac{96ns}{|\mathcal{A}|\kappa a}\right)\frac{\lambda}{4}\right\}, \\
\mathcal{E}_4 &= \left\{\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty \leq \left(1 - \frac{96ns}{|\mathcal{A}|\kappa a}\right)\frac{\mu_0|\mathcal{A}|\lambda}{8snx_{\max}^2}\right\}, \\
\mathcal{E}_5 &= \left\{\|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{C_2\lambda}\right\},
\end{aligned}
$$

where C_2 is a positive constant. Now, we can bound the probability of events $\mathcal{E}_3$, $\mathcal{E}_4$, and $\mathcal{E}_5$ happen simultaneously. From Assumption **A.5** and Hoeffding bound we have the following inequality for $t_1 > 0$

$$
\mathbb{P}\left(\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty \geq t_1\right) = \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n x_{j\mathcal{S}}^T f'(r_j|\boldsymbol{x}_{j,\mathcal{S}}^T\boldsymbol{\beta}^{true})\|_\infty \geq t_1\right) \leq s\exp\left(-\frac{nt_1^2}{2\sigma^2 x_{\max}^2}\right). \qquad (4.7.24)
$$

Similarly for $t_2 > 0$, we have the following result.

$$\mathbb{P}\left(\|\nabla_{\mathcal{S}^c}\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty \geq t_2\right) \leq (d-s)\exp\left(-\frac{nt_2^2}{2\sigma^2 x_{\max}^2}\right). \tag{4.7.25}$$

By setting $t_1 = t_2 = (\frac{1}{4} - \frac{24ns}{|\mathcal{A}|\kappa a})\min\left\{1, \frac{\mu_0|\mathcal{A}|}{8snx_{\max}^2}\right\}\lambda$, we have

$$\mathbb{P}\left((\mathcal{E}_4')^c \cup (\mathcal{E}_5')^c\right) \leq d\exp\left(-\frac{n\lambda^2\left((\frac{1}{4} - \frac{24ns}{|\mathcal{A}|\kappa a})\min\left\{1, \frac{\mu_0|\mathcal{A}|}{8snx_{\max}^2}\right\}\right)^2}{2x_{\max}^2}\right). \tag{4.7.26}$$

We can further bound event $\mathcal{E}_5$ via Lemma A.0.2. We can have the following result by setting t in Lemma A.0.2 satisfying $t \leq \frac{\mu_0|\mathcal{A}|\sqrt{C_2\lambda}}{2n}$.

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{C_2\lambda}\right) \geq 1 - s\exp\left(-\frac{\mu_0|\mathcal{A}|}{8s\sigma_2 x_{\max}^2}\right) - s\exp\left(-\frac{nt^2}{2s\sigma^2 x_{\max}^2}\right). \tag{4.7.27}$$

Moreover, from (A.0.5) in Lemma A.0.2, the following result hold for $|\mathcal{A}| \geq \frac{2s^2x_{\max}^2}{\mu_0}$:

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{\frac{8s^2\sigma^2 x_{\max}^2 n}{\mu_0^2|\mathcal{A}|^2}}\right) \geq 1 - s\exp\left(-\frac{\mu_0|\mathcal{A}|}{8s\sigma_2 x_{\max}^2}\right) - 2\exp\left(-\frac{C_h|\mathcal{A}|\mu_0}{2sx_{\max}^2}\right). \tag{4.7.28}$$

Combining Lemma A.0.7, (4.7.26) and (4.7.27) , we have the following inequality for $\zeta \leq \frac{\mu_0|\mathcal{A}|\sqrt{C_2\lambda}}{2n}$.

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_2 \leq \frac{2n\zeta}{|\mathcal{A}|\mu_0}\right) \geq 1 - \delta_2(n, |\mathcal{A}|, \lambda) - \delta_3(|\mathcal{A}|) - \delta_4(n, |\mathcal{A}|, \zeta). \tag{4.7.29}$$

Similarly, by $|\mathcal{A}| \geq \frac{2s^2x_{\max}^2}{\mu_0}$, the following result comes directly from Lemma A.0.7, (4.7.26) and (4.7.28).

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{\frac{8s^2\sigma^2 x_{\max}^2 n}{\mu_0^2|\mathcal{A}|^2}}\right) \geq 1 - \delta_1(|\mathcal{A}|) - \delta_2(n, |\mathcal{A}|, \lambda) - \delta_3(|\mathcal{A}|). \tag{4.7.30}$$

□

4.7.5 Proof of Proposition 4.5.2

Proof. Directly from Lemma A.0.9.

□

4.7.6 Proof of Proposition 4.5.3

Proof. Since $\{M(i)\}$ is a martingale with bounded difference 1, we can use $M(0)$ to bound the value of $M(T+1)$ with Azuma's inequality as follow:

$$\begin{aligned}\mathbb{P}\left(|M(T+1)-M(0)| \geq \frac{1}{2}M(0)\right) &\leq \exp\left(\frac{-M(0)^2/4}{2(T+2)}\right)\\ \Rightarrow \mathbb{P}\left(M(T+1) \leq \frac{1}{2}M(0)\right) &\leq \exp\left(\frac{-M(0)^2/4}{2(T+2)}\right).\end{aligned}$$

The term $M(0)$ can be expressed as follows

$$\begin{aligned}M(0) &= \mathbb{E}\left[\sum_{i=1}^{T+1} \mathbb{1}(\boldsymbol{x}_i \in U_k, \mathcal{E}_6, \boldsymbol{x} \notin \mathcal{R}_k))\right]\\ &= \sum_{i=1}^{T+1} \mathbb{P}(\boldsymbol{x}_i \in U_k, \mathcal{E}_6, x \notin \mathcal{R}_k). \qquad (4.7.31)\end{aligned}$$

As $\{\boldsymbol{x} \in U_k\}$ is independent of $\{\mathcal{E}_6, \boldsymbol{x} \notin \mathcal{R}_k\}$ and $\{\boldsymbol{x} \notin \mathcal{R}_k\}$ is independent on $\{\mathcal{E}_6\}$, (4.7.31) implies the following inequality

$$\begin{aligned}M(0) &= \sum_{i=1}^{T+1} \mathbb{P}(\boldsymbol{x}_i \in U_k)\mathbb{P}(\mathcal{E}_6)\mathbb{P}(\boldsymbol{x} \notin \mathcal{R}_k)\\ &\geq \sum_{i=1}^{T+1} p^*(1-\frac{7}{T+1})(1-\frac{2C_0|\mathcal{K}|}{T+1}), \qquad (4.7.32)\end{aligned}$$

where (4.7.32) uses assumption **A.3**, Proposition 4.5.2 and Proposition 4.4.3.

When $T \geq \max\{14, 4C_0|\mathcal{K}|\}$, we have

$$\frac{7}{T+1} \leq \frac{1}{2} \qquad (4.7.33)$$

$$\frac{2C_0|\mathcal{K}|}{T+1} \leq \frac{1}{2}, \qquad (4.7.34)$$

which implies that

$$M(0) \geq \sum_{i=1}^{T+1} \frac{p^*}{4} = \frac{p^*(T+1)}{4}. \tag{4.7.35}$$

Therefore, the following inequalities hold

$$\begin{aligned}
&\mathbb{P}\left(M(T+1) \leq \frac{p^*(T+1)}{8}\right) \leq \mathbb{P}\left(M(T+1) \leq \frac{1}{2}M(0)\right) \leq \exp\left(\frac{-(p^*)^2(T+1)^2/64}{2(T+2)}\right) \\
\Rightarrow &\mathbb{P}\left(M(T+1) \leq \frac{p^*(T+1)}{8}\right) \leq \exp\left(-\frac{(p^*)^2((T+2)^2+1-2(T+2))}{128(T+2)}\right) \\
\Rightarrow &\mathbb{P}\left(M(T+1) \leq \frac{p^*(T+1)}{8}\right) \leq \exp\left(-\frac{(p^*)^2T}{128} - \frac{p^*}{128(T+2)}\right) \\
\Rightarrow &\mathbb{P}\left(M(T+1) \leq \frac{p^*(T+1)}{8}\right) \leq \exp\left(-\frac{(p^*)^2T}{128}\right)
\end{aligned} \tag{4.7.36}$$

□

4.7.7 Proof of Proposition 4.5.4

Proof. According to Lemma A.0.10, when event $\mathcal{E}_6$ defined by (4.5.8) happens, the following inequality must hold for any $\boldsymbol{x} \in U_k$,

$$\mathbb{E}(R_k|\boldsymbol{x}, \boldsymbol{\beta}_k^{random}(t)) \geq \max_{j \neq k} \mathbb{E}(R_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{random}(t)) + \frac{h}{2}.$$

Therefore, the lower-level decision-making process of the algorithm, in which the decision-maker will successfully select arm i for x by using the random sample estimator, will maintain the iid property of $\boldsymbol{x}$ since it can be viewed as rejection sampling. From Proposition 4.5.3, we have

$$\mathbb{P}\left(M(T+1) \leq \frac{p^*(T+1)}{8}\right) \leq \exp\left(-\frac{(p^*)^2T}{128}\right). \tag{4.7.37}$$

Since $M(T+1) = \mathbb{E}\left[\sum_{j=1}^{T+1} \mathbb{1}(\boldsymbol{x}_j \in U_k, \mathcal{E}_6, \boldsymbol{x}_j \notin \mathcal{R}_k)|\mathcal{F}_{T+1}\right] = \sum_{j=1}^{T+1} \mathbb{1}(\boldsymbol{x}_j \in U_k, \mathcal{E}_6, \boldsymbol{x}_j \notin \mathcal{R}_k)$, the amount of iid samples among the whole sample for arm k up to time $T+1$ will be lower bounded by $M(T+1)$. Denote $\mathcal{A}$ and n as the set of iid samples belonging to U_K in the whole sample set and size of the whole sample respectively.

The follow inequality holds.

$$\mathbb{P}\left(|\mathcal{A}| \geq \frac{p^*(T+1)}{8}\right) \geq 1 - \exp\left(-\frac{(p^*)^2 T}{128}\right),\ n \leq T+1. \tag{4.7.38}$$

Consider $|\mathcal{A}| \geq \frac{p^*(T+1)}{8}$, $n \leq (T+1)$, $\lambda = C_4\sqrt{\frac{\log(T+1)+\log d}{T+1}}$, and $T \geq T_0$, where $C_4 = \frac{\sqrt{2}x_{\max}}{(\frac{1}{4}-\frac{192s}{p^*\kappa a})\min\{1,\frac{\mu_0}{p^* s x_{\max}^2}\}}$ and $T_0 = \max\left\{14, 4C_0|\mathcal{K}|, \frac{128}{(p^*)^2}, \frac{64}{C_1 p^*)^2}, \frac{256 s^2 x_{\max}^4}{(C_h p^*)^2}, \frac{64 s^2\sigma^2 x_{\max}^4(1+\log s)^2}{(\mu_0 p^*)^2}\right\}$, the following results can be obtained:

$$|\mathcal{A}| \geq \frac{2s^2 x_{\max}^2}{\mu_0},\ a > \frac{96ns}{\kappa|\mathcal{A}|} \text{ and } \beta_{\min} \geq (\frac{96ns}{\kappa|\mathcal{A}|} + a)\lambda.$$

We then have the following result via Proposition 4.5.1.

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\| \geq \sqrt{\frac{512 s^3\sigma^2 x_{\max}^2}{\mu_0^2 (p^*)^2 (T+1)}}\right) \\
&\leq \delta_1\left(\frac{p^*(T+1)}{8}\right) + \delta_2\left(T+1, \frac{p^*(T+1)}{8}, \lambda\right) + \delta_3\left(\frac{p^*(T+1)}{8}\right)
\end{aligned} \tag{4.7.39}$$

Combining $T > T_0$, $\lambda = C_4\sqrt{\frac{\log(T+1)+\log d}{T+1}}$ and the fact $T+1 \geq \sqrt{T+1}\log(T+1)$ for $T > 0$, we have

$$\delta_1\left(\frac{p^*(T+1)}{8}\right) + \delta_2\left(T+1, \frac{p^*(T+1)}{8}, \lambda\right) + \delta_3\left(\frac{p^*(T+1)}{8}\right) \leq \frac{4}{T+1} \tag{4.7.40}$$

$$\mathbb{P}\left(|\mathcal{A}| \leq \frac{p^*(T+1)}{8}\right) \leq \frac{1}{T+1}. \tag{4.7.41}$$

Set $C_\beta = \frac{512\sigma^2 x_{\max}}{\mu_0^2 (p^*)^2}$, and Proposition 4.5.4 directly follows by combining (4.7.40), (4.7.39), (4.7.41) and $\mathbb{P}(\mathcal{E}_6^c) \leq \frac{7}{T+1}$ from Lemma A.0.10.

□

4.7.8 Proof of Theorem 4.4.4

Proof. We divide the time, up to time T, into three groups and derive the cumulative regret bound for each group separately. Consider the following three groups:

1. $x_i \in \mathcal{R}_k, k \in \mathcal{K}$ and $T \leq T_0$.

2. $x_i \notin \mathcal{R}_k, k \in \mathcal{K}$, $T > T_0$ and $\mathcal{E}_6$ doesn't hold,

3. $x_i \notin \mathcal{R}_k, k \in \mathcal{K}$ $T > T_0$ and $\mathcal{E}_6$ holds.

Before going to the detail proof, we first state the choice of T_0 and C_0 such that the requirements of Proposition 4-6 are satisfied.

$$T_0 = \max\left\{\frac{(t_0+1)^2}{e^2} - 1, 14, 4C_0|\mathcal{K}|, \frac{128}{(p^*)^2}, \frac{64}{C_1 p^*)^2}, \frac{256 s^2 x_{\max}^4}{(C_h p^*)^2}, \frac{64 s^2\sigma^2 x_{\max}^4 (1+\log s)^2}{(\mu_0 p^*)^2}\right\}$$

$$C_0 = \max\left\{10, \frac{16}{p^*}, \frac{4}{p^* C_1}, \frac{4x_{\max}^2}{C_5^2}\left((\frac{1}{4} - \frac{576s}{p^*\kappa a})\min\left\{1, \frac{\mu_0 p^*}{192 s x_{\max}^2}\right\}\right)^{-2}, \right.$$
$$\left. \frac{32\sigma_2 s x_{\max}^2(1+\log s)}{p^*\mu_0}, \frac{4\sigma^2 x_{\max}^2(1+\log s)}{t^2}\right\},$$

where $t \leq \min\left\{\frac{\mu_0 p^*\sqrt{\tilde{C}_2\lambda}}{48}, \frac{p^*\mu_0}{48\sigma\sqrt{s}x_{\max}}, \frac{hp^*\mu_0}{192 e\sigma\sqrt{s}R_{\max}x_{\max}}\right\}$, $\tilde{C}_2 = \frac{\mu_0 p^*}{2\sigma_3 s x_{\max}^3(\mu_0 p^* + 48 s x_{\max}^2)}$, $C_1 = \min\left\{1, \kappa^2/\left(192 s\sigma_2 x_{\max}^2(3+2\sqrt{\sigma_2}x_{\max})\right)^2\right\}$ and $C_5 = \frac{\beta_{\min} p^*\kappa}{(2304 s + a p^*\kappa)\sqrt{1+\log d}}$.

Regret in part 1: Denote the regret for the first part as $R_1(T)$.

$$R_1(T) \leq R_{\max}\left(\sum_{i=T_0}^{T} \mathbb{1}(x_i \in \mathcal{R}_k, k \in \mathcal{K}) + T_0\right) \leq R_{\max}\left(\sum_{k\in\mathcal{K}} n_k + T_0\right). \quad (4.7.42)$$

From Proposition 4.4.3, we know that

$$\mathbb{P}\left(n_k \leq \frac{3t_0(1+\log(T)-\log(t_0))}{2|\mathcal{K}|}\right) \geq 1 - \frac{2}{T+1}. \quad (4.7.43)$$

If we require $t_0 = 2C_0|\mathcal{K}|$, $C_0 \geq 10$, and $T \geq \max\{(t_0+1)^2/e^2 - 1, 14\}$, then the above equation can be simplified to

$$\mathbb{P}\left(n_k \leq 6C_0 \log T\right) \geq 1 - \frac{2}{T+1} \Rightarrow \mathbb{P}\left(n_k > 6C_0 \log T\right) \leq \frac{2}{T+1} \quad (4.7.44)$$

which implies

$$\mathbb{P}\left(\sum_{k\in\mathcal{K}} n_k > 6C_0|\mathcal{K}|\log T\right) \leq \mathbb{P}\left(\cup_{k\in\mathcal{K}}(n_k > 6C_0\log T)\right) \leq \sum_{k\in\mathcal{K}} \mathbb{P}\left(n_k > 6C_0\log T\right) \leq \frac{2|\mathcal{K}|}{T+1}, \quad (4.7.45)$$

and

$$
\begin{aligned}
R_1(T) \leq R_{\max}\left(\sum_{k\in\mathcal{K}} n_k + T_0\right) &= R_{\max}\left(\sum_{k\in\mathcal{K}} n_k | \sum_{k\in\mathcal{K}} n_k > 6C_0|\mathcal{K}|\log T\right)\mathbb{P}\left(\sum_{k\in\mathcal{K}} n_k > 6C_0|\mathcal{K}|\log T\right) \\
&+ R_{\max}\left(\sum_{k\in\mathcal{K}} n_k | \sum_{k\in\mathcal{K}} n_k \leq 6C_0|\mathcal{K}|\log T\right)\mathbb{P}\left(\sum_{k\in\mathcal{K}} n_k \leq 6C_0|\mathcal{K}|\log T\right) \\
&+ R_{\max}T_0 \\
&\leq R_{\max}T\frac{2|\mathcal{K}|}{T+1} + R_{\max}6C_0|\mathcal{K}|\log T\left(1-\frac{2|\mathcal{K}|}{T+1}\right) + R_{\max}T_0 \\
&\leq 2R_{\max}|\mathcal{K}| + 6R_{\max}C_0|\mathcal{K}|\log T + R_{\max}T_0 \\
&\leq R_{\max}|\mathcal{K}|(2+6C_0\log T) + R_{\max}T_0. \qquad (4.7.46)
\end{aligned}
$$

Regret in part 2: Denote the regret for the second part as $R_2(T)$.From Lemma A.0.9, we know that

$$
\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\beta}^{random}(t) - \boldsymbol{\beta}^{true}\|_1 \leq \min\left\{\frac{1}{\sigma x_{\max}}, \frac{h}{4e\sigma R_{\max}x_{\max}}\right\}\right) \geq 1-\frac{7}{T+1},\ k\in\mathcal{K} \\
\Rightarrow &\mathbb{P}(\mathcal{E}_6(T)) \geq 1-\frac{7|\mathcal{K}|}{T+1}. \qquad (4.7.47)
\end{aligned}
$$

Therefore, $R_2(T)$ can be bounded as follows

$$
\begin{aligned}
R_2(T) &\leq \mathbb{E}[\sum_{i=1}^{T}\mathbb{1}(\mathcal{E}_6(i)^c)R_{\max}] \\
&= \sum_{i=1}^{T}\mathbb{E}[\mathbb{1}(\mathcal{E}_6(i)^c)]R_{\max} \\
&= \sum_{i=1}^{T}\mathbb{P}(\mathcal{E}_6(i)^c)R_{\max} \\
&\leq 7R_{\max}|\mathcal{K}|\log(T+1). \qquad (4.7.48)
\end{aligned}
$$

Regret in part 3: Denote the regret for the third part as $R_3(T)$. Without loss of generality, we assume that arm i is true optimal arm at time t. Then, the regret at time t can be bounded as follows

$$
r_t = \mathbb{E}\left(\mathbb{1}\left(j = \arg\max_{k\in\mathcal{K}}\mathbb{E}[R_k|\boldsymbol{x}_t, \boldsymbol{\beta}_k^{whole}(t)]\right)(\mathbb{E}[R_i|\boldsymbol{x}_t, \boldsymbol{\beta}_i^{true}] - \mathbb{E}[R_j|\boldsymbol{x}_t, \boldsymbol{\beta}_j^{true}])\right)
$$

$$\leq \mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]>\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]\right)\left(\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{true}]-\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{true}]\right)\right). \tag{4.7.49}$$

Denote $\mathcal{E}(t,\delta)_{8,k}=\{\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{true}]>\mathbb{E}[R_k|\boldsymbol{x}_t,\boldsymbol{\beta}_k^{true}]+\delta\},\ k\neq i,k\in\mathcal{K}$. Then we have the following bound.

$$\begin{aligned}
&r_t\\
\leq&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\left\{\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]>\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]\right\}\cap\mathcal{E}(t,\delta)_{8,j}\right)\right.\\
&\left.\times\left(\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{true}]-\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{true}]\right)\right)\\
+&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\left\{\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]>\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]\right\}\cap\mathcal{E}(t,\delta)_{8,j}^c\right)\right.\\
&\left.\times\left(\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{true}]-\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{true}]\right)\right)\\
\leq&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\left\{\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]>\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]\right\}\cap\mathcal{E}(t,\delta)_{8,j}\right)(2R_{\max})\right) \qquad (4.7.50)\\
+&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\left\{\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]>\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]\right\}\cap\mathcal{E}(t,\delta)_{8,j}^c\right)(\delta)\right). \qquad (4.7.51)
\end{aligned}$$

The term in (4.7.51) can be bounded as follows

$$\begin{aligned}
&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\left\{\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]>\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]\right\}\cap\mathcal{E}(t,\delta)_{8,j}^c\right)(\delta)\right)\\
\leq&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\mathcal{E}(t,\delta)_{8,j}^c\right)(\delta)\right)\\
=&\sum_{j\neq i}\mathbb{P}\left(\mathcal{E}(t,\delta)_{8,j}^c\right)\delta\\
=&(|\mathcal{K}|-1)CR_{\max}\delta^2\leq CR_{\max}|\mathcal{K}|\delta^2, \qquad (4.7.52)
\end{aligned}$$

where the last inequality comes from assumption **A.2**. Now we consider the term

in (4.7.50), which can be bounded as follows

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\left\{\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]>\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]\right\}\cap\mathcal{E}(t,\delta)_{8,j}\right)2R_{\max}\right)\\
\leq&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]-\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{true}]\right.\right.\\
&\left.\left.>\ \mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]-\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{true}]+\delta\right)2R_{\max}\right)\\
\leq&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\left|\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{whole}(t)]-\mathbb{E}[R_j|\boldsymbol{x}_t,\boldsymbol{\beta}_j^{true}]\right|\right.\right.\\
&\left.\left.>\ -\left|\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{whole}(t)]-\mathbb{E}[R_i|\boldsymbol{x}_t,\boldsymbol{\beta}_i^{true}]\right|+\delta\right)(2R_{\max})\right)\\
\leq&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(R_{\max}\sigma e^{2\sigma x_{\max}b}x_{\max}\|\boldsymbol{\beta}_k^{true}-\boldsymbol{\beta}_k^{whole}(t)\|_1\right.\right.\\
&\left.\left.>\ -R_{\max}\sigma e^{2\sigma x_{\max}b}x_{\max}\|\boldsymbol{\beta}_i^{true}-\boldsymbol{\beta}_i^{whole}(t)\|_1+\delta\right)(2R_{\max})\right)\\
\leq&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\|\boldsymbol{\beta}_k^{true}-\boldsymbol{\beta}_k^{whole}(t)\|_1+\|\boldsymbol{\beta}_i^{true}-\boldsymbol{\beta}_i^{whole}(t)\|_1\geq\frac{\delta}{R_{\max}\sigma e^{2\sigma x_{\max}b}x_{\max}}\right)(2R_{\max})\right),
\end{aligned}
\tag{4.7.53}
$$

where the second last inequality comes from the first part of the Lemma A.0.10 and $\|\boldsymbol{\beta}\|_1\leq b$ in assumption **A.1**. From Proposition 4.5.4, we have the following inequality.

$$
\mathbb{P}\left(\|\boldsymbol{\beta}_k^{whole}(t)-\boldsymbol{\beta}_k^{true}\|_2\geq\sqrt{C_\beta\frac{s^2}{T}}\right)\leq\frac{12}{T+1}. \tag{4.7.54}
$$

As $\|\boldsymbol{\beta}_k^{whole}(t)-\boldsymbol{\beta}_k^{true}\|_2\geq\frac{1}{\sqrt{s}}\|\boldsymbol{\beta}_k^{whole}(t)-\boldsymbol{\beta}_k^{true}\|_1$, (4.7.54) implies

$$
\mathbb{P}\left(\|\boldsymbol{\beta}_k^{whole}(t)-\boldsymbol{\beta}_k^{true}\|_1\geq\sqrt{C_\beta\frac{s^3}{T+1}}\right)\leq\frac{12}{T+1}. \tag{4.7.55}
$$

Denote event $\mathcal{E}_9$ as follows

$$
\mathcal{E}_9=\{\|\boldsymbol{\beta}_k^{whole}(t)-\boldsymbol{\beta}_k^{true}\|_1\geq\frac{\delta}{2R_{\max}\sigma e^{2\sigma x_{\max}b}x_{\max}},\ k\in\mathcal{K}\}. \tag{4.7.56}
$$

Combining (4.7.53) and (4.7.55), we have:

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\|\boldsymbol{\beta}_j^{true}-\boldsymbol{\beta}_j^{whole}(t)\|_1\|\boldsymbol{\beta}_i^{true}-\boldsymbol{\beta}_i^{whole}(t)\|_1\geq\frac{\delta}{R_{\max}\sigma e^{2\sigma x_{\max}b}x_{\max}}\right)(2R_{\max})\right)\\
=&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\|\boldsymbol{\beta}_j^{true}-\boldsymbol{\beta}_j^{whole}(t)\|_1+\|\boldsymbol{\beta}_i^{true}-\boldsymbol{\beta}_i^{whole}(t)\|_1\geq\frac{\delta}{R_{\max}\sigma e^{2\sigma x_{\max}b}x_{\max}}\middle|\,\mathcal{E}_9\right)\mathbb{1}(\mathcal{E}_9)(2R_{\max})\right)\\
+&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\|\boldsymbol{\beta}_j^{true}-\boldsymbol{\beta}_j^{whole}(t)\|_1+\|\boldsymbol{\beta}_i^{true}-\boldsymbol{\beta}_i^{whole}(t)\|_1\geq\frac{\delta}{R_{\max}\sigma e^{2\sigma x_{\max}b}x_{\max}}\middle|\,\mathcal{E}_9^c\right)\mathbb{1}(\mathcal{E}_9^c)(2R_{\max})\right)\\
\leq&\mathbb{E}\left(\sum_{j\neq i}\mathbb{1}\left(\frac{1}{2}\delta+\frac{1}{2}\delta\geq\delta\middle|\,\mathcal{E}_9\right)\mathbb{1}\left(\mathcal{E}_9\right)\left(2R_{\max}\right)\right)+0\\
=&\mathbb{E}\left(\mathbb{1}\left(\mathcal{E}_9(t)\right)\left(2R_{\max}\right)\right)\leq 2R_{\max}\mathbb{P}(\mathcal{E}_9).
\end{aligned}
\tag{4.7.57}
$$

Furthermore, by setting $\delta=2R_{\max}\sigma e^{2\sigma x_{\max}b}x_{\max}\sqrt{C_\beta\frac{s^3}{T+1}}$, we have the following result:

$$
r_t\leq 2R_{\max}\mathbb{P}(\mathcal{E}_9)+CR_{\max}|\mathcal{K}|\delta^2\leq\frac{24R_{\max}|\mathcal{K}|}{T+1}+CR_{\max}|\mathcal{K}|\frac{4R_{\max}^2\sigma^2e^{4\sigma x_{\max}b}x_{\max}^2C_\beta s^3}{T+1}=\frac{C_{R_3}}{T+1}
\tag{4.7.58}
$$

where $C_{R_3}=24R_{\max}|\mathcal{K}|+4e^{4\sigma_2 x_{\max}b}CR_{\max}^3|\mathcal{K}|x_{\max}^2C_\beta s^3$. Hence, the third part of the regret can be bounded as follows:

$$
R_3(T)=\sum_{i=1,i\in\mathcal{R}(T)}^{T}r_t\leq\sum_{i=1}^{T}\frac{C_{R_3}}{T}\leq\int_1^T\frac{C_{R_3}}{t}dt\leq C_{R_3}\log(T)
\tag{4.7.59}
$$

Finally, the total regret bound can be obtained by combining the bounds for these three parts:

$$
\begin{aligned}
&R_1(T)+R_2(T)+R_3(T)\\
&\leq R_{\max}[|\mathcal{K}|(2+6C_0\log T)+T_0]+7R_{\max}|\mathcal{K}|\log(T+1)+C_{R_3}\log(T)\\
&\leq R_{\max}(T_0+|\mathcal{K}|)+(6R_{\max}|\mathcal{K}|C_0+31R_{\max}|\mathcal{K}|+4\sigma^2e^{4\sigma_2x_{\max}b}CR_{\max}^3|\mathcal{K}|x_{\max}^2C_\beta s^3)\log(T+1)\\
&=O(|\mathcal{K}|s^2(s+\log d)\log T).
\end{aligned}
$$

□

4.8 Conclusion

In this research, we develop the G-MCP-Bandit algorithm for online learning and decision-making processes in high-dimensional settings under limited samples. We adopt the matrix perturbation technique to derive new oracle inequality for the MCP estimator under non-iid samples and further propose a linear approximation method, the 2sWL procedure, to overcome the computational and statistical challenges associated with solving the MCP estimator (an NP-complete problem) under the bandit setting. We demonstrate that the MCP estimator solved by the 2sWL procedure matches the oracle estimator with high probability and converges to the true parameters with the optimal convergence rate. Further, we show that the cumulative regret of the G-MCP-Bandit algorithm over the sample size T is bounded by $O(\log T)$, which is the lowest theoretical bound for all possible algorithms under both low-dimensional and high-dimensional settings. In the covariate dimension d, the cumulative regret of the G-MCP-Bandit algorithm is bounded by $O(\log d)$, which is also a tighter bound than existing bandit algorithms. Finally, we illustrate that compared to other benchmark algorithms, the G-MCP-Bandit algorithm performs favorably in synthetic-data-based and real-data-based experiments.

Implementing the G-MCP-Bandit algorithm under high-dimensional data with a large decision set in an online setting can be challenging in practice, and addressing these challenges can extend this research to several directions. One of the major challenges is the computation time, especially when the covariate dimension and the decision set are extremely large. In particular, during a collaboration with a leading online marketplace, we adopted the G-MCP-Bandit algorithm, aiming to improve its product recommendation system. Using its datasets (with 5 million covariates and 30 million products), we showed that the G-MCP-Bandit algorithm improved the prediction of the conversion rate by 15% and the expected revenue by 5% on average, but a single server could take hours to execute the algorithm. We can implement the G-MCP-Bandit algorithm in a hybrid online-offline setting, where we recommend products by following the bi-level decision structure for every user but update the parameter vector estimation $\boldsymbol{\beta}^{random}$ and $\boldsymbol{\beta}^{whole}$ in batches every a couple of hours. Yet, in order to implement the G-MCP-Bandit algorithm in *online* settings, where we also update the parameter vector estimation for every incoming user, parallel computation techniques must be developed to tremendously

reduce the computation time. Other challenges for the G-MCP-Bandit algorithm are how to simultaneously recommend multiple products and how to dynamically update the recommendation if the user did not click the recommended products but kept refreshing the recommendation page. Tackling these challenges requires an integration of the assortment optimization and Bayesian learning into the G-MCP-Bandit algorithm.

Chapter 5
Conclusions and Future Research

This dissertation consists of three problems in high-dimensional learning and decision making:

1. A sample average approximation with the folded concave penalty for high-dimensional stochastic programming

2. An accelerated interior point gradient method for large scale linear constrained nonconvex programming

3. A contextual bandit algorithm for online learning and decision making with high-dimensional features

The first work is presented in Chapter 2. We propose the RSAA, a modification to the SAA by incorporating a regularization scheme called the FCP. This modification targets the high-dimensional SP problems with sparsity. We show that when the solution is sparse or can be approximated by a sparse solution, the regularization can significantly reduce the required number of samples in some high-dimensional SP applications. Compared to the conventional SAA approach that requires the sample size to grow polynomially in the number of dimensions, the RSAA requires the number of samples that is only poly-logarithmic in the dimensionality. Future direction includes:

1. (**Development of new solution scheme**) In our current work, we directly adopt the second order interior point algorithm in [92, 93], of which computation complexity is $O(p^6/\epsilon \log(1/\epsilon))$. Although the dependence on error ϵ is promising, the dependence on dimensionality p is not good enough, especially

for the high-dimensional problem. We will explore the solution scheme with better dimensionality and error trade-off, such as coordinate descent method and accelerated gradient descent with eigenvalue checking.

2. (**The possibility of distributed setting**) Our current work focuses on the sample size requirement. In practice, due to storage limitation or data privacy, the dataset needs to be separately stored at different locations. It poses a new challenge on how to design an efficient distributed algorithm to solve the RSAA problem. The communication efficient framework will be considered.

3. (**Perishable data**) In our current setting, we assume that the true model is static and the whole dataset could be stored. But in real world application (e.g. online advertising), the true model might evolve with time and we can only efficiently keep and process the very limit amount of active data, i.e. the historical data is perishing. We will extend our current work with the stochastic approximation (e.g. stochastic gradient descent (SGD) and Stochastic Variance Reduction gradient (SVRG)) to develop a new computational and statistical friendly framework.

In Chapter 3 we discuss the second work. We design an accelerated interior point gradient method (AIP-GM) for non-convex programming with linear constraints. Many important problems (e.g., l_1-minimization, regularized neural network) can be formulated into this form. AIP-GM is guaranteed to reach a ϵ approximated second order solution in $O(\epsilon^{-7/4}\log(1/\epsilon)^2)$ iteration. It improves upon the $O(\epsilon^{-2})$ complexity of the gradient descent methods and provides additional second order guarantee. In each iteration, only gradient calculation and matrix-vector multiplication are required, which makes AIP-GM being suitable for large scale problem arising in machining learning as well as other areas. Future research directions include:

1. (**ADMM with eigenvalue checking**) Our current framework can be summarized as accelerating the first order interior point method with eigenvalue checking. We will research boosting ADMM type of algorithm with the similar technique.

2. (**Statistical property**) Different from the vanilla first order methods, our approach can even attain the second necessary solution. In [47] authors show

that the higher order necessary solution can ensure statistical property for FCP least squared regression problem. We will study whether our approach can attain the solution with the statistical guarantee for general convex loss functions in machine learning or statistical learning.

3. (**Numerical studies**) We will conduct comprehensive numerical examples presentation, solution comparison and discussions on findings/performances.

The last work is shown in Chapter 4. we develop the MCP-Bandit algorithm for online learning and decision-making processes in high-dimensional data settings. To further tackle the computational and statistical challenges associated with solving the MCP estimator under non-i.i.d. samples, we propose a linear approximation method, 2sWL procedure, under the bandit setting and show that the MCP estimator solved by the 2sWL procedure matches the oracle estimator with high probability. We demonstrate that the cumulative regret of the MCP-Bandit algorithm over sample size T is bounded by $O(\log T)$, which is the lowest theoretical bound for all possible algorithms. In covariate dimension d and the number of significant covariate dimension s, the cumulative regret of the MCP-Bandit algorithm is bounded by $O(s^2(s + \log d))$, which is also a tighter bound than the Lasso-Bandit algorithm. We show that the MCP-Bandit algorithm performs favorably in all our experiments, especially when the data sparsity level is high or when the sample size is not too large. We will consider the following future research direction:

1. (**Contextual arm**) We now only focus on the setting that the user has covariates. A more realistic situation would be that arms are also described by covariates and we need to make a decision based on both covariates. We will consider adding arm covariates module into our model to address this issue.

2. (**Perishable data**) The real world problem may involve time-evolving effects. We will combine our current model with the methods in time-series analysis.

Appendix A
Supplement material for Chapter 4

Lemma A.0.1. *Let $\mathcal{A}$ be the set of iid samples. Under assumption A.1 and A.5, there exists a constant $\mu_0 > 0$ such that for all feasible $\boldsymbol{\xi}$ defined in assumption A.4 we have*

$$\mathbb{P}\left(\lambda_{\min}(\nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}(\boldsymbol{\xi})) \geq \frac{|\mathcal{A}|}{2n}\mu_0\right) \leq 1 - s\exp\left(-\frac{|\mathcal{A}|\mu_0}{8s\sigma_2 x^2_{\max}}\right). \tag{A.0.1}$$

Proof. Proof of Lemma A.0.1 Note that $f(\cdot|\cdot)$ is convex and has smooth gradient. We denote $\boldsymbol{z}'_j = \boldsymbol{x}_{j,\mathcal{S}}\sqrt{f''(r_j|\boldsymbol{x}^T_{j,\mathcal{S}}\boldsymbol{\xi}_{\mathcal{S}})}$. Combine with $f(\cdot|\cdot) = -\log g(\cdot|\cdot)$ and we have

$$\nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{i,\mathcal{S}}x^T_{i,\mathcal{S}}f''(r_i|\boldsymbol{x}^T_{i,\mathcal{S}}\boldsymbol{\xi}_{\mathcal{S}}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}'_i(\boldsymbol{z}'_i)^T \succeq \lambda_{\min}\left(\frac{1}{n}\sum_{j\in\mathcal{A}}\boldsymbol{z}'_j(\boldsymbol{z}'_j)^T\right) I.$$

Then, we bound $\lambda_{\min}\left(\frac{1}{n}\sum_{j\in\mathcal{A}^c}\boldsymbol{z}'_j(\boldsymbol{z}'_j)^T\right)$ via Theorem 5.1.1 in [84] with $\epsilon = 1/2$:

$$\mathbb{P}\left(\lambda_{\min}(\frac{1}{n}\sum_{j\in\mathcal{A}}\boldsymbol{z}'_j(\boldsymbol{z}'_j)^T) \leq \frac{1}{2}\lambda_{\min}(\mathbb{E}[\frac{1}{n}\sum_{j\in\mathcal{A}}\boldsymbol{z}'_j(\boldsymbol{z}'_j)^T])\right) \leq s\left(\frac{\exp(-1/2)}{\sqrt{1/2}}\right)^{\frac{\lambda_{\min}(\mathbb{E}[\frac{1}{n}\sum_{j\in\mathcal{A}}\boldsymbol{z}'_j(\boldsymbol{z}'_j)^T])}{s\sigma_2 x^2_{\max}/n}} \tag{A.0.2}$$

$$\Rightarrow \mathbb{P}\left(\lambda_{\min}(\frac{1}{n}\sum_{j\in\mathcal{A}}^n\boldsymbol{z}'_j(\boldsymbol{z}'_j)^T) \leq \frac{|\mathcal{A}|}{2n}\lambda_{\min}(\mathbb{E}[\boldsymbol{z}'_j(\boldsymbol{z}'_j)^T])\right) \leq s\exp\left(-\frac{\log(e/2)n\lambda_{\min}(\frac{|\mathcal{A}|}{n}\mathbb{E}[\boldsymbol{z}'_j(\boldsymbol{z}'_j)^T])}{2s\sigma_2 x^2_{\max}}\right), \tag{A.0.3}$$

where (A.0.2) uses $0 \leq \lambda_{\min}(\frac{1}{n}\boldsymbol{z}_j'(\boldsymbol{z}_j')^T) \leq \lambda_{\max}(\frac{1}{n}\boldsymbol{z}_j'(\boldsymbol{z}_j')^T) \leq \frac{s}{n}(z'_{\max})^2 = \frac{s}{n}\sigma_2 x_{\max}^2$ and the last inequality comes from the assumption **A.1**. As we only consider the significant dimensions, under assumption **A.4**, we can verify that there exists a $\mu_0 > 0$ such that $\mathbb{E}[\boldsymbol{z}_j'(\boldsymbol{z}_j')^T] = \mathbb{E}[\nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi})] \succeq \mu_0 I$. Then, we have $\mathbb{1}\left(\lambda_{\min}(\frac{1}{n}\sum_{j\in\mathcal{A}}^n \boldsymbol{z}_j'(\boldsymbol{z}_j')^T) \leq \frac{|\mathcal{A}|}{2n}\lambda_{\min}(\mathbb{E}[\boldsymbol{z}_j'(\boldsymbol{z}_j')^T])\right) \geq \mathbb{1}\left(\lambda_{\min}(\frac{1}{n}\sum_{j\in\mathcal{A}}^n \boldsymbol{z}_j'(\boldsymbol{z}_j')^T) \leq \frac{|\mathcal{A}|}{2n}\mu_0\right)$. Thus (A.0.3) implies

$$\mathbb{P}\left(\lambda_{\min}(\frac{1}{n}\sum_{j\in\mathcal{A}}^n \boldsymbol{z}_j'(\boldsymbol{z}_j')^T) \leq \frac{|\mathcal{A}|}{2n}\mu_0\right) \leq s\exp\left(-\frac{\log(e/2)|\mathcal{A}|\lambda_{\min}(\mathbb{E}[\boldsymbol{z}_j'(\boldsymbol{z}_j')^T])}{2s\sigma_2 x_{\max}^2}\right).$$

Combining with the fact $\log(e/2)/2 \geq 1/8$, Lemma A.0.1 follows immediately. □

Lemma A.0.2. *Let the whole sample size be n and iid random sample set be $\mathcal{A}$. If assumptions* **A.1**,**A.4** *and* **A.5** *hold, there exist $\mu_0 > 0$ such that for $t > 0$ we have*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\| \geq \frac{2nt}{|\mathcal{A}|\mu_0}\right) \leq s\exp\left(-\frac{|\mathcal{A}|\mu_0}{8s\sigma_2 x_{\max}^2}\right) + s\exp\left(-\frac{nt^2}{2s\sigma^2 x_{\max}^2}\right). \tag{A.0.4}$$

Furthermore, if $|\mathcal{A}| \geq \frac{2s^2 x_{\max}^2}{\mu_0}$ we have

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_2 \geq \sqrt{\frac{8s^2\sigma^2 x_{\max}^2 n}{\mu_0^2|\mathcal{A}|^2}}\right) \leq s\exp\left(-\frac{\mu_0|\mathcal{A}|}{8s\sigma_2 x_{\max}^2}\right) + 2\exp\left(-\frac{C_h|\mathcal{A}|\mu_0}{2s\sigma_2 x_{\max}^2}\right), \tag{A.0.5}$$

where C_h is a positive constant.

Proof. Proof of Lemma A.0.2

From the definition of oracle solution, we know

$$\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{oracle}) = 0. \tag{A.0.6}$$

Expanding (A.0.6) at $\boldsymbol{\beta}^{true}$ we will have the following result for some $\boldsymbol{\xi} \in \{\tau\boldsymbol{\beta}^{oracle} + (1-\tau)\boldsymbol{\beta}^{true}, \tau \in [0,1]\}$.

$$\begin{aligned}\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true}) + \nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}(\xi)(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) &= 0\\ \nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}(\xi)(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) &= -\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})\end{aligned}$$

$$
\begin{aligned}
(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true})^T \nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\xi)(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) &= -(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true})^T \nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}) \\
\lambda_{\min}(\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\xi)) \|(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true})\|_2^2 &\leq \|(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true})\|_2 \|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})\|_2 \\
\lambda_{\min}(\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\xi)) \|(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true})\|_2 &\leq \|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})\|_2. \qquad \text{(A.0.7)}
\end{aligned}
$$

The $\lambda_{\min}(\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\xi))$ term on the left hand side of (A.0.7) can be lower bounded away 0 via Lemma A.0.1 with high probability. Thus we only need to construct the upper bound for right-hand side of (A.0.7) that can be expanded as follows

$$
\|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})\|_2 = \left\| \frac{1}{n} \sum_{j=1}^{n} x_{j\mathcal{S}}^T f^{'}(r_j | \boldsymbol{x}_{j,\mathcal{S}}^T \boldsymbol{\beta}^{true}) \right\|_2 . \qquad \text{(A.0.8)}
$$

Under assumption **A.5**, we have $|f^{'}(r_j | \boldsymbol{x}_{j,\mathcal{S}}^T \boldsymbol{\beta}^{true})| \leq \sigma$. Combining with $\mathbb{E}[f^{'}(r_j | \boldsymbol{x}_{j,\mathcal{S}}^T \boldsymbol{\beta}^{true})] = 0$, we can verify that $f^{'}(r_j | \boldsymbol{x}_{j,\mathcal{S}}^T \boldsymbol{\beta}^{true})$ is a σ-subgaussian random variable. From Hoeffding inequality, there exists a $t > 0$ such that

$$
\mathbb{P}\left(\left| \frac{1}{n} \sum_{j=1}^{n} x_{ji}^T f^{'}(r_j | \boldsymbol{x}_j^T \boldsymbol{\beta}^{true}) \right| \geq t \right) \leq \exp\left(-\frac{nt^2}{2\sigma^2 x_{\max}^2} \right) \quad \forall i \in \mathcal{S}. \qquad \text{(A.0.9)}
$$

Hence, we have

$$
\begin{aligned}
\mathbb{P}\left(\|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})\|_2 \geq t \right) &= \mathbb{P}\left(\left\| \frac{1}{n} \sum_{j=1}^{n} x_{ji}^T f^{'}(r_j | \boldsymbol{x}_j^T \boldsymbol{\beta}^{true}) \right\|_2 \geq t \right) \\
&\leq \mathbb{P}\left(\sqrt{|\mathcal{S}|} \left\| \frac{1}{n} \sum_{j=1}^{n} x_{ji}^T f^{'}(r_j | \boldsymbol{x}_j^T \boldsymbol{\beta}^{true}) \right\|_\infty \geq t \right) \\
&\leq s \exp\left(-\frac{nt^2}{2s\sigma^2 x_{\max}^2} \right), \qquad \text{(A.0.10)}
\end{aligned}
$$

where the inequality in (A.0.10) follows from $|\mathcal{S}| \leq s$. Combining (A.0.10), (A.0.7) and Lemma A.0.1, the statement in (A.0.4) follows.

Now, the first half of Lemma A.0.2 has been proven, and we switch to the second half. Denote $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, ..., \epsilon_n]$ where $\epsilon_j = f^{'}(r_j | \boldsymbol{x}_{j,\mathcal{S}}^T \boldsymbol{\beta}^{true})$, $j = 1, 2, .., n$. Then $\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{true}})$ can be rewritten as $\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{true}}) = \frac{1}{n} \boldsymbol{X}_{\mathcal{S}} \boldsymbol{\epsilon}$ with $\boldsymbol{X}_{\mathcal{S}} = [\boldsymbol{x}_{1,\mathcal{S}}, ..., \boldsymbol{x}_{n,\mathcal{S}}]$. Using the Hanson-Wright inequality (Theorem 1.1 in [72]), we have

$$
\mathbb{P}\{ |\boldsymbol{\epsilon}^T (\frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^T \boldsymbol{X}_{\mathcal{S}}) \boldsymbol{\epsilon} - \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T (\frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^T \boldsymbol{X}_{\mathcal{S}}) \boldsymbol{\epsilon}]| > \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T (\frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^T X_{\mathcal{S}}) \boldsymbol{\epsilon}] \}
$$

$$
\begin{aligned}
&\leq 2\exp\left(-C_h\min\left\{\frac{\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\boldsymbol{\epsilon}]}{\sigma^2\|\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}}\|_2}, \frac{(\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\boldsymbol{\epsilon}])^2}{\sigma^4\|\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}}\|_F^2}\right\}\right)\\
&\leq 2\exp\left(-C_h\min\left\{\frac{\lambda_{\min}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}{\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}\frac{\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}]}{\sigma^2}, \frac{\lambda_{\min}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})^2}{\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})^2}\frac{\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}]^2}{s\sigma^4}\right\}\right)\\
&\leq 2\exp\left(-C_h\min\left\{n\frac{\lambda_{\min}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}{\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}, \frac{n^2}{s}\frac{\lambda_{\min}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})^2}{\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})^2}\right\}\right)\\
&\leq 2\exp\left(-n\frac{C_h\lambda_{\min}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}{\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}\right), \qquad \text{(A.0.11)}
\end{aligned}
$$

where C_h is a positive constant and $\mathbb{E}_{\boldsymbol{\epsilon}}$ denote the expectation with respect to $\boldsymbol{\epsilon}$. The last inequality, (A.0.11), holds when $n \geq s\frac{\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}{\lambda_{\min}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}$. Define the event $\mathcal{E}_1$ as follows

$$
\mathcal{E}_1 = \left\{|\boldsymbol{\epsilon}^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\boldsymbol{\epsilon} - \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\boldsymbol{\epsilon}]| \leq \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\boldsymbol{\epsilon}]\right\}. \qquad \text{(A.0.12)}
$$

Under event $\mathcal{E}_1$, we have

$$
\|\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}\boldsymbol{\epsilon}\|_2 \leq \sqrt{\frac{1}{n}\boldsymbol{\epsilon}^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\boldsymbol{\epsilon}} \leq \sqrt{\frac{2}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\boldsymbol{\epsilon}]}. \qquad \text{(A.0.13)}
$$

Let $\boldsymbol{P}_j = \boldsymbol{X}_{\mathcal{S}}^T(\boldsymbol{X}_{\mathcal{S}}\boldsymbol{X}_{\mathcal{S}}^T)^{-1}\boldsymbol{X}_{\mathcal{S}}$. We have $(\boldsymbol{P}_j\boldsymbol{\epsilon})^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})(\boldsymbol{P}_j\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\boldsymbol{\epsilon}$, and (A.0.13) implies the following result.

$$
\begin{aligned}
\|\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}\boldsymbol{\epsilon}\|_2 &\leq \sqrt{\frac{2}{n}\mathbb{E}_{\boldsymbol{\epsilon}}[(\boldsymbol{P}_j\boldsymbol{\epsilon})^T(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})(\boldsymbol{P}_j\boldsymbol{\epsilon})]}\\
&\leq \sqrt{\frac{2}{n}\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\mathbb{E}_{\boldsymbol{\epsilon}}[\|\boldsymbol{P}_j\boldsymbol{\epsilon}\|_2^2]}\\
&\leq \sqrt{\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})\frac{2s\sigma^2}{n}}, \qquad \text{(A.0.14)}
\end{aligned}
$$

where the last inequality comes the facts that $\mathbb{E}_{\boldsymbol{\epsilon}}[\|\boldsymbol{P}_j\boldsymbol{\epsilon}\|_2^2] = s\sigma^2$ in which $\boldsymbol{P}_j$ can be viewed as a projection matrix from n dimension to s dimension and ϵ_j is a σ-subguassian random variable. Therefore, from $\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{true}}) = \frac{1}{n}\boldsymbol{X}_{\mathcal{S}}\boldsymbol{\epsilon}$,

$\lambda_{\max}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}}) \leq sx_{\max}^2$ and (A.0.11)-(A.0.14), we have the following inequalities.

$$\mathbb{P}\left(\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})\|_2 \leq \sqrt{\frac{2s^2\sigma^2x_{\max}^2}{n}}\right) \geq 1 - 2\exp\left(-n\frac{C_h\lambda_{\min}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})}{sx_{\max}^2}\right). \tag{A.0.15}$$

Since $\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}} = \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_{j,\mathcal{S}}\boldsymbol{x}_{j,\mathcal{S}}^T \succeq \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_{j,\mathcal{S}}\boldsymbol{x}_{j,\mathcal{S}}^T\frac{f''(r_j|\boldsymbol{x}_{j,\mathcal{S}}^T\boldsymbol{\xi})}{\sigma_2} = \frac{1}{\sigma_2}\nabla_{\mathcal{S},\mathcal{S}}^2\mathcal{L}(\boldsymbol{\xi})$. We then may apply the Lemma A.0.1 to further lower bound $\lambda_{\min}(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^T\boldsymbol{X}_{\mathcal{S}})$ by $\frac{|\mathcal{A}|\mu_0}{2n\sigma_2}$ for some $\mu_0 > 0$ with high probability and then (A.0.5) follows. □

Lemma A.0.3. *If there exists K and σ_0 such that $K^2\left(E[\exp(\boldsymbol{z}_{t,i}^2/K^2) - 1]\right) \leq \sigma_0^2$, then the following probability bound will hold for all $t > 0$:*

$$P\left\{\|\frac{1}{n}\sum_{j=1}^n \boldsymbol{z}_j\boldsymbol{z}_j^T - E[\boldsymbol{z}_j\boldsymbol{z}_j^T]\|_\infty \geq 2K^2t + 2K\sigma_0\sqrt{2t} + 2K\sigma_0\lambda\left(\frac{K}{\sigma_0}, n, \binom{d}{2}\right)\right\} \leq \exp\left(-nt\right) \tag{A.0.16}$$

where $\lambda\left(\frac{K}{\sigma_0}, n, \binom{d}{2}\right) = \sqrt{\frac{2\log(d(d-1))}{n}} + \frac{K\log(d(d-1))}{n}$.

Proof. Proof of A.0.3 From the exercise 14.3 in [14]. □

Lemma A.0.4. *If there exist κ_0, $\mathcal{S}$, and $\boldsymbol{z}_j$, $j = 1, 2, .., n$ such that $\|u_S\|_1^2 \leq \frac{|\mathcal{S}|}{\kappa_0}u^T\mathbb{E}[\boldsymbol{z}_j\boldsymbol{z}_j^T]u$ holds for all $u \in \mathcal{U} \doteq \{u : \|u_{S^c}\|_1 \leq 3\|u_S\|\}$ and $\left\|\frac{1}{n}\sum_{j=1}^n \boldsymbol{z}_j\boldsymbol{z}_j^T - \mathbb{E}[\boldsymbol{z}_j\boldsymbol{z}_j^T]\right\| \leq \frac{\kappa}{32|S|}$, then for all $u \in \mathcal{U}$, the follow inequality holds:*

$$\|u_S\|_1^2 \leq \frac{|S|}{\kappa_0/2}u^T\left[\frac{1}{n}\sum_{j=1}^n \boldsymbol{z}_j\boldsymbol{z}_j^T\right]u \tag{A.0.17}$$

Proof. Proof of A.0.4 From Corollary 6.8 in [14]. □

Lemma A.0.5. *Let $\boldsymbol{x}_j$, $j = 1, 2, ..., n$, be random iid samples. Under assumptions* **A.4** *and* **A.5**, *the follow inequality holds for all $\boldsymbol{u}$ such that $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq 3\|\boldsymbol{u}_{\mathcal{S}}\|_1$:*

$$\mathbb{P}\left(\frac{\kappa}{2s}\|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq u^T\nabla^2\mathcal{L}(\boldsymbol{\beta})\boldsymbol{u}\right) \geq 1 - \exp(-C_1n), \tag{A.0.18}$$

where $C_1 = \min\left\{1, \kappa^2/\left(192s\sigma_2x_{\max}^2(3 + 2\sqrt{\sigma_2}x_{\max})\right)^2\right\}$.

Proof. Proof of A.0.5 From the definition of $\mathcal{L}(\boldsymbol{\beta})$, we have $\nabla^2\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j\boldsymbol{x}_j^T f''(r_j, \boldsymbol{x}_j^T\boldsymbol{\xi})$. Under assumption **A. 5**, we know that f is convex with smooth gradient. We may

denote $\boldsymbol{z}_j = \boldsymbol{x}_j\sqrt{f''(r_j, \boldsymbol{x}_j^T\boldsymbol{\xi})}$ and then get $\nabla^2\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{n}\sum_{j=1}^n \boldsymbol{z}_j\boldsymbol{z}_j^T$. Furthermore, under assumption **A.1** and **A.5**, we have $|f''(r_j|\boldsymbol{x}_j^T\boldsymbol{\xi})| \leq \sigma_2$ and $\|\boldsymbol{x}\|_\infty \leq x_{\max}$, which implies that $\boldsymbol{z}_j$ is element-wise bounded by $z_{\max} = \|\boldsymbol{z}_j\|_\infty = \left\|\boldsymbol{x}_j\sqrt{f''(r_j|\boldsymbol{x}_j^T\boldsymbol{\xi})}\right\|_\infty \leq \sqrt{\sigma_2}x_{\max}$. Since $\boldsymbol{z}_j$ is bounded, it will satisfy the definition of the subguassian random variable. We can use the Lemma A.0.3 as a bridge to connect the sample matrix $\frac{1}{n}\sum_{j=1}^n \boldsymbol{z}_j\boldsymbol{z}_j^T$ to its population counterpart $\mathbb{E}[\boldsymbol{z}_j\boldsymbol{z}_j^T]$. Let $K = z_{\max}$ and $\sigma_0 = \sqrt{2}z_{\max}$ and we will have $K^2\left(E[\exp(\boldsymbol{z}_{t,i}^2/K^2) - 1]\right) \leq z_{\max}^2(e-1) \leq \sigma_0^2$ for all $t \geq 0$ and $i = 1, 2, ..., d$. Therefore, under Lemma A.0.3, for $t > 0$, we have

$$P\left\{\left\|\frac{1}{n}\sum_{j=1}^n \boldsymbol{z}_j\boldsymbol{z}_j^T - E[\boldsymbol{z}_j\boldsymbol{z}_j^T]\right\|_\infty \geq 2z_{\max}^2 t + 4z_{\max}^2\sqrt{t} + \sqrt{8}z_{\max}^2\lambda\left(\frac{\sqrt{2}}{2}, n, \binom{d}{2}\right)\right\} \leq \exp(-nt), \tag{A.0.19}$$

where $\lambda\left(\frac{\sqrt{2}}{2}, n, \binom{d}{2}\right) = \sqrt{\frac{2\log(d(d-1))}{n}} + \frac{z_{\max}\log(d(d-1))}{n}$. (A.0.19) indicates that when the sample size is large enough, $\frac{1}{n}\sum_{j=1}^n \boldsymbol{z}_j\boldsymbol{z}_j^T$ will not be far away from $\mathbb{E}[\boldsymbol{z}_j\boldsymbol{z}_j^T]$ element-wise with high probability.

Now we only need to show that if $\frac{1}{n}\sum_{j=1}^n \boldsymbol{z}_j\boldsymbol{z}_j^T$ is close enough to $\mathbb{E}[\boldsymbol{z}_j\boldsymbol{z}_j^T]$, $\nabla^2\mathcal{L}$ satisfies (A.0.18). To this end, we need Lemma A.0.4. We set $n \geq \log d/C_1$ and $t = C_1$ in (A.0.19). Then the following inequalities hold.

$$2z_{\max}^2 t + 4z_{\max}^2\sqrt{t} \leq 2z_{\max}^2\sqrt{C_1} + 4z_{\max}^2\sqrt{C_1} = 6z_{\max}^2\sqrt{C_1} \tag{A.0.20}$$

$$\sqrt{8}z_{\max}^2\lambda\left(\frac{\sqrt{2}}{2}, n, \binom{d}{2}\right) \leq \sqrt{8}z_{\max}^2\left(\sqrt{\frac{2\log(d^2)}{n}} + \frac{z_{\max}\log(d^2)}{n}\right) \leq 8\sqrt{2}z_{\max}^2(1 + z_{\max})\sqrt{C_1}, \tag{A.0.21}$$

where (A.0.20) and (A.0.21) use $\log d/n \leq C_1 \leq 1$. Combining (A.0.20) and (A.0.21), we have

$$\begin{aligned} 2z_{\max}^2 t + 4z_{\max}^2\sqrt{t} + \sqrt{8}z_{\max}^2\lambda\left(\frac{\sqrt{2}}{2}, n, \binom{d}{2}\right) &\leq 2z_{\max}^2\left(3 + 4\sqrt{2}(1 + z_{\max})\right)\sqrt{C_1} \\ &\leq 6z_{\max}^2\left(3 + 2z_{\max}\right))\sqrt{C_1} \leq \frac{\kappa}{32s}, \end{aligned} \tag{A.0.22}$$

where (A.0.22) uses $\sqrt{2} \leq \frac{3}{2}$ and $C_1 \leq \kappa^2/\left(192s\sigma_2 x_{\max}^2\left(3 + 2\sqrt{\sigma_2}x_{\max}\right)\right)^2 \leq$

$\kappa^2/\left(192sz_{\max}^2\left(3+2z_{\max}\right)\right))^2$. Then, (A.0.19) can satisfy the following inequality.

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{z}_j\boldsymbol{z}_j^T - E[\boldsymbol{z}_j\boldsymbol{z}_j^T]\right\|_\infty \leq \frac{\kappa}{32s}\right\} \geq 1-\exp\left(-C_1 n\right). \tag{A.0.23}$$

The statement of Lemma A.0.5 follows by combining (A.0.23) with Lemma A.0.4. □

Lemma A.0.6. *Let $\mathcal{A}_k^{iid}$ be the index set such that for all $i \in \mathcal{A}_k^{iid}$, $\boldsymbol{x}_i$ are random iid samples. If for all $\boldsymbol{u}$ such that $\|\boldsymbol{u}_{S^c}\|_1 \leq 3\|\boldsymbol{u}_S\|_1$, we have $\frac{\kappa}{2s}\|\boldsymbol{u}_S\|_1^2 \leq \boldsymbol{u}^T\nabla^2\mathcal{L}_{\mathcal{A}_k^{iid}}(\boldsymbol{\xi})\boldsymbol{u}$, then the follow inequality holds:*

$$\frac{|\mathcal{A}^{iid}|\kappa}{2ns}\|\boldsymbol{u}_S\|_1^2 \leq \boldsymbol{u}^T\nabla^2\mathcal{L}(\boldsymbol{\xi})\boldsymbol{u}, \tag{A.0.24}$$

where $\mathcal{L}_{\mathcal{A}}(\boldsymbol{\beta})$ denotes the likelihood function with samples only in $\mathcal{A}^{iid}$.

Proof. proof of A.0.6 We can rewrite $\nabla\mathcal{L}(\boldsymbol{\xi})$ with $\boldsymbol{z}_j = \boldsymbol{x}_j\sqrt{f''(r_j|\boldsymbol{x}_j^T\boldsymbol{\xi})}$ as follow.

$$\begin{aligned}
\boldsymbol{u}^T\nabla^2\mathcal{L}(\boldsymbol{\xi})\boldsymbol{u} &= \boldsymbol{u}^T\left[\frac{1}{n}\sum_{j\in\mathcal{A}_k^{iid}}\boldsymbol{z}_j\boldsymbol{z}_j^T\right]\boldsymbol{u} + \boldsymbol{u}^T\left[\frac{1}{n}\sum_{j\in(\mathcal{A}_k^{iid})^c}\boldsymbol{z}_j\boldsymbol{z}_j^T\right] \\
&\geq \frac{|\mathcal{A}_k^{iid}|}{n}^T\left[\frac{1}{|\mathcal{A}_k^{iid}|}\sum_{j\in\mathcal{A}_k^{iid}}\boldsymbol{z}_j\boldsymbol{z}_j^T\right] \\
&\geq \frac{|\mathcal{A}_k^{iid}|}{n}^T\nabla\mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi}) \\
&\geq \frac{|\mathcal{A}_k^{iid}|}{n}\frac{\kappa}{2s}\|\boldsymbol{u}_S\|_1^2 \\
&= \frac{|\mathcal{A}_k^{iid}|\kappa}{2ns}\|\boldsymbol{u}_S\|_1^2.
\end{aligned} \tag{A.0.25}$$

□

Lemma A.0.7. *Let the whole sample size be n and the set for iid random sample in U_k be $\mathcal{A}$, $k \in \mathcal{K}$. If assumptions* **A.4** *and* **A.5** *hold, then the follow result holds.*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true}\|_1 \leq \frac{96ns\lambda}{|\mathcal{A}|\kappa}\right) \geq 1-\exp\left(-C_1|\mathcal{A}|\right)-\exp\left(-\frac{n\lambda^2}{8x_{\max}^2}+\log d\right), \tag{A.0.26}$$

where $C_1 = \min\left\{1, \kappa^2/\left(192s\sigma_2 x_{\max}^2(3+2\sqrt{\sigma_2}x_{\max})\right)^2\right\}$.

Proof. Proof of lemma A.0.7 Let $\mathcal{L}_{\mathcal{A}}(\beta)$ be the loss function only includes samples in $\mathcal{A}$. Under assumption **A.4**, we have

$$\frac{\kappa}{s}\|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \boldsymbol{u}^T\mathbb{E}[\nabla^2\mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi})]\boldsymbol{u}, \tag{A.0.27}$$

for all u such that $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq 3\|\boldsymbol{u}_{\mathcal{S}}\|_1$. The following result follows from (A.0.27) and Lemma A.0.5:

$$\mathbb{P}\left(\frac{\kappa}{2s}\|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \boldsymbol{u}^T\nabla^2\mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi})\boldsymbol{u}\right) \geq 1-\exp(-C_1|\mathcal{A}|). \tag{A.0.28}$$

Moreover, via Lemma A.0.6, for all $\boldsymbol{u}$ such that $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq 3\|\boldsymbol{u}_{\mathcal{S}}\|_1$ the follow inequality holds.

$$\mathbb{P}\left(\frac{|\mathcal{A}|\kappa}{2ns}\|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \boldsymbol{u}^T\nabla^2\mathcal{L}(\boldsymbol{\xi})\boldsymbol{u}\right) \geq 1-\exp\left(-C_1|\mathcal{A}|\right). \tag{A.0.29}$$

Since $\boldsymbol{\beta}^{lasso}$ is the optimal solution to the Lasso problem, we can ensure the following inequality:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}^{lasso}) + \lambda\|\boldsymbol{\beta}^{lasso}\|_1 &\leq \mathcal{L}(\boldsymbol{\beta}^{true}) + \lambda\|\boldsymbol{\beta}^{true}\|_1 \\
\mathcal{L}(\boldsymbol{\beta}^{lasso}) - \mathcal{L}(\boldsymbol{\beta}^{true}) + \lambda\|\boldsymbol{\beta}^{lasso}\|_1 &\leq \lambda\|\boldsymbol{\beta}^{true}\|_1 \qquad \text{(A.0.30)} \\
\nabla\mathcal{L}(\boldsymbol{\beta}^{true})^T(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true}) + \lambda\|\boldsymbol{\beta}^{lasso}\|_1 &\leq \lambda\|\boldsymbol{\beta}^{true}\|_1 \qquad \text{(A.0.31)} \\
-\|\nabla\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty\|\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true}\|_1 + \lambda\|\boldsymbol{\beta}^{lasso}\|_1 &\leq \lambda\|\boldsymbol{\beta}^{true}\|_1, \qquad \text{(A.0.32)}
\end{aligned}$$

where (A.0.31) uses the convexity of $\mathcal{L}(\boldsymbol{\beta}^{lasso})$. Denote event $\mathcal{E}_0$ as follows.

$$\mathcal{E}_0 = \left\{\|\nabla\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty < \frac{1}{2}\lambda\right\}. \tag{A.0.33}$$

Under $\mathcal{E}_0$, (A.0.32) can be further simplified into

$$\begin{aligned}
-\frac{1}{2}\lambda\|\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true}\|_1 + \lambda\|\boldsymbol{\beta}^{lasso}\|_1 &\leq \lambda\|\boldsymbol{\beta}^{true}\|_1 \\
-\frac{1}{2}\|\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true}\|_1 + \|\boldsymbol{\beta}^{lasso}\|_1 &\leq \|\boldsymbol{\beta}^{true}\|_1
\end{aligned}$$

$$-\frac{1}{2}\|\boldsymbol{\beta}_S^{lasso}-\boldsymbol{\beta}_S^{true}\|_1-\frac{1}{2}\|\boldsymbol{\beta}_{S^c}^{lasso}-\boldsymbol{\beta}_{S^c}^{true}\|_1+\|\boldsymbol{\beta}_S^{lasso}\|_1+\|\boldsymbol{\beta}_{S^c}^{lasso}\|_1\leq\|\boldsymbol{\beta}_S^{true}\|_1+\|\boldsymbol{\beta}_{S^c}^{true}\|_1. \tag{A.0.34}$$

As $\boldsymbol{\beta}_{S^c}^{true}=\mathbf{0}$ by definition, we then have

$$\begin{aligned}-\frac{1}{2}\|\boldsymbol{\beta}_S^{lasso}-\boldsymbol{\beta}_S^{true}\|_1-\frac{1}{2}\|\boldsymbol{\beta}_{S^c}^{lasso}-\boldsymbol{\beta}_{S^c}^{true}\|_1+\|\boldsymbol{\beta}_S^{lasso}\|_1+\|\boldsymbol{\beta}_{S^c}^{lasso}-\mathbf{0}\|_1&\leq\|\boldsymbol{\beta}_S^{true}\|_1+0\\ -\frac{1}{2}\|\boldsymbol{\beta}_S^{lasso}-\boldsymbol{\beta}_S^{true}\|_1-\frac{1}{2}\|\boldsymbol{\beta}_{S^c}^{lasso}-\boldsymbol{\beta}_{S^c}^{true}\|_1+\|\boldsymbol{\beta}_S^{lasso}\|_1+\|\boldsymbol{\beta}_{S^c}^{lasso}-\boldsymbol{\beta}_{S^c}^{true}\|_1&\leq\|\boldsymbol{\beta}_S^{true}\|_1+0\end{aligned} \tag{A.0.35}$$

Rearrange (A.0.35) and we may have

$$\|\boldsymbol{\beta}_{S^c}^{lasso}-\boldsymbol{\beta}_{S^c}^{true}\|_1\leq 3\|\boldsymbol{\beta}_S^{lasso}-\boldsymbol{\beta}_S^{true}\|_1 \tag{A.0.36}$$

Denote $\boldsymbol{u}=\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true}$. Then, we have $\|\boldsymbol{u}_{S^c}\|_1\leq 3\|\boldsymbol{u}_S\|_1$. Connecting (A.0.29), we can obtain

$$\mathbb{P}\left((\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})^T\nabla^2\mathcal{L}(\boldsymbol{\xi})(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})\geq\frac{|\mathcal{A}|\kappa}{2ns}\|\boldsymbol{\beta}_S^{lasso}-\boldsymbol{\beta}_S^{true}\|_1^2\right)\geq 1-\exp\left(-C_1|\mathcal{A}|\right). \tag{A.0.37}$$

Now, we turn back to (A.0.30) and use the Taylor expansion on $\mathcal{L}(\boldsymbol{\beta}^{lasso})$ at $\boldsymbol{\beta}^{true}$ the following inequality holds for some $\boldsymbol{\xi}$.

$$\nabla\mathcal{L}(\boldsymbol{\beta}^{true})^T(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})+\frac{1}{2}(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})^T\nabla^2\mathcal{L}(\boldsymbol{\xi})(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})+\lambda\|\boldsymbol{\beta}^{lasso}\|_1\leq\lambda\|\boldsymbol{\beta}^{true}\|_1. \tag{A.0.38}$$

Combining (A.0.32) and (A.0.38), we know that with probability $1-\exp(-C_1n)$, the follow results hold.

$$\begin{aligned}&-\|\nabla\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty\|(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})\|_1+\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}_S^{lasso}-\boldsymbol{\beta}_S^{true}\|_1^2+\lambda\|\boldsymbol{\beta}^{lasso}\|_1\leq\lambda\|\boldsymbol{\beta}^{true}\|_1\\ \Rightarrow&-\|\nabla\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty\|(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})\|_1+\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}_S^{lasso}-\boldsymbol{\beta}_S^{true}\|_1^2\leq\lambda(\|\boldsymbol{\beta}^{true}\|_1-\|\boldsymbol{\beta}^{lasso}\|_1)\\ \Rightarrow&-\|\nabla\mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty\|(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})\|_1+\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}_S^{lasso}-\boldsymbol{\beta}_S^{true}\|_1^2\leq\lambda\|\boldsymbol{\beta}^{true}-\boldsymbol{\beta}^{lasso}\|_1\end{aligned} \tag{A.0.39}$$

Under event $\mathcal{E}_0$, we have

$$
\begin{aligned}
&-\frac{1}{2}\lambda\|(\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true})\|_1+\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}_{\mathcal{S}}^{lasso}-\boldsymbol{\beta}_{S}^{true}\|_1^2\leq\lambda\|\boldsymbol{\beta}^{true}-\boldsymbol{\beta}^{lasso}\|_1\\
\Rightarrow&\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}_{S}^{lasso}-\boldsymbol{\beta}_{S}^{true}\|_1^2\leq\frac{3}{2}\lambda\|\boldsymbol{\beta}^{true}-\boldsymbol{\beta}^{lasso}\|_1\\
\Rightarrow&\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}_{S}^{lasso}-\boldsymbol{\beta}_{S}^{true}\|_1^2\leq 6\lambda\|\boldsymbol{\beta}_{\mathcal{S}}^{true}-\boldsymbol{\beta}_{\mathcal{S}}^{lasso}\|_1 \qquad \text{(A.0.40)}\\
\Rightarrow&\|\boldsymbol{\beta}_{S}^{lasso}-\boldsymbol{\beta}_{S}^{true}\|_1\leq\frac{24ns}{|\mathcal{A}|\kappa}\lambda\\
\Rightarrow&\|\boldsymbol{\beta}^{lasso}-\boldsymbol{\beta}^{true}\|_1\leq\frac{96ns}{|\mathcal{A}|\kappa}\lambda, \qquad \text{(A.0.41)}
\end{aligned}
$$

where (A.0.40) and (A.0.41) use $\|\boldsymbol{\beta}_{S^c}^{lasso}-\boldsymbol{\beta}_{S^c}^{true}\|_1\leq 3\|\boldsymbol{\beta}_{S}^{lasso}-\boldsymbol{\beta}_{S}^{true}\|_1$ in (A.0.36).

Now, we assess the probability of event $\mathcal{E}_0$. The i-th element of $\nabla\mathcal{L}(\boldsymbol{\beta}^{ture})$ is $\frac{1}{n}\sum_{i=1}^n x_{ji}f^{'}(r_i|\boldsymbol{x}_j^T\boldsymbol{\beta}^{true})$. Denote $X_{ji}=x_{ji}f^{'}(r_i|\boldsymbol{x}_j^T\boldsymbol{\beta}^{true})$ for $j=1,2,...n$. Under assumptions **A.1** and **A.5** , X_{ji} are $x_{\max}\sigma$-subguassian random variables with mean 0. We can use Hoeffding inequality to build the following probability bound.

$$
\begin{aligned}
&\mathbb{P}(|\frac{1}{n}\sum_{i=1}^n x_{ji}f^{'}(r_i|\boldsymbol{x}_j^T\boldsymbol{\beta}^{true})|\geq t)\leq\exp\left(-\frac{nt^2}{2\sigma^2x_{\max}^2}\right)\\
\Rightarrow&\mathbb{P}\left(\max_j|\frac{1}{n}\sum_{i=1}^n x_{ji}f^{'}(r_i|\boldsymbol{x}_j^T\boldsymbol{\beta}^{true})|\leq t\right)\geq 1-\sum_{j=1}^p\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^n x_{ji}f^{'}(r_i|\boldsymbol{x}_j^T\boldsymbol{\beta}^{true})|\geq t\right)\\
&\geq 1-d\exp\left(-\frac{nt^2}{2\sigma^2x_{\max}^2}\right) \qquad \text{(A.0.42)}
\end{aligned}
$$

Set $t=\frac{1}{2}\lambda$, and we will have event $\mathcal{E}_0$ defined in (A.0.33) holds with at least probability $1-\exp(-\frac{n\lambda^2}{8x_{\max}^2}+\log d)$. The desirable result follows by (A.0.37) and (A.0.42).

□

Lemma A.0.8. *Let* $t_0=2C_0|\mathcal{K}|$, $C_0=\max\{10,16/p^*\}$, *and* $T\geq\max\{(t_0+1)^2/e^2-1,e\}$. *Under assumptions* **A.3** *and* **A.4**, *the following statements hold.*

1. $\mathbb{P}\left\{n<\frac{1}{2}C_0(T+1)\ \text{or}\ n>6C_0\log(T+1)\right\}\leq\frac{2}{T+1}$
2. $\mathbb{P}\left\{|\mathcal{A}|<\frac{1}{4}p^*C_0\log(T+1)\right\}\leq\frac{1}{T+1}$
3. $\mathbb{P}\left\{|\mathcal{A}|/n<\frac{1}{24}p^*\right\}\leq\frac{3}{T+1}$

Proof. Proof of A.0.8 ***To show statement 1.*** From Proposition 4.4.3, we have

$$\mathbb{P}\left(C_0(1+\log(T+1)-\log(t_0)) \leq n \leq 3C_0(1+\log(T)-\log(t_0))\right) \geq 1-\frac{2}{T+1}. \tag{A.0.43}$$

As we have $T \geq e$, the following result holds.

$$3C_0(1+\log(T)-\log(t_0)) \leq 3C_0(\log(T)+\log(T)-0) \leq 6C_0\log(T) < 6C_0\log(T+1). \tag{A.0.44}$$

From $T \geq (t_0+1)^2/e^2+1 \Rightarrow \frac{1}{2}\log(T+1)-\log(t_0+1) \geq -1$, we have

$$\begin{aligned} C_0(1+\log(T+1)-\log(t_0+1)) &= C_0(1+\frac{1}{2}\log(T+1)+\frac{1}{2}\log(T+1)-\log(t_0+1)) \\ &\geq C_0(1+\frac{1}{2}\log(T+1)-1) \\ &= \frac{1}{2}C_0\log(T+1). \end{aligned} \tag{A.0.45}$$

The statement 1 is obtained by combining (A.0.44),(A.0.45) and (A.0.43).

To show statement 2. In assumption **A.4**, we assume that for $\boldsymbol{x} \in U_k,\ k \in \mathcal{K}$, the restricted eigenvalue condition is held. And under Assumption **A.3**, we have $\mathbb{P}(\boldsymbol{x} \in U_k) \geq p^*$. Thus, among all n samples, the expected number of samples belong to U_k will be lower bounded by:

$$\mathbb{E}[\mathbb{1}(\boldsymbol{x} \in U_k)] \geq p^*C_0(1+\log(T+1)-\log(t_0+1)). \tag{A.0.46}$$

Since $T > (t_0+1)^2/e^2-1$ implies $\frac{1}{2}\log(T+1) > \log(t_0+1)-1$. (A.0.46) can be simplified into the following inequality.

$$\mathbb{E}[\sum_{i=1}^{n}\mathbb{1}(x_i \in U_k)] \geq \frac{1}{2}p^*C_0\log(T+1). \tag{A.0.47}$$

We apply the Chernoff inequality on $\sum_{i=1}^{n}\mathbb{1}(x_i \in U)$:

$$\mathbb{P}\left(\sum_{i=1}^{n}\mathbb{1}(\boldsymbol{x}_i \in U_k) < \frac{1}{2}\mathbb{E}[\sum_{i=1}^{n}\mathbb{1}(\boldsymbol{x}_i \in U_k)]\right) \leq \exp\left(-\frac{1}{8}\mathbb{E}[\sum_{i=1}^{n}\mathbb{1}(\boldsymbol{x}_i \in U_k)]\right)$$

$$\Rightarrow \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}(\boldsymbol{x}_i \in U_k) < \frac{1}{4}p^* C_0 \log(T+1)\right) \leq \exp\left(-\frac{1}{16}p^* C_0 \log(T+1)\right), \tag{A.0.48}$$

where (A.0.48) uses (A.0.47). The statement 2 of Lemma A.0.8 can be proved by (A.0.48) with $C_0 \geq 16/p^*$.

To show statement 3. Notice that the follow result hold.

$$\begin{aligned}\left\{|\mathcal{A}|/n \geq \frac{1}{24}p^*\right\} &\supseteq \left\{|\mathcal{A}| \geq \frac{1}{4}C_0 p^* \log(T+1)\right\} \cap \{n \leq 6C_0 \log(T+1)\} \\ &= \left(\left\{|\mathcal{A}| < \frac{1}{4}C_0 p^* \log(T+1)\right\} \cup \{n > 6C_0 \log(T+1)\}\right)^c.\end{aligned} \tag{A.0.49}$$

Hence we can obtain

$$\begin{aligned}\mathbb{P}\left\{|\mathcal{A}|/n \geq \frac{1}{24}p^*\right\} &\geq \mathbb{P}\left\{\left(\left\{|\mathcal{A}_k| < \frac{1}{4}C_0 p^* \log(T+1)\right\} \cup \{n > 6C_0 \log(T+1)\}\right)^c\right\} \\ &= 1 - \mathbb{P}\left\{\left\{|\mathcal{A}| < \frac{1}{4}C_0 p^* \log(T+1)\right\} \cup \{n > 6C_0 \log(T+1)\}\right\} \\ &= 1 - \mathbb{P}\left\{|\mathcal{A}| < \frac{1}{4}C_0 p^* \log(T+1)\right\} - \mathbb{P}\left\{n > 6C_0 \log(T+1)\right\}.\end{aligned} \tag{A.0.50}$$

The remaining part follows by combining the statement 1 and statement 2 with (A.0.50).

□

Lemma A.0.9. *Let* $t_0 = 2C_0|\mathcal{K}|$, $T \geq \max\{(t_0+1)^2/e^2 - 1, e\}$, $\lambda = C_5\sqrt{1 + \frac{\log d}{\log(T+1)}}$, *and* $a > \frac{2304s}{p^*\kappa}$. *If assumptions* **A.1**,**A.3**,**A.4** *and* **A.5** *hold, we have*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_1 \leq \min\left\{\frac{1}{\sigma x_{\max}}, \frac{h}{4e\sigma R_{\max} x_{\max}}\right\}\right) \geq 1 - \frac{7}{T+1}, \tag{A.0.51}$$

where

$$\begin{aligned}C_0 = \max\Bigg\{10, \frac{16}{p^*}, \frac{4}{p^* C_1}, \frac{4x_{\max}^2}{C_5^2}\left((\frac{1}{4} - \frac{576s}{p^*\kappa a})\min\left\{1, \frac{\mu_0 p^*}{192 s x_{\max}^2}\right\}\right)^{-2}, \\ \frac{32\sigma_2 s x_{\max}^2(1+\log s)}{p^*\mu_0}, \frac{4\sigma^2 x_{\max}^2(1+\log s)}{t^2}\Bigg\},\end{aligned}$$

$t \leq \min\left\{\frac{\mu_0 p^*\sqrt{\tilde{C}_2\lambda}}{48}, \frac{p^*\mu_0}{48\sigma\sqrt{s}x_{\max}}, \frac{hp^*\mu_0}{192e\sigma\sqrt{s}R_{\max}x_{\max}}\right\}$, $\tilde{C}_2 = \frac{\mu_0 p^*}{2\sigma_3 s x_{\max}^3(\mu_0 p^* + 48 s x_{\max}^2)}$ *and* $C_5 = \frac{\beta_{\min} p^* \kappa}{(2304s + a p^* \kappa)\sqrt{1+\log d}}$

Proof. Proof of Lemma A.0.9 Using Lemma A.0.8, $t_0 = 2C_0|\mathcal{K}|$, $T \geq \max\{(t_0+1)^2/e^2 - 1, e\}$, and $C_0 \geq \max\{10, 16/p^*\}$, we have

$$\mathbb{P}\left\{n \geq \frac{1}{2}C_0 \log(T+1)\right\} \leq 1 - \frac{2}{T+1} \tag{A.0.52}$$

$$\mathbb{P}\left\{|\mathcal{A}| \geq \frac{1}{4}p^* C_0 \log(T+1)\right\} \geq 1 - \frac{1}{T+1} \tag{A.0.53}$$

$$\mathbb{P}\left\{\frac{|\mathcal{A}|}{n} \geq \frac{1}{24}p^*\right\} \geq \frac{3}{T+1}. \tag{A.0.54}$$

Thus with probability $1 - \frac{3}{T+1}$ we have

$$\begin{aligned}
\beta_{\min} &= (\frac{2304s}{p^*\kappa} + a)C_5\sqrt{1+\log d} \geq (\frac{2304s}{p^*\kappa} + a)\lambda \geq (\frac{96ns}{\kappa|\mathcal{A}|} + a)\lambda \\
a &> \frac{2304s}{p^*\kappa} \geq \frac{96ns}{\kappa|\mathcal{A}|} \\
\tilde{C}_2 &= \frac{\mu_0 p^*}{2\sigma_3 s x_{\max}^3(\mu_0 p^* + 48 s x_{\max}^2)} \leq \frac{\mu_0|\mathcal{A}|}{2\sigma_3 s x_{\max}^3(\mu_0|\mathcal{A}| + n2s x_{\max}^2)} = C_2
\end{aligned} \tag{A.0.55}$$

If we require $t \leq \frac{\mu_0|\mathcal{A}|\sqrt{\tilde{C}_2\lambda}}{2n} \leq \frac{\mu_0|\mathcal{A}|\sqrt{C_2\lambda}}{2n}$, from (4.5.6) in Proposition 4.5.1 , we can obtain the following inequality.

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_2 \geq \frac{2nt}{|\mathcal{A}|\mu_0}\right) \leq \delta_2(n, |\mathcal{A}|, \lambda) + \delta_3(|\mathcal{A}|) + \delta_4(n, |\mathcal{A}|, t). \tag{A.0.56}$$

Since $\delta_2(n, |\mathcal{A}|, \lambda), \delta_3(|\mathcal{A}|)$ and $\delta_4(n, |\mathcal{A}|, t)$ decrease when we have larger $|\mathcal{A}|$ and n, we may pick proper C_0 such that at given time T we will have enough $|\mathcal{A}|$ and n according to (A.0.52)-(A.0.54). As we require

$$\begin{aligned}
C_0 = \max\Bigg\{&\frac{4}{p^* C_1}, \frac{4x_{\max}^2}{C_5^2}\left((\frac{1}{4} - \frac{576s}{p^*\kappa a})\min\left\{1, \frac{\mu_0 p^*}{192 s x_{\max}^2}\right\}\right)^{-2}, \\
&\frac{32\sigma_2 s x_{\max}^2(1+\log s)}{p^*\mu_0}, \frac{4\sigma^2 x_{\max}^2(1+\log s)}{t^2}\Bigg\}
\end{aligned}$$

and $\lambda = C_5\sqrt{1 + \log d/\log(T+1)}$ one may verify the the follow result hold with

probability $1 - \frac{3}{T+1}$.

$$\delta_2(n, |\mathcal{A}|, \lambda) + \delta_3(|\mathcal{A}|) + \delta_4(n, |\mathcal{A}|, t) \leq \frac{4}{T+1}. \tag{A.0.57}$$

Hence, we have

$$\begin{aligned} &\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_2 \leq \frac{2nt}{|\mathcal{A}|\mu_0}\right) \geq 1 - \frac{7}{T+1} \\ \Rightarrow &\mathbb{P}\left(\|\boldsymbol{\beta}^{MCP} - \boldsymbol{\beta}^{true}\|_1 \leq \frac{2nt\sqrt{s}}{|\mathcal{A}|\mu_0}\right) \geq 1 - \frac{7}{T+1}, \end{aligned} \tag{A.0.58}$$

where (A.0.58) uses $\boldsymbol{\beta}^{MCP}$ being the oracle solution with $\boldsymbol{\beta}^{MCP}_{S^c} = \boldsymbol{\beta}^{true}_{S^c} = \mathbf{0}$. Moreover, combine $t \leq \min\left\{\frac{p^*\mu_0}{48\sigma\sqrt{s}x_{\max}}, \frac{hp^*\mu_0}{192e\sigma\sqrt{s}R_{\max}x_{\max}}\right\}$, (A.0.54) and we have the following results.

$$\frac{2nt\sqrt{s}}{|\mathcal{A}|\mu_0} \leq \frac{2nhp^*\mu_0\sqrt{s}}{192e\sigma\sqrt{s}R_{\max}x_{\max}|\mathcal{A}|\mu_0} = \frac{h}{4e\sigma R_{\max}x_{\max}} \cdot \frac{n}{|\mathcal{A}|} \cdot \frac{p^*}{24} \leq \frac{h}{4e\sigma R_{\max}x_{\max}} \tag{A.0.59}$$

$$\frac{2nt\sqrt{s}}{|\mathcal{A}|\mu_0} \leq \frac{p^*\mu_0\sqrt{s}}{48\sigma\sqrt{s}x_{\max}|\mathcal{A}|\mu_0} = \frac{1}{\sigma_2 x_{\max}} \cdot \frac{n}{|\mathcal{A}|} \cdot \frac{p^*}{24} \leq \frac{1}{\sigma x_{\max}} \tag{A.0.60}$$

Desirable result follows immediately. □

Lemma A.0.10. *Under assumptions* **A.3** *and* **A.5**, *for any* $\boldsymbol{x} \in U_k, i \in \mathcal{K}$, *the following two statements hold.*

1. $\left|\mathbb{E}(R_i|x, \boldsymbol{\beta}_i^{true}) - \mathbb{E}(R_i|x, \boldsymbol{\beta}_i^{MCP})\right| \leq R_{\max}e^{\sigma x_{\max}\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1}\sigma x_{\max}\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1$
2. *Moreover, if* $\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \leq \min\left\{\frac{1}{\sigma x_{\max}}, \frac{h}{4e\sigma R_{\max}x_{\max}}\right\}$, $k \in \mathcal{K}$, *we have* $\mathbb{E}(R_i|\boldsymbol{x}, \beta_i^{MCP}) \geq \max_{j\neq i}\mathbb{E}(R_j|\boldsymbol{x}, \beta_j^{MCP}) + \frac{h}{2}$.

Proof. Proof of Lemma A.0.10 ***To show the part 1.*** We first expand the left-hand-side as follows.

$$\begin{aligned} &\left|\mathbb{E}(R_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{true}) - \mathbb{E}(R_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{MCP})\right| \\ =&\left|\int_{-\infty}^{+\infty} r_i dF(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{true}) - \int_{-\infty}^{+\infty} r_i dF(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{MCP})\right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{true})} dr_i - \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{MCP})} dr_i \right| && \text{(A.0.61)} \\
&= \left| \int_{-\infty}^{+\infty} r_i \left(e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{true})} - e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{MCP})} \right) dr_i \right|. \\
&= \left| \int_{-\infty}^{+\infty} -r_i \left(e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i)} \right)' \Big|_{\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^{true} + \boldsymbol{\delta}} \boldsymbol{x}^T(\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}) dr_i \right|, && \text{(A.0.62)}
\end{aligned}
$$

where (A.0.61) uses f being the negative log density function and $\boldsymbol{\delta}$ is between $\mathbf{0}$ and $\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}$. We then pull $\boldsymbol{x}^T(\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true})$ out of the integral.

$$
\begin{aligned}
&\left| \int_{-\infty}^{+\infty} -r_i \left(e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i)} \right)' \Big|_{\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^{true} + \delta} \boldsymbol{x}^T(\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}) dr_i \right| \\
=& \left| \boldsymbol{x}^T(\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}) \int_{-\infty}^{+\infty} -r_i \left(e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i)} \right)' \Big|_{\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^{true} + \boldsymbol{\delta}} dr_i \right| \\
\leq& \left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T(\boldsymbol{\beta}_i^{true} + \boldsymbol{\delta}))} f'(r_i|\boldsymbol{x}^T(\boldsymbol{\beta}_i^{true} + \boldsymbol{\delta})) dr_i \right| x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1.
\end{aligned} \tag{A.0.63}
$$

As we assume $|f'(\cdot|\cdot)|$ is bounded by σ in assumption **A.5**, (A.0.63) is upper bounded by

$$
\begin{aligned}
&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T(\boldsymbol{\beta}_i^{true} + \boldsymbol{\delta}))} f'(r_i|\boldsymbol{x}^T(\boldsymbol{\beta}_i^{true} + \boldsymbol{\delta})) dr_i \right| x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \\
\leq& \left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T(\boldsymbol{\beta}_i^{true} + \boldsymbol{\delta}))} dr_i \right| \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1. && \text{(A.0.64)}
\end{aligned}
$$

We then expand term $f(r_i|\boldsymbol{x}^T(\boldsymbol{\beta}_i^{true} + \boldsymbol{\delta}))$ in (A.0.64), and there exists a $\boldsymbol{\xi}$ between $\mathbf{0}$ and $\boldsymbol{\beta}^{true} + \boldsymbol{\delta}$ such that

$$
\begin{aligned}
&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T(\boldsymbol{\beta}_i^{true} + \boldsymbol{\delta}))} dr_i \right| \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \\
=& \left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{true}) - f'(r_i|\boldsymbol{x}^T\boldsymbol{\xi})\boldsymbol{x}^T\boldsymbol{\delta}} dr_i \right| \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_j^{true}\|_1 \\
\leq& \left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{true}) + |f'(r_i|\boldsymbol{x}^T\boldsymbol{\xi})| \|\boldsymbol{x}\|_\infty \|\boldsymbol{\delta}\|_1} dr_i \right| \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \\
\leq& \left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^T\boldsymbol{\beta}_i^{true})} dr_i \right| e^{\sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1} \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 && \text{(A.0.65)} \\
=& |\mathbb{E}(R_i|x, \boldsymbol{\beta}_i^{true})| e^{\sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1} \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 && \text{(A.0.66)}
\end{aligned}
$$

where (A.0.65) uses that $\boldsymbol{\delta}$ is between $\mathbf{0}$ and $\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}$, which implies $\|\boldsymbol{\delta}\|_1 \leq$

$\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1$, and (A.0.66) comes from the definition of $\mathbb{E}(R_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{true})$. Combining $|r_i| \leq R_{\max}$, (A.0.66), and (A.0.62), we have:

$$\left|\mathbb{E}(R_i|x, \boldsymbol{\beta}_i^{true}) - \mathbb{E}(R_i|x, \boldsymbol{\beta}_i^{MCP})\right| \leq R_{\max} e^{\sigma x_{\max}\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1} \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1. \tag{A.0.67}$$

To show the part 2. Note that the assumption $\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \leq \frac{1}{\sigma x_{\max}},\ k \in \mathcal{K}$ implies the following inequality:

$$\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \leq \frac{1}{\sigma x_{\max}} \Rightarrow e^{\sigma x_{\max}\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1} \leq e \tag{A.0.68}$$

Combining (A.0.68) and (A.0.67), we obtain

$$\begin{aligned}\left|\mathbb{E}(r_i|x, \boldsymbol{\beta}_i^{true}) - \mathbb{E}(r_i|x, \boldsymbol{\beta}_i^{MCP})\right| \leq & R_{\max} e^{\sigma x_{\max}\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1} \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \\ \leq & R_{\max} e \sigma x_{\max} \|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \end{aligned} \tag{A.0.69}$$

Under assumption **A.3**, for any $x \in U_k$, the following inequalities hold:

$$\begin{aligned} & \mathbb{E}(R_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{true}) \geq \max_{j \neq i} \mathbb{E}(R_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{true}) + h \\ \Rightarrow & \mathbb{E}(r_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{true}) - \mathbb{E}(r_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{MCP}) \geq \max_{j \neq i} \left[\mathbb{E}(r_j|x, \boldsymbol{\beta}_j^{true}) - \mathbb{E}(r_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{MCP})\right] \\ & \qquad + \max_{j \neq i} \mathbb{E}(r_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{MCP}) - \mathbb{E}(r_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{MCP}) + h \\ \Rightarrow & \mathbb{E}(r_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{MCP}) - \max_{j \neq i} \mathbb{E}(r_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{MCP}) \geq - \left|\mathbb{E}(r_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{MCP}) - \mathbb{E}(r_i|x, \boldsymbol{\beta}_i^{true}))\right| \\ & \qquad - \max_{j \neq i} \left|\mathbb{E}(r_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{true}) - \mathbb{E}(r_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{MCP})\right| + h. \end{aligned} \tag{A.0.70}$$

As we assume $\|\boldsymbol{\beta}_k^{MCP} - \boldsymbol{\beta}_k^{true}\|_1 \leq \frac{h}{4e\sigma R_{\max} x_{\max}},\ k \in \mathcal{K}$, we have

$$\|\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true}\|_1 \leq \frac{h}{4e\sigma R_{\max} x_{\max}} \Rightarrow \|R_{\max} e \sigma x_{\max} (\boldsymbol{\beta}_i^{MCP} - \boldsymbol{\beta}_i^{true})\|_1 \leq \frac{h}{4} \tag{A.0.71}$$

Combining (A.0.69),(A.0.71) and (A.0.70), we will have

$$\mathbb{E}(r_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{MCP}) - \max_{j \neq i} \mathbb{E}(r_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{MCP}) \geq -\frac{h}{4} - \frac{h}{4} + h$$

$$\Rightarrow \mathbb{E}(r_i|\boldsymbol{x}, \boldsymbol{\beta}_i^{MCP}) \geq \max_{j \neq i} \mathbb{E}(r_j|\boldsymbol{x}, \boldsymbol{\beta}_j^{MCP}) + \frac{h}{2}. \tag{A.0.72}$$

□

Lemma A.0.11. *Denote events $\mathcal{E}_3, \mathcal{E}_4$, and $\mathcal{E}_5$ as follows*

$$\mathcal{E}_3 = \left\{ \|\nabla_{\mathcal{S}^c} \mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty \leq \left(1 - \frac{96ns}{|\mathcal{A}|\kappa a}\right) \frac{\lambda}{4} \right\} \tag{A.0.73}$$

$$\mathcal{E}_4 = \left\{ \|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty \leq \left(1 - \frac{96ns}{|\mathcal{A}|\kappa a}\right) \frac{\mu_0 |\mathcal{A}| \lambda}{8snx_{\max}^2} \right\} \tag{A.0.74}$$

$$\mathcal{E}_5 = \left\{ \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{C_2 \lambda} \right\}, \tag{A.0.75}$$

where $C_2 \doteq \frac{\mu_0|\mathcal{A}|}{2\sigma_3 s x_{\max}^3 (\mu_0|\mathcal{A}| + 2snx_{\max}^2)}$. Under assumption **A.1** *and* **A.5**, *events $\mathcal{E}_3, \mathcal{E}_4$ and $\mathcal{E}_5$ implies $\mathcal{E}_2$ defined in* (4.7.19).

Proof. Proof of Lemma A.0.11 We first expend $\nabla \mathcal{L}(\boldsymbol{\beta}^{oracle})$ at $\boldsymbol{\beta}^{true}$.

$$\begin{aligned} \nabla \mathcal{L}(\boldsymbol{\beta}^{oracle}) &= \nabla \mathcal{L}(\boldsymbol{\beta}^{true}) + \nabla^2 \mathcal{L}(\boldsymbol{\xi})(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) && \text{(A.0.76)} \\ &= \nabla \mathcal{L}(\boldsymbol{\beta}^{true}) + \nabla^2 \mathcal{L}(\boldsymbol{\beta}^{true})(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) \\ &\quad + (\nabla^2 \mathcal{L}(\boldsymbol{\xi}) - \nabla^2 \mathcal{L}(\boldsymbol{\beta}^{true}))(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}), && \text{(A.0.77)} \end{aligned}$$

where $\boldsymbol{\xi} = \tau \boldsymbol{\beta}^{true} + (1 - \tau)\boldsymbol{\beta}^{oracle}$, $\tau \in [0, 1]$. The last term in (A.0.77) can be further expanded as follows

$$\begin{aligned} &(\nabla^2 \mathcal{L}(\boldsymbol{\xi}) - \nabla^2 \mathcal{L}(\boldsymbol{\beta}^{true}))(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) \\ =&\frac{1}{n} \sum_{j=1}^n \left[f''(r_j|\boldsymbol{x}_j^T \boldsymbol{\xi}) - f''(r_j|\boldsymbol{x}_j^T \boldsymbol{\beta}^{true}) \right] \boldsymbol{x}_j \boldsymbol{x}_j^T (\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) \\ =&\frac{1}{n} \sum_{j=1}^n \left[-f'''(r_j|\boldsymbol{x}_j^T \eta) \boldsymbol{x}_j^T (\boldsymbol{\xi} - \boldsymbol{\beta}^{true}) \right] \boldsymbol{x}_j \boldsymbol{x}_j^T (\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}), \end{aligned} \tag{A.0.78}$$

where (A.0.78) comes from the mean value theorem and the fact that η is on the line of of $\boldsymbol{\xi}$ and $\boldsymbol{\beta}^{true}$. Hence, assumption **A.5** and (A.0.78) imply

$$\begin{aligned} &\|(\nabla^2 \mathcal{L}(\boldsymbol{\xi}) - \nabla^2 \mathcal{L}(\boldsymbol{\beta}^{true}))(\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true})\|_\infty \\ &= \left\| \frac{1}{n} \sum_{j=1}^n \left[-f'''(r_j|\boldsymbol{x}_j^T \eta) \boldsymbol{x}_j^T (\boldsymbol{\xi} - \boldsymbol{\beta}^{true}) \right] \boldsymbol{x}_j \boldsymbol{x}_j^T (\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) \right\|_\infty \end{aligned}$$

$$
\begin{aligned}
&\leq \left\| \frac{1}{n} \sum_{j=1}^{n} \sigma_3 x_{\max} (\boldsymbol{\xi} - \boldsymbol{\beta}^{true}) \boldsymbol{x}_j \boldsymbol{x}_j^T (\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) \right\|_\infty \\
&\leq \left\| \frac{1}{n} \sum_{j=1}^{n} \sigma_3 x_{\max} (\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true})^T \boldsymbol{x}_j \boldsymbol{x}_j^T (\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}) \right\|_\infty \\
&\leq \sigma_3 x_{\max} \lambda_{\max} (\frac{1}{n} X_{\mathcal{S}} X_{\mathcal{S}}^T) \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2^2 \\
&\leq \sigma_3 s x_{\max}^3 \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2^2. \qquad \text{(A.0.79)}
\end{aligned}
$$

Combining (A.0.77), (A.0.79), and the fact $\boldsymbol{\beta}_{\mathcal{S}^c}^{oracle} = \boldsymbol{\beta}_{\mathcal{S}^c}^{true} = 0$, we have

$$
\begin{aligned}
\|\nabla_{\mathcal{S}^c} \mathcal{L}(\boldsymbol{\beta}^{oracle})\|_\infty \leq \|\nabla_{\mathcal{S}^c} \mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty + \|\nabla^2_{\mathcal{S}^c,\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})(\boldsymbol{\beta}_{\mathcal{S}}^{oracle} - \boldsymbol{\beta}_{\mathcal{S}}^{true})\|_\infty \\
+ \sigma_3 s x_{\max}^3 \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2^2. \qquad \text{(A.0.80)}
\end{aligned}
$$

In addition, from $\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{oracle}) = 0$ and (A.0.77), we have

$$
\begin{aligned}
(\boldsymbol{\beta}_{\mathcal{S}}^{oracle} - \boldsymbol{\beta}_{\mathcal{S}}^{true}) = -(\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}))^{-1} (\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}) \\
+ (\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\xi}) - \nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}))(\boldsymbol{\beta}_{\mathcal{S}}^{oracle} - \boldsymbol{\beta}_{\mathcal{S}}^{true})). \qquad \text{(A.0.81)}
\end{aligned}
$$

Under events $\mathcal{E}_3$, $\mathcal{E}_4$, and (A.0.81), the inequality (A.0.80) can be upper bounded as follows.

$$
\begin{aligned}
&\|\nabla_{\mathcal{S}^c} \mathcal{L}(\boldsymbol{\beta}^{oracle})\|_\infty \leq \left(1 - \frac{96ns}{|\mathcal{A}|\kappa a}\right) \frac{\lambda}{4} + \sigma_3 x_{\max} \lambda_{\max} (\frac{1}{n} X_{\mathcal{S}} X_{\mathcal{S}}^T) \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2^2 \\
&\quad + \|\nabla^2_{\mathcal{S}^c,\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})(\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}))^{-1} (\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}) \\
&+ (\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\xi}) - \nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}))(\boldsymbol{\beta}_{\mathcal{S}}^{oracle} - \boldsymbol{\beta}_{\mathcal{S}}^{true}))\|_\infty \\
&\leq \left(1 - \frac{96ns}{|\mathcal{A}|\kappa a}\right) \frac{\lambda}{4} + \sigma_3 s x_{\max}^3 \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2^2 \\
&\quad + \left\|\nabla^2_{\mathcal{S}^c,\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})(\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}))^{-1}\right\| \left(\|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})\|_\infty + \sigma_3 s x_{\max}^3 \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2^2\right) \\
&\leq \left(1 - \frac{96ns}{|\mathcal{A}|\kappa a}\right) \frac{\lambda}{4} + \sigma_3 s x_{\max}^3 \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|_2^2 \\
&+ \left\|\nabla^2_{\mathcal{S}^c,\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true})(\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{true}))^{-1}\right\| \left(\left(1 - \frac{96ns}{|\mathcal{A}|\kappa a}\right) \frac{\mu_0 |\mathcal{A}| \lambda}{8sn x_{\max}^2} + \sigma_3 s x_{\max}^3 \|\boldsymbol{\beta}^{oracle} - \boldsymbol{\beta}^{true}\|^2\right). \\
&\qquad \text{(A.0.82)}
\end{aligned}
$$

Note that the maximum value of $\left\|\nabla^2_{\mathcal{S}^c,\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})(\nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true}))^{-1}\right\|$ can be bounded.

$$\left\|\nabla^2_{\mathcal{S}^c,\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})(\nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true}))^{-1}\right\| \leq \max_{\|v\|=1}\left\|\nabla^2_{\mathcal{S}^c,\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})(\nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true}))^{-1}v\right\|. \tag{A.0.83}$$

From (A.0.83) and Lemma A.0.1, the following inequality holds with probability $1-2s\exp\left(-\frac{|\mathcal{A}|\mu_0}{4s\sigma_2 x^2_{\max}}\right)$.

$$\begin{aligned}\max_{\|v\|=1}\left\|\nabla^2_{\mathcal{S}^c,\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})(\nabla^2_{\mathcal{S},\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true}))^{-1}v\right\| &\leq \frac{2n}{\mu_0|\mathcal{A}|}\max_{\|v\|=1}\left\|\nabla^2_{\mathcal{S}^c,\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{true})v\right\| \\ &\leq \frac{2n}{\mu_0|\mathcal{A}|}\cdot sx^2_{\max} = \frac{2snx^2_{\max}}{\mu_0|\mathcal{A}|}. \end{aligned} \tag{A.0.84}$$

Thus, (A.0.82) can be simplified to:

$$\begin{aligned}\|\nabla_{\mathcal{S}^c}\mathcal{L}(\boldsymbol{\beta}^{oracle})\|_\infty &\leq \left(1-\frac{96ns}{|\mathcal{A}|\kappa a}\right)\frac{\lambda}{4} + \sigma_3 sx^3_{\max}\|\boldsymbol{\beta}^{oracle}-\boldsymbol{\beta}^{true}\|^2 \\ &\quad + \frac{2snx^2_{\max}}{\mu_0|\mathcal{A}|}\left(\left(1-\frac{96ns}{|\mathcal{A}|\kappa a}\right)\frac{\mu_0|\mathcal{A}|\lambda}{8snx^2_{\max}} + \sigma_3 sx^3_{\max}\|\boldsymbol{\beta}^{oracle}-\boldsymbol{\beta}^{true}\|^2\right) \\ &= \left(1-\frac{96ns}{|\mathcal{A}|\kappa a}\right)\frac{\lambda}{2} + \frac{\sigma_3 sx^3_{\max}(\mu_0|\mathcal{A}|+2snx^2_{\max}}{\mu_0|\mathcal{A}|}\|\boldsymbol{\beta}^{oracle}-\boldsymbol{\beta}^{true}\|_2^2. \end{aligned} \tag{A.0.85}$$

Further, conditioning on event $\mathcal{E}_5$ defined in (A.0.75), we have:

$$\begin{aligned}\|\nabla_{\mathcal{S}^c}\mathcal{L}(\boldsymbol{\beta}^{oracle})\|_\infty &\leq \left(1-\frac{96ns}{|\mathcal{A}|\kappa a}\right)\frac{\lambda}{2} + \frac{\sigma_3 sx^3_{\max}(\mu_0|\mathcal{A}|+2snx^2_{\max}}{\mu_0|\mathcal{A}|}\left(\sqrt{C_2\lambda}\right)^2 \\ &\leq \left(1-\frac{96ns}{|\mathcal{A}|\kappa a}\right)\lambda, \end{aligned} \tag{A.0.86}$$

where (A.0.86) uses $C_2 = \frac{\mu_0|\mathcal{A}|}{2\sigma_3 sx^3_{\max}(\mu_0|\mathcal{A}|+2snx^2_{\max})}$. The inequality (A.0.86) directly implies event $\mathcal{E}_2$.

□

Bibliography

[1] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

[2] Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2012). Online-to-confidence-set conversions and application to sparse stochastic bandits. In *AISTATS*, volume 22, pages 1–9.

[3] Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Advances in Neural Information Processing Systems*, pages 1538–1546.

[4] Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. (2016). Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*.

[5] Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pages 127–135.

[6] Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). Network flows: theory, algorithms, and applications.

[7] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

[8] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

[9] Bastani, H. and Bayati, M. (2015). Online decision-making with high-dimensional covariates.

[10] Bian, W., Chen, X., and Ye, Y. (2015). Complexity analysis of interior point algorithms for non-lipschitz and nonconvex minimization. *Mathematical Programming*, 149(1-2):301–327.

[11] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

[12] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

[13] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[14] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

[15] Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

[16] Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215.

[17] Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted âĎŞ1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905.

[18] Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2016). Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*.

[19] Cartis, C., Gould, N. I., and Toint, P. L. (2011a). Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295.

[20] Cartis, C., Gould, N. I., and Toint, P. L. (2011b). On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739.

[21] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

[22] Chen, C., Li, X., Tolman, C., Wang, S., and Ye, Y. (2013). Sparse portfolio selection via quasi-norm regularization. *arXiv preprint arXiv:1312.6350*.

[23] Chen, X., Ge, D., Wang, Z., and Ye, Y. (2014). Complexity of unconstrained l_2-l_p minimization. *Mathematical Programming*, 143(1-2):371–383.

[24] Consortium, I. W. P. et al. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med*, 2009(360):753–764.

[25] Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366.

[26] Deshpande, Y. and Montanari, A. (2012). Linear bandits in high dimension and recommendation systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1750–1754. IEEE.

[27] Elmachtoub, A. N., McNellis, R., Oh, S., and Petrik, M. (2017). A practical method for solving contextual bandit problems using decision trees. In *Proceedings of the Thirty-third Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.

[28] Fan, J., Han, F., and Liu, H. (2014a). Challenges of big data analysis. *National science review*, 1(2):293–314.

[29] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

[30] Fan, J., Liu, H., Sun, Q., and Zhang, T. (2015). Tac for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *arXiv preprint arXiv:1507.01037*.

[31] Fan, J., Liu, H., Sun, Q., and Zhang, T. (2018). I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of statistics*, 46(2):814.

[32] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

[33] Fan, J., Xue, L., and Zou, H. (2014b). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819.

[34] Ge, D., Jiang, X., and Ye, Y. (2011). A note on the complexity of l p minimization. *Mathematical programming*, 129(2):285–299.

[35] Ge, D., Wang, Z., Ye, Y., and Yin, H. (2015a). Strong np-hardness result for regularized *l_q*-minimization problems with concave penalty functions. *arXiv preprint arXiv:1501.00622*.

[36] Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015b). Escaping from saddle points-online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842.

[37] Goldenshluger, A., Zeevi, A., et al. (2013). A linear response bandit problem. *Stochastic Systems*, 3(1):230–261.

[38] Grant, M., Boyd, S., and Ye, Y. (2008). Cvx: Matlab software for disciplined convex programming.

[39] Haeser, G., Liu, H., and Ye, Y. (2017). Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. *arXiv preprint arXiv:1702.04300*.

[40] Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618.

[41] James, G., Witten, D., and Hastie, T. (2014). An introduction to statistical learning: With applications in r.

[42] Jiang, B., Lin, T., Ma, S., and Zhang, S. (2016). Structured nonconvex and nonsmooth optimization: Algorithms and iteration complexity analysis. *arXiv preprint arXiv:1605.02408*.

[43] Kim, S.-J., Koh, K., Lustig, M., Boyd, S., and Gorinevsky, D. (2007). An interior-point method for large-scale \ell_1-regularized least squares. *IEEE journal of selected topics in signal processing*, 1(4):606–617.

[44] Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502.

[45] Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

[46] Kuczyński, J. and Woźniakowski, H. (1992). Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM journal on matrix analysis and applications*, 13(4):1094–1122.

[47] Liu, H., Du, G., Zhang, L., Lewis, M. M., Wang, X., Yao, T., Li, R., and Huang, X. (2016a). Folded concave penalized learning in identifying multimodal mri marker for parkinsonâĂŹs disease. *Journal of neuroscience methods*, 268:1–6.

[48] Liu, H., Wang, X., Yao, T., Li, R., and Ye, Y. (2018). Sample average approximation with sparsity-inducing penalty for high-dimensional stochastic programming. *Mathematical Programming*, pages 1–40.

[49] Liu, H., Yao, T., Li, R., et al. (2016b). Global solutions to folded concave penalized nonconvex learning. *The Annals of Statistics*, 44(2):629–659.

[Liu et al.] Liu, H., Yao, T., Li, R., and Ye, Y. Folded concave penalized sparse linear regression: Sparsity, statistical performance, and algorithmic theory for local solutions. *Mathematical Programming*, pages 1âĂŞ–34.

[51] Liu, H. and Ye, Y. (2019). High-dimensional learning under approximate sparsity: A unifying framework for nonsmooth learning and regularized neural networks. *arXiv preprint arXiv:1903.00616*.

[52] Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484.

[53] Luo, Z.-Q., Pang, J.-S., and Ralph, D. (1996). *Mathematical programs with equilibrium constraints*. Cambridge University Press.

[54] Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1):77–91.

[55] McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman and Hall/CRC.

[56] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

[57] Meinshausen, N., Yu, B., et al. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270.

[58] Mitchell, J. (2012). How google search really works. `https://readwrite.com/2012/02/29/interview_changing_engines_mid-flight_qa_with_goog/#awesm=~oiNkM4tAX3xhbP`. Accessed: Oct 22nd, 2018.

[59] Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

[60] Moulines, E. and Bach, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459.

[61] Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356.

[62] Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

[63] Nesterov, Y. and Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205.

[64] Nievergelt, Y. (2000). A tutorial history of least squares with applications to astronomy and geodesy. *Journal of Computational and Applied Mathematics*, 121(1):37–72.

[65] Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., Konig, R., Haas, S., Eils, R., et al. (2006). Customized oligonucleotide microarray gene expression–based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of clinical oncology*, 24(31):5070–5078.

[66] OxfordDictionaries (2018). How many words are there in the english language? `https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language/`. Accessed: Oct 22nd, 2018.

[67] Qiang, S. and Bayati, M. (2016). Dynamic pricing with demand covariates. *Browser Download This Paper*.

[68] Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*.

[69] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

[70] Rockafellar, R. T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42.

[71] Rosen, J. B. (1960). The gradient projection method for nonlinear programming. part i. linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 8(1):181–217.

[72] Rudelson, M., Vershynin, R., et al. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9.

[73] Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

[74] Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

[75] Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658.

[76] Scott, S. L. (2015). Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31(1):37–45.

[77] Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory*. SIAM.

[78] Shapiro, A. and Xu, H. (2008). Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. *Optimization*, 57(3):395–418.

[79] Shewan, D. (2017). The comprehensive guide to online advertising costs. `https://www.wordstream.com/blog/ws/2017/07/05/online-advertising-costs`. Accessed: Oct 22nd, 2018.

[80] Slivkins, A. (2014). Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568.

[81] Talluri, K. T. and Van Ryzin, G. J. (2006). *The theory and practice of revenue management*, volume 68. Springer Science & Business Media.

[82] Tencent (2012). Predict the click-through rate of ads given the query and user information. `https://www.kaggle.com/c/kddcup2012-track2`. Accessed: Oct 22nd, 2018.

[83] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[84] Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

[85] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166.

[86] Van de Geer, S. A. et al. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

[87] Wang, L., Kim, Y., and Li, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics*, 41(5):2505.

[88] Wang, Y., Yin, W., and Zeng, J. (2015). Global convergence of admm in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*.

[89] Wang, Z., Liu, H., and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164.

[90] WordStream (2017). Average ctr (click-through rate): Learn how your ctr compares. `https://www.wordstream.com/average-ctr`. Accessed: Oct 22nd, 2018.

[91] Ye, Y. (1991). An o (n 3 l) potential reduction algorithm for linear programming. *Mathematical programming*, 50(1):239–258.

[92] Ye, Y. (1992). On affine scaling algorithms for nonconvex quadratic programming. *Mathematical Programming*, 56(1-3):285–300.

[93] Ye, Y. (1998). On the complexity of approximating a kkt point of quadratic programming. *Mathematical programming*, 80(2):195–211.

[94] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942.

[95] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594.

[96] Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, pages 576–593.

[97] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

[98] Zhao, T., Liu, H., and Zhang, T. (2014). Pathwise coordinate optimization for sparse learning: Algorithm and theory. *arXiv preprint arXiv:1412.7477*.

[99] Zhao, T., Liu, H., Zhang, T., et al. (2018). Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics*, 46(1):180–218.

[100] Zhu, Z., Dang, C., and Ye, Y. (2012). A fptas for computing a symmetric leontief competitive economy equilibrium. *Mathematical programming*, 131(1):113–129.

[101] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Vita

Xue Wang

Xue Wang received a B.S. degree in Industrial Engineering from Tsinghua University, Beijing, China, in 2013. He is currently pursuing the Ph.D. degree in Industrial Engineering with a dual title in Operations Research at Pennsylvania State University in University Park, Pennsylvania. His research interests include nonconvex optimization as well as high dimensional statistical learning problems with application interests in online decision making.