

© 2022 Ke Li

TOPICS IN HIGH-DIMENSIONAL LINEAR BANDITS AND APPROXIMATE BAYESIAN SAMPLING

BY

KE LI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Assistant Professor Naveen N. Narisetty, Co-Chair and Co-Director of Research
Assistant Professor Yun Yang, Co-Chair and Co-Director of Research
Professor Feng Liang
Associate Professor Georgios Fellouris

Abstract

In this dissertation, we study the regret lower bound and propose algorithms for general high-dimensional linear problems and propose efficient generative models for Bayesian inference.

In the first project, we consider the multi-armed bandit problem with high-dimensional features. First, we prove a minimax lower bound, $\mathcal{O}((\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}} + \log T)$, for the cumulative regret, in terms of horizon T , dimension d and a margin parameter $\alpha \in [0, 1]$, which controls the separation between the optimal and the sub-optimal arms. This new lower bound unifies existing regret bound results that have different dependencies on T due to the use of different values of margin parameter α explicitly implied by their assumptions. Second, we propose a simple and computationally efficient algorithm inspired by the general Upper Confidence Bound (UCB) strategy that achieves a regret upper bound matching the lower bound. The proposed algorithm uses a properly centered ℓ_1 -ball as the confidence set in contrast to the commonly used ellipsoid confidence set. In addition, the algorithm does not require any forced sampling step and is thereby adaptive to the practically unknown margin parameter. Simulations and a real data analysis are conducted to compare the proposed method with existing ones in the literature.

In the second project, we propose an Upper Confidence Bound (UCB) based algorithm with variable selection. One main contribution of the project is that our proposed algorithm has feature (or variable) selection consistency in bandit settings and facilitates the construction of a confidence region for the true parameter vector. In particular, our proposed algorithm constructs a properly centered ellipsoid confidence set for selected features, and achieves a non-asymptotic regret bound of $\mathcal{O}(\log^2 T + \log d)$ in terms of horizon T and dimension d . We show through a matching minimax lower bound that our proposed algorithm is nearly optimal. Finally, we utilize regularized estimators such as SCAD and MCP as examples of variable selection methods and demonstrate the effectiveness of our proposed algorithm using synthetic and real datasets.

In the third project, we propose an efficient sampling method, which borrows the ideas from generative models with techniques from the optimal transport theory, for Bayesian inference. Specifically, we construct a transport map that transforms a simple reference distribution into the target distribution. The new approach can produce independent and exact random samples from the target distribution while maintaining similar

computational efficiency as variation approximation. In particular, we characterize the optimal transport maps separately for two common cases in Bayesian inference: 1. the target distribution is continuous; 2. the target distribution contains both discrete and continuous random variables. Finally, we use the characterizations of the optimal transport map to develop a finite approximation map family to construct rich posterior approximations.

To my family.

Acknowledgments

First, I would like to express my sincerest appreciation to my advisors, Professor Naveen N. Narisetty and Professor Yun Yang, for their guidance, patience and encouragement over these years. Their great passion towards research, broad vision and generosity in time helped me to become a better researcher. I feel extremely fortunate to have worked under their supervision which pushed me farther than I could imagine.

Second, I would like to give a special thanks to my thesis committee members Professor Feng Liang and Professor Georgios Fellouris for their constructive comments on my thesis research. It was a pleasure to benefit from their collective insights. In addition, I would like to thank all the faculty and staff members in Department of Statistics at University of Illinois Urbana-Champaign for providing such a friendly environment. I am also grateful to all my friends who have shared great memories with me in life and study.

Finally, I would like to thank my family for their unconditional love and continuous encouragement. I would never go through this wonderful and challenging journey without them.

Table of Contents

Chapter 1	Introduction	1
Chapter 2	Regret Lower Bound and Optimal Algorithm for High-Dimensional Contextual Linear Bandit	3
2.1	Introduction	3
2.2	Problem Formulation	5
2.3	ℓ_1 -Confidence Ball Algorithm for High-dimensional Linear Bandit	12
2.4	Experimental Results	22
2.5	Supplementary Material	25
Chapter 3	High-dimensional Linear Bandits with Variable Selection	47
3.1	Introduction	47
3.2	Problem Formulation	50
3.3	UCB Algorithm with Variable Selection	51
3.4	Theoretical Properties	57
3.5	Empirical Results	65
3.6	Discussion	70
3.7	Supplementary Material	72
Chapter 4	Efficient Bayesian Sampling Using Generative Models	105
4.1	Introduction	105
4.2	Background and Problem Setup	107
4.3	Generative Model Learning for Bayesian Inference	111
4.4	Computation	114
4.5	Numerical Results	123
4.6	Supplementary Material	125
References		133

Chapter 1

Introduction

This dissertation consists of three projects that I did through my PhD study. The first two projects focus on the high-dimensional linear bandit problems. The third project aims to provide an efficient sampling method with generative model for Bayesian inference.

In modern era, the increasing availability of massive personalized datasets enables decision-makers to tailor decisions specifically suited to individuals. For example, doctors can personalize treatments for patients based on their medical records and demographics; news recommendation system can post articles according to users' personal interests and browsing history. In these situations, decision-makers need to adapt their decisions to the sequentially collected data and take actions to maximize their total reward within a time period. Mathematically, this problem can be formulated as multi-armed bandit problems [1]–[3]. In this dissertation, we focus on the high-dimensional linear bandit problems where the dimension d of the feature vectors and be comparable or even larger than the total number of rounds T , which plays the role of the “sample size” in the statistical perspective. In the first chapter, we first prove a minimax regret lower bound for policies in term of the dimension d and the time horizon T , which is useful to assess the quality of decision-rules. Secondly, we propose a simple and computationally efficient algorithm inspired by the Upper Confidence Ball (UCB) strategy. The proposed algorithm uses a properly centered ℓ_1 -confidence ball and does not require forced sampling step [4], [5], which can be costly and wasteful in practical scenarios such as medical and financial applications. In the second chapter, we propose a general framework for bandit algorithms to incorporate variable selection in the UCB type algorithms. In particular, our proposed algorithm has feature selection consistency and facilitates the construction of *sparse* confidence sets for true parameter vectors. As a consequence, the proposed algorithm achieves better cumulative regret in terms of dimension d and horizon T compared to the ℓ_1 -confidence ball algorithm proposed in the first project. In addition, we instantiate

the general framework with nonconvex regularized estimation methods such as MCP [6] and SCAD [7] and illustrate the effectiveness of the proposed algorithm in both synthetic data and real applications. Finally, our proof of regret upper bound for proposed algorithms is based on non-asymptotic analysis and worth highlighting on its own. In particular, we utilize a novel strategy to prove the *restricted eigenvalue condition* for sample covariance matrix under dependent data due to the exploitation nature of the bandit problems. Our analysis shows that the randomness in the feature space facilitates the implicit exploration of the feature space and therefore the random sampling steps can be avoided.

A central problem in modern Bayesian statistics is how to conduct valid statistical inference on model parameters in a computationally efficient manner. The main hurdle in implementing the Bayesian inference is the computation of the normalization constant in the posterior distributions, which is mathematically intractable unless conjugate priors are employed. Two main approaches, which only requires the knowledge of the target distribution density up to a constant and thus avoids the explicit computation of the normalization constant, are widely employed in Bayesian inference. One is based on (approximate) sampling from the posterior distribution [8] via, for example, Markov Chain Monte Carlo (MCMC) algorithms; and another is through variational approximation [9] by turning the integration problem into an optimizing problem. Yet these two approaches have their respective limitations. MCMC may suffer from slow-mixing due to complicated model structures, while the variation approximation is not guaranteed to generate (almost) exact samples from the target distribution and therefore may lead to estimation inconsistency. In the third chapter, we propose a new sampling approach with generative models, which uses the techniques from the optimal transport theory. The new approach can produce *independent and exact* samples from the target distribution while maintaining similar computational efficiency as variation approximation. Specifically, we construct a non-linear and invertible transport map T that transforms a simple reference distribution μ such as standard normal to the target distribution π_n , which can be mathematically described as the push-forward distribution $T_{\#}\mu$ of μ through map T . In addition, we impose extra constraint by requiring T to be the optimal transport map sending μ to the target posterior π_n in the Monge problem [10], [11] of minimizing the ℓ_2 transportation cost $\mathbb{E}_{\mu}[\|T(\theta) - \theta\|_2^2]$ subject to the constraint that $\pi_n = T_{\#}\mu$. This extra constraint based on transportation cost minimizing leads to a *parsimonious* model characterizing the posterior and makes our estimated transport map *more interpretable*. We provide the explicit characterizations of the optimal transport map T^* for two common cases in Bayesian inference: 1. the target distribution is continuous; 2. the target distribution contains both discrete and continuous random variables. Based on this theoretical guidance, we propose a finite approximation family for the optimal transport map T^* with diminishing approximation error as the family capacity increases.

Chapter 2

Regret Lower Bound and Optimal Algorithm for High-Dimensional Contextual Linear Bandit

2.1 Introduction

In the big data era, the abundance of personalized information enables decision-makers to make individualized decisions for improving the long term reward by incorporating this contextual information. For example, internet marketing companies may utilize users' searching history and demographics to display personalized online advertisements to improve the conversion rate [1]. In personalized medicine, doctors assign treatments tailored to the individual patient based on the context of the patient's own medical records and genetic information such as biomarkers [4]. In these examples, data are often collected sequentially and decision-makers need to pick the best action to maximize the long term reward based on the current predict response in a sequential fashion. Mathematically, this problem can be formulated as a contextual bandit problem [1], [3], where an agent sees a d -dimensional feature vector and has to choose among K actions (arms) at each of the T rounds to maximize the cumulative reward. When the expected reward is a linear function of the features, this problem is known as the linear bandit problem, or the multi-armed bandit problem with linear payoff functions [1], [2]. Under this setting, [12], [3] and [13] showed a polynomial dependence of the cumulative regret on dimension d and time horizon T . Precisely, [3] proved a regret upper bound scaling as $\mathcal{O}(\sqrt{dT})$, while [12] and [13] proved regret upper bounds scaling as $\mathcal{O}(d\sqrt{T})$.

We focus on the high-dimensional regime where the dimension d of the feature vector can be comparable, or even much larger than the total number of rounds T that plays the role of “sample size” in the statistical perspective. The high-dimensionality of the features and the dependence between the samples induced by the bandit policy make the high dimensional linear bandit problem very challenging both for methodological development and theoretical analysis. In particular, the ordinary least squares (OLS), a traditional statistical method for linear regression and its variants serve as the cornerstones for most linear bandit algorithms, they require a substantial number of samples to accurately estimate the parameters, incurring the polynomial- d dependence of the regret. In the statistical literature, it is well-known that imposing extra lower-dimensional structures such as sparsity on the model improves the minimax risk, or the sample complexity of the best learning algorithm. For example, under the sparsity assumption that the response, or reward, only depends on a small subset of all features with size $s_0 \ll d$, the minimax risk reduces from $\sqrt{d/T}$ to $\sqrt{s_0 \log d/T}$, where s_0 now plays the role of the “effective dimension” and the extra $\log d$ term is due to the uncertainty in feature selection. Unfortunately, sparsity generally does not help improve the regret in linear bandit. Roughly speaking, this failure of sparsity adaptation is due to the exploitation nature of a good bandit problem that tends to choose an optimal arm as often as possible, which prevents the exploration of different directions in the feature space and results in an ill-conditioned design matrix. Interestingly, we find that sparsity adaptation is still possible in the high-dimensional linear bandit when features associated with different arms have a sufficient separation. In particular, we show that under the stochastic assumption that feature vectors are random whose population-level covariance matrices are well-conditioned (see Assumption 2 for a precise definition), the randomness in the features prevents the selected arm to collapse and leads to a regret bound that scales logarithmically in the feature dimension d . Moreover, this randomness in the feature implicitly encourages explorations in the feature space, due to which no *forced exploration* is needed. A simple and computationally efficient algorithm combining the Upper Confidence Bound (UCB) technique [2], [14] with the LASSO [15] estimator induces sparsity in the estimated regression parameter and enjoys the improved regret bound.

In this chapter, we make a number of contributions to the multi-armed bandit literature. Theoretically, we introduce a relaxation of the margin condition (Assumption 1(b)) from [5] and [4], which precisely captures the hardness of the linear bandit problem via a margin parameter $\alpha \in [0, +\infty]$ that controls the “separation” between optimal and sub-optimal arms. With this additional assumption, we prove a regret lower bound of $\mathcal{O}(\log d + \log T)$ for $\alpha = 1$ and $\mathcal{O}((\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}})$ for $\alpha \in [0, 1)$. The dependence of our lower bound on the margin parameter α unifies both the polynomial and the logarithmic regret bound dependencies on T in the literature. More details are provided in the discussion of related works in Section 2.3.4. Methodologically, we propose the use of a properly centered ℓ_1 -ball as the confidence set in contrast to the commonly used ellipsoid

confidence sets in the UCB algorithm. We show that the ℓ_1 -ball confidence set captures the uncertainty and implicitly manages the trade-off between exploration versus exploitation. We also prove that the algorithm is optimal up to a logarithmic factor in T .

Technically, our proof of the upper bound is based on non-asymptotic analysis for the LASSO estimator for dependent data based on a novel proof strategy for verifying the *restricted eigenvalue condition* for the sample covariance matrix under dependent data. In particular, our analysis uses an anti-concentration technique to show that the randomness in the feature vectors facilitates an automatic feature space exploration and prevents feature collapse. Computationally, optimizing the expected reward jointly over the ℓ_1 -ball confidence set and the covariate set is equivalent to finding the arm that maximizes the estimated reward plus a term proportional to the ℓ_∞ -norm (dual norm to the ℓ_1 -norm) of the corresponding feature, which is easy to implement and computationally efficient.

The rest of the chapter is organized as follows. In Section 2.2, we review the background and discuss our assumptions. After that, we present the regret lower bound for policies. In section 2.3, we propose ℓ_1 -confidence ball based algorithm for the high-dimensional linear bandit problem. We present a novel non-asymptotic analysis of the LASSO estimator under the bandit setting, and provide an regret upper bound for the proposed algorithm. Section 2.4 reports experimental results comparing with some existing bandit algorithms. Detailed proofs of theorems are deferred to the appendix section.

2.2 Problem Formulation

Let T be the total time number of steps, which is allowed to be unknown, and K be the number of possible actions (arms). At each round $t = 1, \dots, T$, for each arm $a \in [K]$, the learner observes one of K feature vectors $X_{a,t} \in \mathbb{R}^d$. We consider the high-dimensional regime where the dimension d of the feature space can be comparable or even much larger than the time horizon T . We adopt a standard random-design framework by assuming that for each arm $a \in [K]$, the observed sequence of feature vectors $\{X_{a,t}\}_{t \geq 0}$ are i.i.d. sub-Gaussian random vectors drawn from an unknown distribution $\mathbb{P}_a$ in $\mathbb{R}^d$. The distributions $\mathbb{P}_a$'s for $a \in [K]$ are allowed to have dependence across different arms. More precisely, recall the following notion of sub-Gaussian random variables.

Definition 1 (Sub-Gaussian random variable). *A real-valued random variable z is σ -sub-Gaussian if $\mathbb{E}[e^{tz}] \leq e^{\sigma^2 t^2 / 2}$ for every $t \in \mathbb{R}$. This definition implies $\mathbb{E}[z] = 0$ and $\text{Var}[z] \leq \sigma^2$.*

Let $\mathcal{D}_t = \{X_{a,t}\}_{a=1}^K$ denote the covariate set consisting of feature vectors corresponding to all the arms at time $t \in [T]$. The feature vectors are allowed to be dependent across arms but are assumed to be independent

across different time points t . The decision-maker has access to K arms and each arm yields a random reward, the expected value of which follows a linear function of the associated feature. All arms share the same unknown parameter $\beta_* \in \mathbb{R}^d$ in the linear reward function $\langle \cdot, \beta_* \rangle$. In our setup, if the learner pulls arm $a \in [K]$ at time t , then the following random reward is observed $Y_{a,t} := \langle X_{a,t}, \beta_* \rangle + \epsilon_{a,t}$, where $\epsilon_{a,t}$ are independent σ -sub-Gaussian random variables that are also independent of the sequence $\{X_{a,t}\}_{t \geq 0}$ for all $a \in [K]$. Define σ -algebra $\mathcal{F}_t = \sigma(X_{a,1}, \epsilon_{a,1}, \dots, X_{a,t-1}, \epsilon_{a,t-1}, X_{a,t} : \forall a \in [K])$. Then the sequence of $\{\mathcal{F}_t\}_{t=0}^\infty$ is a *filtration*, and the errors $\{\epsilon_{a,t}\}_{t=1}^\infty$ for each arm $a \in [K]$ forms a *martingale difference sequence* relative to this filtration.

Our goal is to design a sequential decision-making policy π that learns the parameter β_* over time in order to maximize the expected cumulative reward over the time horizon. Let $\hat{a}_t \in [K]$ denote the arm chosen by the policy π at time $t \in [T]$. Then, an admissible policy π is a sequence of random variables $\hat{a}_1, \hat{a}_2, \dots$ taking values in the set $\{1, \dots, K\}$ such that $\hat{a}_t$ is measurable with respect to the σ -algebra $\mathcal{F}_t^+$ generated by the previous feature vectors from each arm and observed rewards of the chosen arms $\{X_{a,s}, Y_{\hat{a}_s, s} : s = 1, \dots, t-1; a = 1, \dots, K\}$ and by the current feature vectors $\{X_{1,t}, \dots, X_{K,t}\}$,

$$\mathcal{F}_t^+ = \sigma(X_{a,1}, Y_{\hat{a}_1, 1}, \dots, X_{a,t-1}, Y_{\hat{a}_{t-1}, t-1}, X_{a,t} : \forall a \in [K]).$$

To characterize the quality of the policy π , we compare it with the oracle policy π^* that uses the knowledge of β_* to choose the best arm $a_t^* = \arg \max_a \langle X_{a,t}, \beta_* \rangle$ maximizing the expected reward at each round $t \in [T]$. Note that under our random-design assumption, a_t^* is also random. To summarize, any policy incurs at each time $t \in [T]$ an expected regret

$$r_t = \mathbb{E} \left[\max_{a \in [K]} \langle X_{a,t}, \beta_* \rangle - \langle X_{\hat{a}_t, t}, \beta_* \rangle \right], \quad (2.1)$$

where the expectation is taken with respect to the randomness in the feature vector $\{X_{a,t}\}$ and the stochastic reward through $\{\epsilon_{a,t}\}$. Our goal is to seek a policy that minimizes the expected cumulative regret $R_T = \sum_{t=1}^T r_t$.

In the high-dimensional regime, the regression coefficient β_* , which is a high-dimensional parameter vector, is assumed to admit a sparse structure, namely, the number of nonzero entries in β_* is much smaller than the ambient dimension d . We denote $S = \{j : \beta_{*,j} \neq 0\}$ as the unknown index set of the nonzero entries of true parameter β_* that collects all influential feature components. Let $s_0 = |S|$ be the number of the nonzero entries in β_* , which satisfies $s_0 \ll d$. For each $t \in [T]$, let $X_t \in \mathbb{R}^{t \times d}$ be the time t -observed design matrix whose rows correspond to all the selected feature vectors $X_{\hat{a}_s, s}$ for $s = 1, \dots, t$ after $\hat{a}_s$ being pulled, and

vector $Y_t = (Y_1, \dots, Y_t)^T \in \mathbb{R}^t$ collects the observed rewards.

2.2.1 Assumptions

Let $\hat{\Sigma}_t = X_t^T X_t / t$ denote the sample covariance matrix at the end of round $t \in [T]$. For each matrix $A \in \mathbb{R}^{n \times d}$ and index set $I \subset [d]$, we use A_I to denote the submatrix of A that collects all columns whose indices are in I . Let $\mathbb{S}^{d \times d}$ denote the cone of all $d \times d$ positive semidefinite matrices. We now define the *sparse eigenvalue condition* for a matrix in $\mathbb{S}^{d \times d}$.

Definition 2 (Sparse eigenvalues). *For a $d \times d$ matrix $A \in \mathbb{S}^{d \times d}$ and $1 \leq m \leq d$, define $\rho_{\min}(m, A)$ and $\rho_{\max}(m, A)$ as the minimum and maximum m -sparse eigenvalues of A if*

$$\rho_{\min}(m, A) \leq \frac{\Delta^T A \Delta}{\|\Delta\|_2^2} \leq \rho_{\max}(m, A) \text{ for any } \Delta \neq 0 \text{ with } \|\Delta\|_0 \leq m.$$

Sparse eigenvalue condition is an important requirement in high-dimensional estimation problems [16], [17]. In particular, it often serves as a bridge for proving the *restricted eigenvalue condition* (c.f. Definition 3 In Section 2.3.5), which is nearly necessary for accurately learning the unknown parameter β_* according to high-dimensional statistics literature [18].

We now state two key assumptions for our theoretical analysis.

Assumption 1 (On the true reward).

(a) *Sparsity condition: There exist positive constants b and s_0 such that*

$$\|\beta_*\|_1 \leq b \text{ and } \|\beta_*\|_0 \leq s_0.$$

(b) *Margin condition: There exist positive constants Δ_* , C_1 and $\alpha \in [0, +\infty]$, such that for $h \in$*

$$\left[C_1 \sqrt{\frac{\log d}{T}}, \Delta_* \right], P(\langle X_{a_t^*, t}, \beta_* \rangle \leq \max_{b \neq a_t^*} \langle X_{b, t}, \beta_* \rangle + h) \leq \frac{1}{2} \left(\frac{h}{\Delta_*} \right)^\alpha.$$

The first part of the assumption requires boundedness and sparsity of the true parameter β_* for its estimability, which are standard requirements in the literature [4], [13].

The second part of the assumption controls the probability of the features falling into an h -neighborhood of the decision boundary. As α increases, the margin condition becomes stronger since the feature vectors are less likely to fall close to the decision boundary. As a result, it will be easier for a bandit policy to distinguish the optimal arm from sub-optimal arms. In particular, in the most extreme case with $\alpha = +\infty$, there is a deterministic positive gap between the rewards of the optimal arm and sub-optimal arms, making the bandit problem easiest. On the contrary, the margin condition becomes weaker as α decreases as it is satisfied by a large class of distributions. For example, with $\alpha = 0$, the feature vectors can have arbitrary distributions

around the decision boundary and do not need to exhibit any separation, making the bandit problem hardest. In particular, our condition under $\alpha = 0$ corresponds to the setup considered in [13], where at each round t , the decision-maker is given an arbitrary set $\mathcal{D}_t$ of feature vectors with no separation assumptions on the feature vectors. As a consequence, the cumulative regret of their algorithm scales $\mathcal{O}(\sqrt{T})$ in horizon T .

The margin condition with $\alpha = 1$ has been assumed by [19], [5] and [4] and will be satisfied when the density of $\langle X_{a,t}, \beta_* \rangle$ is upper bounded uniformly for $\forall a \in [K]$. Assumption 1(b) is a probabilistic relaxation of the usual gap assumption made for problem dependent bounds in the multi-armed bandit literature. For example, [13] assume that there exists some gap $\Delta > 0$ between the rewards of the best arm and the “second best” arm in the covariate set $\mathcal{D}_t$ such that $\min_{b \neq a_t^*} \langle X_{a_t^*,t} - X_{b,t}, \beta_* \rangle > \Delta$, which corresponds to our margin condition with $\alpha = +\infty$ and $\Delta_* = \Delta$. Assumption 1(b) relaxes this by allowing the gap $\langle X_{a_t^*,t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b,t}, \beta_* \rangle$ to be arbitrarily close to zero. This assumption resembles the usual margin condition in the classification literature [20]. The probability of features falling into an h -neighborhood of the decision boundary, $P(\{\langle X_{a_t^*,t}, \beta_* \rangle \leq \langle X_{b,t}, \beta_* \rangle + h, \forall b \neq a_t^*\})$, diminishes to zero as $h \rightarrow 0_+$, where a_t^* is the optimal arm at time t . In our regret analysis, we will choose a diminishing sequence of h values $\{h_t : t \in [T]\}$ to control the probability of pulling sub-optimal arms for deriving the regret bound in Theorem 2. The constant Δ_* in this assumption is related to the effective number of candidates for optimal arms. For example, if only m many arms have strictly positive probability to be optimal, then under the assumption that the probability density function of $\langle X_{a,t}, \beta_* \rangle$ is bounded for each $a \in [K]$, a simple union argument verifies Assumption (b) with $1/\Delta_* = O(m)$.

We provide a concrete example of distributions for which the margin condition holds with different values of α . We consider a 2-armed bandit problem with the covariate set $\mathcal{D}_t = \{X_t, X_t + Z_t\} \subseteq \mathbb{R}^d$ in each round t , where X_t can follow any distribution and $Z_t = (\zeta_t, 0, \dots, 0)^T \in \mathbb{R}^d$ has a random variable ζ_t at the first entry and zero otherwise. In addition, we assume that the first entry of the parameter vector β_* is 1. The margin condition for the 2-armed bandit problem can be expressed as

$$P(|\langle Z_t, \beta_* \rangle| \leq h) = P(|\zeta_t| \leq h) \leq \frac{1}{2} \left(\frac{h}{\Delta_*} \right)^\alpha.$$

Then, for any random variable ζ_t with distribution satisfying the above inequality around 0, the margin condition will hold with α . For example, ζ_t can be drawn from a random signed Beta distribution, i.e., $\zeta_t \sim \varepsilon \text{Beta}(\alpha, 1)$, where ε equals to ± 1 with equal probabilities. In this example, parameter $\alpha > 0$ corresponds to the margin parameter. This 2-armed bandit problem is also applicable in real world applications, in which the competing two candidates are very similar and only a few features has marginal difference between them. In such cases, the margin condition can be viewed as an assumption on the difference, e.g., Z_t , between the

feature vectors and the parameter α characterizes the level of similarity among different feature vectors.

Assumption 2 (On the features).

- (a) *Boundedness*: For some positive constant $x_{\max}$, $\|X_{a,t}\|_{\infty} \leq x_{\max}$ for all $a \in [K]$ and $t \in [T]$.
- (b) *Anti-concentration*: There exists a positive constant ζ such that for each $a \in [K]$, $t \in [T]$, $v \in \mathbb{R}^d$ and $h \in \mathbb{R}_+$, $P(\langle X_{a,t}, v \rangle^2 \leq h\|v\|_2^2) \leq \zeta h$.
- (c) *Sparse eigenvalues*: There exist constants $\Lambda_0 \geq \lambda_0 > 0$ and $C^* > 0$ such that for each $a \in [K]$, the minimum and maximum C^*s_0 -sparse eigenvalues of the matrix $\Sigma_a := \mathbb{E}[X_{a,t}X_{a,t}^T]$ are bounded below and above by λ_0^2 and Λ_0^2 , respectively. In addition, there exist positive constants $\Lambda_1^2 \geq \phi_0^2 > 0$ such that the minimum and maximum C^*s_0 -sparse eigenvalues of $\mathbb{E}[X_{a_t^*}X_{a_t^*}^T|\Gamma_t]$ are bounded below and above by ϕ_0^2 and Λ_1^2 , respectively. The event Γ_t is defined as $\Gamma_t := \{\langle X_{a_t^*,t}, \beta_* \rangle \geq \max_{b \neq a_t^*} \langle X_{b,t}, \beta_* \rangle + \Delta_*\}$.

Assumption 2(a) together with Assumption 1(a) ensures that the maximum regret at each time is bounded, since $|\langle X_{a,t}, \beta_* \rangle| \leq x_{\max}b$ by Cauchy-Schwarz inequality. Assumption 2(b) is an anti-concentration condition that plays a critical role in controlling the estimation accuracy for the regression coefficient β_* . In particular, it ensures directions of random feature vectors $X_{a,t}$ from each arm to be uniformly scattered so that the sample covariance matrix $\hat{\Sigma}_t$ will not concentrated in a single direction such as the direction of β_* . This anti-concentration assumption is mild. For example, it is implied by the boundedness of the probability density function of $\langle X_{a,t}, v \rangle$ for each $v \in \mathbb{R}^d$. Moreover, the parameter ζ only appears in the regret upper bound via a burn-in term and does not affect the leading term. More details can be found in the discussion after Theorem 2.

Assumption 2(c) is made to ensure that the design matrix corresponding to each arm is well-behaved so that the sample covariance matrix $\hat{\Sigma}_t$ satisfies the restricted eigenvalue condition (Definition 3) with a high probability (c.f. Proposition 2 in Section 2.3.5). The restricted eigenvalue condition is required for the ℓ_1 -error bound analysis of LASSO estimator (Proposition 1) and can be implied by the sparse eigenvalue condition (Definition 2) through the transfer principle (Lemma 5). Details of the proof can be found in Section 2.3.5. Here, C^* is a sufficiently large constant that could depend on K and can scale as $\mathcal{O}(K \log K)$ in the worst case. Similar assumptions are commonly adopted in existing work in the high dimensional bandit setting, e.g. [4] and more generally in the high-dimensional statistics literature, e.g. [21] and [22]. This assumption is critical for controlling the estimation accuracy for the regression coefficient β_* and in turn the regret analysis.

When samples are independent, several existing results in the literature (e.g. [21] and [22]) show a similar condition, *Restricted Isometry Property* (RIP), on the sample covariance matrix $\hat{\Sigma}_t$. However, these results on RIP and the corresponding proofs are not applicable in our case as the independence between

samples is violated in the bandit setting. In a nutshell, our proof shows that each vector in the restricted cone $\mathbb{C}(I) = \{v \in \mathbb{R}^d : \|v_{I^c}\|_1 \leq 3\|v_I\|_1\}$ with $I = S$ can be well approximated by a C^*s_0 -sparse vector in $\mathbb{R}^d$ (c.f. Lemma 6 in the appendix), from which we can construct a covering set of this restricted cone with C^*s_0 -sparse vectors of controlled cardinality. Then, we apply concentration inequalities after properly decoupling the dependence structure induced by the bandit policy (c.f. the proof of Proposition 2 in the appendix) to obtain a lower bound on the quadratic form $v^T \hat{\Sigma}_t v$ for each sparse vector $v \in \mathbb{R}^d$ in the covering set. Finally, we apply a union bound to prove a lower bound on $v^T \hat{\Sigma}_t v$ uniformly over all $v \in \mathbb{R}^d$ belonging to the restricted cone $\mathbb{C}(S)$, which implies the restricted eigenvalue condition for $\hat{\Sigma}_t$. Our actual proof is even more delicate than the proof outline described here and guarantees that the restricted eigenvalue does not depend on K . The description after Proposition 2 illustrates the high level idea of the actual proof.

To conclude this subsection, we provide a simple example where all our assumptions are satisfied with (ζ, Δ_*) not depending on the dimension d . For each arm, suppose the feature vector follows a truncated multivariate normal distribution truncated on the set $\|X_{a,t}\|_\infty \leq 1$ with different mean vectors μ_a and the same identity matrix as the covariance matrix. Under this setup, the inner product $\langle X_{a,t}, v \rangle$ follows a truncated normal distribution with bounded density function, so the first part of Assumption 2 is satisfied. Assumption 1(b) with $\alpha \in [0, 1]$ can be verified similarly as the order statistics of $\{\langle X_{a,t}, \beta_* \rangle : a \in [K]\}$ have bounded density functions. For truncated normal distributions, the optimal region Γ_t is well-spread, so the eigenvalue condition (Assumption 2(b) and 2(c)) is also satisfied. In the same example, non-identity covariance matrix can also be allowed if the eigenvalues of the covariance matrix are lower and upper bounded by some absolute positive constants.

2.2.2 Regret Lower Bound

In this section we provide a regret lower bound for any policy π in the linear bandit environment under the assumptions made in Section 2.2.1. We also provide a sketched proof for Theorem 1 later in Section 2.3.5.

Theorem 1. *Suppose the margin parameter $\alpha \in [0, 1]$. Let $\mathcal{P}_\alpha$ be the class of distributions of $(X_{a,t}, Y_{a,t} : a \in [d])$, where the feature vector $X_{a,t}$ is drawn from $\mathbb{P}_a$ and the reward $Y_{a,t} = \langle X_{a,t}, \beta_* \rangle + \epsilon_{a,t}$ for $a \in [K]$ satisfies Assumptions 1–2 with parameter α in Assumption 2(b). Then for large enough horizon T ,*

$$\inf_{\pi} \sup_{P_{X,Y} \in \mathcal{P}_\alpha} R_T(\pi, \pi^*) \geq C_L \left((\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}} + \log T \right), \quad (2.2)$$

where C_L is a constant independent of T and d .

Our Theorem 1 provides a more general result compared to existing results in the literature considering

both the dimension d and horizon T by leveraging the margin condition (Assumption 1(b)). The theorem shows a poly-logarithmic dependence on dimension d . In addition, it unifies the polynomial and logarithmic dependence on horizon T through the margin parameter α , which describes the hardness of distinguishing the optimal arm from the sub-optimal arms. In particular, when $\alpha = 1$, the lower bound scales as $\Omega(\log d + \log T)$, which is logarithmic in both d and T ; when $\alpha = 0$, the lower bound scales as $\Omega(\sqrt{T \log d})$, which is poly-logarithmic in d and polynomial in T .

For comparison, [23] prove a lower bound of $\Omega(T^{\frac{1-\alpha}{2}})$ for one-dimensional one-armed linear bandit problem, where they assume a margin condition corresponding to Assumption 1(b) with $\alpha \in (0, 1]$. In addition, they propose a policy with forced sampling, which achieves optimal regret in horizon T , i.e., $\mathcal{O}(T^{\frac{1-\alpha}{2}})$, when $\alpha \in (0, 1)$. Later, [19] prove a lower bound of $\Omega(\log T)$ for low-dimensional linear bandit problem, where they also assume a margin condition which corresponds to the $\alpha = 1$ case in Assumption 1(b). Meanwhile, they propose the OLS-bandit algorithm with $\mathcal{O}(d^3 \log T)$ regret upper bound, which is sub-optimal in d according to Theorem 1. [24] also prove a regret lower bound of $\Omega(d\sqrt{T})$ for low-dimensional linear bandit problem, where the covariate set $\mathcal{D}_t$ is a compact set of infinitely many feature vectors, e.g., the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$. Later, [3] prove a lower bound of $\Omega(\sqrt{Td})$ for low-dimensional contextual bandit problem, where no margin condition is assumed and the regret lower bound therein corresponds to the worst case when $\alpha = 0$, and propose the LinUCB algorithm with $\tilde{\mathcal{O}}(\sqrt{dT})$ regret upper bound, which is near-optimal according to Theorem 1. Our lower bound result unifies the existing results through the parameter α and applies to the high-dimensional linear bandit problems. More details of the comparison are deferred to Section 2.3.4.

Remark 1. *The problem formulation considered in [19] and [4] is slightly different from ours, where they assume that at each round, there is one common feature vector shared by all arms and each arm has its own parameter vector. Their formulation can be mathematically reparametrized into our formulation. In particular, they assume that each of the K arms has its own parameter vector $\beta_{*,a} \in \mathbb{R}^d$ for $a \in [K]$. The common feature vector X_t in each round is drawn i.i.d. from a distribution. Then, this formulation can be embedded into our formulation. Specifically, the feature vector $X_{a,t}$ for the a -th arm is a Kd -dimensional vector defined as $X_{a,t} := (0, \dots, 0, X_t^T, 0, \dots, 0)^T$, where the a -th block of $X_{a,t}$ is X_t and zero elsewhere. The common parameter $\beta_* := (\beta_{*,1}^T, \dots, \beta_{*,K}^T)^T$ in our setting is a concatenation of all vectors $\beta_{*,a}$ for $a \in [K]$. Due to the reparametrization, the regret lower bound result, i.e., $\Omega(\log T)$, proved in [19] can also be applied in our formulation.*

2.3 ℓ_1 -Confidence Ball Algorithm for High-dimensional Linear Bandit

In this section we propose a conceptually simple and computationally efficient bandit algorithm. The proposed algorithm is nearly optimal—it matches the regret lower bound when margin parameter α belongs to $[0, 1]$ and is optimal up to a factor of $\log T$ when $\alpha = 1$.

2.3.1 Optimism in the Face of Uncertainty

In this section, we motivate our ℓ_1 -confidence ball based method under the general framework of optimism in the face of uncertainty (OFU) introduced in [2] and [13]. More precisely, suppose that at the beginning of each round, say $t + 1$, we can construct a confidence set $\mathcal{C}_{t+1} \subseteq \mathbb{R}^d$ for the parameter β_* using the past selected features $X_{\hat{a}_1,1}, \dots, X_{\hat{a}_t,t}$ and observed rewards $Y_1, \dots, Y_t$ up to round t , so that β_* belongs to $\mathcal{C}_{t+1}$ with high probability. Then a OFU-based algorithm chooses the action-estimate pair

$$(X_{\hat{a}_{t+1,t+1}}, \tilde{\beta}_{t+1}) = \arg \max_{(x,\beta) \in \mathcal{D}_{t+1} \times \mathcal{C}_{t+1}} \langle x, \beta \rangle, \quad (2.3)$$

which jointly maximizes the expected reward within the covariate set $\mathcal{D}_{t+1} = \{X_{a,t+1}\}_{a=1}^K$ and the confidence set $\mathcal{C}_{t+1}$ of β_* . The shape and size of the confidence set $\mathcal{C}_{t+1}$ reflects our uncertainty on the unknown parameter β_* . In high-dimensional linear bandits, we consider to use the LASSO [15] estimator as the center of $\mathcal{C}_{t+1}$, since it adapts the sparsity structure of the parameter. As a remark, other penalized estimators such as MCP [25] and SCAD [26] can also be considered as the center for the confidence set, though need more sophisticated regret analysis.

At the beginning of round $t + 1$, we form the LASSO estimator $\hat{\beta}_t$ for β_* by viewing the observations as coming from the linear model $Y = X\beta_* + \epsilon$, with design matrix $X = X_t \in \mathbb{R}^{t \times d}$, response vector $Y = Y_t \in \mathbb{R}^t$, both defined in Section 2.2, and noise vector $\epsilon \in \mathbb{R}^t$ whose entries are independent σ -sub-Gaussian random variables. Given a regularization parameter $\lambda \geq 0$, the LASSO estimator is the solution of

$$\hat{\beta}_t \in \arg \min_{\beta} \{\mathcal{L}_t(\beta) + \lambda \|\beta\|_1\}, \quad (2.4)$$

where $\mathcal{L}_t$ denotes the empirical loss function at round t . In our regression framework, the loss function is the least squares objective $\mathcal{L}_t(\beta) = \frac{1}{2t} \|Y - X\beta\|_2^2$. Proposition 1 in Section 2.3.5 suggests a choice of the regularization parameter as $\lambda = \lambda_t = 2\sigma x_{\max} \sqrt{(2 \log t + 2 \log d)/t}$, where constant $x_{\max}$ (as well as the ϕ_0 in the following display) is defined in Section 2.2.1.

Having determined the center of the confidence set $\mathcal{C}_t$, it remains to determine its shape. Our proposal of the ℓ_1 ball as the shape of the confidence set is motivated by both theoretical and computational considerations: (i) Theoretically, the shape of the confidence set is important because the cumulative regret can be roughly characterized by the sum of the “volumes” of the confidence sets $\mathcal{C}_t$ over the time horizon. In the low-dimensional regime where $d \ll T$, since all norms in the Euclidean space are equivalent (up to some constant depending on the dimension), the shape of the ball induced by a norm has limited impact on the regret. In contrast, as the dimension d grows, unit balls under different norms have drastically different volumes making the choice of the shape crucial. For example, the volume ratio between unit ℓ_1 -ball and unit ℓ_2 -ball is $O(d^{-d/2})$, indicating the importance of using the right “norm” for constructing $\mathcal{C}_t$. As we will show in Corollary 1 in the appendix, the use of ℓ_1 -ball improves the dependence of the regret on the feature dimension from d to s_0 modulo $\log d$ terms where recall that s_0 is the sparsity level of the model. (ii) Computationally, we want to maintain the convexity of the optimization problem (2.3) by requiring the confidence set $\mathcal{C}_t$ to be convex. These two factors together motivate us to use a properly centered ℓ_1 -ball for $\mathcal{C}_t$, since $q = 1$ is the smallest number maintains the convexity of the ℓ_q -ball over $q \geq 0$.

Corollary 1 in the appendix provides a high probability upper bound to the ℓ_1 -distance $\|\hat{\beta}_t - \beta_*\|_1$ between $\hat{\beta}_t$ and β_* , motivating us to use the following confidence set

$$\mathcal{C}_{t+1} = \left\{ \beta \in \mathbb{R}^d : \|\hat{\beta}_t - \beta\|_1 \leq \tau_t \right\}, \quad (2.5)$$

where $\tau_t = \frac{384s_0\sigma x_{\max}}{\phi_0^2} \sqrt{\frac{2\log t + 2\log d}{t}}$.

Note that the size of the confidence set scales poly-logarithmically in feature dimension d (i.e. $\mathcal{O}(\sqrt{\log d})$), which improves upon the $\mathcal{O}(\sqrt{d})$ size of the confidence ellipsoid in [13] that has a polynomial dependence on dimension d . This is because we have leveraged on the sparsity structure of β_* captured by the LASSO estimator.

2.3.2 ℓ_1 -Confidence Ball Based Algorithm

In this subsection, we formally describe our algorithm for the high-dimensional linear bandit problem, which is summarized in Algorithm 1 below. The algorithm takes as inputs an initial regularization parameter λ_0 and an initial diameter of the confidence set τ_0 . Recall from the previous subsection that the ℓ_1 -confidence ball at time t is $\mathcal{C}_t = \{\beta : \|\beta - \hat{\beta}_{t-1}\|_1 \leq \tau_{t-1}\}$, where $\hat{\beta}_{t-1}$ is the LASSO estimator given by (2.4), and τ_{t-1}

is given in (2.5). The action selection rule (2.3) of the algorithm can be reformulated as

$$\begin{aligned} X_{\hat{a}_t, t} &= \arg \max_{x \in \mathcal{D}_t} \left\{ \max_{\beta \in \mathcal{C}_t} \langle x, \beta \rangle \right\} = \arg \max_{x \in \mathcal{D}_t} \left\{ \max_{\|u\|_1 \leq \tau_{t-1}} \langle x, \hat{\beta}_{t-1} + u \rangle \right\} \\ &= \arg \max_{x \in \mathcal{D}_t} \left\{ \langle x, \hat{\beta}_{t-1} \rangle + \tau_{t-1} \|x\|_\infty \right\}. \end{aligned} \quad (2.6)$$

The criterion function in the second line in equation (2.6) is composed of a point estimate of the expected reward plus the confidence width of x through its ℓ_∞ norm, making the optimization simple and the algorithm computationally efficient. The confidence width, also known as the exploration bonus, encourages the learner to visit new areas in the feature space to attain an optimal exploration versus exploitation trade-off. Computation-wise, the OLS-bandit algorithm of [19] and the LASSO-bandit algorithm of [4] require a pre-processing step to select a subset of arms based on forced-sample estimation; the OFUL-LS algorithm of [13] needs optimization over an ellipsoid that involves matrix inversion, which makes it computationally expensive.

Remark 2. *Our proposed algorithm reduces to a greedy algorithm when all $X_{a,t}$'s have the same ℓ_∞ -norm. However, in cases when $X_{a,t}$'s have very different magnitudes across arms, their sup-norms are able to reflect the prediction uncertainty in the point estimates $\langle X_{a,t}, \hat{\beta}_{t-1} \rangle$'s. As a consequence, our algorithm encourages explorations on those arms with high uncertainty while the greedy algorithm completely ignores this information and tends to exhibit poor performance.*

Algorithm 1 ℓ_1 -Confidence Ball Based Algorithm

Require: Initial regularization parameter λ_0 , and initial diameter τ_0 for the confidence set.

- 1: Initialize $\hat{\beta}_0$ by 0 in $\mathbb{R}^d$.
 - 2: **for** $t = 1$ **to** T **do**
 - 3: Observe $\mathcal{D}_t = \{X_{a,t}\}_{a=1}^K$.
 - 4: Obtain $\hat{\beta}_{t-1}$ by solving equation (2.4), i.e. $\hat{\beta}_{t-1} \in \arg \min_{\beta} \{\mathcal{L}_{t-1}(\beta) + \lambda_{t-1} \|\beta\|_1\}$.
 - 5: Construct confidence set $\mathcal{C}_t = \{\beta \in \mathbb{R}^d : \|\hat{\beta}_{t-1} - \beta\|_1 \leq \tau_{t-1}\}$.
 - 6: Select feature $X_{\hat{a}_t, t}$ from equation (2.6), and observe $Y_t = \langle X_{\hat{a}_t, t}, \beta_* \rangle + \epsilon_t$.
 - 7: Set $\lambda_t = \lambda_0 \sqrt{\frac{\log d + \log t}{t}}$ and $\tau_t = \tau_0 \sqrt{\frac{\log d + \log t}{t}}$.
 - 8: **end for**
-

2.3.3 Regret Analysis of Proposed Algorithm

In this subsection, we provide our theoretical regret analysis. The following theorem provides an upper bound on the cumulative regret for the proposed ℓ_1 -confidence ball based algorithm.

Theorem 2. *Suppose that Assumptions 1–2 hold and $K \geq 2$, $d \geq 1$, $\lambda_0 = 2\sqrt{2}\sigma x_{\max}$ and $\tau_0 = \frac{384\sqrt{2}s_0\sigma x_{\max}}{\phi_0^2}$.*

The expected cumulative regret of our algorithm at time horizon T can be bounded as

$$R_T \leq I_1^\alpha + \overbrace{C_3 x_{\max} b \Lambda_0^2 (K \log K) s_0 \log d}^{I_2} + \overbrace{C_4 \sigma^2 x_{\max}^5 b s_0^2 \zeta^2 K^2 \log d / \Delta_*^2}^{I_3} + \overbrace{C_5 x_{\max} b \log T}^{I_4}, \quad (2.7)$$

where

$$I_1^\alpha = \begin{cases} C_2 \frac{s_0^{\alpha+1} \sigma^{\alpha+1} x_{\max}^{2(\alpha+1)}}{\Delta_*^\alpha \phi_0^{2(\alpha+1)}} (\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}}, & \text{when } \alpha \in [0, 1), \\ C_2 \frac{s_0^2 \sigma^2 x_{\max}^4}{\Delta_* \phi_0^4} [\log d + \log T] \log T, & \text{when } \alpha = 1, \\ C_2 \frac{s_0^2 \sigma^2 x_{\max}^4}{(\alpha-1) \Delta_* \phi_0^4} \log d, & \text{when } \alpha \in (1, +\infty), \\ C_2 \frac{s_0^2 \sigma^2 x_{\max}^4}{\Delta_* \phi_0^4} \log d, & \text{when } \alpha = +\infty, \end{cases}$$

which is the leading term of the regret bound that depends on the margin parameter α defined in Assumption 1(b), and C_2, C_3, C_4, C_5 are positive constants not depending on T, d or α .

The regret bound provided by Theorem 2 for our algorithm shows that the regret of the algorithm grows poly-logarithmically in d , i.e., $R_T = \mathcal{O}((\log d)^{\frac{\alpha+1}{2}})$, when $\alpha \in [0, 1)$; logarithmically in d , i.e., $\mathcal{O}(\log d)$, when $\alpha \in [1, +\infty]$. Meanwhile, the cumulative regret grows polynomially in T , i.e., $R_T = \mathcal{O}(T^{\frac{1-\alpha}{2}})$, when $\alpha \in [0, 1)$; poly-logarithmically in T , i.e., $R_T = \mathcal{O}((\log T)^2)$, when $\alpha = 1$; and logarithmically in T , i.e., $R_T = \mathcal{O}(\log T)$, when $\alpha \in (1, +\infty]$. In addition, for $\alpha \in (1, +\infty)$, the term I_1^α will increase as $\alpha \rightarrow 1_+$ and scale as $\mathcal{O}(\log d \log T)$ when $\alpha = 1$ in the extreme case, which fills the gap between $\alpha = 1$ and $\alpha \in (1, +\infty)$ cases. Comparing the regret upper bound for our proposed method with the minimax regret lower bound established in Section 2.2.2, we can see that our method is optimal up to a factor of $\log T$ when $\alpha \in [0, 1]$.

Intuitively, $I_2 + I_3$ on the right-hand side of display (2.7) describes the regret caused by the “burn-in” period of exploring the space of the feature space which does not contribute to the asymptotic regret growth, while the last term I_4 is the cumulative regret when $\mathcal{C}_{t+1}$ does not contain β_* in Corollary 1. In the regret bound, constant Δ_* (defined in Assumption 1(b)) plays the role of the gap parameter as appeared in a typical problem-dependent regret bound [13]. The first term is due to the risk of selecting sub-optimal arms when the features fall near the decision boundary, which is controlled by the margin condition (Assumption 1(b)). From the result of Theorem 2, we can see that when α is larger, the feature vectors are less likely to fall into the close neighborhood of the decision boundary and the bandit environment will become easier to learn. In particular, for the $\alpha = +\infty$ case when there is a deterministic positive gap Δ_* between the rewards of the optimal arm and sub-optimal arms, the regret bound for the algorithm scales polynomially in both dimension

d and horizon T , i.e., $\mathcal{O}(\log d + \log T)$. We discuss how our regret bound compares with and provides a unified view of the existing results in Section 2.3.4.

2.3.4 Comparisons with Existing Literature

Problem Formulation

The stochastic linear bandit problem was first introduced by [2], and was subsequently studied by [12], [24] and [3]. Later, [13] and [27] proposed the OFUL algorithm for both low-dimensional and high-dimensional settings. The model parametrization considered in [13], [12] and [24] is the same as ours. However, there is one major difference between the two formulations that the covariate set in their problem consists of infinitely many feature vectors, e.g., the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$, while the covariate set $\mathcal{D}_t$ in our paper consists of finitely many feature vectors drawn from some underlying distribution. Moreover, the covariate set in [12] and [24] does not change over time, therefore the optimal arm will be the same in different rounds. In contrast, the optimal arm in our setting varies along the time due to the randomness in the feature vectors. In addition, the feature space can be explored implicitly in our setting and no explicit exploration phase is needed as in [24].

In another parallel strand of linear bandit research, [19] proposed an OLS-bandit algorithm for low-dimensional multi-armed bandit problems, where the formulation is different from ours. In particular, they consider the K -armed bandit problem with bounded linear response, where feature vector at each time is common to all arms and parameters are arm-specific. After that, [4] proposed a Lasso-bandit algorithm for the high-dimensional linear bandit problem which requires forced sampling and prior knowledge on their gap parameter. Later, in the same setting as [4], [5] introduced an MCP-bandit algorithm which also requires forced sampling and knowledge of gap parameter. As discussed in the remarks after Theorem 1, our problem formulation is also different from that in [4]. The formulation in our paper is meaningful on its own in real world applications. In some real applications, many baseline features are available which have the same effect in the sense of sharing the same regression coefficient β_* across different arms. In such cases, our formulation can help estimate the parameters associated with the common features more accurately by using all the data. However, the formulation in [4] does not incorporate such structure, which will cause loss of information. For example, in clinical trials or advertisement problems, we can map the interactions between the features of patients (or customers) and treatments (or advertisements) as a feature vectors $X_{a,t}$ for each arm, with one common parameter vector β_* across arms. In these applications, the parameters corresponding to the main effect of patients (or customers) such as age and gender are the same for all arms. Our formulation facilitates sample-efficient estimation of these parameters by combining data from multiple

arms. In comparison, the parameters under the formulation in [4] are estimated separately for different arms. In addition, the difference between the two formulations has been considered in [28], where they observe that our formulation is advantageous in such practical applications. In particular, when the number of arms is large, it is not practical to apply the formulation in [4] due to the large number of parameters. When the set of arms changes over time such as in online advertisement, it is also not feasible to assign a different parameter for every new incoming item as in [4].

Apart from the above benefits, the motivation for our setting comes from applications such as online advertisements, where there are finitely many available products (arms) for each incoming customer. In addition, available products are usually different for different customers and drawn from some underlying distribution, where each arm is a category of products that is associated with a distribution over the set of products.

Regret Bound Analysis

In terms of the regret bound analysis, [24] proved a lower bound of $\Omega(d\sqrt{T})$ for both cumulative regret and Bayes risk in low-dimensional linear bandit problem, where the covariate set $\mathcal{D}_t$ is a compact set of infinitely many feature vectors, e.g., the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$. They also proposed an algorithm with matching regret upper bound that alternates between exploration and exploitation phases. [3] proposed the LinUCB algorithm, which has the regret upper bound $\mathcal{O}(\sqrt{Td\log(T\log(T)/\delta)})$ with probability $1 - \delta$, and proved a regret lower bound of $\Omega(\sqrt{dT})$ for the low-dimensional contextual bandit problem with linear payoff, which corresponds to our $\alpha = 0$ case. [13] proved that the expected cumulative regret of the OFUL algorithm scales as $\mathcal{O}(d\sqrt{T})$ in both low-dimensional and high-dimensional settings. In the above literature, the feature vectors are allowed to be chosen arbitrarily by an adversary and the analysis is based on the worst case, leading to the polynomial dependence on dimension d . In contrast, [19] proved a cumulative regret of $\mathcal{O}(d^3 \log T)$ for their proposed OLS-bandit algorithm and a regret lower bound of $\Omega(\log T)$ corresponding to the margin condition (Assumption 1(b)) with $\alpha = 1$. However, the cumulative regret of the above algorithms scales polynomially in dimension d , which can be sub-optimal in the high-dimensional setting. The Lasso-bandit algorithm in [4] and the MCP-bandit algorithm in [5] are shown to have an improved regret upper bound $\mathcal{O}(\log d)$ in dimension d under the margin condition with $\alpha = 1$, but require forced sampling, which could be costly in some practical cases such as medical and marketing applications. In comparison, our proposed algorithm (Algorithm 1) does not require any forced sampling or knowledge of the gap parameter. Meanwhile, our theoretical results on regret upper bound can be applied to any margin parameter $\alpha \in [0, +\infty]$. On the other hand, the setting of [5] and [4] corresponds only to the case of $\alpha = 1$, for which our algorithm achieves the same regret bound dependence on d , i.e., $\mathcal{O}(\log d)$.

2.3.5 Proof Sketch for the Main Results

In this section, we outline the proofs for Theorem 1 and Theorem 2.

Proof Sketch for Theorem 1

In this section, we provide a sketch of the proof for Theorem 1. In [19], the author already proved a lower bound of $\Omega(\log T)$ under $\alpha = 1$ case. In this paper, we prove a lower bound on dimension d and also the polynomial dependence on horizon T when margin condition parameter $\alpha < 1$. In order to prove the lower bound result, we follow the standard proof technique by designing a class $\mathcal{P}_{\mathcal{M}}$ of distributions, where $\mathcal{M}$ denotes the parameter space for the regression coefficient vector β in the linear model $Y_{a,t} = \langle X_{a,t}, \beta \rangle + \epsilon_{a,t}$, and then reducing the worst case of cumulative regret in $\mathcal{P}_{\mathcal{M}}$ to the average case by introducing a prior distribution over the class $\mathcal{P}_{\mathcal{M}}$. The novelty of our proof comes from the design of the class $\mathcal{P}_{\mathcal{M}}$, which is a hard-to-learn subset of $\mathcal{P}_{\alpha}$ that yields the lower bound. Specifically, $\mathcal{P}_{\mathcal{M}}$ is composed of bandit problems with two arms, the first of which is $X^{(1)} = (X_0, X_1, \dots, X_d)$ and the second is $X^{(2)} = (0, -X_1, \dots, -X_d)$. Here X_0 is a random variable taking values in $\{-1, 0, 1\}$, while X_i for $i \in \{1, \dots, d\}$ follows truncated normal distribution of $N(0, 1)$. We take $\mathcal{M} = \{\beta = (\beta_0, \beta_1, \dots, \beta_d) \in \mathbb{R}^{d+1} : \beta_0 = 1, \beta_u = \theta, \beta_j = 0 \text{ for } j \neq u, u \in [d]\}$ as a set of true parameter β_* , where θ is a tuning parameter controlling the signal strength of the second largest component of β . Here we assume that $\beta_0 = 1$ is known and does not need to be estimated, and the other nonzero entry $\beta_u = c\sqrt{\frac{\log d}{T}}$ with a small constant c . The choice of $\mathcal{M}$ is to make sure that assumptions in Section 2.2.1 are satisfied and the magnitude of β_u is small so that it is hard for policy to distinguish it from 0. Moreover, we specify a uniform prior distribution on the parameter vector β_* over the set $\mathcal{M}$. According to Bayesian decision rule, we prove that the average cumulative regret can be lower bounded by the regret incurred by the Bayesian optimal policy. By following the common proof technique of reducing the problem of obtaining the estimation lower bound to a multiple hypothesis testing problem, we show that learning β from space $\mathcal{M}$ reduces to a variable selection problem that causes the poly-logarithmic term in d and the polynomial term in T in our new lower bound. Specifically, the multiple hypothesis testing problem is to determine the index u of the nonzero entry β_u from $\{1, \dots, d\}$ and the optimal Bayes rule is $\hat{u}_{t-1} = \arg \max_{j \in [d]} |\hat{\beta}_{j,t-1}|$, where $\hat{\beta}_{t-1}$ is defined according to the Bayesian policy. Finally, by applying Fano's Lemma (Lemma 1), we derive the lower bound for the probability of error $P(\hat{u}_{t-1} \neq u)$ in multiple hypothesis testing problem that leads to the regret lower bound.

Proof Sketch for Theorem 2

In this section, we provide the proof strategy for Theorem 2. In order to prove Theorem 2, we need to prove in Corollary 1 that the ℓ_1 -confidence set $\mathcal{C}_t$ in (2.5) contains β_* with high probability for properly selected regularization parameters. Since we choose LASSO estimator as the center of the confidence set $\mathcal{C}_t$, we need to analyze the error bound of the LASSO estimator. This is challenging since the observed data are highly dependent from each other due to bandit policy, thus we cannot directly apply the standard error analysis of the LASSO estimator. In particular, the restricted eigenvalue condition for $\hat{\Sigma}_t$, which is required for the analysis of LASSO estimator, is hard to prove for dependent data. In this paper, we introduce a convergence result (Proposition 1) on LASSO estimator for dependent data and a novel method to prove the eigenvalue condition (Proposition 2). Finally, we derive the cumulative regret taking into account of the regret incurred under different circumstances.

Firstly, we provide an ℓ_1 -error bound for the LASSO estimator. Specifically, Proposition 1 shows that the LASSO estimator falls into an $\mathcal{O}(\sqrt{\log d/t})$ neighborhood around the true parameter β_* with high probability. Before the statement of Proposition 1, we introduce the definition of *restricted eigenvalue condition* for positive semidefinite matrices.

Definition 3 (Restricted eigenvalue condition). *For any set of indices $I \subseteq [d]$ and a positive constant ϕ , define the set*

$$\mathcal{C}(I, \phi) = \{M \in \mathbb{S}^{d \times d} \mid \forall v \in \mathbb{R}^d \text{ s.t. } \|v_{I^c}\|_1 \leq 3\|v_I\|_1, \text{ we have } \|v\|_1^2 \leq |I|(v^T M v)/\phi^2\}.$$

A matrix that belongs to this set is said to satisfy the restricted eigenvalue condition over index set I with parameter ϕ .

On the event that the sample covariance matrix $\hat{\Sigma}_t$ at time t belongs to $\mathcal{C}(S, \phi)$ for some suitable constant ϕ , Proposition 1 extends the usual LASSO analysis for i.i.d. observations to our bandit setting where observations are dependent due to the sequential decision policy.

Proposition 1. *Let X_i denote the i^{th} row of X and $Y(i)$ denote the i^{th} entry of Y . The sequence $\{X_i, i = 1, \dots, t\}$ forms an adapted sequence of observations, i.e., X_i depends on past features and their rewards $\{X_{i'}, Y(i')\}_{i'=1}^{i-1}$. Also, assume that all X_i satisfy $\|X_i\|_\infty \leq x_{\max}$ and the regularization parameter*

$\lambda = 2\sigma x_{\max} \sqrt{(2\log t + 2\log d)/t}$. Then for any $\phi > 0$, we have

$$\begin{aligned} & P \left[\|\hat{\beta} - \beta_*\|_1 \leq \frac{6s_0\sigma x_{\max}}{\phi^2} \sqrt{\frac{2\log t + 2\log d}{t}} \right] \\ & \geq 1 - \frac{2}{t} - P[\hat{\Sigma}(X) \notin \mathcal{C}(\text{supp}(\beta_*), \phi)], \end{aligned}$$

where $\hat{\beta}$ is the LASSO estimator by solving equation (2.4) with X, Y and regularization parameter λ .

Proposition 1 is a more general version of the LASSO oracle inequality. It is an adaption of Proposition 1 in [4], where we plug-in specific values for λ and the ℓ_1 -error bound for $\hat{\beta}$. For self-containedness, we also include a proof for Proposition 1 in the appendix. It allows for the adapted sequence of observations and errors that are σ -sub-Gaussian conditional on all past observations. Moreover, note that the performance of the LASSO estimator depends on the choice of parameter ϕ and the structure of the sample covariance matrix. This is one of the reasons why multi-armed bandit algorithms such as [4] has forced sampling step to ensure that the sample covariance matrix $\hat{\Sigma}_t$ is positive definite in some directions with high probability.

Secondly, we prove the restricted eigenvalue condition for $\hat{\Sigma}_t$ based on dependent data. Proposition 2 shows that with the selection rule induced by our method and the construction of the confidence set, the sample covariance matrix will satisfy the restricted eigenvalue condition with high probability.

Proposition 2. Suppose we construct the confidence set $\mathcal{C}_{t+1} = \{\beta \in \mathbb{R}^d : \|\hat{\beta}_t - \beta\|_1 \leq \frac{6s_0\sigma x_{\max}}{\phi^2} \sqrt{\frac{2\log t + 2\log d}{t}}\}$ with some constant $\phi > 0$ that is large enough. Then if the time horizon t exceeds a certain threshold (i.e. $t \geq \mathcal{O}(\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$), the sample covariance matrix $\hat{\Sigma}_t$ are guaranteed to satisfy the compatibility condition with high probability, i.e.

$$P[\hat{\Sigma}_t \in \mathcal{C}(\text{supp}(\beta_*), \phi_*)] \geq 1 - \mathcal{O}(1/t) - \mathcal{O}(e^{-t+s_0 \log K \log d}),$$

where $\phi_* = \frac{\phi_0}{8}$.

The detailed proof of Proposition 2 is given in Section 2.5.3. Here we only provide some intuition about how we prove Proposition 2. Firstly, we provide a crude result of the restricted eigenvalue condition for the sample covariance matrix in Proposition 3, i.e., the restricted eigenvalue condition constant is of order $1/K$ where K is the number of arms. The proof of Proposition 3 consists of three main steps:

- (i) First, we discretize $\mathcal{H} = \{v \in S^{d-1} : \|v\|_0 \leq Cs_0\}$, where constant $C = \Theta(\zeta\Lambda_0^2 K \log K)$, with a ϵ -net $\mathcal{N}_\epsilon$ (Definition 5). Then, we prove the restricted eigenvalue condition for $\hat{\Sigma}_t$ over vectors in $\mathcal{N}_\epsilon$. In detail, We apply concentration inequalities after properly decoupling the dependence structure induced by the bandit policy to obtain a lower bound of order $1/K$ on the quadratic form $\langle v, \hat{\Sigma}_t v \rangle$ for all $v \in \mathcal{N}_\epsilon$

with high probability. This step is guaranteed by Assumption 2(b), which ensures that the probability of feature vectors locating at the original point is small.

- (ii) We prove the restricted eigenvalue condition over $\mathcal{H}$ through approximation from $\mathcal{N}_\epsilon$.
- (iii) We prove the eigenvalue condition over the cone $\mathbb{C}_3(S) = \{v \in \mathbb{R}^d : \|v_{S^c}\|_1 \leq 3\|v_S\|_1\}$ (Definition 4), which implies the restricted eigenvalue condition. Here we apply the Transfer Principle (Lemma 5) and carefully choose the constant C as discussed earlier in step (i).

Then in Proposition 2, we prove a refined restricted eigenvalue condition based on Proposition 3. The crucial part is to eliminate the dependence on K which is required in Proposition 3. This is challenging since the observed samples are high dependent and in Proposition 3, we need to consider the worst case across K arms. Instead of applying forced sampling to guarantee the eigenvalue condition of $\hat{\Sigma}_t$ as in [4], we apply a novel technique to prove the result in Proposition 2. In particular, with the crude restricted eigenvalue condition in Proposition 3, we can show that the optimal arm will be pulled sufficient times, i.e., a positive fraction of time horizon. As a result, the restricted eigenvalue condition in Proposition 2 is satisfied since Assumption 2(c), which ensures that the feature vectors are randomly diverse, i.e., the minimum eigenvalue of the population covariance matrix is positive. Intuitively, this can guarantee that the space of feature vectors can be automatically explored without forced sampling. The main proof of Proposition 2 can be summarized in two steps:

- (i) In Lemma 10, we show that based on Proposition 3, the algorithm will pull the optimal arm at a positive fraction of times after some time point. In particular, Assumption 1(b) and 2(c) guarantee that $P(\Gamma_t) \geq 1/2$. Under event Γ_t , there is a positive gap Δ_* between the reward of optimal arm and sub-optimal arms. According to Proposition 1 and Proposition 3, for each arm $X_{a,t}$, the optimal reward of $\tilde{\beta}$ within $\mathcal{C}_t$ in (2.3) is close to the true reward of $X_{a,t}$. Thus, the algorithm will only select the optimal arm under Γ_t after some time point according to equation (2.3).
- (ii) Assumption 2(c) guarantees the sparse eigenvalue condition with constant ϕ_0 for optimal arm under event Γ_t . We have shown in Lemma 10 that the optimal arm will be pulled frequently. Then, combining the above results, we prove the restricted eigenvalue condition for $\hat{\Sigma}_t$ with constant $\phi_* = \phi_0/8$, which does not depend on K , even though the observed samples are dependent.

Thirdly, we provide the following corollary which shows that the true parameter β_* lies in the ℓ_1 -confidence set $\mathcal{C}_{t+1}$ with high probability. Corollary 1 follows from a direct application of Proposition 1 and Proposition 2, and will play a crucial role in analyzing the cumulative regret of our algorithm.

Corollary 1. *Suppose that Assumptions 1–2 hold and the regularization parameter is chosen as $\lambda_t = 2\sigma x_{\max} \sqrt{(2\log t + 2\log d)/t}$. Then, there exist absolute constants $C, c_1, c_2 > 0$, such that for all $t \geq C(\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 K^2 \log d}{\Delta_*^2})$, with probability at least $1 - c_1 t^{-1} - c_2 e^{-t + s_0 \log K \log d}$, β_* lies in the confidence set $\mathcal{C}_{t+1}$ defined by Equation (2.5).*

The constants c_1, c_2 in Corollary 1 are absolute constants and can be chosen as $c_1 = 1/(2e)$ and $c_2 = 2e^2$. The size of the confidence set $\mathcal{C}_{t+1}$ scales as $\mathcal{O}(\sqrt{\frac{\log d + \log t}{t}})$ in dimension d and time t . In comparison, the size of the ellipsoid confidence set of [13] centered at the ridge regression estimator scales as $\mathcal{O}(\sqrt{\frac{d \log t}{t}})$, which has a polynomial dependence on the dimension d , and is exponentially larger than the size of our ℓ_1 confidence set.

Finally, we apply the result in Corollary 1 to compute the cumulative regret of the proposed algorithm. As discussed after Theorem 2, for the instant regret incurred in “burn-in” period or when $\mathcal{C}_t$ does not contain β_* , we simply bound it with worst-case regret $2bx_{\max}$. When $\mathcal{C}_t$ contains β_* , we consider two complementary cases based on event $\Gamma_{\xi_t} = \{\omega \in \Omega : \langle X_{a_t^*, t}, \theta_* \rangle \geq \langle X_{b, t}, \theta_* \rangle + \xi_t, \forall b \neq a_t^*\}$ with $\xi_t = 2x_{\max} \tau_t$. We show that when Γ_{ξ_t} holds, the algorithm will only select optimal arm and incurs no regret. Then according to Assumption 1(b), we bound the regret incurred when Γ_t does not hold. In the end, we sum up the regret from different parts to obtain the cumulative regret.

2.4 Experimental Results

We compare our ℓ_1 -confidence ball based method with (i) the OFUL-LS method proposed in [13], (ii) the OLS-bandit introduced in [19], and (iii) the LASSO-bandit algorithm in [4] in both synthetic data and real data experiments. The first two methods are not specifically designed for high-dimensional settings. We note that the parametrization of OLS-bandit and LASSO-bandit methods is slightly different from ours, because they treat different arms to have different parameter vectors and one common feature vector. We apply these methods after converting their parametrization into ours.

2.4.1 Synthetic Data

In the synthetic data experiment, we consider three scenarios for K, d and s_0 : (1) $K = 5, d = 100, s_0 = 5$; (2) $K = 5, d = 1000, s_0 = 5$; (3) $K = 50, d = 20, s_0 = 2$. In each case, a randomly chosen subset of s_0 features is predictive of the reward. At each time t , feature vectors for K arms are i.i.d. generated by truncating a Gaussian distribution $N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ so that $\|X_{a, t}\|_{\infty} \leq 1$. The error term $\epsilon_{a, t}$ follows a zero mean normal distribution with variance $\sigma^2 = 1$. For this setting, our margin condition holds with $\alpha = 1$ so

that we expect a regret growing logarithmically in T . For our method, we choose the initial regularization parameter $\lambda_0 = 0.5$ and diameter $\tau_0 = 1$ and the method is robust to the choice of these tuning parameters. For LASSO and OLS, we choose the forced sampling parameter $q = 1$, the localization parameter $h = 5$ for LASSO-bandit and $h = 1$ for OLS-bandit. For LASSO-bandit, we further set the initial parameters $\lambda_1 = \lambda_{2,0} = 0.5$ and for OFUL-LS, we set $\lambda = 1$ and $\delta = 10^{-4}$.

Figure 2.1 compares our method with competitors on the aforementioned synthetic data with a time horizon of $T = 2000$. The curves are the average cumulative regrets over 5 trials. We observe that the proposed ℓ_1 -confidence ball based algorithm outperforms all the three competing algorithms in all the cases. In cases (1) and (2) where dimension d is from moderately large to high, OLS-bandit and OFUL-LS algorithms do not perform well since they are not designed for high-dimensional settings and fail to capture the sparsity structure. In case (3) where K is large and the feature vector is low-dimensional, our ℓ_1 -confidence ball algorithm still outperforms the competitors, while the performance of LASSO-bandit is no longer competitive. One possible reason is that since the number of arms is larger, it may need much more exploration in the feature space than it is allowed. LASSO-bandit does forced sampling only at a limited number of pre-specified times due to which it may not have sufficient exploration to accommodate the large number of arms. In contrast, our algorithm performs an implicit exploration and does not require forced sampling.

2.4.2 Warfarin Dosing Data

We now consider data from a real experiment from the healthcare context where a physician needs to determine the optimal warfarin dosage for each patient. The warfarin dataset [29] has experimental data on 5528 patients, and contains information on patients' demographics, diagnosis and genetic information. The same dataset is also used by [4], where they demonstrated the benefits of applying LASSO-bandit algorithms over OFUL-LS, OLS-bandit algorithms, and other low-dimensional methods. [4] partitioned the dataset into three arms based on the optimal dosage for each patient. However, their partitions are highly unbalanced with one arm having only 13% and another arm having 54% of the patients. Therefore, in our analysis, we evenly partition the dataset into four arms based on the quantiles of the optimal warfarin dosages in the dataset: (1) Level 1: $(0.0, 19.5]$ mg/week, (2) Level 2: $(19.5, 28.0]$ mg/week, (3): $(28.0, 38.5]$ mg/week, and (4) Level 4: $(38.5, \infty)$ mg/week. The proportion of each arm is nearly 25% after this partition. Following [4], we construct 94 patient-specific covariates, including intercept and indicators for categorical variables.

We apply our ℓ_1 -confidence ball based method to the dataset along with other methods. We include a Doctor's policy for comparison, which always assigns a level 2 dose that has the highest percentage (27.9%) of patients in the warfarin data. We also include oracle policies that assign the optimal dose given the true

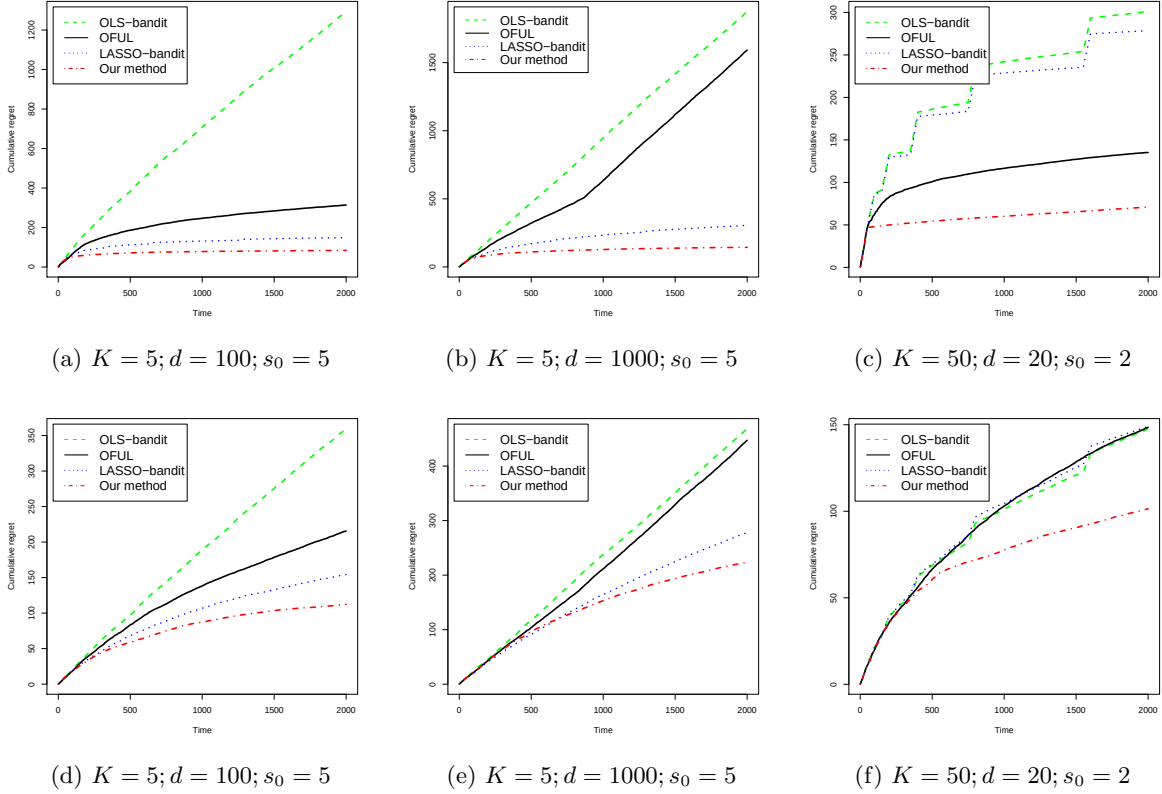


Figure 2.1: Comparison of the cumulative regret of the ℓ_1 -confidence ball algorithm against competitors. Nonzero components in β_* are generated from $\text{Unif}(0, 1)$ in (a)-(c), and $\text{Unif}(0, 0.2)$ in (d)-(f). Lower value indicates better performance.

parameter β_* . Similar to [4], the “true” parameter vector is estimated using all patient outcomes. We consider two versions of optimal policy: Linear Oracle that estimates β_* using linear regression, and Logit Oracle that estimates β_* using multinomial logistic regression (with four categories).

For the real dataset, the expected regret (2.1) is not computable, since we do not know the true parameter vector β_* . Instead, we use a surrogate measure which calculates the misclassification rate of assigning optimal dosage to patients. The misclassification rate is calculated as the number of incorrect decisions divided by the number of patients. The lower the expected regret, the lower the misclassification rate. We consider 10 random permutations of patients and take the average of the misclassification rate of 10 permutations. Figure 2.2 illustrates the average fraction of incorrect dosing decisions under different policies. We observe that our method outperforms other competitors except for the oracle policies. Especially, when the number of observations is smaller than 1000, the performance of our method is clearly better than other algorithms, which indicates the benefits of our method when the sample size is relatively small compared to the dimension.

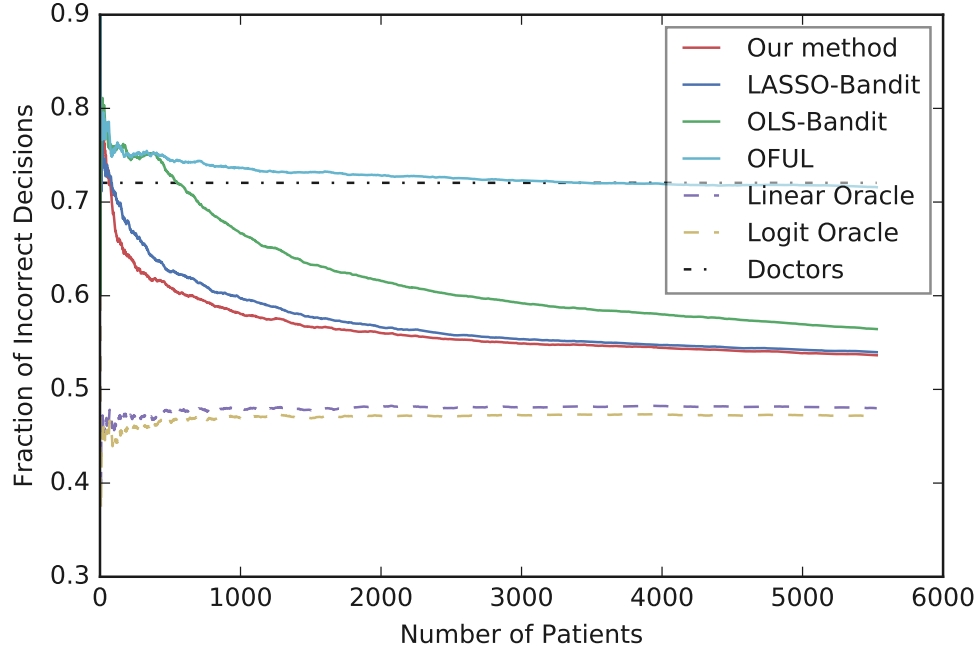


Figure 2.2: Comparison of the fraction of incorrect dosing decisions under the oracle, LASSO Bandit, OLS Bandit, OFUL, the ℓ_1 -confidence ball algorithm, and doctor policies in the warfarin data. Lower value indicates better performance.

2.5 Supplementary Material

2.5.1 Proof of Theorem 1

We will apply the Bayesian decision theory to show the lower bound of the cumulative regret up to horizon time T . To prove the $\log T$ term in Theorem 1, we can follow the procedure in [19]. For simplicity, we will focus on the terms in lower bound involving the dimension d .

Suppose we have two arms, the first of which is $X^{(1)} = (X_0, X_1, \dots, X_d)$ and the second is $X^{(2)} = (0, -X_1, \dots, -X_d)$. Here the two arms are not independent from each other. Entry X_0 in the arm vector follows a discrete distribution which will be specified later. The other entries $\{X_1, \dots, X_d\}$ are i.i.d. truncated normal distribution of $N(0, 1)$. For simplicity, we first consider the standard normal case, later the proof can be easily applied to the truncated normal case. The common parameter vector is β is from a set $\mathcal{M} = \{\beta \in \mathbb{R}^{d+1} : \beta_0 = 1, \beta_u = \theta, \beta_j = 0 \text{ for } j \neq u, u \in [d]\}$. The sparsity of parameters in set $\mathcal{M}$ is then $s_0 = 2$. Within the set $\mathcal{M}$, the first entry β_0 of the parameter vector is assumed to be known, so we do not need to estimate it. The remaining d entries of the parameter vector has exactly one nonzero entry with value

θ . We assume that the parameter vector β is uniformly distributed within the set $\mathcal{M}$. In order to guarantee that the assumptions in section 2.2.1 hold for the configuration of the parameters in $\mathcal{M}$ and data, we define $\beta_{\min} := \sqrt{\frac{\log d}{T}}$ and $\theta := c\beta_{\min}$, where the constant c is sufficiently small and will be defined momentarily, and

$$\begin{cases} P(X_0 = 0) = C\beta_{\min}^\alpha \\ P(X_0 = a) = \frac{1}{2}(1 - C\beta_{\min}^\alpha), \quad \text{when } a = \pm 1 \end{cases} \quad (2.8)$$

Define $\mathcal{P}_\alpha$ as the bandit environment where Assumptions 1–2 hold with constant α in Assumption 1(b), and $\mathcal{P}_\mathcal{M}$ be the set of environments defined as above with β from $\mathcal{M}$. We have $\mathcal{P}_\mathcal{M} \subseteq \mathcal{P}_\alpha$ according to the configuration of $\mathcal{P}_\mathcal{M}$. Moreover, we define $\mathcal{F}_{t-1}^+ = \sigma(X_1^{(1)}, X_1^{(2)}, Y_1^{(\pi_1)}, \dots, X_{t-1}^{(1)}, X_{t-1}^{(2)}, Y_{t-1}^{(\pi_t)}, X_t^{(1)}, X_t^{(2)})$ as a σ -algebra. Then for any policy π , the supreme of the cumulative regret of π at horizon T within $\mathcal{P}_\alpha$ can be lower bounded by:

$$\begin{aligned} & \sup \{R_T(\pi, \pi^*) : P_{X,Y} \in \mathcal{P}_\alpha\} \\ & \geq \sup_{P_{X,Y} \in \mathcal{P}_\mathcal{M}} \mathbb{E} \sum_{t=1}^T |\beta^T \Delta_{x,t}| [I\{\beta^T \Delta_{x,t} \geq 0, \pi_t = 2\} + I\{\beta^T \Delta_{x,t} < 0, \pi_t = 1\}] \\ & \geq \mathbb{E} \left\{ \sum_{t=1}^T (\mathbb{E}_\beta [\beta^T \Delta_{x,t} I\{\beta^T \Delta_{x,t} \geq 0\} | \mathcal{F}_{t-1}^+] I\{\pi_t = 2\} \right. \\ & \quad \left. - \mathbb{E}_\beta [\beta^T \Delta_{x,t} I\{\beta^T \Delta_{x,t} < 0\} | \mathcal{F}_{t-1}^+] I\{\pi_t = 1\}) \right\}, \end{aligned} \quad (2.9)$$

where $P_{X,Y} := (P_{X^{(1)}, Y^{(1)}}, P_{X^{(2)}, Y^{(2)}})$ is the joint distribution of the feature vectors and the response variables, and $\Delta_{x,t} := X_t^{(1)} - X_t^{(2)} = (\Delta_{0,t}, \dots, \Delta_{d,t})$. Here $\mathbb{E}_\beta [\cdot | \mathcal{F}_{t-1}^+]$ denotes the expectation of β in $\mathcal{M}$ conditioned on $\mathcal{F}_{t-1}^+$, which is the σ -algebra generated by feature vectors up to time t and observed response variables of the chosed arms up to time $t-1$. Then, according to the Bayesian decision rule in [19], the optimal policy is $\hat{\pi}_t = 1$ if

$$\begin{aligned} \mathbb{E}_\beta [\beta^T \Delta_{x,t} I\{\beta^T \Delta_{x,t} \geq 0\} | \mathcal{F}_{t-1}^+] & \geq -\mathbb{E}_\beta [\beta^T \Delta_{x,t} I\{\beta^T \Delta_{x,t} < 0\} | \mathcal{F}_{t-1}^+] \\ \iff \mathbb{E}_\beta [\beta^T \Delta_{x,t} | \mathcal{F}_{t-1}^+] & \geq 0 \end{aligned} \quad (2.10)$$

and $\hat{\pi}_t = 2$ otherwise. Define $\mathcal{F}_{t-1} = \sigma(X_1^{(1)}, X_1^{(2)}, Y_1^{(\pi_1)}, \dots, X_{t-1}^{(1)}, X_{t-1}^{(2)}, Y_{t-1}^{(\pi_t)})$ as the σ -algebra generated by feature vectors and observed responses up to time $t-1$. We set $\hat{\beta}_{t-1} = \mathbb{E}[\beta | \mathcal{F}_{t-1}^+]$. Since $\Delta_{x,t}$ is independent from $\mathcal{F}_{t-1}$ and β , we have $\mathbb{E}[\beta | \mathcal{F}_{t-1}^+] = \mathbb{E}[\beta | \mathcal{F}_{t-1}]$. The Bayesian policy is then $\hat{\pi} = \{\hat{\pi}_t : t \in [T]\}$, where $\hat{\pi}_t = I\{\hat{\beta}_{t-1}^T \Delta_{x,t} \geq 0\} + 2I\{\hat{\beta}_{t-1}^T \Delta_{x,t} < 0\}$. If we define $u \in [d]$ as the location of the nonzero entry θ in

parameter vector β , then according to the distribution of β within $\mathcal{M}$, u is uniform distributed in $[d]$. Then, we have

$$\begin{aligned}
& \sup \{R_T(\pi, \pi^*) : P_{X,Y} \in \mathcal{P}_\alpha\} \\
& \geq \sup \{R_T(\hat{\pi}, \pi^*) : P_{X,Y} \in \mathcal{P}_\mathcal{M}\} \\
& \geq \sum_{t=1}^T \mathbb{E}_{\beta,u} \mathbb{E}_{X,Y|\beta} \left[|\beta^T \Delta_{x,t}| I \left\{ \text{sign}(\beta^T \Delta_{x,t}) \neq \text{sign}(\hat{\beta}_{t-1}^T \Delta_{x,t}) \right\} \right] \\
& = \sum_{t=1}^T \mathbb{E}_{\beta,u} \mathbb{E} \left[|X_{0,t} + 2\theta X_{u,t}| I \left\{ \text{sign}(\beta^T \Delta_{x,t}) \neq \text{sign}(\hat{\beta}_{t-1}^T \Delta_{x,t}) \right\} \right] \\
& \geq \sum_{t=1}^T \mathbb{E}_{\beta,u} P(X_0 = 0) \mathbb{E}_{X,Y|\beta} \left[|2\theta X_{u,t}| I \left\{ \text{sign}(\beta^T \Delta_{x,t}) \neq \text{sign}(\hat{\beta}_{t-1}^T \Delta_{x,t}) \right\} | X_0 = 0 \right] \\
& \geq \sum_{t=1}^T \mathbb{E}_{\beta,u} P(X_0 = 0) \frac{1}{4} \mathbb{E}_{X,Y|\beta} \left[|2\theta X_{u,t}| I \left\{ \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right\} | X_0 = 0, \beta^T \Delta_{x,t} \geq 0 \right] \\
& \geq \sum_{t=1}^T \mathbb{E}_{\beta,u} \frac{1}{2} C c \beta_{\min}^{\alpha+1} \mathbb{E}_{X,Y|\beta} \left[I \left\{ 1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right\} | X_0 = 0, \beta^T \Delta_{x,t} \geq 0 \right] \\
& = \sum_{t=1}^T \frac{1}{2} C c \beta_{\min}^{\alpha+1} \mathbb{E}_{\beta,u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right]. \tag{2.11}
\end{aligned}$$

The third inequality above is by the conditional expectation with respect to the event $\{X_0 = 0\}$, and the fourth inequality is by the symmetric distribution of $\beta^T \Delta_{x,t}$ when $X_0 = 0$. The last inequality is by the fact that when $X_{u,t} \in (1, 2)$, then $|\theta X_{u,t}| \geq c \beta_{\min}$.

Then, it suffices to prove the lower bound for the expectation term in the last equation, where

$$\begin{aligned}
& \mathbb{E}_{\beta,u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right] \\
& = \mathbb{E}_{\beta,u} \left[P(1 < X_{u,t} < 2) P_{X,Y|\beta} \left(\hat{\beta}_{u,t-1} X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2 \right) \right] \\
& \geq \tilde{C} P \left(\hat{\beta}_{u,t-1} X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2 \right). \tag{2.12}
\end{aligned}$$

Here $\hat{\beta}_{-u,t-1} \in \mathbb{R}^{d-1}$ is the subvector of $\hat{\beta}_{t-1}$ without the 0-th and u -th entries, and $X_{-u,t}$ is defined as

$$X_{-u,t} := (X_{1,t}, \dots, X_{u-1,t}, X_{u+1,t}, \dots, X_d)^T \in \mathbb{R}^{d-1}.$$

The probability $\tilde{C} := P(1 < X_{u,t} < 2)$ is positive, since $X_{u,t}$ is standard normal by definition.

In order to lower bound the probability in (2.12), we use the Fano's inequality to reduce the probability to the false discovery rate in a multiple hypothesis testing problem. Specifically, we define a test function $\hat{u}_{t-1} := \arg \max_{j \in [d]} |\hat{\beta}_{j,t-1}|$. We then show that the probability in (2.12) is lower bounded by $P(\hat{u}_{t-1} \neq u)$.

In fact, if event $\{\hat{u}_{t-1} \neq u\}$ holds, then we have $|\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2$ by the definition of $\hat{u}_{t-1}$. In addition, we have

$$\begin{aligned}
& P\left(\hat{\beta}_{u,t-1}X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2, |\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2\right) \\
& \geq P\left(|2\hat{\beta}_{u,t-1}| + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2, |\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2\right) \\
& = P\left(2 + N\left(0, \frac{\|\hat{\beta}_{-u,t-1}\|_2^2}{|\hat{\beta}_{u,t-1}|^2}\right) < 0 \mid 1 < X_{u,t} < 2, |\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2\right) \\
& \geq P(N(2, 1) < 0) = C'
\end{aligned} \tag{2.13}$$

The last equation is because that $X_{-u,t}$ is independent from $\hat{\beta}_{t-1}$ and $X_{u,t}$. In addition, the random variable $\hat{\beta}_{-u,t-1}^T X_{-u,t}$ conditioned on $\hat{\beta}_{t-1}$ follows normal distribution with mean 0 and variance $\|\hat{\beta}_{-u,t-1}\|_2^2$. Then, by applying the above inequalities to eq. (2.12), we have

$$\mathbb{E}_{\beta,u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right] \geq \tilde{C} C' P(\hat{u}_{t-1} \neq u). \tag{2.14}$$

It suffices to lower bound $P(\hat{u}_{t-1} \neq u)$ for each time t . We first state a variant of Fano's lower bound in multiple hypothesis testing, the details of which can be found in Chapter 15 of [22].

Lemma 1. *Assume that U is uniform on $\mathcal{U}$. For any Markov chain $U \rightarrow (X, Y) \rightarrow \hat{U}$,*

$$P(\hat{U} \neq U) \geq 1 - \frac{I(X, Y; U) + \log 2}{\log(|\mathcal{U}|)}. \tag{2.15}$$

Here $I(X, Y; U)$ is the mutual information between (X, Y) and U , and $|\mathcal{U}|$ is the cardinality of set $\mathcal{U}$.

According to the definition of $\mathcal{P}_{\mathcal{M}}$, u is uniform in $[d]$, therefore $\log(|\mathcal{U}|) = \log d$. Then, we prove an upper bound for $I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u)$ to guarantee that the probability $P(\hat{u}_{t-1} \neq u)$ in (2.14) is bounded away from 0. Here $X_{\{1:t-1\}}$ and $Y_{\{1:t-1\}}$ represent the set of feature vectors and response variables up to time $t-1$. Based on the chain rule of the mutual information, we have

$$I(X, Y; u) = I(X; u) + I(Y; u|X) = I(Y; u|X), \tag{2.16}$$

where the last equality is due to the independence between X and the uniformly distributed u . Moreover, according to Chapter 15 of [22], the conditional mutual information of Y and u conditioned on X can be

upper bounded by the Kullback-Leibler divergence, i.e.,

$$I(Y; u|X) \leq \frac{1}{|\mathcal{U}|^2} \sum_{j,k=1}^{|\mathcal{U}|} D(P_{\beta^j|X} \| P_{\beta^k|X}), \quad (2.17)$$

where $P_{\beta^j|X}$ is the distribution of Y conditioned on X and the parameter vector $\beta^j \in \mathcal{M}$ corresponding to $u^j \in \mathcal{U}$. The KL-divergence of between two normal distributions, i.e., $P_{\beta^j|X} = N(X\beta^j, \sigma^2)$ and $P_{\beta^k|X} = N(X\beta^k, \sigma^2)$, can be upper bounded as

$$\begin{aligned} & D(P_{\beta^j|X} \| P_{\beta^k|X}) \\ &= \frac{1}{2\sigma^2} \|X(\beta^j - \beta^k)\|_2^2 = \frac{1}{2\sigma^2} \|(X_j - X_k)\theta\|_2^2 \\ &\leq \frac{1}{2\sigma^2} \theta^2 4x_{\max}^2 = \frac{2x_{\min}^2}{\sigma^2} \cdot \frac{c^2 \log d}{T}. \end{aligned} \quad (2.18)$$

Summing up the mutual information up to time $t-1$, and according to eq. (2.16), we have that

$$I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u) = I(Y_{\{1:t-1\}}; u) \leq \sum_{s=1}^{t-1} \frac{2x_{\min}^2}{\sigma^2} \cdot \frac{c^2 \log d}{T} \leq \frac{2x_{\max}^2 c^2 \log d}{\sigma^2}. \quad (2.19)$$

Define constant $c := \frac{\sigma}{2x_{\max}}$. Then, if $\log d \geq 4 \log 2$, we have that

$$\begin{aligned} P(\hat{u}_{t-1} \neq u) &\geq 1 - \frac{I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u) + \log 2}{\log |\mathcal{U}|} \\ &\geq 1 - \frac{\frac{1}{2} \log d + \log 2}{\log d} \geq \frac{1}{4}. \end{aligned} \quad (2.20)$$

By applying eq. (2.20) and eq. (2.14) to eq. (2.11), it can be derived that

$$\begin{aligned} & \sup \{R_T(\pi, \pi^*) : P_{X,Y} \in \mathcal{P}_\alpha\} \\ &\geq \sum_{t=1}^T \frac{1}{2} C c \beta_{\min}^{\alpha+1} \mathbb{E}_{\beta,u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right] \\ &\geq \sum_{t=1}^T \frac{1}{2} C c \beta_{\min}^{\alpha+1} \tilde{C} C' P(\hat{u}_{t-1} \neq u) \\ &\geq \frac{\tilde{C} C' C c}{8} \left(\frac{\log d}{T} \right)^{\frac{\alpha+1}{2}} T. \end{aligned}$$

Therefore, the supreme cumulative regret incurred by any policy π can be lower bounded as

$$\sup \{R_T(\pi, \pi^*) : P_{X,Y} \in \mathcal{P}_\alpha\} \geq \begin{cases} \tilde{C}_L (\log d), & \alpha = 1 \\ \tilde{C}_L \left((\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}} \right), & \alpha \in [0, 1), \end{cases} \quad (2.21)$$

where $\tilde{C}_L$ is some positive constant.

Combining the regret lower bound $\Omega(\log T)$ from [19] on the horizon T , we prove the regret lower bound in Theorem 1.

2.5.2 Proof of Proposition 1

For consistency with Proposition 1, let X_i be the i^{th} row of X and $Y(i)$ be the i^{th} entry of Y . The sequence $\{X_i : i = 1, \dots, t\}$ forms an adapted sequence of observations, i.e. X_t may depend on the history $\{X_s, Y(s)\}_{s=1}^{t-1}$. And let $\epsilon \in \mathbb{R}^t$ be the σ -sub-Gaussian errors.

Before proving Proposition 1, we first state the following lemmas for adapted sequences.

Lemma 2. (*Bernstein Concentration*). *Let $\{D_k, \mathcal{G}_k\}_{k=1}^\infty$ be a martingale difference sequence, and suppose that D_k is σ -sub-Gaussian in an adapted sense, i.e., for all $\alpha \in \mathbb{R}$, $\mathbb{E}[e^{\alpha D_k} | \mathcal{G}_{k-1}] \leq e^{\alpha^2 \sigma^2 / 2}$ almost surely. Then, for all $t \geq 0$, $P[|\sum_{k=1}^n D_k| \geq t] \leq 2 \exp[-t^2 / 2n\sigma^2]$.*

Proof of Lemma 2 follows from Theorem 2.3 of [22] when $\alpha_* = \alpha_k = 0$ and $\nu_k = \sigma$ for all k .

Lemma 3. *Define the event*

$$\mathcal{F}(\lambda_0(\gamma)) = \{\max_{r \in [d]} (2|\epsilon^T X^{(r)}|/t) \leq \lambda_0(\gamma)\},$$

where $X^{(r)}$ is the r^{th} column of X and $\lambda_0(\gamma) = 2\sigma x_{\max} \sqrt{(\gamma^2 + 2 \log d)/t}$. Then we have $P[\mathcal{F}(\lambda_0(\gamma))] \geq 1 - 2 \exp[-\gamma^2/2]$.

Proof of Lemma 3 can be found in Lemma EC.2. of [4].

Lemma 4. *For any $\lambda_0 \in \mathbb{R}^+$, when $\lambda \geq \lambda_0$, on event $\mathcal{F}(\lambda_0)$, we have*

$$\|X(\hat{\beta} - \beta_*)\|_2^2/t \leq 3\lambda \|\hat{\beta}_S - \beta_{*,S}\|_1 - \lambda \|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1.$$

Proof. According to the definition of the LASSO estimator (2.4), we have

$$\frac{1}{2t} \|Y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1 \leq \frac{1}{2t} \|Y - X\beta_*\|_2^2 + \lambda \|\beta_*\|_1. \quad (2.22)$$

Since $\lambda \geq \lambda_0$, then if event $\mathcal{F}(\lambda_0)$ holds, we have $\lambda \geq 2\|\epsilon^T X\|_\infty/t$. Thus,

$$\begin{aligned}
\frac{1}{2t}\|X(\hat{\beta} - \beta_*)\|_2^2 &\leq \frac{1}{t}\|\epsilon^T X\|_\infty\|\hat{\beta} - \beta_*\|_1 + \lambda(\|\beta_*\|_1 - \|\hat{\beta}\|_1) \\
&\leq \frac{\lambda}{2}(\|\hat{\beta}_S - \beta_{*,S}\|_1 + \|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1) + \lambda(\|\beta_{*,S}\|_1 - \|\hat{\beta}\|_1) \\
&\leq \frac{\lambda}{2}(\|\hat{\beta}_S - \beta_{*,S}\|_1 + \|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1) \\
&\quad + \lambda(\|\hat{\beta}_S - \beta_{*,S}\|_1 - \|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1) \\
&= \frac{3}{2}\lambda\|\hat{\beta}_S - \beta_{*,S}\|_1 - \frac{\lambda}{2}\|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1.
\end{aligned} \tag{2.23}$$

□

Now we can prove Proposition 1. From Lemma 4, we have

$$\|\hat{\beta}_{S^c} - \beta_{*,S^c}\|_1 \leq 3\|\hat{\beta}_S - \beta_{*,S}\|_1. \tag{2.24}$$

Then we choose $\gamma = 2 \log t$ for event $\mathcal{F}(\lambda_0(\gamma))$. By the definition of λ , we have $\lambda \geq \lambda_0$. Thus, if both events $\mathcal{F}(\lambda_0(\gamma))$ and $\{\hat{\Sigma}(X) \in \mathcal{C}(\text{supp}(\beta_*), \phi)\}$ hold, we have

$$\begin{aligned}
\|\hat{\beta} - \beta_*\|_1^2 &\leq \frac{s_0}{\phi^2}\|X(\hat{\beta} - \beta_*)\|_2^2/t \leq \frac{s_0}{\phi^2}\left(\frac{2}{t}\|\epsilon^T X\|_\infty + 2\lambda\right)\|\hat{\beta} - \beta_*\|_1 \\
&\leq \frac{3s_0\lambda}{\phi^2}\|\hat{\beta} - \beta_*\|_1 = \frac{6s_0\sigma x_{\max}}{\phi^2}\sqrt{\frac{2 \log t + 2 \log d}{t}}\|\hat{\beta} - \beta_*\|_1.
\end{aligned} \tag{2.25}$$

Thus,

$$\begin{aligned}
&P\left[\|\hat{\beta} - \beta_*\|_1 \leq \frac{6s_0\sigma x_{\max}}{\phi^2}\sqrt{\frac{2 \log t + 2 \log d}{t}}\right] \\
&\geq P[\mathcal{F}(\lambda_0(\gamma))] - P\left[\{\hat{\Sigma}(X) \notin \mathcal{C}(\text{supp}(\beta_*), \phi)\}\right] \\
&\geq 1 - \frac{2}{t} - P\left[\{\hat{\Sigma}(X) \notin \mathcal{C}(\text{supp}(\beta_*), \phi)\}\right].
\end{aligned} \tag{2.26}$$

2.5.3 Proof of Proposition 2

Definition 4. For a constant $\alpha \geq 1$ and index set I , define the set

$$\mathbb{C}_\alpha(I) = \{\Delta \in \mathbb{R}^d : \|\Delta_{I^c}\|_1 \leq \alpha\|\Delta_I\|_1\}.$$

Our goal is to prove that with high probability, for $\forall v \in \mathbb{C}_3(S)$

$$v^T \hat{\Sigma}_t v = \frac{1}{t} \|X_t v\|_2^2 = \frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, v \rangle^2 \geq \kappa \|v\|_2^2,$$

for some constant $\kappa > 0$. To prove Proposition 2, we first state a weaker version about the compatibility condition of the sample covariance matrix.

Proposition 3. *For the adapted sequence $\{X_{\hat{a}_s, s} : s = 1, \dots, t\}$ induced by the bandit policy, the sample covariance matrix $\hat{\Sigma}_t$ are guaranteed to satisfy the compatibility condition uniformly with high probability, i.e.*

$$P\left(\hat{\Sigma}_t \in \mathcal{C}(\text{supp}(\beta_*), \frac{1}{\sqrt{32\zeta K}})\right) \geq 1 - \mathcal{O}(e^{s_0 K \log K \log d} [e^{-c_0 t} + e^{-\tilde{C} \log(K)t}]),$$

where $c_0, \tilde{C}$ are constants. In addition, we can derive a uniform bound for the compatibility condition over t that exceeds a certain threshold (i.e. $t \geq \mathcal{O}(\log T + s_0 K \log K \log d)$) with high probability, i.e.

$$P\left(\forall t \geq T', \hat{\Sigma}_t \in \mathcal{C}(\text{supp}(\beta_*), \frac{1}{\sqrt{32\zeta K}})\right) \geq 1 - \frac{2}{T},$$

where $T' = \mathcal{O}(\log T + s_0 K \log K \log d)$.

We first provide the outline of the proof for Proposition 3.

- (i) Discretize the unit sphere $S^{d-1} \in \mathbb{R}^d$, and show eigenvalue condition of $\hat{\Sigma}_t$ on a finite set $\mathcal{N}_0$.
- (ii) Show eigenvalue condition for all m -sparse vectors within sphere S^{d-1} .
- (iii) Transfer eigenvalue condition of $\hat{\Sigma}_t$ to vectors in $\mathbb{C}_3(S)$, which implies compatibility condition.

In the following, we state a result from [16], which is used for step (iii) in the proof.

Lemma 5 (Transfer Principle). *Suppose $\hat{\Sigma}_t$ and Σ are matrix with non-negative diagonal entries, and assume $\eta \in (0, 1)$, $m \in \{1, \dots, d\}$ are such that*

$$\forall v \in \mathbb{R}^d \text{ with } \|v\|_0 \leq m, v^T \hat{\Sigma}_t v \geq (1 - \eta) v^T \Sigma v. \quad (2.27)$$

Assume D is a diagonal matrix whose elements $D_{j,j}$ are non-negative and satisfy $D_{j,j} \geq [\hat{\Sigma}_t]_{j,j} - (1 - \eta) \Sigma_{j,j}$. Then

$$\forall x \in \mathbb{R}^d, x^T \hat{\Sigma}_t x \geq (1 - \eta) x^T \Sigma x - \frac{\|D^{1/2} x\|_1^2}{m - 1}. \quad (2.28)$$

The proof of Lemma 5 can be found in Lemma 5.1 of [16].

According to Lemma 5, it suffices to prove the eigenvalue condition (2.27) for m -sparse vectors with sufficient large $m = Cs_0$, where constant C will be specified later. To prove this, we first introduce ϵ -net on unit sphere for approximation.

Definition 5. (*Nets, covering numbers*). Let (X, d) be a metric space and let $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ of X is called an ϵ -net of X if every point $x \in X$ can be approximated to within ϵ by some point $y \in \mathcal{N}_\epsilon$, i.e. so that $d(x, y) \leq \epsilon$. The minimal cardinality of an ϵ -net of X , if finite, is denoted $\mathcal{N}(X, \epsilon)$ and is called the covering number of X (at scale of ϵ).

The following provides a bound for the cardinality of the ϵ -net that can approximate the points on the unit Euclidean sphere within range of ϵ .

Lemma 6. (*Covering numbers of the sphere*). The unit Euclidean sphere S^{d-1} equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that

$$\mathcal{N}(S^{d-1}, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^d.$$

Denote $E_J = \text{span}\{e_j : j \in J\}$, then we define $\mathcal{N}_0$ as the ϵ -net for $S^{d-1} \cap (\cup_{|J|=Cs_0} E_J)$. For each subset $J \subseteq [d]$ with $|J| = Cs_0$, the set $S^{d-1} \cap E_J$ can be viewed as a unit sphere in $\mathbb{R}^{Cs_0}$. Then according to Lemma 6, we have a bound for the covering number N of $\mathcal{N}_0$ when $\epsilon \leq 1$.

$$\begin{aligned} \log N &= \log \left(\binom{d}{Cs_0} \left(1 + \frac{2}{\epsilon}\right)^{Cs_0} \right) \\ &\leq \log \left(\left(\frac{ed}{Cs_0}\right)^{Cs_0} \left(\frac{3}{\epsilon}\right)^{Cs_0} \right) \\ &\lesssim Cs_0 [\log d + \log(\frac{1}{\epsilon})]. \end{aligned} \tag{2.29}$$

According to Assumption (b), when $h > 0$ is small, i.e. $h = \mathcal{O}(\frac{1}{K})$, we have for $\forall v \in \mathbb{R}^d$ that:

$$P(\min_a \langle X_{a,t}, v \rangle^2 \leq h \|v\|_2^2) \leq \zeta Kh.$$

From now on we fix a sparse vector v with $\|v\|_2 = 1$ and $\|v\|_0 \leq Cs_0$. Define

$$N_t = \sum_{s=1}^t \mathbb{1}_{\{\min_a \langle X_{a,s}, v \rangle^2 \leq h \|v\|_2^2\}}$$

Then N_t is the sum of i.i.d. indicator random variables which have expectation greater than $(1 - \zeta Kh)$. Now

we define $S_t = \sum_{s=1}^t Z_s$ where

$$Z_s = \mathbb{1}_{\{\min_a \langle X_{a,s}, v \rangle^2 \leq h \|v\|_2^2\}} - \mathbb{E}[\mathbb{1}_{\{\min_a \langle X_{a,s}, v \rangle^2 \leq h \|v\|_2^2\}}],$$

are centered random variables. Then by Hoeffding's Lemma, we have for $\forall t \in \mathbb{N}_+$ and $\delta \in \mathbb{R}$,

$$P\left(\frac{1}{t}S_t \leq \delta\right) \leq \exp(-2\delta^2 t).$$

Let $\delta = -\frac{1}{2}(1 - \zeta Kh)$, we have

$$P\left(\frac{1}{t}N_t \leq \frac{1}{2}(1 - \zeta Kh)\right) \leq \exp(-\frac{1}{2}(1 - \zeta Kh)^2 t) = \exp(-c_0 t)$$

where $c_0 = \frac{1}{2}(1 - \zeta Kh)^2$. Then taking the union of the event over all vectors in $\mathcal{N}_0$, we can obtain

$$\begin{aligned} & P\left(\forall v \in \mathcal{N}_0 : \frac{1}{t} \sum_{s=1}^t \langle X_{a_s, s}, v \rangle^2 \leq \frac{1}{2}h(1 - \zeta Kh)\right) \\ & \leq N \exp(-c_0 t) \leq \exp(-c_0 t + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon})). \end{aligned} \quad (2.30)$$

In the above inequalities, we have proved the minimum eigenvalue condition for all vectors in ϵ -net $\mathcal{N}_0$. In the following, we will show that the restricted eigenvalue condition for $C s_0$ -sparse vectors can also be implied by the eigenvalue condition on $\mathcal{N}_0$. Before that, we state a result on the spectral norm of symmetric matrices.

Lemma 7 (Computing the spectral norm on a net). *Let A be a symmetric $d \times d$ matrix, and let $\mathcal{N}_\epsilon$ be an ϵ -net of S^{d-1} for some $\epsilon \in [0, 1)$. The spectral norm of A can be computed via the associated quadratic form $\|A\| = \sup_{x \in S^{d-1}} |\langle Ax, x \rangle|$. Then*

$$\|A\| = \sup_{x \in S^{d-1}} |\langle Ax, x \rangle| \leq (1 - 2\epsilon)^{-1} \sup_{x \in \mathcal{N}_\epsilon} |\langle Ax, x \rangle|.$$

The proof of Lemma 7 can be found in Lemma 5.4 of [30].

According to the definition of $\mathcal{N}_0$, for any $C s_0$ -sparse vector v and symmetric $d \times d$ matrix A , there is a vector $u \in \mathcal{N}_0$, such that $\|u - v\|_2 \leq \epsilon$ and $\text{supp}(v) \subseteq \text{supp}(u)$. Then

$$\begin{aligned} v^T A v &= u^T A u + (u - v)^T A (u - v) + 2u^T A (v - u) \\ &\geq u^T A u - 2\|u - v\|_2 \|A\|_{op} \|u\|_2, \end{aligned} \quad (2.31)$$

where $\|A\|_{op}$ is the spectral norm of A constrained on sparse vectors with ℓ_0 -norm not greater than $C s_0$, i.e.

$$\|A\|_{op} = \max_{v \in S^{d-1}, \|v\|_0 \leq Cs_0} \|Av\|_2.$$

We have proved a uniform lower bound of $u^T \hat{\Sigma}_t u$ for $u \in \mathcal{N}_0$ and $t \geq T_1$. Then to prove a lower bound of $v^T \hat{\Sigma}_t v$ for $v \in S^{d-1}$ and $\|v\|_0 \leq Cs_0$, based on (2.31) and the fact that there exist $u \in \mathcal{N}_0$ such that $\|u - v\|_2 \leq \epsilon$, it suffices to show that $\|\hat{\Sigma}_t\|_{op}$ is upper bounded. And by Lemma 7, we only need to bound the spectral norm of $\hat{\Sigma}_t$ on $\mathcal{N}_0$.

Here we have for a fixed vector $v \in \mathcal{N}_0$ and $\forall t \in \mathbb{N}_+$,

$$\frac{1}{t} \sum_{s=1}^t \langle X_{a,s}, v \rangle^2 \leq \frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2.$$

Define $Z_s^a = \langle X_{a,s}, v \rangle^2 - v^T \Sigma_a v$, $S_t^a = \sum_{s=1}^t Z_s^a$ and $S_t = \sum_{a=1}^K S_t^a$ for $a \in [K]$. Then Z_s^a is centered sub-exponential random variables. Before proving the upper bound of $\|\hat{\Sigma}_t\|_{op}$, we will state some properties of sub-exponential random variables. Firstly, we introduce some parameters of sub-gaussian and sub-exponential random variables.

Definition 6. *The sub-gaussian norm of a sub-gaussian random variable X , denoted $\|X\|_{\psi_2}$, is defined as*

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

Similarly, the sub-exponential norm of a sub-exponential random variable X , denoted $\|X\|_{\psi_1}$, is defined as

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}|X|^p)^{1/p}.$$

The details of the definition and related properties can be found in Section 5.2.3 and Section 5.2.4 in [30]. The following is a relationship between sub-gaussian random variables and sub-exponential random variables.

Lemma 8. *(Sub-exponential is sub-gaussian squared). A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.$$

Proof of Lemma 8 can be found in Lemma 5.14 of [30].

Lemma 9. *(MGF of sub-exponential random variables). Let X be a centered sub-exponential random variable. Then for t such that $|t| \leq c_1/\|X\|_{\psi_1}$, one has*

$$\mathbb{E} \exp(tX) \leq \exp(c_2 t^2 \|X\|_{\psi_1}^2),$$

where $c_1, c_2 > 0$ are absolute constants.

Proof of Lemma 9 can be found in Lemma 5.15 of [30].

By the result in Lemma 8, we have for sub-exponential random variable Z_s^a ,

$$\kappa_a = \|Z_s^a\|_{\psi_1} \leq 2\|\langle X_{a,s}, v \rangle\|_{\psi_1} \leq 4\|\langle X_{a,s}, v \rangle\|_{\psi_2}^2 \leq 4\lambda_{\max}^2(\Sigma_a).$$

Then for any $\mu > 0$ and $\delta > 0$, we have

$$\begin{aligned} P\left(\frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta\right) &= P\left(\sum_{s=1}^t [\max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2] \geq \delta t\right) \\ &\leq e^{-\mu \delta t} \mathbb{E}[\exp(\mu \sum_{s=1}^t [\max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2])] \\ &= e^{-\mu \delta t} \prod_{s=1}^t \mathbb{E}[\exp(\mu [\max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2])] \end{aligned} \quad (2.32)$$

According to Lemma 9, for μ such that $|\mu| \leq \frac{c_1}{\max_a \kappa_a}$, we have

$$\begin{aligned} \mathbb{E}[\exp(\mu [\max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2])] &\leq \sum_{a=1}^K \mathbb{E}[\exp\{\mu (\langle X_{a,s}, v \rangle^2 - \mathbb{E}[\langle X_{a,s}, v \rangle^2])\}] \\ &\leq \sum_{a=1}^K e^{c_2 \mu^2 \kappa_a^2} \leq e^{c_2 \mu^2 \max_a \kappa_a^2 + \log(K)}. \end{aligned}$$

Applying the above inequality to inequality (2.32), we have for μ such that $|\mu| \leq \frac{c_1}{\max_a \kappa_a}$,

$$P\left(\frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta\right) \leq \exp(-\mu \delta t + t c_2 \mu^2 \max_a \kappa_a^2 + t \log K). \quad (2.33)$$

The right hand side of the above inequality achieves the minimum value at $\mu = \frac{\delta}{2c_2 \max_a \kappa_a^2}$, where the minimum value is $\exp(-\frac{\delta^2 t}{4c_2 \max_a \kappa_a^2} + t \log K)$.

Now if $\frac{\delta}{2c_2 \max_a \kappa_a^2} > \frac{c_1}{\max_a \kappa_a}$, the right hand side of the equation (2.33) obtain the minimum value at $\mu = \frac{c_1}{\max_a \kappa_a}$ and

$$\begin{aligned} &P\left(\frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta\right) \\ &\leq \exp\left(-\frac{c_1 \delta t}{\max_a \kappa_a} + t c_2 \frac{c_1^2}{\max_a \kappa_a^2} \max_a \kappa_a^2 + t \log K\right). \end{aligned}$$

Moreover, if $\frac{\delta}{2c_2 \max_a \kappa_a^2} > \frac{c_1}{\max_a \kappa_a}$, we apply this to the above inequality and obtain

$$P\left(\frac{1}{t} \sum_{s=1}^t \max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta\right) \leq \exp\left(-\frac{c_1 \delta t}{2 \max_a \kappa_a} + t \log K\right).$$

Now we have for $\forall \delta > 0$ and $t \in \mathbb{N}_+$

$$\begin{aligned} & P\left(\frac{1}{t} \sum_{s=1}^t \max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta\right) \\ & \leq \exp\left(-\min\left\{\frac{\delta^2}{4c_2 \max_a \kappa_a^2}, \frac{c_1 \delta}{2 \max_a \kappa_a}\right\}t + t \log K\right). \end{aligned}$$

If we set $\delta = c_3 \log(K) \Lambda_0^2$ where c_3 is a sufficient large positive constant, then we have for a fixed vector $v \in \mathcal{N}_0$

$$P\left(\frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta\right) \leq \exp\left(-\tilde{C} \log(K)t\right),$$

where $\tilde{C} = \min\left\{\frac{c_3^4 \log K}{64c_2}, \frac{c_1 c_3}{8}\right\} - 1 \leq \min\left\{\frac{c_3^2 \Lambda_0^4 \log K}{4c_2 \max_a \kappa_a^2}, \frac{c_1 c_3 \Lambda_0^2}{2 \max_a \kappa_a}\right\} - 1$ is a positive constant independent on K and t . Taking the union of the probability over all vectors in $\mathcal{N}_0$, we have

$$\begin{aligned} & P\left(\forall v \in \mathcal{N}_0 : \langle v, \hat{\Sigma}_t v \rangle \geq \tilde{c} \log(K) \Lambda_0^2\right) \\ & \leq N \exp\left(-\tilde{C} \log(K)t\right) \\ & \lesssim \exp\left(-\tilde{C} \log(K)t + Cs_0 \log d + Cs_0 \log\left(\frac{1}{\epsilon}\right)\right). \end{aligned} \tag{2.34}$$

Now we provide the proof of Proposition 3.

Proof. We define events $\mathcal{G}_1$ and $\mathcal{G}_2$ as

$$\begin{aligned} \mathcal{G}_1 &= \left\{ \forall u \in \mathcal{N}_0 : \frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, u \rangle \geq \frac{1}{2} h(1 - \zeta K h) \right\}; \\ \mathcal{G}_2 &= \left\{ \forall u \in \mathcal{N}_0 : \langle u, \hat{\Sigma}_t u \rangle \leq \tilde{c} \log(K) \Lambda_0^2 \right\}. \end{aligned}$$

If event $\mathcal{G}_2$ holds, then according to Lemma 7 and definition of $\|\cdot\|_{op}$ in (2.31), we have

$$\|\hat{\Sigma}_t\|_{op} \leq (1 - 2\epsilon)^{-1} \max_{v \in \mathcal{N}_0} \langle v, \hat{\Sigma}_t v \rangle \leq (1 - 2\epsilon)^{-1} \tilde{c} \log(K) \Lambda_0^2. \tag{2.35}$$

If both event $\mathcal{G}_1$ and $\mathcal{G}_2$ hold, based on (2.31), we can prove the restricted eigenvalue condition for all

Cs_0 -sparse vectors in unit sphere S^{d-1} , i.e. for $\forall v \in S^{d-1}$ and $\|v\|_0 \leq Cs_0$ that

$$\begin{aligned}\langle v, \hat{\Sigma}_t v \rangle &\geq \frac{1}{2}h(1 - \zeta Kh) - 2\epsilon \|\hat{\Sigma}_t\|_{op} \\ &\geq \frac{1}{2}h(1 - \zeta Kh) - \frac{2\epsilon}{1 - 2\epsilon} \tilde{c} \log(K) \Lambda_0^2.\end{aligned}\tag{2.36}$$

Now if we set $h = \frac{1}{2\zeta K} = \mathcal{O}(\frac{1}{K})$ and $\epsilon = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\log(K)\Lambda_0^2}\}$, then for $\forall v \in S^{d-1}$ and $\|v\|_0 \leq Cs_0$, we have

$$\langle v, \hat{\Sigma}_t v \rangle \geq \frac{1}{4}h(1 - \zeta Kh) = \frac{1}{16\zeta K}.$$

Now we proved part (i) and (ii) in Proposition 3, i.e. the minimum eigenvalue condition for all Cs_0 -sparse vectors in S^{d-1} . To finished the proof, we need to apply Lemma 5 to prove part (iii).

To apply Lemma 5, we set $\Sigma = \frac{1}{8\zeta K} I_{d \times d}$ as a $d \times d$ diagonal matrix, $m = Cs_0$ and $\eta = \frac{1}{2}$. Moreover, we set $D_{j,j} = [\hat{\Sigma}_t]_{j,j}$ to satisfy the condition on diagonal matrix D . By greedy method for constructing $\mathcal{N}_0$, we can let $e_j \in \mathcal{N}_0$ for $\forall j \in [d]$, where e_j is unit vector in $\mathbb{R}^d$ having exactly one entry equal to 1 and 0 otherwise. Then under event $\mathcal{G}_2$, we have

$$D_{j,j} = [\hat{\Sigma}_t]_{j,j} = \langle e_j, \hat{\Sigma}_t e_j \rangle \leq \tilde{c} \log(K) \Lambda_0^2.$$

Now according to (2.28), if both events $\mathcal{G}_1$ and $\mathcal{G}_2$ hold, we have for $\forall x \in S^{d-1}$ and $x \in \mathbb{C}_3(S)$ that

$$\begin{aligned}x^T \hat{\Sigma}_t x &\geq \frac{1}{2}x^T \Sigma x - \frac{\|D^{1/2}x\|_1^2}{Cs_0 - 1} \\ &\geq \frac{1}{16\zeta K} - \frac{\tilde{c} \log(K) \Lambda_0^2 \|x\|_1^2}{Cs_0 - 1} \\ &\geq \frac{1}{16\zeta K} - \frac{\tilde{c} \log(K) \Lambda_0^2 \cdot 16s_0}{Cs_0 - 1},\end{aligned}\tag{2.37}$$

where the last inequality above is because for $x \in S^{d-1}$ and $x \in \mathbb{C}_3(S)$, we have

$$\|x\|_1 \leq (1 + 3)\|x_S\|_1 \leq 4\sqrt{s_0}\|x_S\|_2 \leq 4\sqrt{s_0}.$$

Then, if we set $C = 2 \cdot 16^2 \tilde{c} \log(K) \Lambda_0^2 \zeta K + \frac{1}{s_0} = \Theta(\zeta \Lambda_0^2 K \log K)$, we can have $x^T \hat{\Sigma}_t x \geq \frac{1}{32\zeta K}$ for unit vector $x \in \mathbb{C}_3(S)$, which implies compatibility condition.

Now, we prove the uniform bound on the compatibility condition in Proposition 3 by showing that events $\mathcal{G}_1$ and $\mathcal{G}_2$ hold with high probability.

At first, for any $T_1 \in \mathbb{N}_+$ we have

$$\begin{aligned}
& P \left(\forall t \geq T_1, \forall v \in \mathcal{N}_0 : \frac{1}{t} \sum_{s=1}^t \langle X_{a_s, s}, v \rangle^2 \leq \frac{1}{2} h(1 - \zeta K h) \right) \\
& \leq N \sum_{t=T_1}^{\infty} e^{-c_0 t} = \frac{1}{1 - e^{-c_0}} \exp(-c_0 T_1 + \log N) \\
& \lesssim \exp(-c_0 T_1 + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon})) = \delta_2,
\end{aligned}$$

where $c_0 = \frac{1}{2}(1 - \zeta K h)^2 = \frac{1}{8}$. If we set $\delta_2 = \frac{1}{T}$, then with the choice of $\epsilon = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\log(K)\Lambda_0^2}\}$ and C from above, we have

$$\begin{aligned}
T_1 &= \mathcal{O}(\log T + C s_0 [\log d + \log(\frac{1}{\epsilon})]) \\
&= \mathcal{O}(\log T + s_0 \zeta \Lambda_0^2 K (\log K) \log d + \log(\log K) + \log \Lambda_0^2) \\
&= \mathcal{O}(\log T + s_0 K (\log K) \log d).
\end{aligned} \tag{2.38}$$

The above inequality provides a uniform lower bound for the eigenvalue of sample covariance matrix $\hat{\Sigma}_t$ over $\mathcal{N}_0$ for $t \geq T_1$.

Moreover, we can obtain an union probability bound for all vectors in $\mathcal{N}_0$ and some $T_2 \in \mathbb{N}_+$ in (2.34).

$$\begin{aligned}
& P \left(\forall t \geq T_2, \forall v \in \mathcal{N}_0 : \langle v, \hat{\Sigma}_t v \rangle \geq \tilde{c} \log(K) \Lambda_0^2 \right) \\
& \leq N \sum_{t=T_2}^{\infty} \exp(-\tilde{C} \log(K) t) \\
& \lesssim \exp \left(-T_2 \tilde{C} \log(K) + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon}) \right) = \delta_3.
\end{aligned}$$

If we set $\delta_3 = \frac{1}{T}$, then we have

$$T_2 = \frac{\log(T) + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon})}{\tilde{C} \log(K)} = \mathcal{O} \left(\log T + C s_0 \log d + C s_0 \log(\frac{1}{\epsilon}) \right).$$

Now, we can show that compatibility condition holds on $\mathbb{C}_3(S)$ uniformly for all $t \geq \max(T_1, T_2) = \mathcal{O}(\log T + s_0 K \log K \log d)$ with high probability, i.e.

$$P \left(\forall t \geq \max(T_1, T_2), v \in \mathbb{C}_3(S) : v^T \hat{\Sigma}_t v \geq \frac{1}{32\zeta K} \right) \geq 1 - \delta_2 - \delta_3 = 1 - \frac{2}{T}.$$

□

With the crude compatibility condition from Proposition 3, we can prove Proposition 2. Firstly, we prove that the optimal arm will be pulled a positive fraction of time by the ℓ_1 -confidence ball based algorithm after some time point.

Lemma 10. *Suppose we construct the confidence set $\mathcal{C}_t$ in Proposition 2 with $\phi = \frac{1}{\sqrt{32\zeta K}}$, then when time horizon t exceeds a certain threshold (i.e. $t \geq \mathcal{O}(\log T + s_0 K \log K \log d + \frac{s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$), decision-makers will pull the optimal arms a positive fraction of time.*

Proof. Before the proof, we first define the event in Proposition 1 with $\phi^2 = \frac{1}{32\zeta K}$,

$$\mathcal{G}_{3,t+1} = \left\{ \|\hat{\beta}_t - \beta_*\|_1 \leq \frac{6s_0\sigma x_{\max}}{1/(32\zeta K)} \sqrt{\frac{2\log t + 2\log d}{t}} \right\}.$$

According to Proposition 3, $\mathcal{G}_{3,t}$ will hold uniformly over $t \geq T'$ with high probability.

From Assumption 2(c), we have event Γ_t such that $P(\Gamma_t) \geq \frac{1}{2}$. Then with sufficient large $\phi \geq \frac{1}{\sqrt{32\zeta K}}$ in Proposition 2, if both Γ_t and $\mathcal{G}_{3,t}$ hold, we can only select the arm $b \in [K]$ instead of the optimal arm a_t^* when

$$\begin{aligned} \max_{\beta \in \mathcal{C}_t} \langle X_{b,t}, \beta \rangle &\geq \max_{\beta \in \mathcal{C}_t} \langle X_{a_t^*,t}, \beta \rangle \\ \implies \langle X_{b,t}, \beta_* \rangle + 2\|X_{b,t}\|_\infty \frac{6s_0\sigma x_{\max}}{1/(32\zeta K)} \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}} &\geq \langle X_{a_t^*,t}, \beta_* \rangle. \end{aligned}$$

The second inequality is because for all $\beta \in \mathcal{C}_t$, by triangle inequality we have

$$\begin{aligned} \|\beta - \beta_*\|_1 &\leq \|\beta - \hat{\beta}_t\|_1 + \|\hat{\beta}_t - \beta_*\|_1 \\ &\leq \left(\frac{1}{\phi^2} + 32\zeta K\right)(6s_0\sigma x_{\max}) \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}} \\ &\leq 384\zeta K s_0\sigma x_{\max} \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}} \end{aligned}$$

Then we will always select the optimal arm a_t^* if

$$\begin{aligned} \Delta_* &\geq 384s_0\sigma x_{\max}^2 \zeta K \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}} \\ &\geq 2\|X_{b,t}\|_\infty \frac{6s_0\sigma x_{\max}}{1/(32\zeta K)} \sqrt{\frac{2\log(t-1) + 2\log d}{t-1}}. \end{aligned}$$

By solving the above inequality, we have that for $t \geq \mathcal{O}(\frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$, we will always select the optimal

arm a_t^* when both Γ_t and $\mathcal{G}_{3,t}$ hold. $\square$

Proof. (Proposition 2) Now we consider time such that $t \geq T'' = \mathcal{O}(\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$.

From the above proof, we know that we will select the optimal arm with high probability.

Suppose $t \geq 2T''$ and $v \in \mathbb{R}^d$, then we have

$$\begin{aligned} \frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, v \rangle^2 &\geq \frac{1}{t} \sum_{s=T''+1}^t \langle X_{\hat{a}_s, s}, v \rangle^2 \\ &\geq \frac{1}{t} \sum_{s=T''+1}^t \langle X_{\hat{a}_s, s}, v \rangle^2 \mathbb{1}_{\{\Gamma_s \cap \mathcal{G}_{3,t}\}} \\ &= \frac{1}{t} \sum_{s=T''+1}^t \langle X_{a_s^*, s}, v \rangle^2 \mathbb{1}_{\{\Gamma_s\}} \mathbb{1}_{\{\mathcal{G}_{3,t}\}} \end{aligned} \quad (2.39)$$

Since event $\mathcal{G}_{3,t}$ only depends on the history up to time $t-1$, the random variables $\{X_{a_s^*, t} | \mathcal{G}_{3,t}\}$ are i.i.d. Moreover, since $\{\mathbb{1}_{\{\Gamma_s\}} : s \geq T''\}$ are i.i.d Bernoulli random variables, then according the Hoeffding's Lemma, we have

$$\sum_{s=T''+1}^t \mathbb{1}_{\{\Gamma_s\}} \leq \frac{1}{2} \cdot \frac{1}{2} (t - T'') \quad \text{w.p.} \leq \mathcal{O}(e^{-c_4 t}), \quad (2.40)$$

where c_4 is some constant.

Now we consider the sum of $\langle X_{a_s^*, s}, v \rangle^2$ over times when event Γ_t holds

$$\sum_{s: \Gamma_s \text{ holds}} \langle X_{a_s^*, s}, v \rangle^2 - \mathbb{E}[\langle X_{a_s^*, s}, v \rangle^2 | \Gamma_t].$$

Define $Z_s = \langle X_{a_s^*, s}, v \rangle^2 - \mathbb{E}[\langle X_{a_s^*, s}, v \rangle^2 | \Gamma_t]$, then Z_s are i.i.d. sub-exponential random variables. Moreover, $\mathbb{E}[\langle X_{a_s^*, s}, v \rangle^2 | \Gamma_t] \geq \phi_0^2 \|v\|_2^2$ according to Assumption (c), and from Lemma 8, we have condition on event Γ_t

$$\|Z_s\|_{\psi_1} \leq 2\|\langle X_{a_s^*, s}, v \rangle^2\|_{\psi_1} \leq 4\|\langle X_{a_s^*, s}, v \rangle\|_{\psi_2}^2 = 4v^T \mathbb{E}[X_{a_t^*} X_{a_t^*}^T | \Gamma_t] v \leq 4\Lambda_1^2.$$

Then applying Bernstein inequality for sub-exponential random variables, we can have a lower bound for a fixed vector v and $\epsilon = \frac{\phi_0^2}{2}$.

$$\begin{aligned} &P\left(\left|\sum_{s: \Gamma_s \text{ holds}} Z_s\right| \geq \epsilon m \mid \sum_{s=T''+1}^t \mathbb{1}_{\{\Gamma_s\}} = m\right) \\ &\leq 2 \exp\left(-\tilde{c}_5 \min\left(\frac{\epsilon^2}{\|Z_s\|_{\psi_1}^2}, \frac{\epsilon}{\|Z_s\|_{\psi_1}}\right) m\right) \\ &= 2 \exp\left(-\tilde{c}_5 \min\left(\frac{\phi_0^4}{4\|Z_s\|_{\psi_1}^2}, \frac{\phi_0^2}{2\|Z_s\|_{\psi_1}}\right) m\right) = \mathcal{O}(e^{-c_5 m}), \end{aligned} \quad (2.41)$$

where $\tilde{c}_5, c_5$ are constants.

Now combining inequalities (2.39)–(2.41), we have for any v in ϵ' -net $\mathcal{N}_1$ for $S^{d-1} \cap (\cup_{|J|=C's} E_J)$, where ϵ and C' need to be specified, and $t \geq 2T''$ that

$$\frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, v \rangle^2 \geq \frac{1}{t} \sum_{s=T''+1}^t \langle X_{a_s^*, s}, v \rangle^2 \mathbb{1}_{\{\Gamma_s\}} \geq \frac{1}{t} \frac{\phi_0^2}{2} \cdot \frac{(t - T'')}{2} \|v\|_2^2 \geq \frac{\phi_0^2}{16},$$

with probability greater than $1 - \mathcal{O}(e^{-c_4(t-T'')} + e^{-c_5 \cdot \frac{1}{4}(t-T'')})$.

Similar as before, we can extend the above inequality to all vectors in $\mathbb{C}_3(S)$ by using the approximation of $\mathcal{N}_1$, where $C = \mathcal{O}(\log K)$ and $\epsilon = \mathcal{O}(\frac{1}{\log K})$. Then for any $v \in \mathbb{C}_3(S)$ and $t \geq 2T''$, we have

$$\hat{\Sigma}_t \in \mathcal{C}(\text{supp}(\beta_*), \frac{\phi_0}{8})$$

with probability greater than $1 - \mathcal{O}(e^{s_0 \log K \log d} [e^{-c_4(t-T'')} + e^{-c_5 \cdot \frac{1}{4}(t-T'')}] = 1 - \mathcal{O}(e^{-t+s_0 \log K \log d})$.

Then combine Proposition 1 on $\mathcal{G}_{3,t}$ and the above inequality, we can prove the Proposition 2. $\square$

2.5.4 Proof of Theorem 2

Define T'' as the threshold in Proposition 2, then we divide the cumulative regret into three groups:

- (a) Initialization when $t \leq T'' = \mathcal{O}(\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2})$.
- (b) Times $t > T''$ when $\hat{\Sigma}_t \notin \mathcal{C}(\text{supp}(\beta_*), \phi_*)$, where $\phi_* = \frac{\phi_0}{8}$.
- (c) Times $t > T''$ when $\hat{\Sigma}_t \in \mathcal{C}(\text{supp}(\beta_*), \phi_*)$, where $\phi_* = \frac{\phi_0}{8}$.

The cumulative regret J_1 from time periods in group (a) at time T is bounded by at most $2bx_{\max}T'' = \mathcal{O}(2bx_{\max}[\log T + s_0 K \log K \log d + \frac{\sigma^2 x_{\max}^4 s_0^2 \zeta^2 K^2 \log d}{\Delta_*^2}])$.

According to Proposition 2, we can bound the cumulative regret in group (b) at time T .

$$\begin{aligned} J_2 &= \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{\hat{\Sigma}_t \notin \mathcal{C}(\text{supp}(\beta_*), \phi_*)\}}] \leq \sum_{t=T''}^T 2bx_{\max} P(\hat{\Sigma}_t \notin \mathcal{C}(\text{supp}(\beta_*), \phi_*)) \\ &\lesssim \sum_{t=T''}^T 2bx_{\max} \left[\frac{1}{t} + e^{-t+s_0 \log k \log d} \right] = \mathcal{O}(x_{\max} b \log T). \end{aligned}$$

To prove the bound of the cumulative regret in group (c), we first define events:

$$\Gamma_{\xi_t} = \{\omega \in \Omega : \langle X_{a_t^*, t}, \theta_* \rangle \geq \langle X_{b, t}, \theta_* \rangle + \xi_t, \forall b \neq a_t^*\}.$$

In addition, we define $A_t := \{\hat{\Sigma}_t \in \mathcal{C}(\text{supp}(\beta_*), \phi_*)\}$ for simplicity. Then the cumulative regret at time T in group (c) can be written as:

$$J_3 = \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}\}}] + \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}].$$

Now we set $\xi_t = 2x_{\max}\tau_{t-1} = 2x_{\max}\tau_0 \sqrt{\frac{\log d + \log(t-1)}{t-1}}$ with $\tau_0 = \frac{384\sqrt{2}s_0\sigma x_{\max}}{\phi_0^2}$, then when both Γ_{ξ_t} and A_t hold, we have for any $k \neq a_t^*$ that

$$\begin{aligned} & \max_{\beta \in \mathcal{C}_t} \langle X_{a_t^*, t}, \beta \rangle - \max_{\beta \in \mathcal{C}_t} \langle X_{k, t}, \beta \rangle \\ & \geq \langle X_{a_t^*, t}, \beta_* \rangle - 2x_{\max}\tau_{t-1} - \langle X_{k, t}, \beta_* \rangle \geq 0. \end{aligned}$$

So we will always select the optimal arm a_t^* under the event $A_t \cap \Gamma_{\xi_t}$, and the first part of the cumulative regret in J_3 will be zero.

Now we consider the time when both A_t and $\Gamma_{\xi_t}^c$ hold. Then if a sub-optimal arm $k \neq a_t^*$ is selected, the regret incurred will be

$$\langle X_{a_t^*, t} - X_{k, t}, \beta_* \rangle \leq \max_{\beta \in \mathcal{C}_t} \langle X_{a_t^*, t}, \beta \rangle - \max_{\beta \in \mathcal{C}_t} \langle X_{k, t}, \beta \rangle + 2x_{\max}\tau_{t-1} \leq 2x_{\max}\tau_{t-1} = \xi_t.$$

Moreover, from Assumption 1(b), we have

$$P\left(\langle X_{a_t^*, t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b, t}, \beta_* \rangle \leq \xi_t\right) \leq \frac{1}{2} \left(\frac{\xi_t}{\Delta_*}\right)^\alpha.$$

(i) For the $\alpha \in [0, 1]$ case, we can bound the second term in J_3 as

$$\begin{aligned} & \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] \leq \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{\text{select arm } i \neq s_t^*\}}] \\ & \leq \sum_{t=T''}^T \xi_t P\left(\langle X_{a_t^*, t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b, t}, \beta_* \rangle \leq \xi_t\right) \\ & \leq \sum_{t=T''}^T \frac{1}{2\Delta_*^\alpha} \xi_t^{\alpha+1} \\ & \leq \sum_{t=1}^T \frac{2^\alpha (x_{\max}\tau_0)^{\alpha+1}}{\Delta_*^\alpha} \left(\frac{\log d + \log t}{t}\right)^{\frac{\alpha+1}{2}} \\ & = \begin{cases} C_2 \frac{s_0^{\alpha+1} \sigma^{\alpha+1} x_{\max}^{2(\alpha+1)}}{\Delta_*^\alpha \phi_0^{2(\alpha+1)}} (\log d)^{\frac{\alpha+1}{2}} T^{\frac{1-\alpha}{2}}, & \text{when } \alpha \in [0, 1), \\ C_2 \frac{s_0^2 \sigma^2 x_{\max}^4}{\Delta_* \phi_0^4} [\log d + \log T] \log T, & \text{when } \alpha = 1, \end{cases} \end{aligned}$$

(ii) For the $\alpha \in (1, +\infty)$ case, the second term in J_3 can be bounded as

$$\begin{aligned}
& \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] \\
& \leq \sum_{t=T''}^T \xi_t P\left(\langle X_{a_t^*, t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b, t}, \beta_* \rangle \leq \xi_t\right) \\
& \leq \sum_{t=T''}^T \xi_t \cdot \min\left(1, \frac{\xi_t^\alpha}{2\Delta_*^\alpha}\right).
\end{aligned}$$

Moreover, according to the definition of ξ_t , we have that

$$\frac{\xi_t^\alpha}{2\Delta_*^\alpha} \geq 1 \iff t \lesssim T''' := \frac{4^{1-1/\alpha} x_{\max}^2 \tau_0^2 \log d}{\Delta_*^2}.$$

Then, the above summation can be decomposed into two parts:

$$\begin{aligned}
& \sum_{t=T''}^T \xi_t \cdot \min\left(1, \frac{\xi_t^\alpha}{2\Delta_*^\alpha}\right) \\
& \leq \sum_{t=T''}^{T'''} \xi_t + \sum_{t=T'''}^T \frac{1}{2\Delta_*^\alpha} \xi_t^{\alpha+1} \\
& \leq x_{\max} b T''' + \frac{2^\alpha (x_{\max} \tau_0)^{\alpha+1}}{\Delta_*^\alpha} \sum_{t=T'''}^T \left(\frac{\log d + \log t}{t}\right)^{\frac{\alpha+1}{2}}.
\end{aligned} \tag{2.42}$$

Here for $t \geq T'''$, we have that

$$\frac{\log d + \log t}{t} \leq \frac{\log d + \log T'''}{T'''} = \mathcal{O}\left(\frac{\Delta_*^2}{4^{1-1/\alpha} x_{\max}^2 \tau_0^2}\right).$$

We define the constant $M := \frac{\Delta_*^2}{4^{1-1/\alpha} x_{\max}^2 \tau_0^2}$. Then, the second term in the right-hand side of inequality (2.42)

can be bounded by

$$\begin{aligned}
& \sum_{t=T'''}^T \left(\frac{\log d + \log t}{t} \right)^{\frac{\alpha+1}{2}} \\
& \leq \sum_{t=1}^T \left(\frac{\log d + \log t}{t} \right)^{\frac{\alpha+1}{2}} \mathbb{1}_{\left\{ \frac{\log d + \log t}{t} \leq M \right\}} \\
& \lesssim \sum_{t=1}^T \left(\frac{\log d}{t} \right)^{\frac{\alpha+1}{2}} \mathbb{1}_{\left\{ \frac{\log d}{t} \leq M \right\}} \\
& \lesssim \int_1^T \left(\frac{\log d}{t} \right)^{\frac{\alpha+1}{2}} \mathbb{1}_{\left\{ \frac{\log d}{t} \leq M \right\}} dt \\
& = \int_{\log d/T}^{\min(\log d, M)} \log d \cdot u^{\frac{\alpha+1}{2}-2} du \\
& \leq \frac{2}{\alpha-1} \log d \cdot M^{\frac{\alpha-1}{2}} \\
& = \frac{2\Delta_*^{\alpha-1} \log d}{(\alpha-1)4^{(\alpha-1)^2/2\alpha} x_{\max}^{\alpha-1} \tau_0^{\alpha-1}}.
\end{aligned}$$

The equality in the above is due to the change of variable by taking $u = \frac{\log d}{t}$. Therefore, the second term in J_3 can be bounded for $\alpha \in (1, +\infty)$ as

$$\begin{aligned}
& \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] \\
& \leq x_{\max} b T''' + \frac{2^\alpha (x_{\max} \tau_0)^{\alpha+1}}{\Delta_*^\alpha} \cdot \frac{2\Delta_*^{\alpha-1} \log d}{(\alpha-1)4^{(\alpha-1)^2/2\alpha} x_{\max}^{\alpha-1} \tau_0^{\alpha-1}} \\
& \leq x_{\max} b T''' + \frac{2^{3-1/\alpha} x_{\max}^2 \tau_0^2 \log d}{(\alpha-1)\Delta_*} \\
& \leq C_2 \left[\frac{bs_0^2 \sigma^2 x_{\max}^5 \log d}{\Delta_*^2 \phi_0^4} + \frac{s_0^2 \sigma^2 x_{\max}^4 \log d}{(\alpha-1)\Delta_* \phi_0^4} \right].
\end{aligned}$$

(iii) For the $\alpha = +\infty$ case, Assumption 1(b) implies that for $\xi_t < \Delta_*$, the following inequality holds.

$$P \left(\langle X_{a_t^*, t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b, t}, \beta_* \rangle \leq \xi_t \right) = 0.$$

Therefore, the second term in J_3 can be bounded as

$$\begin{aligned}
\sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] &\leq \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{\text{select arm } i \neq s_t^*\}}] \\
&\leq \sum_{t=T''}^T \xi_t P\left(\langle X_{a_t^*, t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b, t}, \beta_* \rangle \leq \xi_t\right) \\
&\leq \sum_{t=T''}^T \xi_t \mathbb{1}_{\{\xi_t \geq \Delta_*\}} \\
&\stackrel{(a)}{\leq} \sum_{t=T''}^T \frac{\xi_t^{\gamma+1}}{\Delta_*^\gamma} \\
&= \sum_{t=T''}^T \frac{2^{\gamma+1} (x_{\max} \tau_0)^{\gamma+1}}{\Delta_*^\gamma} \left(\frac{\log d + \log t}{t} \right)^{\frac{\gamma+1}{2}}
\end{aligned}$$

where constant $\gamma \geq 0$. The inequality (a) above is due to $\xi_t/\Delta_* \geq 1$. By letting $\gamma \rightarrow 1_+$, we can obtain an upper bound for the $\alpha = \infty$ case, i.e.,

$$\sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] \leq C_2 \frac{s_0^2 \sigma^2 x_{\max}^4 \log d}{\Delta_* \phi_0^4}.$$

By summing the cumulative regret in three groups, we can obtain the upper bound for the total expected cumulative regret of our ℓ_1 -confidence ball based method up to time T .

Chapter 3

High-dimensional Linear Bandits with Variable Selection

3.1 Introduction

The increasing availability of massive personalized datasets provides opportunities for decision-makers to tailor decisions specifically suited to individuals. For example, doctors can personalize treatments for patients based on the electronic medical records and demographics; web recommendation systems can provide advertisements according to customer profiles and browsing histories. In these situations, decision-makers need to adapt their decisions to the sequentially collected data and choose the best action to maximize their cumulative reward over a specified time period, say from round 1 to T . A major challenge with this problem is the need to balance the “exploration” of new actions and the “exploitation” of the best action based on the available information. This exploration versus exploitation trade-off in sequential decision-making can be formulated as a contextual bandit problem [1], [3], where the agent can assess action-specific contextual information described as d -dimensional feature vectors and select among K actions (arms) at each of the T rounds to maximize the cumulative reward. When the expected reward and the associated feature vector of each arm are linearly related, this problem is known as the linear bandit problem, or the multi-armed bandit problem with linear payoff functions [1], [2]. Several algorithms to solve the multi-armed bandit problem have been proposed in the literature [3], [12], [13]. In this chapter, we focus on the setting where the dimension d of the feature vector can be comparable, or even much larger than the total number of rounds T , which plays the role of “sample size” from a statistical perspective. Due to the curse of dimensionality, it is common to assume the sparsity structure on the data generating model that the response, or the reward, depends

only on a small subset of the features. The major contributions made in this work can be summarized in the following.

Methodologically, we propose a novel and general framework for designing bandit algorithms that incorporates variable selection within the Upper Confidence Bound (UCB) type algorithms [2], [14]. The UCB technique has been extensively studied in recent years [3], [13], [27], [31], where the algorithm constructs a confidence set (interval) for the true parameter as the accompanied searching space for candidate parameters when optimizing over actions in order to promote exploration based on the current estimation uncertainty. However, the constructed confidence sets in the existing UCB algorithms do not possess sparse structure, so the size of the confidence set will be too large to honestly reflect the estimation uncertainty in high-dimensional settings. As a consequence, the cumulative regret (difference between the best achievable and the attained cumulative rewards) of the UCB algorithms will be sub-optimal, since the cumulative regret can be roughly characterized by the sum of the size of the confidence sets, which is discussed in [13] and [3]. We also provide some intuitive descriptions in Section 3.3.1. Our proposed algorithm makes use of high-dimensional variable selection techniques to construct a refined confidence set having sparsity structure. We introduce a set of desirable properties (Definition 8) such that any variable selection procedure satisfying these properties, when combined with our algorithm, will enjoy the same theoretical guarantee in terms of an upper bound to the cumulative regret. In a nutshell, these properties require the variable selection procedure to: 1. the probability of including any unimportant feature variables should diminish as data accumulates; 2. gradually identify all important feature variables as data accumulates according to their relative importance in terms of a properly scaled signal-to-noise ratio. Equipped with an eligible variable selection procedure, we propose to use a sparse ellipsoid as the confidence set for the true parameter vector, which is restricted in the low-dimensional subspace corresponding to the currently selected important variables. In addition, we show that variable selection procedures induced by commonly used regularized estimation methods such as MCP [6] and SCAD [7] are instances meeting our requirement. We prefer the use of nonconvex penalties such as MCP [6] and SCAD [7] over LASSO [15] in our algorithm since the former possesses variable selection consistency properties under mild conditions on the design matrix [32]–[34] while the latter requires more stringent ones such as the incoherence condition [35]. Computationally, our proposed algorithm is more efficient to implement than existing UCB algorithms with non-sparse confidence sets, since the sparsity in our confidence set also restricts the searching space in the accompanied optimization problem for selecting the best arms.

Our theoretical results show that the proposed algorithm achieves an improved regret upper bound due to the utilization of sparsity in constructing the confidence sets. Under some common assumptions made in the literature [4], [5], [31], we prove that our proposed algorithm achieves a cumulative regret that scales as $\mathcal{O}(\log^2 T + \log d)$, which decouples the time horizon T and dimension d and improves the $\mathcal{O}((\log T + \log d)^2)$

regret from Bastani and Bayati [4] and the $\mathcal{O}(\log d \log T)$ regret from Wang, Wei, and Yao [5], where the interaction term of order $\mathcal{O}(\log d \log T)$ can be substantially large. Specifically, we show that any reasonably good variable selection procedure that is able to identify relatively important variables helps the construction of a sparse confidence set that possesses good properties for approximating the true parameter vector (Theorem 3). In consequence, our proposed algorithm has a finer control of the regret incurred during the transition phase from the initial period (high-dimensional regime) to later period (low-dimensional regime), which ensures that the impact of dimension d on the regret does not deteriorate as T increases and leads to an improved cumulative regret. We also present a matching regret lower bound for stochastic linear bandit problems, which demonstrates the hardness of the problem and effectiveness of our proposed algorithm. From our numerical study, we also find that the use of variable selection can substantially improve the empirical performance in both synthetic and real data analysis.

Our proofs towards our theoretical findings are also highly non-trivial and worth highlighting. Technically, we introduce a novel regret analysis strategy for analyzing the proposed bandit algorithm. Majority of the existing studies on high-dimensional statistics are limited to the setting of independent data which do not easily generalize to bandit problems due to the data dependence arising from sequential decision making. In particular, due to the exploitation nature of the bandit problem, the feature vectors of the selected arms tend to concentrate around a small region in the feature space, leading to an ill-conditioned design matrix. In the conventional bandit literature, this ill-conditioning issue is usually resolved through forced sampling steps [4], [5], which requires fine tuning and can be costly and wasteful in practical scenarios such as medical and marketing applications. In comparison, we introduce a novel technique to prove the *restricted eigenvalue condition* for the sample covariance matrix under dependent data based on a non-asymptotic analysis. Our analysis shows that randomness of feature vectors (characterized by Assumption 4) facilitates exploration in the feature space and therefore prevents the design matrix from collapsing into certain directions (without the need of forced sampling).

The remainder of the chapter is organized as follows. In Section 3.2, we briefly introduce the necessary background knowledge and describe the problem formulation. In Section 3.3, we first motivate and introduce a set of desirable properties for a good variable selection procedure to meet, and then propose the general UCB algorithm with variable selection (Algorithm 2). After that, we describe the UCB algorithm using variable selection criterion induced by regularized estimation procedures (Algorithm 3). In Section 3.4, we present our main theoretical results regarding the proposed algorithms. Specifically, in Section 3.4.1, we discuss the main assumptions required for the general UCB algorithm with variable selection; moreover, we prove a regret upper bound (Theorem 4) as well as a nearly matching regret lower bound for any policy. In Section 3.4.2, we introduce the additional assumptions for the nonconvex regularized methods, demonstrate

that these desirable properties (Corollary 2) are satisfied by the nonconvex regularized methods, and prove a regret upper bound for the corresponding algorithm (Theorem 8). Empirical results on simulation and real data analysis are presented in Section 3.5. The chapter is concluded with a discussion in Section 3.6. Detailed proofs of theorems and other technical results are deferred to appendices.

3.2 Problem Formulation

We now provide a formulation for the linear bandit problem. We start by introducing some notations which will be used throughout the chapter.

Notations. For a vector $\beta \in \mathbb{R}^d$, define its support set to be $\text{supp}(\beta) = \{j : \beta_j \neq 0\}$. We use β_V to denote both a vector in $\mathbb{R}^{|V|}$ containing the components of β corresponding to the subset $V \subseteq \{1, \dots, d\}$ and a vector in $\mathbb{R}^d$ with $\text{supp}(\beta_V) = V$, where $(\beta_V)_j = \beta_j$ for $j \in V$ and zero otherwise, but this will be clear depending on the context. For a matrix A , we use A_V to denote the submatrix of A restricted to the columns within the subset $V \subseteq \{1, \dots, d\}$, and use $\|A\|_1, \|A\|_2$ to denote the matrix ℓ_1 norm and spectral norm, respectively. For function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we write ∇f and $\nabla^2 f$ to denote the gradient/subgradient and second-order gradient/subgradient of f respectively, if they exist. We use $\|x\|_p$ to denote the ℓ_p -norm of a vector $x \in \mathbb{R}^d$. For a set A in $\mathbb{R}^d$, the diameter of the set with ℓ_p -norm is defined as $\text{diam}_p(A) = \sup\{\|x - y\|_p : x, y \in A\}$. The cardinality of a set $I \subseteq \{1, \dots, K\}$ is denoted as $|I|$.

Let T be the total time number of steps that is allowed to be unknown, and K be the number of possible actions (arms). At each round $t \in [T]$, a feature vector $\mathbf{X}_{a,t} \in \mathbb{R}^d$ associated with each arm $a \in [K]$ is presented. The decision-maker has access to the action set $\mathcal{D}_t = \{\mathbf{X}_{a,t}\}_{a=1}^K$ consisting of features from all arms, which are allowed to have dependence, and selects one out of K arms to maximize his/her expected reward.

In our setup, each arm yields a random reward, the expected value of which follows a linear function of the associated feature variables. All arms share the same unknown parameter $\beta^* \in \mathbb{R}^d$ in the linear reward function $\langle \cdot, \beta^* \rangle$. If the decision-maker pulls arm $a \in [K]$ at time t , then the random reward is observed as $\langle \mathbf{X}_{a,t}, \beta^* \rangle + \omega_{a,t}$. We adopt a standard random-design framework by assuming the observed sequence of features $\{\mathbf{X}_{a,t}\}_{t \geq 0}$ to be marginally i.i.d. copies of a sub-Gaussian random vector from $\mathbb{P}_a$ for each arm $a \in [K]$, while errors $\omega_{a,t}$ are independent σ -sub-Gaussian random variables (see Definition 7 below), which are also independent of the sequence $\{\mathbf{X}_{a,t}\}_{t \geq 0}$. We consider the high-dimensional regime where the dimension d of the feature space can be comparable or even much larger than the time horizon T . Despite the high-dimensional appearance, we assume the true parameter vector β^* to admit a sparse structure, namely, the number of nonzero entries in β^* is much smaller than the ambient dimension d .

We denote $S^* = \{j \in [d] : \beta_j^* \neq 0\}$ as the unknown index set of the nonzero entries of β^* that collects all important feature components in predicting the reward. Let $s_0 = |S^*|$ be the number of the nonzero entries in β^* , which satisfies $s_0 \ll d$. For each $t \in [T]$, let $\mathbf{X}_t \in \mathbb{R}^{t \times d}$ be the time t -observed design matrix, where each row corresponds to the feature vector $\mathbf{X}_{\hat{a}_s, s}$ of the selected arm $\hat{a}_s$ for $s = 1, \dots, t$, and vector $\mathbf{Y}_t = (Y_1, \dots, Y_t)^T \in \mathbb{R}^t$ collects the observed rewards up to time t .

Definition 7 (Sub-Gaussian random variable). *A real-valued random variable z is σ -sub-Gaussian if $\mathbb{E}[e^{tz}] \leq e^{\sigma^2 t^2/2}$ for every $t \in \mathbb{R}$.*

In particular, this definition implies $\mathbb{E}[z] = 0$ and $\text{Var}[z] \leq \sigma^2$. Our objective is to design a sequential decision-making policy π that learns the parameter β^* over time in order to maximize the expected cumulative reward over the time horizon. Let $\hat{a}_t \in [K]$ denote the arm chosen by the policy at time $t \in [T]$. To characterize the quality of a policy π , we compare it with the oracle policy π^* that uses the knowledge of β^* to choose the best arm $a_t^* = \arg \max_a \langle \mathbf{X}_{a,t}, \beta^* \rangle$ which maximizes the expected reward at each round $t \in [T]$. Note that under our random-design assumption, a_t^* is also random. To summarize, any policy will incur an expected regret at each time $t \in [T]$

$$r_t = \mathbb{E} \left[\max_{a \in [K]} \langle \mathbf{X}_{a,t}, \beta^* \rangle - \langle \mathbf{X}_{\hat{a}_t, t}, \beta^* \rangle \right], \quad (3.1)$$

where the expectation is taken with respect to the randomness in the feature vector $\{\mathbf{X}_{a,t}\}_{a=1}^K$ and the stochastic reward through $\{\omega_{a,t}\}_{a=1}^K$. Our objective is then equivalent to seeking a policy that minimizes the expected cumulative regret $R_T = \sum_{t=1}^T r_t$.

3.3 UCB Algorithm with Variable Selection

In this section, we describe our proposed UCB (Upper Confidence Bound) algorithm with variable selection for the linear bandit problem with high-dimensional features. In Section 3.3.1, we first provide the motivation of incorporating variable selection into bandit algorithms and then describe some desirable properties on the variable selection procedure that will improve the regret. In Section 3.3.2, we formally introduce the proposed UCB algorithm with any generic variable selection procedure with the desirable properties. In Section 3.3.3, we present an instance of variable selection procedure by regularized estimation and describe the corresponding bandit algorithm.

3.3.1 Variable Selection for Linear Bandit

Before motivating the importance of variable selection in high-dimensional bandit problems, we first describe the general framework of *optimism in the face of uncertainty* (OFU) introduced in [2] and [13] for solving

bandit problems. In a typical OFU-based algorithm, a confidence set $\mathcal{C}_t$ is constructed at the beginning of round t using feature vectors from the previously selected arms, i.e., $\mathbf{X}_{\hat{a}_{1,1}}, \mathbf{X}_{\hat{a}_{2,2}} \dots, \mathbf{X}_{\hat{a}_{t-1,t-1}}$, and the observed rewards $Y_1, Y_2, \dots, Y_{t-1}$ up to time $(t-1)$. In a nutshell, the confidence set $\mathcal{C}_t$ is constructed so that it will contain true parameter β^* with high probability. With the constructed confidence set $\mathcal{C}_t$, the OFU-based algorithm chooses the action-estimate pair,

$$(\mathbf{X}_{\hat{a}_t,t}, \tilde{\beta}_{\hat{a}_t,t}) = \arg \max_{(\mathbf{x}, \beta) \in \mathcal{D}_t \times \mathcal{C}_t} \langle \mathbf{x}, \beta \rangle, \quad (3.2)$$

by jointly maximizing the expected reward over feature $\mathbf{x}$ and parameter β within the feasible set $\mathcal{D}_t \times \mathcal{C}_t$. Heuristically, the instant regret $\tilde{r}_t$ incurred by the algorithm at time t can be roughly characterized by the product of the maximum norm of the feature vectors in $\mathcal{D}_t$ and the diameter of $\mathcal{C}_t$ if β^* belongs to $\mathcal{C}_t$, that is,

$$\begin{aligned} \tilde{r}_t &:= \langle \mathbf{X}_{a_t^*,t}, \beta^* \rangle - \langle \mathbf{X}_{\hat{a}_t,t}, \beta^* \rangle \stackrel{(i)}{\leq} \langle \mathbf{X}_{a_t^*,t}, \beta^* - \tilde{\beta}_{a_t^*,t} \rangle - \langle \mathbf{X}_{\hat{a}_t,t}, \beta^* - \tilde{\beta}_{\hat{a}_t,t} \rangle \\ &\stackrel{(ii)}{\leq} \|\mathbf{X}_{a_t^*,t}\|_p \cdot \|\beta^* - \tilde{\beta}_{a_t^*,t}\|_q + \|\mathbf{X}_{\hat{a}_t,t}\|_p \cdot \|\beta^* - \tilde{\beta}_{\hat{a}_t,t}\|_q \stackrel{(iii)}{\leq} 2 \sup\{\|\mathbf{x}\|_p : \mathbf{x} \in \mathcal{D}_t\} \times \text{diam}_q(\mathcal{C}_t), \end{aligned}$$

where step (i) follows by the optimality of solution $\mathbf{X}_{\hat{a}_t,t}$ given $\tilde{\beta}_{\hat{a}_t,t}$ in problem (3.2) so that $\langle \mathbf{X}_{a_t^*,t}, \tilde{\beta}_{a_t^*,t} \rangle \leq \langle \mathbf{X}_{\hat{a}_t,t}, \tilde{\beta}_{\hat{a}_t,t} \rangle$, step (ii) follows by applying the Hölder inequality, and step (iii) is because the feasibility of $\tilde{\beta}_{\hat{a}_t,t}$ and the pre-assumption $\beta^* \in \mathcal{C}_t$ together would imply $\max\{\|\beta^* - \tilde{\beta}_{a_t^*,t}\|_q, \|\beta^* - \tilde{\beta}_{\hat{a}_t,t}\|_q\} \leq \text{diam}_q(\mathcal{C}_t)$. Here, the ℓ_p -norm is conjugate to the ℓ_q -norm, or $\frac{1}{p} + \frac{1}{q} = 1$. This simple regret analysis helps explain the motivation, but tighter regret bounds can be obtained by more refined analysis, as in the proofs of our theoretical results (c.f. Section 3.4). Based on the preceding display, the cumulative regret $\tilde{R}_T$ of the algorithm can be upper bounded by

$$\tilde{R}_T = \sum_{t=1}^T \tilde{r}_t \leq 2 \sum_{t=1}^T \sup\{\|\mathbf{x}\|_p : \mathbf{x} \in \mathcal{D}_t\} \times \text{diam}_q(\mathcal{C}_t) \lesssim \sum_{t=1}^T \text{diam}_q(\mathcal{C}_t).$$

From this simple regret analysis, we can observe that the cumulative regret of the OFU-based algorithm is conceptually proportional to the sum of the diameters of $\mathcal{C}_t$'s. Therefore, reducing the size of confidence set $\mathcal{C}_t$ while maintaining $\beta^* \in \mathcal{C}_t$ will help improve the cumulative regret performance of the bandit algorithm.

The OFU-based algorithm of [13] is based on an ellipsoid confidence set $\mathcal{C}_t$ centered around the ridge estimator of β^* . The diameter, $\text{diam}_2(\mathcal{C}_t)$, of $\mathcal{C}_t$ scales as $\mathcal{O}(\sqrt{d/t})$ in dimension d and time t , which makes the cumulative regret to scale as $\mathcal{O}(\sqrt{dT})$ in the total time horizon T . [31] proposed a UCB-type algorithm with an ℓ_1 -confidence set $\mathcal{C}_t$ as an ℓ_1 -ball centered around the LASSO estimator of β^* whose ℓ_1 -diameter $\text{diam}_1(\mathcal{C}_t)$ scales as $\mathcal{O}(\sqrt{\log d/t})$. Under some mild conditions on the separation between optimal and sub-optimal

arms and the underlying random design distribution, [31] show that if the true parameter β^* is sparse, then the cumulative regret incurred by their algorithm is $\mathcal{O}(\log d \log T)$, which only scales logarithmically in d . However, since the confidence set $\mathcal{C}_t$ from [31] is still a subset with full dimension d in the feature space $\mathbb{R}^d$, the dependence of the diameter $\text{diam}_1(\mathcal{C}_t)$ on dimension d cannot be eliminated. Due to the coupling between feature dimensionality and statistical accuracy inherited from an ℓ_1 -confidence set whose diameter should be at least $\mathcal{O}(\sqrt{\log d/t})$ in order to cover β^* , the cumulative regret $\mathcal{O}(\log d \log T)$ of the algorithm from [31] cannot be avoided. Theoretically, this cumulative regret, as a product between $\log d$ and $\log T$, is strictly sub-optimal compared to our additive regret minimax lower bound $\log T + \log d$ established in Theorem 5.

The large volume of a non-sparse confidence set and the resulting sub-optimal regret motivates us to consider confidence sets with sparse structures. In particular, we propose a confidence set $\mathcal{C}_t$ restricted in a low-dimensional subspace of $\mathbb{R}^d$, where the subspace is adaptively chosen by variable selection procedures. Intuitively, if the support S^* of the true parameter β^* is known, the original bandit problem reduces to a low-dimensional problem and the constructed confidence set $\mathcal{C}_t^*$ lies in the subspace corresponding to S^* . Then, the diameter of $\mathcal{C}_t^*$ should only depend on the sparsity level s_0 instead of feature dimension d . Through variable selection, if we can select a set of important variables and construct a sparse confidence set based on the selected variables, it might be possible to achieve an optimal regret bound. This indicates the importance of doing variable selection in high-dimensional bandit problems.

In this chapter, we introduce a sequential selection consistency property (see Definition 8 below) for a variable selection procedure $\mathcal{V}$ to be adequate for bandit problems. In a nutshell, variable selection procedure $\mathcal{V}$ is required to sequentially identify only entries of β^* with relatively large magnitude, where the magnitude threshold decreases in time t so that eventually all signals will be identified. Because of the sequential selection property, the estimation error for the reward of each arm due to those variables not selected by $\mathcal{V}$ decreases in time and only contributes to the “burn-in” term in the cumulative regret.

The precise definition of sequential selection consistency for bandit problems is given as follows. Let $\mathcal{F}_t = \sigma(\mathbf{X}_{\hat{a}_1,1}, Y_1, \mathbf{X}_{\hat{a}_2,2}, Y_2, \dots, \mathbf{X}_{\hat{a}_t,t}, Y_t)$ denote the σ -algebra of history up to time t and $2^{[d]}$ is the power set of $[d]$.

Definition 8 ((M_0, t_0, M_1, M_2) -Sequential Selection Consistency). *Let $\mathcal{V}_t : \mathcal{F}_t \rightarrow 2^{[d]}$ and $\pi_t : \mathcal{F}_t \rightarrow [d]$ denote the index set selected by variable selection procedure $\mathcal{V}$ and the arm chosen by bandit policy π at time t , respectively. We say that the selection-policy pair $(\mathcal{V}, \pi)$ is (M_0, t_0, M_1, M_2) -sequentially selection consistent if the followings hold,*

- (a) *At each time t that is sufficiently large, the probability that the variable selection procedure incorrectly*

selects a zero entry of β^* is bounded as follows:

$$P(\{\exists j \in [d] : \beta_j^* = 0 \text{ and } j \in \mathcal{V}_t(\mathcal{F}_t)\}) \leq M_0/t, \text{ for all } t > t_0,$$

where t_0 is a large enough time point.

(b) For a given $\varepsilon > 0$, define $I_\varepsilon = \{i \in [d] : |\beta_i^*| > \varepsilon\}$, the index set of coefficients having magnitude larger than ε . For any $\varepsilon > 0$, there exists a time point $t_\varepsilon = \max\left(\frac{M_1(d)}{\varepsilon^2}, M_2(d)\right)$ such that for time points $t \geq t_\varepsilon$, the probability of selecting all the variables of I_ε is large as follows:

$$P(\{I_\varepsilon \subseteq \mathcal{V}_t(\mathcal{F}_t)\}) \geq 1 - M_0/t.$$

In Definition 8, M_0 is a constant that does not depend on dimension d or horizon T , while constants M_1 and M_2 are functions of dimension d . The form of $M_1(d)$ depends on the distribution of error $\omega_{a,t}$ in the reward. For instance, if the error $\omega_{a,t}$ is sub-Gaussian as assumed in this paper, the constant $M_1(d)$ scales as $\mathcal{O}(\log d)$. The form of $M_2(d)$ depends on the variable selection procedure. For example, $M_2(d)$ scales as $\mathcal{O}(\log d)$ for our proposed algorithm with regularized estimation methods for variable selection, which will be introduced in Section 3.3.3. The $\mathcal{O}(\varepsilon^2)$ scaling of t_ε is because the typical sample complexity of achieving accuracy ε scales as $\mathcal{O}(\varepsilon^2)$ where t_ε can be viewed as the effective sample size in the bandit setting.

According to condition (a) in Definition 8, a desirable variable selection procedure in bandit settings is less likely to select zero entries of β^* when t is sufficiently large. In condition (b), the entries of β^* with magnitude greater than ε will be selected after observing t_ε samples with a high probability. As the constant ε decreases to zero, the variable selection procedure $\mathcal{V}$ will eventually select all nonzero entries of β^* . Definition 8 allows the possibility of incorrect selection of variables with probability converging to zero at a rate inversely proportional to the number of observed samples.

In the next subsection, we will describe the general UCB algorithm with variable selection procedure. After that, we introduce the regularized methods as instances of variable selection procedures and use MCP and SCAD estimators as examples for illustration.

3.3.2 Framework of the Proposed UCB Algorithm

In this subsection, we provide details of the proposed algorithm based on a generic sequential selection consistent procedure $\mathcal{V}$. At the beginning of each round $(t + 1)$, we have observations from the previous t steps, where $\mathbf{Y}_t = \mathbf{X}_t \beta^* + \boldsymbol{\omega}_t$ with design matrix $\mathbf{X}_t \in \mathbb{R}^{t \times d}$, response vector $\mathbf{Y}_t \in \mathbb{R}^t$, and error vector $\boldsymbol{\omega}_t \in \mathbb{R}^t$. The algorithm consists of following main steps:

- (i) At time $(t + 1)$, apply variable selection procedure $\mathcal{V}$ to obtain a subset $\hat{S}_t = \mathcal{V}_t(\mathcal{F}_t)$ as an estimator of S^* .
- (ii) Given the subset $\hat{S}_t$, we solve the constraint least squares optimization problem, i.e.,

$$\hat{\beta}_t^{ols} \in \arg \min_{\beta \in \mathbb{R}^d, \text{supp}(\beta) \subseteq \hat{S}_t} \mathcal{L}_t(t) = \frac{1}{2t} \|\mathbf{Y}_t - \mathbf{X}_t \beta\|_2^2. \quad (3.3)$$

Here the support of the solution $\hat{\beta}_t^{ols} \in \mathbb{R}^d$ is within $\hat{S}_t$.

- (iii) Construct a confidence set $\mathcal{C}_{t+1}$ for the true parameter $\beta_{\hat{S}_t}^*$ centered around $\hat{\beta}_t^{ols}$ as follows:

$$\mathcal{C}_{t+1} = \left\{ \beta \in \mathbb{R}^d : \|\beta - \hat{\beta}_t^{ols}\|_{\hat{\Sigma}_t} \leq \tau_t, \text{supp}(\beta) \subseteq \hat{S}_t \right\}, \quad (3.4)$$

where $\tau_t = \tau_0 \sqrt{|\hat{S}_t|(\log |\hat{S}_t| + \log t)/t}$ is the diameter of $\mathcal{C}_{t+1}$, and τ_0 is an initial parameter. Here $\hat{\Sigma}_t = \mathbf{X}_t^T \mathbf{X}_t / t$ (see in Section 3.4.1) is the sample covariance matrix up to time t , and $\|\beta - \hat{\beta}_t^{ols}\|_{\hat{\Sigma}_t}$ is the weighted ℓ_2 -norm of the vector $\beta - \hat{\beta}_t^{ols}$, where $\|\mathbf{u}\|_{\hat{\Sigma}_t}^2 := \mathbf{u}^T \hat{\Sigma}_t \mathbf{u}$ for any $\mathbf{u} \in \mathbb{R}^d$.

- (iv) Apply the confidence set $\mathcal{C}_{t+1}$ to the optimization problem (3.2), i.e.,

$$(\mathbf{X}_{\hat{a}_{t+1}, t+1}, \tilde{\beta}_{\hat{a}_{t+1}, t+1}) = \arg \max_{(\mathbf{x}, \beta) \in \mathcal{D}_{t+1} \times \mathcal{C}_{t+1}} \langle \mathbf{x}, \beta \rangle,$$

and obtain the optimal arm $\hat{a}_{t+1}$ at time $t + 1$.

Through steps (i) and (ii), we obtain an estimation $\hat{\beta}_t^{ols}$ for the subvector $\beta_{\hat{S}_t}^*$ of the true parameter β^* , where subset $\hat{S}_t$ is an approximation of the support S^* of β^* . In step (iii), the constructed confidence set $\mathcal{C}_t$ is centered around $\hat{\beta}_t^{ols}$ and restricted in the subspace corresponding to $\hat{S}_t$. Therefore, the diameter τ_t of $\mathcal{C}_t$ only depends on $|\hat{S}_t|$ instead of dimension d . In step (iv), the optimization problem (3.2) reduces to a low-dimensional problem due to the sparse structure of $\mathcal{C}_t$, which also facilitate scalable computation.

3.3.3 UCB Algorithm with Regularized Estimation

In this subsection, we describe how we utilize regularized estimation methods as special instances of variable selection procedure which satisfy our desirable properties in Definition 8. There has been extensive literature [32]–[34] on support recovery in high dimensional linear regression based on regularized methods such as LASSO [15], MCP [6] and SCAD [7]. In particular, regularized estimators using nonconvex penalties, including the MCP and SCAD estimators, have the appealing theoretical property that the selected support tends to agree with the target support S^* of the true parameter β^* with high probability under milder condition on the

Algorithm 2 UCB Algorithm with Variable Selection (UCB-VS)

Require: Initial diameter τ_0 for the confidence set $\mathcal{C}_t$.

- 1: Initialize $\hat{\beta}_0^{ols}$ as 0 in $\mathbb{R}^d$.
 - 2: **for** $t \in [T]$ **do**
 - 3: Observe $\mathcal{D}_t = \{\mathbf{X}_{a,t}\}_{a=1}^K$ and $\mathcal{F}_{t-1}$.
 - 4: Obtain the subset $\hat{S}_{t-1} = \mathcal{V}_{t-1}(\mathcal{F}_{t-1})$ by applying variable selection.
 - 5: Obtain $\hat{\beta}_{t-1, \hat{S}_{t-1}}^{ols}$ by solving equation (3.3), i.e., $\hat{\beta}_{t-1}^{ols} \in \arg \min_{\beta \in \mathbb{R}^d, \text{supp}(\beta) \subseteq \hat{S}_{t-1}} \mathcal{L}_{t-1}(\beta)$.
 - 6: Construct ellipsoid confidence set $\mathcal{C}_t$:
 - 7: $\mathcal{C}_t = \{\beta \in \mathbb{R}^d : \|\beta - \hat{\beta}_{t-1}^{ols}\|_{\hat{\Sigma}_{t-1}} \leq \tau_{t-1}, \text{supp}(\beta) \subseteq \hat{S}_{t-1}\}$,
 - 8: where $\hat{\Sigma}_{t-1} = (\mathbf{X}_{t-1}^T \mathbf{X}_{t-1} / (t-1))$ is the sample covariance matrix up to time t .
 - 9: Select optimal arm $\hat{a}_t$ and corresponding feature vector $\mathbf{X}_{\hat{a}_t, t}$ from equation (3.2), and observe reward $Y_t = \langle \mathbf{X}_{\hat{a}_t, t}, \beta^* \rangle + \omega_t$.
 - 10: Set $\tau_t = \tau_0 \sqrt{\frac{|\hat{S}_t|(\log |\hat{S}_t| + \log t)}{t}}$.
 - 11: **end for**
-

design matrix [33], [34] than those using convex penalties. Therefore, we also adopt nonconvex regularization methods for sequential variable selection in our bandit setting, since our theoretical results in Section 3.4 show that they will satisfy the desired sequential selection consistency of Definition 8 without the need of imposing overly restrictive conditions on the model.

Recall that at the beginning of each round $(t+1)$, we have observations from a linear model $\mathbf{Y}_t = \mathbf{X}_t \beta^* + \omega_t$, where $\mathbf{X}_t \in \mathbb{R}^{t \times d}$, $\mathbf{Y}_t \in \mathbb{R}^t$ and $\omega_t \in \mathbb{R}^t$. We consider the following generic form of a regularized estimator as the solution of the optimization problem

$$\hat{\beta}_t \in \arg \min_{\|\beta\|_1 \leq R} \{\mathcal{L}_t(\beta) + \rho_\lambda(\beta)\}, \quad (3.5)$$

where $\mathcal{L}_t$ denotes the empirical loss function at round t and ρ_λ denotes a regularization function to be described momentarily. In our regression framework, the loss function is defined as the square loss, i.e.,

$$\mathcal{L}_t(\beta) = \frac{1}{2t} \|\mathbf{Y}_t - \mathbf{X}_t \beta\|_2^2. \quad (3.6)$$

Moreover, ρ_λ is chosen as a coordinate-wise regularization function, meaning that ρ_λ can be expressed as:

$$\rho_\lambda(\beta) = \sum_{j=1}^d \rho_\lambda(\beta_j). \quad (3.7)$$

We note here that the functions $\rho_\lambda : \mathbb{R} \rightarrow \mathbb{R}$ on the right-hand side of the above equation are univariate functions acting upon each coordinate, whereas the function $\rho_\lambda : \mathbb{R}^d \rightarrow \mathbb{R}$ on the left-hand side is acting on the entire regression vector. For variable selection based on a regularization function to possess selection consistency property, we will require certain assumptions on the regularization function which will be discussed

in Section 3.4.2.

We use MCP and SCAD estimators for variable selection in our proposed UCB algorithm (Algorithm 3). In the algorithm, the regularization parameter λ_t at time t is in the order of $\Theta(\sqrt{\log d/t})$. The heuristics of the choice of λ_t can be summarized in two parts. Firstly, at time t , the variable selection procedure $\mathcal{V}$ will select the entries of β^* with magnitude greater than a threshold scaling as $\Theta(\sqrt{\log d/t})$ with high probability (Theorem 6). Therefore, the selection-policy pair $(\mathcal{V}, \pi)$ satisfies Definition 8. Secondly, the estimation error for those entries of β^* with magnitude greater than the threshold is $\mathcal{O}(\sqrt{1/t})$, which is smaller than the $\mathcal{O}(\sqrt{\log d/t})$ error without variable selection. As a result, the diameter of $\mathcal{C}_t$ does not depend on the dimension d , leading to an improved cumulative regret for the algorithm.

Remark: In Algorithm 3, we need to apply the regularized estimation for variable selection at each round. According to Definition 8, for any entry of β^* with magnitude ε , the variable selection procedure $\mathcal{V}$ needs $\mathcal{O}(1/\varepsilon^2)$ rounds to select it with high probability. Due to this observation, a more computationally efficient approach is to conduct variable selection every several rounds. Moreover, since smaller signals requires substantially longer time to identify due to the $\mathcal{O}(1/\varepsilon^2)$ scaling, we may conduct variable selection more frequently at the beginning while less frequent in later periods, for example, conduct variable selection at round $a, a^2, a^3, \dots$ for some integer $a > 1$. Alternatively, the step-wise variable selection method [36] can be an computationally-efficient greedy approach for variable selection without the need of refitting a regularized regression for variable selection, where one includes or excludes one variable at each round until no further variables can be added or deleted according to certain criterion. The theoretical analysis of the step-wise selection method is more involved and we leave it for future research. Empirically, we find from our numerical study in Section 3.5 that our proposed algorithm outperforms the ℓ_1 -confidence ball algorithm from [31] which does not incorporate variable selection. In addition, we also apply our proposed algorithm with cross-validation on optimizing initial parameters for real datasets, which performs comparably to our proposed algorithm with optimal initial parameters. This indicates the effectiveness of our proposed algorithm in real applications.

3.4 Theoretical Properties

In this section, we present the theoretical results for regret analysis of the proposed algorithm (Algorithm 2). In Section 3.4.1, we discuss the main assumptions that are required for the theoretical analysis under which we prove a regret upper bound for Algorithm 2. In addition, we provide a regret lower bound for linear bandit problems under our assumptions as a benchmark for assessing the proposed algorithm. In Section 3.4.2, we demonstrate that the regularized estimation methods using non-convex penalties described in Section 3.3.3

Algorithm 3 UCB-VS Algorithm with MCP (or SCAD)

Require: Initial regularization parameter λ_0 for regularization function ρ_λ , initial diameter τ_0 for the confidence set $\mathcal{C}_t$ and initial parameter R for constraint in (3.5).

- 1: Initialize $\hat{\beta}_0^{ols}$ as 0 in $\mathbb{R}^d$.
 - 2: **for** $t \in [T]$ **do**
 - 3: Observe $\mathcal{D}_t = \{X_{a,t}\}_{a=1}^K$ and $\mathcal{F}_{t-1}$.
 - 4: Obtain $\hat{\beta}_{t-1}$ by solving equation (3.5), $\hat{\beta}_{t-1} \in \arg \min_{\|\beta\|_1 \leq R} \{\mathcal{L}_{t-1}(\beta) + \rho_{\lambda_{t-1}}(\beta)\}$.
 - 5: Obtain the subset $\hat{S}_{t-1} = \text{supp}(\hat{\beta}_{t-1})$.
 - 6: Obtain $\hat{\beta}_{t-1}^{ols}$ by solving equation (3.3), i.e., $\hat{\beta}_{t-1}^{ols} \in \arg \min_{\beta \in \mathbb{R}^d, \text{supp}(\beta) \subseteq \hat{S}_t} \mathcal{L}_{t-1}(\beta)$.
 - 7: Construct ellipsoid confidence set $\mathcal{C}_t$:
 - 8: $\mathcal{C}_t = \{\beta \in \mathbb{R}^d : \|\hat{\beta}_{t-1}^{ols} - \beta\|_{\hat{\Sigma}_{t-1}} \leq \tau_{t-1}, \text{supp}(\beta) \subseteq \text{supp}(\hat{\beta}_{t-1}^{ols})\}$.
 - 9:
 - 10: Select optimal arm $\hat{a}_t$ and corresponding feature vector $\mathbf{X}_{\hat{a}_t, t}$ from equation (3.2), and observe reward $Y_t = \langle \mathbf{X}_{\hat{a}_t, t}, \beta^* \rangle + \omega_t$.
 - 11: Set $\lambda_t = \lambda_0 \sqrt{\frac{\log d + \log t}{t}}$, $\tau_t = \tau_0 \sqrt{\frac{|\hat{S}_t|(\log |\hat{S}_t| + \log t)}{t}}$.
 - 12: **end for**
-

satisfy Definition 8, and thus can be used for variable selection to attain the derived regret upper bound. It is worthy highlighting that the derived regret upper bound for Algorithm 3 matches the regret lower bound up to a logarithmic term in horizon T , i.e., $\mathcal{O}(\log T)$, and is strictly better than existing regret upper bounds in the high dimensional linear bandit literature.

3.4.1 Regret Analysis

In this section, we present theoretical results for the proposed UCB algorithm (Algorithm 2) with variable selection and a regret lower bound for policies. We first describe the assumptions made for the analysis.

Let $\hat{\Sigma}_t = \mathbf{X}_t^T \mathbf{X}_t / t$ denote the sample covariance matrix at the end of round $t \in [T]$. Let $\mathbb{S}^{d \times d}$ denote the set of all $d \times d$ positive semidefinite matrices. Consider the following subset of $\mathbb{S}^{d \times d}$ composed of all matrices that satisfy the *sparse eigenvalue condition* defined as below.

Definition 9 (Sparse eigenvalues). *For a $d \times d$ matrix $A \in \mathbb{S}^{d \times d}$ and $1 \leq m \leq d$, define $\rho_{\min}^2(m, A)$ and $\rho_{\max}^2(m, A)$ as the minimum and maximum m -sparse eigenvalues of A if*

$$\rho_{\min}^2(m, A) \leq \frac{\Delta^T A \Delta}{\|\Delta\|_2^2} \leq \rho_{\max}^2(m, A) \quad \text{for any } \Delta \neq 0 \text{ with } \|\Delta\|_0 \leq m.$$

We now provide the two main assumptions for our theoretical analysis.

Assumption 3 (On the true reward).

- (a) *Sparsity condition:* There exist positive constants b and s_0 such that $\|\beta^*\|_1 \leq b$ and $\|\beta^*\|_0 \leq s_0$.

(b) *Margin condition:* There exist positive constants Δ_* and C_1 , such that for $h \in \left[C_1 \sqrt{\frac{\log d}{T}}, \Delta_* \right]$, we have $P(\langle \mathbf{X}_{a_t^*, t}, \boldsymbol{\beta}^* \rangle \leq \max_{b \neq a_t^*} \langle \mathbf{X}_{b, t}, \boldsymbol{\beta}^* \rangle + h) \leq h/(2\Delta_*)$.

Assumption 3(a) requires the boundedness and sparsity of the true parameter $\boldsymbol{\beta}^*$. This is a standard requirement made in bandit literature [4], [13]. Assumption 3(b) controls the probability of the features falling into an h -neighborhood of the decision boundary. This assumption has been made by [19], [4] and [31]. The assumption will be satisfied if the density of $\langle \mathbf{X}_{a, t}, \boldsymbol{\beta}^* \rangle$ are upper bounded uniformly for all $a \in [K]$. This assumption is a probabilistic relaxation of the usual gap assumption made for problem dependent bounds in multi-armed bandit problems. For example, in [13], they assume a positive gap Δ_* between the rewards of optimal arm and sub-optimal arms, i.e., $\min_{b \neq a_t^*} \langle \mathbf{X}_{a_t^*, t} - \mathbf{X}_{b, t}, \boldsymbol{\beta}^* \rangle \geq \Delta_*$. Assumption 3(b) relaxes the assumption by allowing the gap to be arbitrarily close to zero.

Assumption 4 (On the features).

(a) *Boundedness:* There exists a positive constant $x_{\max}$ such that $\|\mathbf{X}_{a, t}\|_{\infty} \leq x_{\max}$ for all $a \in [K]$ and $t \in [T]$.

(b) *Anti-concentration:* There exists a positive constant ζ such that for each $a \in [K]$, $t \in [T]$, $v \in \mathbb{R}^d$ and $h \in \mathbb{R}_+$, $P(\langle \mathbf{X}_{a, t}, \mathbf{v} \rangle^2 \leq h \|\mathbf{v}\|_2^2) \leq \zeta h$.

(c) *Sparse eigenvalues:* There exist constants $\Lambda_0 \geq \lambda_0 > 0$ and $C^* > 0$ such that for each $a \in [K]$, the minimum and maximum $C^* s_0$ -sparse eigenvalues of the matrix $\boldsymbol{\Sigma}_a := \mathbb{E}[\mathbf{X}_{a, t} \mathbf{X}_{a, t}^T]$ are bounded below and above by λ_0^2 and Λ_0^2 , respectively. In addition, the minimum and maximum $C^* s_0$ -sparse eigenvalues of $\mathbb{E}[\mathbf{X}_{a_t^*} \mathbf{X}_{a_t^*}^T | \Gamma_t]$ are also bounded by ϕ_0^2 and Λ_1^2 , respectively, where the event Γ_t is defined as $\Gamma_t := \{\langle \mathbf{X}_{a_t^*, t}, \boldsymbol{\beta}_* \rangle \geq \max_{b \neq a_t^*} \langle \mathbf{X}_{b, t}, \boldsymbol{\beta}_* \rangle + \Delta_*\}$.

Assumption 4(a) together with Assumption 3(a) ensures that the maximum regret at any time can be upper bounded, since $|\langle \mathbf{X}_{a, t}, \boldsymbol{\beta}^* \rangle| \leq x_{\max} b$ by Cauchy-Schwarz inequality. Assumption 4(b) is an anti-concentration condition which is important for the regret analysis of the proposed algorithms. In particular, it ensures that the feature vector $\mathbf{X}_{a, t}$ from each arm scatters uniformly in all directions so that the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_t$ will not concentrate on one direction such as the direction of $\boldsymbol{\beta}^*$. As a result, it guarantees the estimation accuracy for $\boldsymbol{\beta}^*$, which will be reflected in the regret upper bound of the proposed algorithm. The anti-concentration assumption is mild. For example, it can be implied by the boundedness of the density function of $\langle \mathbf{X}_{a, t}, \mathbf{v} \rangle$ for each $\mathbf{v} \in \mathbb{R}^d$. In addition, the constant ζ in the assumption only appears in the “burn-in” term in the regret upper bound of Algorithm 2 (Theorem 4) and does not affect the leading term. Assumption 4(c) ensures that covariance matrix $\hat{\boldsymbol{\Sigma}}_t$ satisfies the sparse eigenvalue condition (Definition 9),

which is required for the selection consistency analysis for the regularized estimators, e.g., MCP and SCAD estimators in Algorithm 3, and estimation error analysis for the OLS estimator after variable selection.

In the following, we present the main results on the proposed UCB algorithm (Algorithm 2) with variable selection. Firstly, we provide a result showing that the center $\hat{\beta}_{t, \hat{S}_t}^{ols}$ of the ellipsoid confidence set $\mathcal{C}_{t+1}$ constructed in Algorithm 2 lies close to $\beta_{\hat{S}_t}^*$ of the true parameter vector β^* with high probability. This result plays an critical role in proving the regret upper bound (Theorem 4) for Algorithm 2.

Theorem 3. *Suppose that Assumptions 3–4 hold and the selection-policy pair $(\mathcal{V}, \pi)$ is (M_0, t_0, M_1, M_2) -sequentially selection consistent (Definition 8). Then, for time t such that $t \geq 2 \max\left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2(d)\right)$, Algorithm 2 with initial diameter $\tau_0 = \frac{8\sigma x_{\max}}{\phi_0}$ satisfies the following property,*

$$\|\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^*\|_{\hat{\Sigma}_t} \leq \tau_0 \sqrt{\frac{|\hat{S}_t|(\log |\hat{S}_t| + \log t)}{t}} + \frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 \|\beta_{S^* \setminus \hat{S}_t}^*\|_1,$$

with probability at least $1 - C_1 M_0 t^{-1} - C_2 e^{s_0 \log(\Lambda_0^2 \zeta K)} [e^{-c_0 t} + e^{-\tilde{C} \log(K)t}]$. The set $\hat{S}_t = \mathcal{V}_t(\mathcal{F}_t)$ and the constant χ_g is given by $\chi_g = C_3 \left[\sigma x_{\max}^3 \zeta K s_0^{5/2} \phi_0^{-1} \sqrt{M_1(d)/\log d} \right]$. Here c_0, C_1, C_2, C_3 and $\tilde{C}$ are absolute constants.

Theorem 3 shows that with the construction of the confidence set $\mathcal{C}_{t+1}$ in Algorithm 2, the subvector $\beta_{\hat{S}_t}^*$ of true parameter β^* lies close to $\mathcal{C}_{t+1}$ with high probability. Comparing with the form of confidence set $\mathcal{C}_{t+1}$, we can observe that for any $\beta \in \mathcal{C}_{t+1}$, the ℓ_2 -distance between β and $\beta_{\hat{S}_t}^*$ scales as $\mathcal{O}(\sqrt{s_0(\log s_0 + \log t)/t} + \|\beta_{S^* \setminus \hat{S}_t}^*\|_1)$, where $\beta_{S^* \setminus \hat{S}_t}^*$ corresponds to the entries of β^* with small magnitude according to Definition 8. In comparison, the ℓ_2 -distance between β^* and β in $\mathcal{C}_t$ from [13] scales as $\mathcal{O}(\sqrt{d \log t/t})$; the ℓ_1 -norm between β^* and β in $\mathcal{C}_t$ from [31] scales as $\mathcal{O}(\sqrt{(\log d + \log t)/t})$, both of which have dependence on dimension d . The confidence set $\mathcal{C}_{t+1}$ in our paper provides a finer control of distance between $\beta_{\hat{S}_t}^*$ and β in $\mathcal{C}_{t+1}$, which depends on the sparsity s_0 and an additional term $\|\beta_{S^* \setminus \hat{S}_t}^*\|_1$. According to Definition 8(b), the ℓ_1 -norm of $\beta_{S^* \setminus \hat{S}_t}^*$ decreases to zero along the time with high probability. Therefore, the ℓ_2 -distance between $\beta_{\hat{S}_t}^*$ and β in $\mathcal{C}_{t+1}$ scales as $\mathcal{O}(\sqrt{s_0(\log s_0 + \log t)/t})$ asymptotically. As a consequence, the leading term in the cumulative regret (Theorem 4) of Algorithm 2 does not dependent on dimension d .

Based on the result in Theorem 3, we provide the following theorem which shows the regret upper bound for Algorithm 2.

Theorem 4. *Suppose that Assumption 1–2 hold and selection-policy pair $(\mathcal{V}, \pi)$ satisfies (M_0, t_0, M_1, M_2) -sequential selection consistency property. When $K \geq 2, d \geq 1$, if we set $\tau_0 = \frac{8\sigma x_{\max}}{\phi_0}$ as in Theorem 3, the*

expected cumulative regret of Algorithm 2 at horizon T can be bounded as

$$\begin{aligned}
R_T \leq & \overbrace{\frac{512x_{\max}^2 s_0^2 \tau_0^2}{\Delta_* \phi_0^2} (\log s_0 + \log T) \log T}^{I_1} + \overbrace{\frac{4s_0 M_1(d)}{\Delta_*} \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right)^2}^{I_2} \\
& + \overbrace{4bx_{\max} \max \left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2(d) \right)}^{I_3} + \overbrace{C_4 x_{\max} b \log T}^{I_4}. \tag{3.8}
\end{aligned}$$

where the constant χ_g is given by $\chi_g = C_3 \left[\sigma x_{\max}^3 \zeta K s_0^{5/2} \phi_0^{-1} \sqrt{M_1(d)/\log d} \right]$. Here C_3 and C_4 are absolute constants independent of T and d .

Theorem 4 shows that the cumulative regret of Algorithm 2 grows poly-logarithmically in both d and T , i.e., $R_T = \mathcal{O}(\log^2 T + \log d)$. Intuitively, $I_2 + I_3$ on the right-hand side of the display (3.8) corresponds to the regret incurred during the “burn-in” period of exploring the feature space and does not affect the asymptotic cumulative regret of the algorithm, while I_4 corresponds to the regret incurred when the confidence set $\mathcal{C}_t$ does not satisfy the property described in Theorem 3. The leading term I_1 describes the regret incurred when selecting the sub-optimal arms and scales as $\mathcal{O}((\log T)^2)$, which is independent from dimension d .

Our proposed algorithm improves the cumulative regret of the algorithm in [4], i.e., $\mathcal{O}((\log d + \log T)^2)$, due to the incorporation of variable selection. Specifically, the selected set $\hat{S}_t = \mathcal{V}_t(\mathcal{F}_t)$ in Algorithm 2 belongs to the support S^* of β^* with high probability according to Definition 8. Therefore, the diameter of $\mathcal{C}_t$ in Algorithm 2 is independent from dimension d with high probability, because $\mathcal{C}_t$ lies in a low-dimensional subspace corresponding to $\hat{S}_t \subseteq S^*$. Since the cumulative regret of Algorithm 2 can be characterized as the sum of the diameters of the confidence sets, the leading term I_1 in the regret upper bound is also independent from dimension d . In addition, the selected subset $\hat{S}_t$ will recover the whole support S^* of β^* asymptotically according to Definition 8. As a result, the OLS estimator $\hat{\beta}_t^{\text{ols}}$ restricted on $\hat{S}_t$ will perform as the oracle OLS estimator when S^* is known. The original bandit problem then reduces to a low-dimensional problem.

In addition to the regret upper bound for the proposed algorithm, we provide a regret lower bound for any policy π in linear bandit environment satisfying the assumptions made in Section 3.4.1. This demonstrates the hardness of the linear bandit problem in this paper and the effectiveness of our proposed algorithm.

Theorem 5. *Let $\mathcal{P}$ be the class of distributions of $(\mathbf{X}_{a,t}, Y_{a,t} : a \in [d])$ which satisfies Assumptions 1–2. Then for large enough horizon T ,*

$$\sup_{P_{X,Y} \in \mathcal{P}_\alpha} \inf_{\pi} R_T(\pi, \pi^*) \geq C_{LB} (\log d + \log T), \tag{3.9}$$

where C_{LB} is a constant independent of T and d .

Our Theorem 5 generalizes the existing results in the literature for stochastic linear bandit problem by considering both the horizon T and dimension d . In particular, the theorem proves a poly-logarithmic dependence on dimension d by leveraging the margin condition (Assumption 3(b)). For comparison, [19] proved a lower bound of $\Omega(\log T)$ for low-dimensional stochastic linear bandit problem under similar assumptions for both feature vectors and true parameter vector β^* . [3] also proved a lower bound of $\Omega(\sqrt{Td})$ for low-dimensional contextual bandit problem, where no margin condition is assumed.

By comparing the regret lower bound in Theorem 5 with the regret upper bound (Theorem 4) for Algorithm 2, we can observe that our proposed algorithm is optimal up to a factor of $\log T$ with respect to horizon T and optimal with respect to dimension d . In particular, our upper bound of order $\mathcal{O}(\log^2 T + \log d)$ also decouples the time horizon T and dimension d , while the bound in [4] has the extra interaction term of order $\mathcal{O}(\log d \log T)$ that can be substantially larger. In other words, thanks to the variable selection, the impact of dimension d on the regret does not deteriorate as T increases.

3.4.2 Application to UCB Algorithm with Regularized Estimator

In this subsection, we show that the variable selection procedure induced by the regularized estimation procedures described in Algorithm 3 satisfies our Definition 8, and discuss the consequence of applying the regret analysis to Algorithm 3. To start with, we introduce an assumption on the regularization functions.

Assumption 5 (On the regularizers). *For a positive constant $\epsilon_0 = o(\|\nabla \mathcal{L}_t(\beta^*)\|_\infty)$, define $\mathbf{\Gamma} = \nabla^2 \mathcal{L}_t(\beta^*)$ and $r = 2\|(\mathbf{\Gamma}_{S_2 S_2})^{-1}\|_\infty (\|\nabla \mathcal{L}_t(\beta^*)\|_\infty + \epsilon_0)$, where subset $S_2 = \{i \in [d] : |\beta_i^*| > r_1\}$ is defined in (3.13) and $r_1 > r$. We assume that the regularization function ρ_λ from equation (3.5) satisfies*

$$\begin{cases} \lim_{t \rightarrow 0^+} \rho'_\lambda(t) = \lambda, \\ \rho'_\lambda(r_1 - r) = \epsilon_0, \\ \inf_{t > 0} \rho''_\lambda(t) = -u, \end{cases} \quad (3.10)$$

where u is an absolute positive constant.

The derivative of $\rho'_\lambda(t)$ of a regularization function is well known to play a critical role as the shrinkage effect [6], [7]. The first two conditions in (3.10) are known as the *selection* and *unbiasedness* properties. In particular, the *selection* condition ensures that the regularized estimator automatically sets small estimated parameters to zero to reduce model complexity, while the *unbiasedness* condition guarantees that the estimator is nearly unbiased when the true parameter is large. More discussion on the regularization functions can be found in [6], [7], [33], [34].

For instance, the univariate function ρ_λ can be the MCP [6] and SCAD [7] regularizers with properly defined parameter λ , for which we provide definitions via their derivatives as follows:

$$\rho'_{\text{MCP}}(\beta) = \begin{cases} \left(\lambda - \frac{|\beta|}{\gamma}\right) \text{sign}(\beta) & \text{if } |\beta| \leq \gamma\lambda, \\ 0 & \text{if } |\beta| \geq \gamma\lambda. \end{cases} \quad (3.11)$$

$$\rho'_{\text{SCAD}}(\beta) = \begin{cases} \lambda \text{sign}(\beta) & \text{if } |\beta| \leq \lambda, \\ \frac{\gamma\lambda - |\beta|}{\gamma - 1} \text{sign}(\beta) & \text{if } \lambda < |\beta| < \gamma\lambda, \\ 0 & \text{if } |\beta| \geq \gamma\lambda. \end{cases} \quad (3.12)$$

In order to apply the results from Theorem 4 to Algorithm 3, we prove that the variable selection procedure induced by the nonconvex regularized estimation procedures satisfies sequential selection consistency property (Definition 8).

Theorem 6. *Suppose Assumptions 3–5 hold, and the parameter λ_t for the regularization function ρ_{λ_t} at time t is defined as $\lambda_t = C_L \sqrt{(\log d + \log t)/t}$ with a constant C_L . Define set $\hat{S}_t = \text{supp}(\hat{\beta}_t)$ at time t , where $\hat{\beta}_t$ is the regularized estimator in Algorithm 3. Divide the true support S^* into the following two sets at time t :*

$$S_1 = \{i \in [d] : |\beta_i^*| \leq r_1, \beta_i^* \neq 0\} \quad \text{and} \quad S_2 = \{i \in [d] : |\beta_i^*| > r_1\}, \quad (3.13)$$

where $r_1 = C_g \sqrt{(\log d + \log t)/t}$ with a constant C_g . There exist positive constants C_0 such that for all $t \geq C_0 s_0 \Lambda_0^4 \zeta K \log^2 K \log d$, we have

$$P\left(\left\{S_2 \subseteq \hat{S}_t \subseteq S^*\right\}\right) \geq 1 - \frac{2}{t} - \mathcal{O}\left(e^{-\frac{1}{2}c_0 t} + e^{-\frac{1}{2}\tilde{C} \log(K)t + s_0 \log(\Lambda_0^2 \zeta K)}\right), \quad (3.14)$$

where c_0 and $\tilde{C}$ are absolute constants.

The parameter R in (3.5) can be chosen as a sufficiently large constant. For example, we can choose a constant R such that $R \geq 2b$, since $\|\beta^*\|_1 \leq b$ from Assumption 3(a). According to (3.14) in the theorem, parameter β_i^* within subset S_2 will be selected by the regularized estimator $\hat{\beta}_t$ with high probability, while the zero entries of β^* are not selected by $\hat{\beta}_t$. In the following corollary, we show that the estimator $\hat{\beta}_t$ satisfying the inequality (3.14) can induce a variable selection procedure that satisfies the sequential selection consistency property (Definition 8).

Corollary 2. *Suppose Assumptions 3–5 hold and the parameter λ_t is defined as in Theorem 6. Define the variable selection procedure $\mathcal{V}$ with $\mathcal{V}_t(\mathcal{F}_t) = \text{supp}(\hat{\beta}_t)$ at time t , where $\hat{\beta}_t$ is the regularized estimator*

in Algorithm 3. The selection-policy pair $(\mathcal{V}, \pi)$ satisfies (M_0, t_0, M_1, M_2) -sequentially selection consistency property (Definition 8). The parameters are defined as $t_0 = M_2(d, T) = C_0 s_0 \Lambda_0^2 \zeta K \log K \log d$ and $M_1(d) = C'_g \log d$, where C_0 , M_0 and C'_g are positive constants.

According to Corollary 2, the variable selection procedure $\mathcal{V}$ induced by regularized estimation procedures possesses sequential selection consistency property (Definition 8), which makes it eligible for our proposed UCB algorithm with variable selection. In the following, we show that equipped with this eligible variable selection procedure, the constructed confidence set $\mathcal{C}_t$ performs well in approximating the true parameter β^* . The proof of Theorem 7 follows the theoretical analysis of Theorem 3 and applies to Algorithm 3 with regularized estimation methods.

Theorem 7. Suppose that Assumptions 3–5 hold. The variable selection procedure $\mathcal{V}$ is induced by the regularized estimator $\hat{\beta}_t$ as in Corollary 2, where the regularization parameter $\lambda_t = C_L \sqrt{(\log t + \log d)/t}$ is defined as in Theorem 6. Then, for time t such that $t \geq 2 \max \left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, C_0 s_0 \Lambda_0^2 \zeta K \log K \log d \right)$, Algorithm 3 with initial diameter $\tau_0 = \frac{8\sigma x_{\max}}{\phi_0}$ satisfies the following property,

$$\|\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^*\|_{\hat{\Sigma}_t} \leq \tau_0 \sqrt{\frac{|\hat{S}_t|(\log |\hat{S}_t| + \log t)}{t}} + \frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 \|\beta_{S^* \setminus \hat{S}_t}^*\|_1,$$

with probability at least $1 - C_1 M_0 t^{-1} - C_2 e^{s_0 \log(\Lambda_0^2 \zeta K)} [e^{-c_0 t} + e^{-\tilde{C} \log(K)t}]$. The set $\hat{S}_t = \mathcal{V}_t(\mathcal{F}_t)$ and the constant χ_g is given by $\chi_g = C_3 \sigma x_{\max}^3 \zeta K s_0^{5/2} \phi_0^{-1} \sqrt{C'_g}$. Here c_0 , M_0 , C_0 , C_1 , C_2 , C_3 and $\tilde{C}$ are absolute constants.

Theorem 7 shows that the confidence set $\mathcal{C}_t$ in Algorithm 3, while being sparse, is still capable of capturing the estimation uncertainty about the true parameter β^* as accurate as those based on non-sparse confidence sets in the existing UCB algorithms [3], [13], [31]. In particular, $\beta_{\hat{S}_t}^*$ of the true parameter β^* lies close to the center $\hat{\beta}_{t, \hat{S}_t}^{ols}$ of the constructed confidence set $\mathcal{C}_t$ in Algorithm 3 with high probability. In the below, we provide an upper bound for the cumulative regret of Algorithm 3.

Theorem 8. Suppose that Assumptions 3–5 hold and the variable selection procedure $\mathcal{V}$ is defined as in Theorem 7. When $K \geq 2$, $d \geq 1$, if we define $\tau_0 = \frac{8\sigma x_{\max}}{\phi_0}$ as in Theorem 3, the expected cumulative regret of Algorithm 3 at horizon T can be bounded as

$$\begin{aligned} R_T \leq & \frac{512x_{\max}^2 s_0^2 \tau_0^2}{\Delta_*^2 \phi_0^2} (\log s_0 + \log T) \log T + \frac{4s_0 C'_g \log d}{\Delta_*} \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right)^2 \\ & + 4bx_{\max} \max \left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, C_0 s_0 \Lambda_0^2 \zeta K \log K \log d \right) + C_4 x_{\max} b \log T. \end{aligned} \quad (3.15)$$

where χ_g is given by $\chi_g = C_3 \left[\sigma x_{\max}^3 \zeta K s_0^{5/2} \phi_0^{-1} \sqrt{C'_g} \right]$. Here C_0 , C_3 , C_4 and C'_g are absolute constants independent of T and d .

Theorem 8 is an application of the regret analysis of Theorem 4. According to Theorem 8, the cumulative regret of the UCB algorithms (Algorithm 3) with regularized estimation methods scales as $\mathcal{O}(\log^2 T + \log d)$ in horizon T and dimension d , which improves that of the existing algorithms for high-dimensional bandit problems that incorporate regularized estimators [4], [5], [31]. The distinguishing point between Algorithm 3 and the existing algorithms with regularized estimators is that Algorithm 3 leverages the variable selection property of the regularized estimators. This indicates that incorporating variable selection in the design of algorithms is beneficial for high-dimensional bandit problems.

3.5 Empirical Results

In this section, we compare our proposed UCB-VS algorithms with MCP and SCAD (Algorithm 3) with the ℓ_1 -confidence ball based method proposed by [31]. We note that the original ℓ_1 -confidence ball based method relies on the LASSO estimator instead of MCP or SCAD estimators considered in this paper. In order to seek answers for whether the variable selection improves the performance of the UCB algorithms, we replace the LASSO estimator with MCP and SCAD estimators for the ℓ_1 -confidence ball based method. Therefore, the major differences between our proposed algorithm and the ℓ_1 -confidence ball based algorithm are whether the variable selection is incorporated and the structure of the induced confidence set $\mathcal{C}_{t+1}$. In the following, we start with a synthetic data experiment in Section 3.5.1 to compare algorithms under different scenarios and conduct a real data experiment in Section 3.5.2 to compare UCB algorithms with and without variable selection.

3.5.1 Synthetic Data

In the synthetic data experiment, we consider three scenarios for K , d and s_0 : (1) $K = 5$, $d = 100$, $s_0 = 5$; (2) $K = 5$, $d = 1000$, $s_0 = 5$; (3) $K = 50$, $d = 20$, $s_0 = 2$. In each case, the first s_0 features are selected to be predictive of the reward. The nonzero entries of β^* corresponding to s_0 features will be generated from $\text{Unif}(0.5, 1)$. At each time t , feature vectors for K arms are i.i.d. generated by truncating a Gaussian distribution $N(0, \Sigma)$ with $\Sigma_{ij} = 0.1^{|i-j|}$ so that $\|X_{a,t}\|_\infty \leq 5$. The error term $\omega_{a,t}$ follows a zero mean normal distribution with variance $\sigma^2 = 1$. For all methods, we tune the initial parameters on evenly separated values within range $[0.5, 1.5]$ to achieve the optimal cumulative regret under each scenario. For ℓ_1 -confidence ball based algorithm with MCP and SCAD estimators, we tune the initial regularization parameter λ_0 and the

diameter τ_0 for the ℓ_1 -confidence ball. For our UCB-VS algorithm with MCP and SCAD estimators, we tune the initial regularization parameter λ_0 and diameter τ_0 for the ellipsoid confidence set. The bound R for $\|\beta\|_1$ in equation (3.5) from Algorithm 3 is to ensure that the minimizer of the optimization problem (3.5) exists and does not diverge. In our empirical study, we do not apply the ℓ_1 bound on β in (3.5) and minimize the unconstrained objective function of (3.5).

Figure 3.1 compares our UCB-VS algorithms with MCP and SCAD estimators with competitors on the aforementioned synthetic data with a time horizon of $T = 2000$. The curves are the average cumulative regrets over 5 trials. We observe that in $d = 100$ and $d = 1000$ cases, the proposed UCB-VS algorithms outperforms the ℓ_1 -confidence ball based algorithms, which does not incorporate variable selection, if they use the same regularized estimation, e.g., MCP and SCAD. While in low-dimensional case, i.e., $d = 20$, our proposed algorithms does not perform better than ℓ_1 -confidence ball based algorithms. In particular, the UCB-VS algorithm with SCAD has comparable performance to ℓ_1 -confidence ball based algorithm with SCAD, while the UCB-VS with MCP performs worse than the ℓ_1 -confidence ball based algorithm with MCP. The simulation result coincides with our theoretical result that variable selection improves the performance of UCB algorithms in the high-dimensional cases. Specifically, in the high-dimensional case where d is moderately large to high, the confidence set $\mathcal{C}_t$ constructed by the UCB-VS algorithms possesses a sparse structure, therefore the confidence set has a much smaller “volume” compared to that of the ℓ_1 -confidence ball and leads to a better cumulative regret which is proportional to the sum of the “volume” of confidence sets along the time. While in the low-dimensional case where the sparsity, e.g., $s_0 = 2$, is comparable to the dimension, e.g., $d = 20$, the sparsity of the confidence set induced by variable selection does not reduce the “volume” of the confidence set much compared to the ℓ_1 -confidence set, therefore both UCB algorithms with and without variable selection have similar performance.

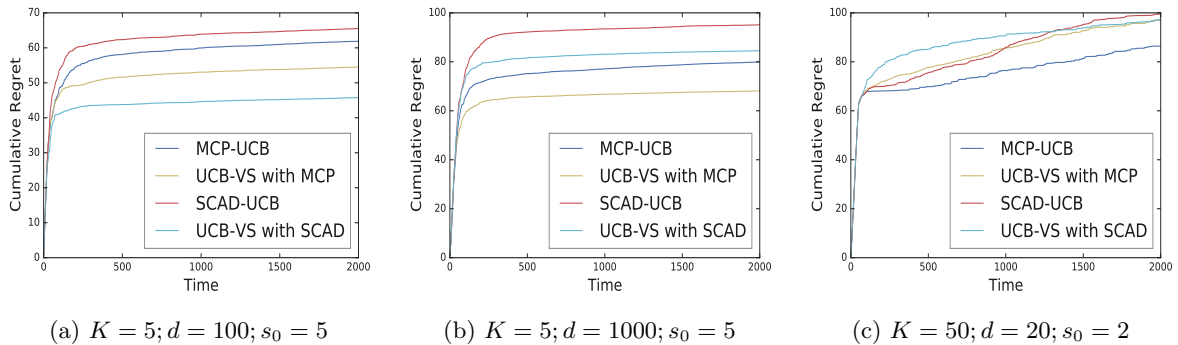


Figure 3.1: Comparison of the cumulative regret of UCB-VS algorithms with MCP and SCAD with ℓ_1 -confidence ball based algorithm with MCP (MCP-UCB) and SCAD (SCAD-UCB). Nonzero components in β_* are generated from $\text{Unif}(0.5, 1)$. Lower value indicates better performance.

3.5.2 Africa Soil Property Data

In the real-data experiment, we consider the Africa soil property dataset [37], which contains representative soil samples that are collected across non-desert area in Africa during 2009–2012. The training dataset contains 1157 soil samples and 3576 covariates for each sample, including mid-infrared absorbance measurements, depth of the soil samples (2 categories: “Topsoil” and “Subsoil”) and some spatial predictors from remote sensing data sources such as satellite images, which are used to predict functional properties of soil such as pH values (pH) and Mehlich-3 extractable Calcium (Ca). For our bandit experiment, we evenly partition the training dataset into two arms according to the categorical covariate depth of the soil samples with each arm having approximately 570 samples. We choose the value of the functional property Ca as the reward for each soil sample. Since the true parameter β^* for predicting rewards is not available, we cannot compute the expected regret of the algorithms. Therefore, we use the collected Ca values from the dataset to compute the surrogate instant regret, which is the difference between the optimal reward value and the one observed from the selected arm at each time.

We compare our proposed UCB-VS algorithm with MCP and SCAD estimators with ℓ_1 -confidence ball based algorithm with MCP and SCAD estimators. Since there exist highly correlated covariates in the dataset, we implement all regularized estimation methods with additional ℓ_2 -penalty. Specifically, the regularized estimator is obtained through

$$\hat{\beta}_{t-1} \in \arg \min_{\|\beta\|_1 \leq R} \{\mathcal{L}_{t-1}(\beta) + \rho_{\lambda_{t-1}}(\beta) + \alpha \lambda_{t-1} \|\beta\|_2^2\},$$

where α controls the relative weight of the ℓ_2 -penalty compared to the MCP (or SCAD) penalty. In the experiment, we consider three different values, i.e., $\{\frac{1}{9}, 1, 9\}$, for α . For UCB-VS algorithms (Algorithm 3), we also construct an ellipsoid confidence set $\mathcal{C}_{t+1}$ centered around the constraint ridge regression estimator $\hat{\beta}_t^r$, which is defined as

$$\hat{\beta}_t^r \in \arg \min_{\beta \in \mathbb{R}^d, \text{supp}(\beta) \subseteq \hat{S}_t} \mathcal{L}_t(t) = \frac{1}{2t} \|\mathbf{Y}_t - \mathbf{X}_t \beta\|_2^2 + \gamma_t \|\beta\|_2^2,$$

where γ_t scales as $\gamma_t = \gamma_0/\sqrt{t}$ such that the estimation bias induced by the ridge regression decays along the time. In consequence, the ellipsoid confidence set $\mathcal{C}_t$ is defined as

$$\mathcal{C}_t = \{\beta \in \mathbb{R}^d : \|\beta_{\hat{S}_{t-1}} - \hat{\beta}_{t-1, \hat{S}_{t-1}}^r\|_{\hat{\mathbf{V}}_{t-1}} \leq \tau_{t-1}, \text{supp}(\beta) \subseteq \hat{S}_{t-1}\},$$

where $\hat{\mathbf{V}}_t = \frac{1}{t}(\mathbf{X}_{t, \hat{S}_t}^T \mathbf{X}_{t, \hat{S}_t} + \gamma_t \mathbf{I})(\mathbf{X}_{t, \hat{S}_t}^T \mathbf{X}_{t, \hat{S}_t})^{-1}(\mathbf{X}_{t, \hat{S}_t}^T \mathbf{X}_{t, \hat{S}_t} + \gamma_t \mathbf{I}) \subseteq \mathbb{R}^{|\hat{S}_t| \times |\hat{S}_t|}$, which is the inverse of the

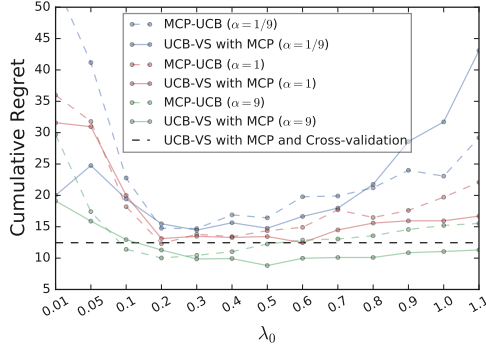
variance of $\hat{\beta}_t^r$ when $\hat{S} = S$.

Moreover, we also conduct the UCB-VS algorithm with cross-validation on parameters $\{\lambda_0, \alpha, \gamma_0\}$ which are often unknown in practice. Specifically, we divide the horizon T into N epochs and the parameters are updated at pre-specified times $\{\frac{T}{2(N-1)}, \frac{T}{2(N-2)} \dots, \frac{T}{2}\}$ so that the cross-validation will only be conducted $\mathcal{O}(\log T)$ times. Within cross-validation, the parameters are optimized based on the predictive performance, e.g., the mean squared error (MSE), with all available data up to the time point. In the experiment, we choose $\alpha = 9$ during the first epoch for the stability due to the high-dimensionality in the initial period.

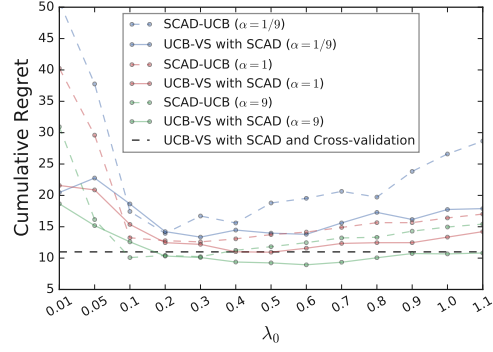
We apply aforementioned UCB algorithms with different initial parameters, including regularization parameter λ_0 for MCP and SCAD estimation, diameter τ_0 for the confidence set and γ_0 for the ridge regression. In the experiment, we alternatively select between two arms for the first 70 iterations as a warm start for all methods to reduce the fluctuation of the performance of methods at the initial stage, after which we apply different bandit algorithms for $T = 500$ horizon to compare the cumulative regret across different methods. In addition, the accuracy of each method is calculated after the warm start as the proportion of times when the optimal arm is selected. We consider 5 permutations of the soil samples and take the average of the cumulative regret over the permutations. As a remark, including a warm start in our bandit experiment is reasonable, since it applies to all methods considered in the experiment and the decision-maker in practice often has some historical data before applying the bandit algorithms. Figure 3.2 illustrates the cumulative regret of all UCB algorithms over horizon $T = 500$ with different initial regularization parameter λ_0 . We observe that the UCB-VS algorithms outperforms the ℓ_1 -confidence ball based algorithms with the same α at most values of the initial parameter λ_0 . Figure 3.3 and Figure 3.4 present the cumulative regret and the accuracy of all algorithms with their optimal initial parameters respectively, where the UCB-VS algorithm with $\alpha = 9$ performs the best in both MCP and SCAD cases. The UCB-VS algorithm with cross-validation performs comparably to the UCB-VS algorithm with $\alpha = 1$ and optimal initial parameters. Moreover, for MCP case, there is a clear gap between the performance of UCB-VS algorithms with $\alpha = 9$ and other algorithms when the time is small, i.e., $t \leq 200$. The results on both the cumulative regret and the accuracy coincide with our theoretical result that variable selection facilitates the UCB algorithms in high-dimensional bandit problems.

3.5.3 Riboflavin Production Data

In the second experiment, we consider a genomic dataset on riboflavin (vitamin B_2) production by bacillus subtilis, which is made publicly available by [38]. The original dataset contains 71 samples and 4088 covariates measuring the logarithm of the expression levels of genes. The response variable indicates the logarithm of the

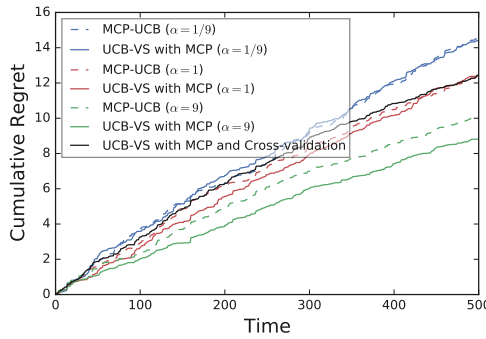


(a) UCB Algorithms with MCP

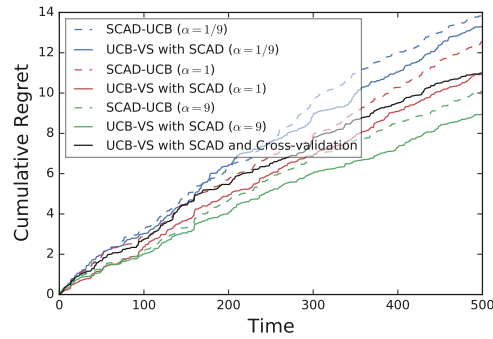


(b) UCB Algorithms with SCAD

Figure 3.2: Comparison of the cumulative regret of UCB-VS algorithms with MCP and SCAD and ℓ_1 -confidence ball based algorithm with MCP (MCP-UCB) and SCAD (SCAD-UCB) in Africa soil data with different initial regularization parameter λ_0 .

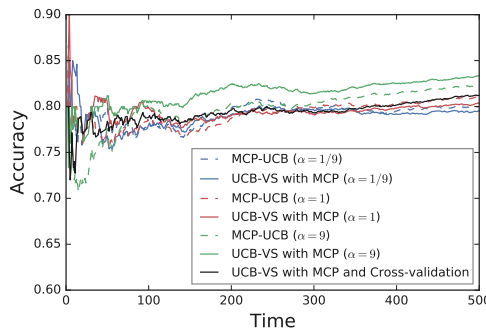


(a) UCB Algorithms with MCP

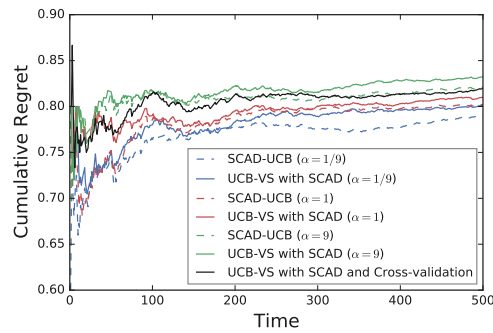


(b) UCB Algorithms with SCAD

Figure 3.3: Comparison of the cumulative regret of UCB-VS algorithms with MCP and SCAD and ℓ_1 -confidence ball based algorithm with MCP (MCP-UCB) and SCAD (SCAD-UCB) in Africa soil data with optimal initial regularization parameter λ_0 .



(a) UCB Algorithms with MCP



(b) UCB Algorithms with SCAD

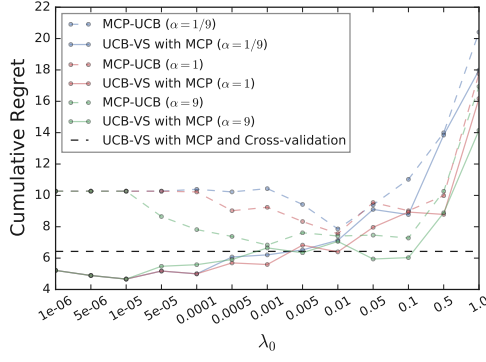
Figure 3.4: Comparison of the accuracy of UCB-VS algorithms with MCP and SCAD and ℓ_1 -confidence ball based algorithm with MCP (MCP-UCB) and SCAD (SCAD-UCB) in Africa soil data with optimal initial regularization parameter λ_0 .

riboflavin production rate. Before the experiment, we implement a pre-screening on the covariates and select a subset of 1075 covariates whose magnitude of correlation with the response variable is greater than 0.2. We conduct a two-armed bandit problem for illustration. In particular, in order to increase the horizon T for asymptotic performance, we run the experiment using bootstrap samples (random samples with replacement), which is also considered in [4]. The results using the bootstrap samples may be optimistic, since algorithms may overfit the samples drawn multiple times from the original data with replacement. Nevertheless, this is still a reasonable comparison among different bandit algorithms, since they are conducted under the same scenario. Specifically, we first expand the original dataset by sampling with replacement to 210 samples. After that, we randomly partitioned the expanded dataset into two groups each having 105 samples. Similar to the experiment on Africa soil data, we compute the cumulative regret of algorithms with the horizon $T = 100$ after 5 iterations of warm start. The final cumulative regret result for each method is the average over 5 bootstrap samplings of the riboflavin production dataset.

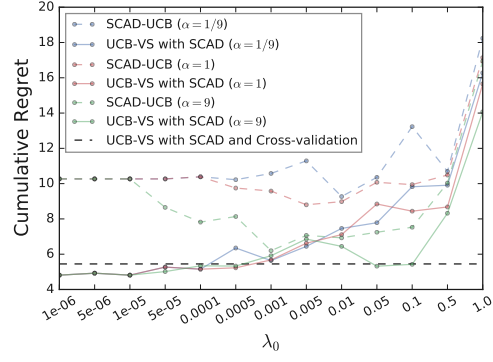
Similar to the Africa soil property data, we compare our proposed UCB-VS algorithm with the ℓ_1 -confidence ball algorithm with different weights α of the additional ℓ_2 -penalty. Figure 3.5 illustrates the cumulative regret of UCB algorithms on riboflavin production data with different initial regularization parameter λ_0 . We observe that the UCB-VS algorithms outperform the ℓ_1 -confidence ball algorithms with the same α at most values of initial λ_0 . In addition, the UCB-VS algorithm with different α have similar performance, which indicates that the performance gap between the UCB-VS algorithm and ℓ_1 -confidence ball algorithm is mainly due to the variable selection. Figure 3.6 and Figure 3.7 present the cumulative regret and the accuracy of all algorithms with their optimal initial parameters respectively. We can observe that the UCB-VS algorithms with different α have the same performance. The UCB-VS algorithm with cross-validation outperforms all ℓ_1 -confidence ball based algorithm with MCP and performs comparably to the UCB-VS algorithms with optimal initial parameters with SCAD, which indicates the efficiency of the cross-validation for selecting parameters. Similar to the Africa soil property dataset, there is a gap between the UCB-VS algorithm with $\alpha = 9$ and other ℓ_1 -confidence ball algorithms at the initial stage, which falls into a high-dimensional regime where the sample size is relatively small compared to the dimension.

3.6 Discussion

In this chapter, we propose a general framework to incorporate variable selection into Upper Confidence Bound (UCB) algorithms in the context of high-dimensional linear bandit problems and prove a regret upper bound (Theorem 4) scaling as $\mathcal{O}(\log^2 T + \log d)$, which is optimal up to a factor of $\log T$ to horizon T and optimal to dimension d . Additionally, we instantiate the framework with the nonconvex regularized estimation

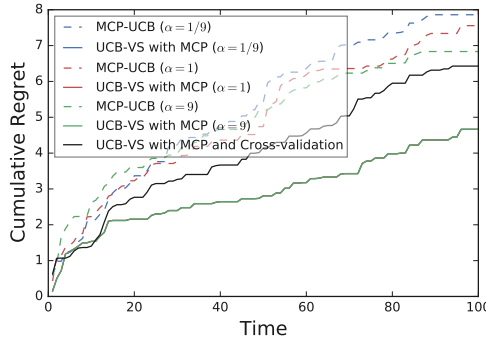


(a) UCB Algorithms with MCP

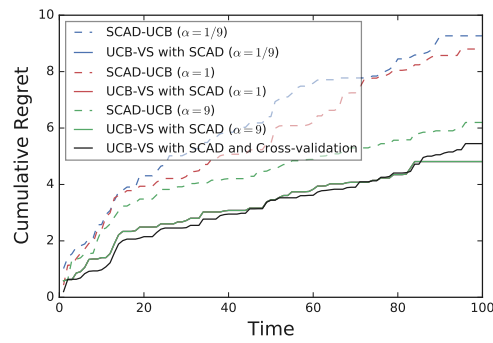


(b) UCB Algorithms with SCAD

Figure 3.5: Comparison of the cumulative regret of UCB-VS algorithms with MCP and SCAD and ℓ_1 -confidence ball based algorithm with MCP (MCP-UCB) and SCAD (SCAD-UCB) in riboflavin production data with different initial regularization parameter λ_0 .

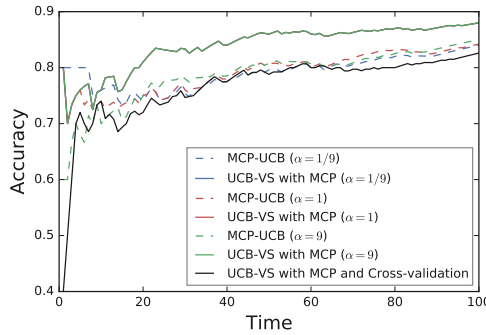


(a) UCB Algorithms with MCP

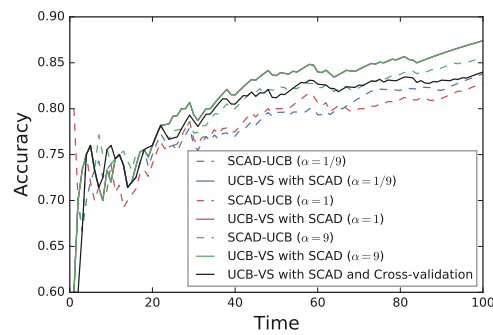


(b) UCB Algorithms with SCAD

Figure 3.6: Comparison of the cumulative regret of UCB-VS algorithms with MCP and SCAD and ℓ_1 -confidence ball based algorithm with MCP (MCP-UCB) and SCAD (SCAD-UCB) in riboflavin production data with optimal initial regularization parameter λ_0 .



(a) UCB Algorithms with MCP



(b) UCB Algorithms with SCAD

Figure 3.7: Comparison of the accuracy of UCB-VS algorithms with MCP and SCAD and ℓ_1 -confidence ball based algorithm with MCP (MCP-UCB) and SCAD (SCAD-UCB) in riboflavin production data with optimal initial regularization parameter λ_0 .

methods, e.g., MCP and SCAD, for variable selection. It may also be interesting to study the performance of other variable selection methods that possess the sequential selection consistency property (Definition 8). For example, as discussed in Section 3.3.3, the step-wise selection method [36] can be computationally more efficient, where the method includes or excludes one variable at each round until no variable can be added or deleted. The step-wise selection method needs more sophisticated analysis and can be considered for future research.

3.7 Supplementary Material

3.7.1 Proof of Theorem 3

In this section, we prove the coverage of the confidence set $\mathcal{C}_t$ for the true parameter vector β^* . We first describe a result (Proposition 4) on the eigenvalue condition for the submatrix $\hat{\Sigma}_{t,SS}$ of sample covariance matrix $\hat{\Sigma}_t$ in the bandit setting, which is necessary for accurately estimating the true parameter vector β^* . Especially, we introduce a novel method to prove the eigenvalue condition adapting to the dependence among data.

Definition 10 (Eigenvalue condition). *For a $d \times d$ matrix $A \in \mathbb{S}^{d \times d}$, define $\rho_{\min}^2(A)$ and $\rho_{\max}^2(A)$ as the minimum and maximum eigenvalues of A if*

$$\rho_{\min}^2(A) \leq \frac{\Delta^T A \Delta}{\|\Delta\|_2^2} \leq \rho_{\max}^2(A) \quad \text{for any } \Delta \neq 0.$$

Proposition 4. *For the adapted sequence $\{X_{\hat{a}_s, s} : s = 1, \dots, t\}$ induced by the bandit policy, the submatrix of sample covariance matrix restricted on true sparsity S , i.e., $\hat{\Sigma}_{t,SS} = \frac{1}{t} X_{t,S}^T X_{t,S}$, is guaranteed to satisfy the eigenvalue condition with high probability, i.e.,*

$$\begin{aligned} P\left(\rho_{\min}^2(\hat{\Sigma}_{t,SS}) \geq \frac{1}{16\zeta K}\right) &\geq 1 - \mathcal{O}\left(e^{s_0 \log(\Lambda_0^2 \zeta K)} \left[e^{-c_0 t} + e^{-\tilde{C} \log(K)t}\right]\right), \\ P\left(\rho_{\max}^2(\hat{\Sigma}_{t,SS}) \leq 2\tilde{c} \log(K) \Lambda_0^2\right) &\geq 1 - \mathcal{O}\left(e^{-\tilde{C} \log(K)t + s_0 \log(\Lambda_0^2 \zeta K)}\right), \end{aligned}$$

where $\tilde{c} = c_3 + 1/\log K$, and $c_0, c_3, \tilde{C}$ are positive constants. In addition, we can derive a uniform bound for the minimum eigenvalue condition over t that exceeds a certain threshold with high probability, i.e.,

$$P\left(\forall t \geq T', \rho_{\min}^2(\hat{\Sigma}_{t,SS}) \geq \frac{1}{16\zeta K}\right) \geq 1 - \frac{2}{T},$$

where $T' = \mathcal{O}(\log T + s_0 \log(\Lambda_0^2 \zeta K))$.

We first provide the outline of the proof for Proposition 4:

- (i) Discretize the unit sphere $S^{s_0-1} \in \mathbb{R}^{s_0}$, and prove eigenvalue condition of $\hat{\Sigma}_{t,SS}$ on a finite set $\mathcal{N}_0$.
- (ii) Prove the eigenvalue condition of $\hat{\Sigma}_{t,SS}$ for all vectors in sphere S^{s_0-1} through approximation, which implies minimum eigenvalue condition in $\mathbb{R}^{s_0}$.

Proof of Proposition 4

In this subsection, we provide the proof for Proposition 4. To begin with, we introduce the definition of ϵ -nets which are used as approximation for the unit sphere S^{d-1} in $\mathbb{R}^d$.

Definition 11. (*Nets, covering numbers*). Let (X, d) be a metric space and let $\epsilon > 0$. A subset $\mathcal{N}_\epsilon$ of X is called an ϵ -net of X if every point $x \in X$ can be approximated to within ϵ by some point $y \in \mathcal{N}_\epsilon$, i.e. so that $d(x, y) \leq \epsilon$. The minimal cardinality of an ϵ -net of X , if finite, is denoted $\mathcal{N}(X, \epsilon)$ and is called the covering number of X (at scale of ϵ).

Lemma 11 provides a bound for the cardinality of the ϵ -net that can approximate the points on the unit Euclidean sphere within range of ϵ .

Lemma 11. (*Covering numbers of the sphere*). The unit Euclidean sphere S^{d-1} equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that

$$\mathcal{N}(S^{d-1}, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^d.$$

The proof can be found in [30].

We first construct a ϵ -net for $S^{d-1} \cap E_S$, where $E_J = \text{span}\{e_j : j \in J\}$ as the subspace of $\mathbb{R}^d$ which corresponds to the covariates of set J . Here e_j is a d -dimensional vector with one at the j -th entry and zero otherwise. Let $\mathcal{N}_0$ be a ϵ -net on $S^{d-1} \cap E_S$, where set $S^{d-1} \cap E_S$ can be viewed as a unit sphere in low-dimensional space $\mathbb{R}^{s_0-1}$. According to Lemma 11, we have a bound for the covering number N of $\mathcal{N}_0$ when $\epsilon \leq 1$.

$$\log N = \log \mathcal{N}(S^{s_0-1}, \epsilon) \leq s_0 \log(1 + \frac{2}{\epsilon}) \leq s_0 [\log(\frac{1}{\epsilon}) + \log 3] \lesssim s_0 \log(\frac{1}{\epsilon}). \quad (3.16)$$

According to Assumption (b), when h is small, i.e., $h = \mathcal{O}(\frac{1}{K})$, we have for $\forall v \in \mathbb{R}^d$ that

$$P(\min_a \langle X_{a,t}, v \rangle^2 \leq h \|v\|_2^2) \leq \zeta K h.$$

Suppose we fix a vector $v \in \mathcal{N}_0$, then $\text{supp}(v) \subseteq S$ according to the definition of $\mathcal{N}_0$. Define

$$N_t = \sum_{s=1}^t \mathbb{I}_{\{\min_a \langle X_{a,s}, v \rangle^2 \leq h \|v\|_2^2\}}.$$

Here N_t is the sum of i.i.d. indicator random variables, which have expectation greater than $(1 - \zeta Kh)$. Now we define $S_t = \sum_{s=1}^t Z_s$, where

$$Z_s = \mathbb{I}_{\{\min_a \langle X_{a,s}, v \rangle^2 \leq h \|v\|_2^2\}} - \mathbb{E}[\mathbb{I}_{\{\min_a \langle X_{a,s}, v \rangle^2 \leq h \|v\|_2^2\}}]$$

are centered random variables. By Hoeffding's Lemma, we have for $\forall t \in \mathbb{N}_+$ and $\delta \in \mathbb{R}$,

$$P\left(\frac{1}{t}S_t \leq \delta\right) \leq \exp(-2\delta^2 t).$$

Let $\delta = -\frac{1}{2}(1 - \zeta Kh)$, we have

$$P\left(\frac{1}{t}N_t \leq \frac{1}{2}(1 - \zeta Kh)\right) \leq \exp(-\frac{1}{2}(1 - \zeta Kh)^2 t) = \exp(-c_0 t),$$

where $c_0 = \frac{1}{2}(1 - \zeta Kh)^2$. Taking the union of the event over all vectors in $\mathcal{N}_0$, we can obtain

$$\begin{aligned} & P\left(\forall v \in \mathcal{N}_0 : \frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, v \rangle^2 \leq \frac{1}{2}h(1 - \zeta Kh)\right) \\ & \leq N \exp(-c_0 t) \leq \exp(-c_0 t + s_0 \log\left(\frac{1}{\epsilon}\right) + s_0 \log 3). \end{aligned} \quad (3.17)$$

In the above inequalities, we have proved the minimum eigenvalue condition with high probability for all vectors in the ϵ -net $\mathcal{N}_0$. In the following, we will show that the minimum eigenvalue condition for all vectors in $S^{d-1} \cap E_S$ can also be implied by the eigenvalue condition on $\mathcal{N}_0$. Before that, we state a result on the spectral norm of symmetric matrices.

Lemma 12 (Computing the spectral norm on a net). *Let A be a symmetric $d \times d$ matrix, and let $\mathcal{N}_\epsilon$ be an ϵ -net of S^{d-1} for some $\epsilon \in [0, 1)$. The spectral norm of A can be computed via the associated quadratic form $\|A\|_2 = \sup_{x \in S^{d-1}} |\langle Ax, x \rangle|$. Then*

$$\|A\|_2 = \sup_{x \in S^{d-1}} |\langle Ax, x \rangle| \leq (1 - 2\epsilon)^{-1} \sup_{x \in \mathcal{N}_\epsilon} |\langle Ax, x \rangle|.$$

The proof of Lemma 12 can be found in Lemma 5.4 of [30].

According to the definition of $\mathcal{N}_0$, for any vector $v \in S^{s_0-1}$ and symmetric $s_0 \times s_0$ matrix A , there is a

vector $u \in \mathcal{N}_0$ such that $\|u_S - v\|_2 \leq \epsilon$, where $u_S \in \mathbb{R}^{s_0}$ is the subvector of u corresponding to S . We have the following inequality

$$\begin{aligned} v^T A v &= u_S^T A u_S + (u_S - v)^T A (u_S - v) + 2u_S^T A (v - u_S) \\ &\geq u_S^T A u_S - 2\|u_S - v\|_2 \|A\|_2 \|u_S\|_2. \end{aligned} \quad (3.18)$$

In inequality (3.17), we prove a uniform lower bound of $u^T \hat{\Sigma}_t u$ for $u \in \mathcal{N}_0$. Then to prove a lower bound of $v^T \hat{\Sigma}_t v$ for $v \in S^{d-1} \cap E_S$. Based on inequality (3.18) and the fact that $v^T \hat{\Sigma}_t v = v_S^T (\hat{\Sigma}_{t,SS}) v_S$ since $\text{supp}(v) \subseteq S$ and there exists $u \in \mathcal{N}_0$ such that $\|u_S - v\|_2 \leq \epsilon$, it suffices to show that $\|\hat{\Sigma}_{t,SS}\|_2$ is upper bounded. According to Lemma 12, we only need to bound the spectral norm of $\hat{\Sigma}_{t,SS}$ on the ϵ -net $\mathcal{N}_0$.

For any fixed vector $v \in \mathcal{N}_0$ and $\forall t \in \mathbb{N}_+$, we have

$$\frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, v \rangle^2 \leq \frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a, s}, v \rangle^2.$$

Define $Z_s^a = \langle X_{a, s}, v \rangle^2 - v^T \Sigma_a v$, $S_t^a = \sum_{s=1}^t Z_s^a$ for $a \in [K]$ and $S_t = \sum_{a=1}^K S_t^a$. Then Z_s^a are centered sub-exponential random variables. Before proving the upper bound for $\|\hat{\Sigma}_{t,SS}\|_2$, we state some properties of sub-exponential random variables. Firstly, we introduce some parameters of sub-gaussian and sub-exponential random variables.

Definition 12. The sub-gaussian norm of a sub-gaussian random variable X , denoted by $\|X\|_{\psi_2}$, is defined as

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

Similarly, the sub-exponential norm of a sub-exponential random variable X , denoted by $\|X\|_{\psi_1}$, is defined as

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}|X|^p)^{1/p}.$$

The details of the definition and related properties can be found in Section 5.2.3 and Section 5.2.4 in [30]. The following is a relationship between sub-gaussian random variables and sub-exponential random variables.

Lemma 13. (Sub-exponential is sub-gaussian squared). A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.$$

Proof of Lemma 13 can be found in Lemma 5.14 of [30].

Lemma 14. (MGF of sub-exponential random variables). Let X be a centered sub-exponential random variable. Then for t such that $|t| \leq c_1/\|X\|_{\psi_1}$, one has

$$\mathbb{E} \exp(tX) \leq \exp(c_2 t^2 \|X\|_{\psi_1}^2),$$

where $c_1, c_2 > 0$ are absolute constants.

Proof of Lemma 14 can be found in Lemma 5.15 of [30].

From the result in Lemma 13, we have for sub-exponential random variable Z_s^a that

$$\kappa_a = \|Z_s^a\|_{\psi_1} \leq 2\|\langle X_{a,s}, v \rangle\|_{\psi_1} \leq 4\|\langle X_{a,s}, v \rangle\|_{\psi_2}^2 \leq 4\lambda_{\max}^2(\Sigma_a). \quad (3.19)$$

Then for any $\mu > 0$ and $\delta > 0$, we have

$$\begin{aligned} P\left(\frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta\right) &= P\left(\sum_{i=1}^t [\max_a \langle X_{a,t}, v \rangle^2 - \Lambda_0^2] \geq \delta t\right) \\ &\leq e^{-\mu \delta t} \mathbb{E}[\exp(\mu \sum_{s=1}^t [\max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2])] \\ &= e^{-\mu \delta t} \prod_{s=1}^t \mathbb{E}[\exp(\mu [\max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2])]. \end{aligned} \quad (3.20)$$

According to Lemma 14, for μ satisfying $|\mu| \leq \frac{c_1}{\max_a \kappa_a}$, we have

$$\begin{aligned} \mathbb{E}[\exp(\mu [\max_a \langle X_{a,s}, v \rangle^2 - \Lambda^2])] &\leq \sum_{a=1}^K \mathbb{E}[\exp\{\mu (\langle X_{a,s}, v \rangle^2 - \mathbb{E}[\langle X_{a,s}, v \rangle^2])\}] \\ &\leq \sum_{a=1}^K e^{c_2 \mu^2 \kappa_a^2} \leq e^{c_2 \mu^2 \max_a \kappa_a^2 + \log K}. \end{aligned}$$

Applying above inequality to inequality (3.20), we have for μ satisfying $|\mu| \leq \frac{c_1}{\max_a \kappa_a}$ that

$$P\left(\frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta\right) \leq \exp(-\mu \delta t + t c_2 \mu^2 \max_a \kappa_a^2 + t \log K). \quad (3.21)$$

The right-hand side of the above inequality achieves the minimum value at $\mu = \frac{\delta}{2c_2 \max_a \kappa_a^2}$, where the minimum value is $\exp(-\frac{\delta^2 t}{4c_2 \max_a \kappa_a^2} + t \log K)$.

Now if $\frac{\delta}{2c_2 \max_a \kappa_a^2} > \frac{c_1}{\max_a \kappa_a}$, the right-hand side of the equation (3.21) obtain the minimum value at

$\mu = \frac{c_1}{\max_a \kappa_a}$ and

$$P \left(\frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta \right) \leq \exp \left(-\frac{c_1 \delta t}{\max_a \kappa_a} + t c_2 \frac{c_1^2}{\max_a \kappa_a^2} \max_a \kappa_a^2 + t \log K \right).$$

Moreover, since $\frac{\delta}{2c_2 \max_a \kappa_a^2} > \frac{c_1}{\max_a \kappa_a}$, we can simplify the above inequality and obtain

$$P \left(\frac{1}{t} \sum_{s=1}^t \max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta \right) \leq \exp \left(-\frac{c_1 \delta t}{2 \max_a \kappa_a} + t \log K \right).$$

Now we have for $\forall \delta > 0$ and $t \in \mathbb{N}_+$ that

$$P \left(\frac{1}{t} \sum_{s=1}^t \max_a \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta \right) \leq \exp \left(-\min \left\{ \frac{\delta^2}{4c_2 \max_a \kappa_a^2}, \frac{c_1 \delta}{2 \max_a \kappa_a} \right\} t + t \log K \right).$$

If we set $\delta = c_3 \log(K) \Lambda_0^2$ where c_3 is a sufficient large positive constant, then we have for a fixed vector $v \in \mathcal{N}_0$ that

$$\begin{aligned} & P \left(\frac{1}{t} \sum_{s=1}^t \max_{a \in [K]} \langle X_{a,s}, v \rangle^2 - \Lambda_0^2 \geq \delta \right) \\ & \leq \exp \left(-\log K \left[\min \left\{ \frac{c_3^2 \Lambda_0^4 \log K}{4c_2 \max_a \kappa_a^2}, \frac{c_1 c_3 \Lambda_0^2}{2 \max_a \kappa_a} \right\} - 1 \right] t \right) \\ & \leq \exp \left(-\tilde{C} \log(K) t \right), \end{aligned}$$

where $\tilde{C} = \min \left\{ \frac{c_3^2 \log K}{64c_2}, \frac{c_1 c_3}{8} \right\} - 1$. The last inequality is due to $\kappa_a \leq 4\lambda_{\max}^2(\Sigma_a) \leq 4\Lambda_0^2$ from inequality (3.19).

When c_3 is large, $\tilde{C} = \frac{c_1 c_3}{8} - 1$ is positive and independent of K . Taking the union of the probability over all vectors in $\mathcal{N}_0$, we have

$$\begin{aligned} & P \left(\forall v \in \mathcal{N}_0 : \langle v_S, (\hat{\Sigma}_{t,SS}) v_S \rangle \geq \tilde{c} \log(K) \Lambda_0^2 \right) \\ & = P \left(\forall v \in \mathcal{N}_0 : \langle v, \hat{\Sigma}_t v \rangle \geq \tilde{c} \log(K) \Lambda_0^2 \right) \\ & \leq N \exp \left(-\tilde{C} \log(K) t \right) \\ & = \exp \left(-\tilde{C} \log(K) t + s_0 \log\left(\frac{1}{\epsilon}\right) + s_0 \log 3 \right), \end{aligned} \tag{3.22}$$

where $\tilde{c} = c_3 + 1/\log(K)$.

Now we prove the minimum eigenvalue condition for matrix $\hat{\Sigma}_{t,SS}$. Define events $\mathcal{G}_{1,t}$ and $\mathcal{G}_{2,t}$ as

$$\begin{aligned}\mathcal{G}_{1,t} &= \{\forall u \in \mathcal{N}_0 : \frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, u \rangle \geq \frac{1}{2}h(1 - \zeta Kh)\}; \\ \mathcal{G}_{2,t} &= \{\forall u \in \mathcal{N}_0 : \langle u, \hat{\Sigma}_t u \rangle \leq \tilde{c} \log(K) \Lambda_0^2\}.\end{aligned}$$

If event $\mathcal{G}_{2,t}$ holds, according to Lemma 12, we have

$$\|\hat{\Sigma}_{t,SS}\|_2 \leq (1 - 2\epsilon)^{-1} \max_{v \in \mathcal{N}_0} \langle v, \hat{\Sigma}_t v \rangle \leq (1 - 2\epsilon)^{-1} \tilde{c} \log(K) \Lambda_0^2. \quad (3.23)$$

If both event $\mathcal{G}_{1,t}$ and $\mathcal{G}_{2,t}$ hold, based on inequality (3.18), we can prove eigenvalue condition for all vector v in unit sphere S^{s_0-1} that

$$\begin{aligned}\langle v, (\hat{\Sigma}_{t,SS})v \rangle &\geq \frac{1}{2}h(1 - \zeta Kh) - 2\epsilon \|\hat{\Sigma}_{t,SS}\|_2 \\ &\geq \frac{1}{2}h(1 - \zeta Kh) - \frac{2\epsilon}{1 - 2\epsilon} \tilde{c} \log(K) \Lambda_0^2.\end{aligned} \quad (3.24)$$

Let $h = \frac{1}{2\zeta K} = \mathcal{O}(\frac{1}{K})$ and $\epsilon = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\Lambda_0^2\zeta K \log K}\}$, then for $\forall v \in S^{s_0-1}$, we have

$$\langle v, (\hat{\Sigma}_{t,SS})v \rangle \geq \frac{1}{4}h(1 - \zeta Kh) = \frac{1}{16\zeta K}.$$

Based on the choice of ϵ and inequality (3.16), the cardinality of the ϵ -net $\mathcal{N}_0$ has the bound

$$\begin{aligned}\log N &\leq \max\{s_0 [\log (64\tilde{c}\Lambda_0^2\zeta K \log K) + \log 3], s_0 \log 12\} \\ &= \mathcal{O}(s_0 [\log K + \log \zeta + \log \Lambda_0^2]).\end{aligned}$$

Now with the choice of h and ϵ , we can prove the second inequality in Proposition 4. According to the definitions of events $\mathcal{G}_{1,t}$ and $\mathcal{G}_{2,t}$, we have

$$\begin{aligned}P\left(\rho_{\min}^2(\hat{\Sigma}_{t,SS}) \geq \frac{1}{16\zeta K}\right) &\geq P(\mathcal{G}_{1,t} \cap \mathcal{G}_{2,t}) \\ &\geq 1 - N \left[\exp(-c_0 t) + \exp(-\tilde{C} \log(K)t) \right] \\ &\geq 1 - \mathcal{O}\left(e^{s_0 \log(\Lambda_0^2 \zeta K)} [e^{-c_0 t} + e^{-\tilde{C} \log(K)t}]\right).\end{aligned}$$

Additionally, based on the definition of $\epsilon = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\Lambda_0^2\zeta K \log K}\}$ and Lemma 12, we can prove the second

inequality in Proposition 4, which bounds the maximum eigenvalue of matrix $\hat{\Sigma}_{t,SS}$.

$$\begin{aligned}
& P\left(\rho_{\max}^2(\hat{\Sigma}_{t,SS}) \geq 2\tilde{c}\log(K)\Lambda_0^2\right) \\
& \leq P\left(\forall v \in \mathcal{N}_0 : \langle v_S, (\hat{\Sigma}_{t,SS})v_S \rangle \geq \tilde{c}\log(K)\Lambda_0^2\right) \\
& \leq \exp\left(-\tilde{C}\log(K)t + s_0\log\left(\frac{1}{\epsilon}\right) + s_0\log 3\right) \\
& \leq \exp\left(-\tilde{C}\log(K)t + s_0\log(64\tilde{c}\zeta K\Lambda_0^2) + s_0\log 3\right) \\
& \leq \mathcal{O}\left(e^{-\tilde{C}\log(K)t + s_0\log(\Lambda_0^2\zeta K)}\right)
\end{aligned}$$

In the end, we prove the last inequality in Proposition 4. For any $T_1 \in \mathbb{N}_+$ we have

$$\begin{aligned}
P\left(\cup_{t \geq T_1} \mathcal{G}_{1,t}^c\right) & \leq N \sum_{t=T_1}^{\infty} e^{-c_0 t} \\
& = \frac{1}{1 - e^{-c_0}} \exp(-c_0 T_1 + \log N) \\
& \lesssim \exp(-c_0 T_1 + s_0 \log(\Lambda_0^2 \zeta K)) = \delta_2,
\end{aligned}$$

where $c_0 = \frac{1}{2}(1 - \zeta K h)^2 = \frac{1}{8}$. Let $\delta_2 = \frac{1}{T}$, then we have $T_1 = \mathcal{O}(\log T + s_0 \log(\Lambda_0^2 \zeta K))$. The above inequality provides a uniform lower bound on the minimum eigenvalue of submatrix of sample covariance matrix, i.e., $\hat{\Sigma}_{t,SS}$, over $\mathcal{N}_0$ for $t \geq T_1$. Similarly, we can obtain an union probability bound for all events $\mathcal{G}_{s,t}$ where t is greater than threshold $T_2 \in \mathbb{N}_+$.

$$\begin{aligned}
P\left(\cup_{t \geq T_2} \mathcal{G}_{2,t}^c\right) & \leq N \sum_{t=T_2}^{\infty} \exp(-\tilde{C}\log(K)t) \\
& \lesssim \exp\left(-T_2 \tilde{C}\log(K) + s_0 \log(\Lambda_0^2 \zeta K)\right) = \delta_3.
\end{aligned}$$

Let $\delta_3 = \frac{1}{T}$, then we have

$$T_2 = \frac{\log(T) + s_0 \log(\Lambda_0^2 \zeta K)}{\tilde{C}\log(K)} = \mathcal{O}(\log T + s_0 \log(\Lambda_0^2 \zeta K)).$$

With the above two union bounds, we can show that minimum eigenvalue condition holds for $\hat{\Sigma}_{t,SS}$ uniformly for all $t \geq \max(T_1, T_2) = \mathcal{O}(\log T + s_0 \log(\Lambda_0^2 \zeta K))$ with high probability, i.e.,

$$P\left(\forall t \geq \max(T_1, T_2) : \rho_{\min}^2(\hat{\Sigma}_{t,SS}) \geq \frac{1}{16\zeta K}\right) \geq 1 - \delta_2 - \delta_3 = 1 - \frac{2}{T}.$$

Proof of Theorem 3

In this subsection, we prove Theorem 3. Before the proof, we first provide the following lemmas (Lemma 15–17) for adapted sequences.

Lemma 15. (*Bernstein Concentration*) Let $\{D_k, \mathcal{G}_k\}_{k=1}^\infty$ be a martingale difference sequence, and suppose that D_k is σ -sub-Gaussian in an adapted sense, i.e., for all $\alpha \in \mathbb{R}$, the inequality $\mathbb{E}[e^{\alpha D_k} | \mathcal{G}_{k-1}] \leq e^{\alpha^2 \sigma^2 / 2}$ holds almost surely. Then, for all $t \geq 0$, $P[|\sum_{k=1}^n D_k| \geq t] \leq 2 \exp[-t^2 / 2n\sigma^2]$.

Proof of Lemma 15 follows Theorem 2.3 of [22] when $\alpha_* = \alpha_k = 0$ and $\nu_k = \sigma$ for all k .

Lemma 16. Define the event

$$\mathcal{F}(\lambda_0(\gamma)) = \{\max_{r \in [d]} (2|\omega^T X^{(r)}|/t) \leq \lambda_0(\gamma)\},$$

where $X^{(r)}$ is the r^{th} column of X and $\lambda_0(\gamma) = 2\sigma x_{\max} \sqrt{(\gamma^2 + 2 \log d)/t}$. Then we have $P[\mathcal{F}(\lambda_0(\gamma))] \geq 1 - 2 \exp[-\gamma^2/2]$.

Proof of Lemma 16 can be found in Lemma EC.2. of [4].

Lemma 17. For $\|\nabla \mathcal{L}_t(\beta^*)\|_\infty$, define $\tilde{\lambda} = \sigma x_{\max} \sqrt{(2 \log t + 2 \log d)/t}$, then we have

$$P\left(\|\nabla \mathcal{L}_t(\beta^*)\|_\infty \leq \tilde{\lambda}\right) \geq 1 - \frac{2}{t}.$$

Proof. According to the definition of $\mathcal{L}_t(\beta) = \frac{1}{2t} \|Y_t - X_t \beta\|_2^2$ in (3.6), we have

$$\nabla \mathcal{L}_t(\beta^*) = \frac{1}{t} X_t^T (Y_t - X_t \beta^*) = \frac{1}{t} \omega_t^T X_t,$$

where $\omega_t \in \mathbb{R}^t$ are vectors of σ -sub-Gaussian noise. Thus, the lemma can be derived from Lemma 16. $\square$

In Lemma 18, we demonstrate that Algorithm 2 will select the optimal arm at a positive fraction of times.

Lemma 18. Suppose we construct the confidence set $\mathcal{C}_t$ in Algorithm 2 with initial parameter τ_0 , then when $t \geq \max\left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2(d, T)\right)$ with constant χ_g , policy π in Theorem 3 will select the optimal arms at a positive fraction of times. The constant χ_g is defined as $\chi_g = \mathcal{O}\left(\sigma x_{\max}^3 \zeta K s_0^{5/2} \phi_0^{-1} \sqrt{M_1(d)/\log d}\right)$.

Proof. We denote design matrix $X_t \in \mathbb{R}^{t \times d}$, reward vector $Y_t \in \mathbb{R}^t$ and noise vector $\omega_t \in \mathbb{R}^t$. According to Definition 8, $\hat{S}_t = \mathcal{V}_t(\mathcal{F}_t)$ is (M_0, t_0, M_1, M_2) -sequential selection consistent. Define constant $C_g = \sqrt{M_1(d)/\log d}$ and $\varepsilon_t = C_g \sqrt{(\log d + \log t)/t}$. Then, for set $I_{\varepsilon_t} = \{i \in [d] : |\beta_i^*| > \varepsilon_t\}$ (Definition 8) and time

$t \geq M_2(d, T)$, we have $t_{\varepsilon_t} = \max\left(\frac{M_1(d)}{\varepsilon_t^2}, M_2(d, T)\right) \leq t$. This implies that for $I_{\varepsilon_t} = \{i \in [d] : |\beta_i^*| > \varepsilon_t\}$ and $t \geq \max(t_0, M_2(d, T))$,

$$P\left(I_{\varepsilon_t} \subseteq \hat{S}_t \subseteq S\right) \geq 1 - 2M_0/t. \quad (3.25)$$

From the optimization problem (3.3), we have OLS estimator $\hat{\beta}_t^{ols}$ which is supported on $\hat{S}_t$, then the subvector $\hat{\beta}_{t, \hat{S}_t}^{ols} \in \mathbb{R}^{|\hat{S}_t|}$ can be expressed as

$$\hat{\beta}_{t, \hat{S}_t}^{ols} = (X_{t, \hat{S}_t}^T X_{t, \hat{S}_t})^{-1} X_{t, \hat{S}_t}^T Y. \quad (3.26)$$

Define $\hat{V}_t = X_{t, \hat{S}_t}^T X_{t, \hat{S}_t} / t \in \mathbb{R}^{|\hat{S}_t| \times |\hat{S}_t|}$, then we can bound the weighted ℓ_2 -norm between $\hat{\beta}_{\hat{S}_t}$ and $\beta_{\hat{S}_t}^*$.

$$\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^* = (X_{t, \hat{S}_t}^T X_{t, \hat{S}_t})^{-1} X_{t, \hat{S}_t}^T Y_t - \beta_{\hat{S}_t}^* \quad (3.27)$$

$$\begin{aligned} &= \beta_{\hat{S}_t}^* + (X_{t, \hat{S}_t}^T X_{t, \hat{S}_t})^{-1} X_{t, \hat{S}_t}^T X_{t, S \setminus \hat{S}_t} \beta_{S \setminus \hat{S}_t}^* + (X_{t, \hat{S}_t}^T X_{t, \hat{S}_t})^{-1} X_{t, \hat{S}_t}^T \omega - \beta_{\hat{S}_t}^* \\ &= (X_{t, \hat{S}_t}^T X_{t, \hat{S}_t})^{-1} X_{t, \hat{S}_t}^T X_{t, S \setminus \hat{S}_t} \beta_{S \setminus \hat{S}_t}^* + (X_{t, \hat{S}_t}^T X_{t, \hat{S}_t})^{-1} X_{t, \hat{S}_t}^T \omega_t. \end{aligned} \quad (3.28)$$

Then we have

$$\|\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^*\|_{\hat{V}_t} \leq \underbrace{\|(X_{t, \hat{S}_t}^T X_{t, \hat{S}_t})^{-1} X_{t, \hat{S}_t}^T X_{t, S \setminus \hat{S}_t} \beta_{S \setminus \hat{S}_t}^*\|_{\hat{V}_t}}_{J_1} + \underbrace{\|(X_{t, \hat{S}_t}^T X_{t, \hat{S}_t})^{-1} X_{t, \hat{S}_t}^T \omega_t\|_{\hat{V}_t}}_{J_2}. \quad (3.29)$$

According to (3.25), the event $\{I_{\varepsilon_t} \subseteq \hat{S}_t \subseteq S\}$ holds with high probability. From Proposition 4, the inequality $\rho_{\min}^2(X_{t, \hat{S}_t}^T X_{t, \hat{S}_t} / t) \geq \rho_{\min}^2(\hat{\Sigma}_{t, SS}) \geq \frac{1}{16\zeta K}$ holds with high probability. This implies that $\rho_{\max}^2(\hat{V}_t^{-1}) \leq 16\zeta K$ with high probability. The first term J_1 can then be bounded as

$$\begin{aligned} J_1 &= \left\| \frac{1}{t} X_{t, \hat{S}_t}^T X_{t, S \setminus \hat{S}_t} \beta_{S \setminus \hat{S}_t}^* \right\|_{\hat{V}_t^{-1}} \\ &\leq \rho_{\max}(\hat{V}_t^{-1}) \left\| \frac{1}{t} X_{t, \hat{S}_t}^T X_{t, S \setminus \hat{S}_t} \beta_{S \setminus \hat{S}_t}^* \right\|_2 \\ &\leq \rho_{\max}(\hat{V}_t^{-1}) \left\| \frac{1}{t} X_{t, \hat{S}_t}^T X_{t, S \setminus \hat{S}_t} \beta_{t, S \setminus \hat{S}_t}^* \right\|_1 \\ &\leq \rho_{\max}(\hat{V}_t^{-1}) \left\| \frac{1}{t} X_{t, \hat{S}_t}^T X_{t, S \setminus \hat{S}_t} \right\|_1 \|\beta_{S \setminus \hat{S}_t}^*\|_1 \\ &\leq 4\sqrt{\zeta K} |\hat{S}_t| x_{\max}^2 \|\beta_{S \setminus \hat{S}_t}^*\|_1 \\ &\leq 4\sqrt{\zeta K} |\hat{S}_t| x_{\max}^2 |S \setminus \hat{S}_t| \varepsilon_t, \end{aligned}$$

where $\varepsilon_t = C_g \sqrt{(\log d + \log t)/t}$. The last inequality uses the bound for the ℓ_1 -norm of matrices, i.e.,

$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$ for matrix $A = (a_{ij})_{i=1, \dots, m, j=1, \dots, n}$. The second term J_2 can be bounded as

$$\begin{aligned}
J_2 &= \left\| \frac{1}{t} X_{t, \hat{S}_t}^T \omega_t \right\|_{\hat{V}_t^{-1}} \\
&\leq \rho_{\max}(\hat{V}_t^{-1}) \left\| \frac{1}{t} X_{t, \hat{S}_t}^T \omega_t \right\|_2 \\
&\leq \rho_{\max}(\hat{V}_t^{-1}) \sqrt{|\hat{S}_t|} \left\| \frac{1}{t} X_{t, \hat{S}_t}^T \omega_t \right\|_\infty \\
&\leq 4\sqrt{s_0 \zeta K} \left\| \frac{1}{t} X_t^T \omega_t \right\|_\infty \\
&\leq 4\sqrt{s_0 \zeta K} \left\| \frac{1}{t} X_t^T \omega_t \right\|_\infty.
\end{aligned}$$

According to Lemma 17, $\left\| \frac{1}{t} X_t^T \omega_t \right\|_\infty \leq \tilde{\lambda}$ with probability greater than $1 - 2/t$, where $\tilde{\lambda}$ takes the value as $\tilde{\lambda} = \sigma x_{\max} \sqrt{(2 \log t + 2 \log d)/t}$. Consequently, term J_2 can be bounded by $4\tilde{\lambda}\sqrt{s_0 \zeta K}$ with high probability.

Combining the bounds for J_1 and J_2 together, we have

$$\begin{aligned}
\|\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^*\|_{\hat{V}_t} &\leq J_1 + J_2 \\
&\leq 4\sqrt{\zeta K} |\hat{S}_t| x_{\max}^2 |S \setminus \hat{S}_t| \varepsilon_t + 4\tilde{\lambda} \sqrt{s_0 \zeta K} \\
&\leq 4\sqrt{\zeta K} [x_{\max}^2 s_0^2 \varepsilon_t + \tilde{\lambda} \sqrt{s_0}] \\
&\leq 4\sqrt{\zeta K} \underbrace{[x_{\max}^2 s_0^2 C_g + \sigma x_{\max} \sqrt{2s_0}]}_{\bar{C}_g} \sqrt{\frac{\log d + \log t}{t}}
\end{aligned}$$

Now, we define event $\mathcal{G}_{3, t+1}$ as below:

$$\mathcal{G}_{3, t+1} = \left\{ \|\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^*\|_{\hat{V}_t} \leq 4\sqrt{\zeta K} \bar{C}_g \sqrt{\frac{\log d + \log t}{t}} \right\}.$$

According to the above argument, when $t \geq \max(t_0, M_2(d, T))$, event $\mathcal{G}_{3, t+1}$ holds with high probability, i.e.,

$$P(\mathcal{G}_{3, t+1}) \geq 1 - \mathcal{O}\left(M_0/t + e^{s_0 \log(\Lambda_0^2 \zeta K)} \left[e^{-c_0 t} + e^{-\bar{C} \log(K)t}\right]\right).$$

Suppose we fix a feature vector $x \in \mathbb{R}^d$. For any $\tilde{\beta} \in \mathcal{C}_{t+1}$, $\text{supp}(\tilde{\beta}) \subseteq \hat{S}_t$ based on the construction of $\mathcal{C}_{t+1}$.

Then we have

$$\begin{aligned}
|\langle x, \tilde{\beta} - \beta^* \rangle| &\leq |\langle x_{\hat{S}_t}, \tilde{\beta}_{\hat{S}_t} - \beta_{\hat{S}_t}^* \rangle| + |\langle x_{S \setminus \hat{S}_t}, \beta_{S \setminus \hat{S}_t}^* \rangle| \\
&\leq |\langle x_{\hat{S}_t}, \tilde{\beta}_{\hat{S}_t} - \hat{\beta}_{t, \hat{S}_t}^{ols} \rangle| + |\langle x_{\hat{S}_t}, \hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^* \rangle| + |\langle x_{S \setminus \hat{S}_t}, \beta_{S \setminus \hat{S}_t}^* \rangle| \\
&\leq \|x_{\hat{S}_t}\|_{\hat{V}_t^{-1}} \left\{ \|\tilde{\beta}_{\hat{S}_t} - \hat{\beta}_{t, \hat{S}_t}^{ols}\|_{\hat{V}_t} + \|\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^*\|_{\hat{V}_t} \right\} + x_{\max} \varepsilon_t |S \setminus \hat{S}_t| \\
&\leq 4x_{\max} \sqrt{\zeta K s_0} \left\{ \tau_t + 4\sqrt{\zeta K \bar{C}_g} \sqrt{\frac{\log d + \log t}{t}} \right\} + x_{\max} s_0 C_g \sqrt{\frac{\log d + \log t}{t}}.
\end{aligned}$$

According to the definition of $\tau_t = \tau_0 \sqrt{\frac{|\hat{S}_t|(\log |\hat{S}_t| + \log t)}{t}} \leq \tau_0 \sqrt{\frac{s_0(\log s_0 + \log t)}{t}}$, we have

$$|\langle x, \tilde{\beta} - \beta^* \rangle| \leq \chi_g \sqrt{\frac{\log d + \log t}{t}},$$

where the constant χ_g is given by

$$\begin{aligned}
\chi_g &= 4x_{\max} \sqrt{\zeta K s_0} \{ \tau_0 \sqrt{s_0} + 4\sqrt{\zeta K \bar{C}_g} \} + x_{\max} s_0 C_g \\
&= \mathcal{O} \left(\sigma x_{\max}^3 \zeta K s_0^{5/2} \phi_0^{-1} \sqrt{M_1(d)/\log d} \right).
\end{aligned}$$

In addition, sub-optimal arms $X_{a,t+1}$ can only be selected when

$$\max_{\beta \in \mathcal{C}_{t+1}} \langle X_{a,t+1}, \beta \rangle \geq \max_{\beta \in \mathcal{C}_{t+1}} \langle X_{a_{t+1}^*, t+1}, \beta \rangle.$$

Combined with the bound on $|\langle x, \tilde{\beta} - \beta^* \rangle|$, this implies that we can select sub-optimal arm $X_{a,t+1}$ only when

$$\langle X_{a_{t+1}^*, t+1} - X_{a,t+1}, \beta^* \rangle \leq 2\chi_g \sqrt{\frac{\log d + \log t}{t}}.$$

According to Assumption (b), the event $\Gamma_t = \{ \langle X_{a_t^*, t}, \beta^* \rangle > \max_{b \neq a_t^*} \langle X_{b,t}, \beta^* \rangle + \Delta_* \}$ holds with probability $P(\Gamma_t) \geq \frac{1}{2}$. Then, if both events Γ_{t+1} and $\mathcal{G}_{3,t+1}$ hold, policy π will always select the optimal arm a_{t+1}^* when

$$\Delta_* \geq 2\chi_g \sqrt{\frac{\log d + \log t}{t}}.$$

By solving the above inequality, we have that for $t \geq \max \left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2(d, T) \right)$, policy π will always select the optimal arm a_t^* when both event Γ_{t+1} and $\mathcal{G}_{3,t+1}$ hold. Therefore,

$$P(\pi_t = a_t^*) \geq P(\Gamma_t \cap \mathcal{G}_{3,t+1}) \geq \frac{1}{2} - \mathcal{O} \left(M_0/t + e^{s_0 \log(\Lambda_0^2 \zeta K)} \left[e^{-c_0 t} + e^{-\tilde{C} \log(K)t} \right] \right).$$

□

In order to prove Theorem 3, we now provide a refined result on the minimum eigenvalue condition for $\hat{\Sigma}_{t,SS}$. In Proposition 4, the minimum eigenvalue of $\hat{\Sigma}_{t,SS}$ scales as $\Theta(1/K)$, which will lead to the dependence on K in the leading term of the regret upper bound. In Proposition 5, we show that the minimum eigenvalue of $\hat{\Sigma}_{t,SS}$ is independent of the number of arms K by applying the result from Lemma 18.

Proposition 5. *Suppose we construct the confidence set $\mathcal{C}_{t+1}$ according to Algorithm 2 with some constant τ_0 . Then when $t \geq 2 \max\left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2\right)$, the submatrix of sample covariance matrix restricted on true sparsity S , i.e., $\hat{\Sigma}_{t,SS} = \frac{1}{t} X_{t,S}^T X_{t,S}$, is guaranteed to satisfy the minimum eigenvalue condition uniformly with high probability, i.e.,*

$$P\left(\rho_{\min}^2(\hat{\Sigma}_{t,SS}) \geq \frac{\phi_0^2}{32}\right) \geq 1 - \mathcal{O}\left(M_0/t + e^{s_0 \log(\Lambda_0^2 \zeta K)} \left[e^{-c_0 t} + e^{-\tilde{C} \log(K)t}\right]\right).$$

Proof. We consider time t such that $t \geq T'' = \max\left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2(d, T)\right)$. From Lemma 18, it is shown that we will select the optimal arm at a positive fraction of times with high probability. Suppose $t \geq 2T''$ and $v \in S^{s_0-1}$, then we have

$$\begin{aligned} \frac{1}{t} \sum_{s=1}^t \langle (X_{\hat{a}_s, s})_S, v \rangle^2 &\geq \frac{1}{t} \sum_{s=T''+1}^t \langle (X_{\hat{a}_s, s})_S, v \rangle^2 \\ &\geq \frac{1}{t} \sum_{s=T''+1}^t \langle (X_{\hat{a}_s, s})_S, v \rangle^2 \mathbb{1}_{\{\Gamma_s \cap \mathcal{G}_{3,s}\}} \\ &= \frac{1}{t} \sum_{s=T''+1}^t \langle (X_{a_s^*, s})_S, v \rangle^2 \mathbb{1}_{\{\Gamma_s\}} \mathbb{1}_{\{\mathcal{G}_{3,s}\}}. \end{aligned} \quad (3.30)$$

Since event $\mathcal{G}_{3,t}$ defined in Lemma 18 only depends on the history up to time $t-1$, the random variables $\{X_{a_t^*, t} | \mathcal{G}_{3,t}\}$ are i.i.d. Moreover, since $\{\mathbb{1}_{\{\Gamma_s\}} : s \geq T''\}$ are i.i.d Bernoulli random variables, then according to Hoeffding's Lemma, we have

$$\sum_{s=T''+1}^t \mathbb{1}_{\{\Gamma_s\}} \leq \frac{1}{2} \cdot \frac{1}{2} (t - T'') \quad \text{w.p.} \leq \mathcal{O}(e^{-c_4 t}), \quad (3.31)$$

where c_4 is an absolute constant. Next, we consider the sum of $\langle (X_{a_s^*, s})_S, v \rangle^2$ over times when event Γ_s holds

$$\sum_{s: \Gamma_s \text{ holds}} \langle (X_{a_s^*, s})_S, v \rangle^2 \mathbb{1}_{\Gamma_s} - \mathbb{E} [\langle (X_{a_s^*, s})_S, v \rangle^2 \mathbb{1}_{\Gamma_s}].$$

Define $Z_s = \langle (X_{a_s^*, s})_S, v \rangle^2 \mathbb{1}_{\Gamma_s} - \mathbb{E} [\langle (X_{a_s^*, s})_S, v \rangle^2 \mathbb{1}_{\Gamma_s}]$, where Z_s are i.i.d. sub-exponential random variables.

Moreover, $\mathbb{E}[\langle (X_{a_s^*, s})_S, v \rangle^2 \mathbb{1}_{\Gamma_t}] \geq \phi_0^2 \|v\|_2^2$ according to Assumption 4(c). Moreover, according to Lemma 13, we have condition on event Γ_t that

$$\|Z_s\|_{\psi_1} \leq 2\|\langle (X_{a_s^*, s})_S, v \rangle^2 \mathbb{1}_{\Gamma_s}\|_{\psi_1} \leq 4\|\langle (X_{a_s^*, s})_S, v \rangle \mathbb{1}_{\Gamma_s}\|_{\psi_2}^2 \leq 4\Lambda_1^2.$$

Applying Bernstein's inequality for sub-exponential random variables, we have a lower bound for a fixed vector v that

$$\begin{aligned} & P\left(\left|\sum_{s:\Gamma_s \text{ holds}} Z_s\right| \geq \frac{\phi_0^2}{2} \cdot m \mid \sum_{s=T''+1}^t \mathbb{1}_{\{\Gamma_s\}} = m\right) \\ &= 2 \exp\left(-\tilde{c}_5 \min\left(\frac{\phi_0^4}{4\|Z_s\|_{\psi_1}^2}, \frac{\phi_0^2}{2\|Z_s\|_{\psi_1}}\right)m\right) = \mathcal{O}(e^{-c_5 m}), \end{aligned} \quad (3.32)$$

where $\tilde{c}_5, c_5$ are constants. Combining inequalities (3.30) – (3.32), we have for any v in ϵ' -net $\mathcal{N}_3$ for $S^{d-1} \cap E_S$, where $E_S = \text{span}\{e_j : j \in S\}$ and ϵ' will be specified below, and $t \geq 2T''$ that

$$\frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, v \rangle^2 \geq \frac{1}{t} \sum_{s=T''+1}^t \langle X_{a_s^*, s}, v \rangle^2 \mathbb{1}_{\{\Gamma_s\}} \geq \frac{1}{t} \frac{\phi_0^2}{2} \cdot \frac{(t - T'')}{2} \|v\|_2^2 \geq \frac{\phi_0^2}{16},$$

with probability greater than $1 - \mathcal{O}(e^{-c_4(t-T'')} + e^{-c_5 \cdot \frac{1}{4}(t-T'')})$.

Similar to the proof for Proposition 4, we can extend the above inequality to all vectors in $S^{d-1} \cap E_S$ through the approximation of ϵ' -net $\mathcal{N}_3$, where $\epsilon' = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\Lambda_1^2 K \log K}\} = \mathcal{O}(\frac{1}{\log K})$. Then for any $v \in S^{d-1} \cap E_S$ and $t \geq 2T''$, we have

$$\langle v, \hat{\Sigma}_t v \rangle \geq \frac{\phi_0^2}{32}.$$

with probability greater than

$$1 - \mathcal{O}(e^{s_0 \log(\Lambda_1^2 \zeta K)} [e^{-c_4(t-T'')} + e^{-c_5 \cdot \frac{1}{4}(t-T'')})] = 1 - \mathcal{O}(e^{-t+s_0 \log(\Lambda_1^2 \zeta K)}).$$

Then combined with the probability of event $\mathcal{G}_{3,t}$, we can prove the Proposition 5 for $t \geq 2T''$, i.e.,

$$P\left(\rho_{\min}^2(\hat{\Sigma}_{t, SS}) \geq \frac{\phi_0^2}{32}\right) \geq 1 - \mathcal{O}(M_0/t) - \mathcal{O}\left(e^{s_0 \log(\Lambda_0^2 \zeta K)} \left[e^{-c_0 t} + e^{-\tilde{C} \log(K)t}\right]\right).$$

□

Theorem 3 can then be derived from Lemma 18 and Proposition 5. We assume the time t satisfies $t \geq 2 \max\left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2(d, T)\right)$. According to Proposition 5, the inequality $\rho_{\min}^2(\hat{\Sigma}_{t, SS}) \geq \frac{\phi_0^2}{32}$ holds with

high probability. Applying this property to (3.29) in the proof of Lemma 18, we have

$$\begin{aligned}
J_1 &\leq \rho_{\max}(\hat{V}_t^{-1}) \left\| \frac{1}{t} X_{t, \hat{S}_t}^T X_{t, S \setminus \hat{S}_t} \right\|_1 \|\beta_{S \setminus \hat{S}_t}^*\|_1 \\
&\leq \frac{4\sqrt{2}}{\phi_0} |\hat{S}_t| x_{\max}^2 \|\beta_{S \setminus \hat{S}_t}^*\|_1 \\
&\leq \frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 \|\beta_{S \setminus \hat{S}_t}^*\|_1.
\end{aligned}$$

Also,

$$\begin{aligned}
J_2 &\leq \rho_{\max}(\hat{V}_t^{-1}) \sqrt{|\hat{S}_t|} \left\| \frac{1}{t} X_{t, \hat{S}_t}^T \omega_t \right\|_{\infty} \\
&\leq \frac{4\sqrt{2}}{\phi_0} \sqrt{|\hat{S}_t|} \left\| \frac{1}{t} X_{t, \hat{S}_t}^T \omega_t \right\|_{\infty} \\
&\leq \frac{4\sqrt{2}}{\phi_0} \sqrt{|\hat{S}_t|} \cdot \sigma x_{\max} \sqrt{(2 \log t + 2 \log |\hat{S}_t|)/t}.
\end{aligned}$$

Thus, it can be derived that

$$\begin{aligned}
\|\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^*\|_{\hat{V}_t} &\leq J_1 + J_2 \\
&\leq \frac{8\sigma x_{\max}}{\phi_0} \sqrt{\frac{|\hat{S}_t|(\log |\hat{S}_t| + \log t)}{t}} + \frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 \|\beta_{S \setminus \hat{S}_t}^*\|_1 \\
&= \tau_0 \sqrt{\frac{|\hat{S}_t|(\log |\hat{S}_t| + \log t)}{t}} + \frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 \|\beta_{S \setminus \hat{S}_t}^*\|_1.
\end{aligned}$$

The last equality is due to the definition of $\tau_0 = \frac{8\sigma x_{\max}}{\phi_0}$ in Theorem 3. Thus, the statement in Theorem 3 is proved.

3.7.2 Proof of Theorem 4

Define $T' = 2 \max\left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2(d, T)\right)$, then the cumulative regret can be divided into three groups:

- (a) Initialization when $t \leq T'$.
- (b) Time $t > T'$ when $\rho_{\min}^2(\hat{\Sigma}_{t, SS}) < \phi_*^2$, where $\phi_*^2 = \frac{\phi_0^2}{32}$.
- (c) Time $t > T'$ when $\rho_{\min}^2(\hat{\Sigma}_{t, SS}) \geq \phi_*^2$, where $\phi_*^2 = \frac{\phi_0^2}{32}$.

The cumulative regret I_1 from time periods in group (a) at time T is bounded by at most $2bx_{\max}T' = 4bx_{\max} \max\left(\frac{4\chi_g^2 \log d}{\Delta_*^2}, t_0, M_2(d, T)\right)$. According to Proposition 5, the cumulative regret in group (b) at time

T can be bounded as

$$\begin{aligned} I_2 &= \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}\{\rho_{\min}^2(\hat{\Sigma}_{t,SS}) < \frac{\phi_0^2}{32}\}] \leq \sum_{t=T''}^T 2bx_{\max} P\left(\rho_{\min}^2(\hat{\Sigma}_{t,SS}) < \frac{\phi_0^2}{32}\right) \\ &\lesssim \sum_{t=T''}^T 2bx_{\max} \left[\frac{M_0}{t} + e^{s_0 \log(\Lambda_0^2 \zeta K)} \left(e^{-c_0 t} + e^{-\tilde{C} \log(K)t} \right) \right] = \mathcal{O}(x_{\max} b \log T). \end{aligned}$$

To prove the upper bound of the cumulative regret in group (c), we define event Γ_{ξ_t} as

$$\Gamma_{\xi_t} = \{\omega \in \Omega : \langle X_{a_t^*, t}, \beta^* \rangle \geq \langle X_{b, t}, \beta^* \rangle + \xi_t, \forall b \neq a_t^*\},$$

where ξ_t will be specified momentarily. In addition, we define event $A_t = \{\rho_{\min}^2(\hat{\Sigma}_{t,SS}) \geq \phi_*^2 = \frac{\phi_0^2}{32}\}$ for simplicity. Then, the cumulative regret at time T in group (c) can be written as:

$$I_3 = \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_{t-1} \cap \Gamma_{\xi_t}\}}] + \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_{t-1} \cap \Gamma_{\xi_t}^c\}}].$$

Similar to the proof in Lemma 18, for each observed feature vector x at time $t+1$, i.e., $X_{a, t+1}$ and the corresponding $\tilde{\beta}$, i.e., $\tilde{\beta}_{a, t+1}$ within the confidence set $\mathcal{C}_{t+1}$, we have

$$\begin{aligned} |\langle x, \tilde{\beta} - \beta^* \rangle| &\leq |\langle x_{\hat{S}_t}, \tilde{\beta}_{\hat{S}_t} - \beta_{\hat{S}_t}^* \rangle| + |\langle x_{S \setminus \hat{S}_t}, \beta_{S \setminus \hat{S}_t}^* \rangle| \\ &\leq |\langle x_{\hat{S}_t}, \tilde{\beta}_{\hat{S}_t} - \hat{\beta}_{t, \hat{S}_t}^{ols} \rangle| + |\langle x_{\hat{S}_t}, \hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^* \rangle| + |\langle x_{S \setminus \hat{S}_t}, \beta_{S \setminus \hat{S}_t}^* \rangle| \\ &\leq \|x_{\hat{S}_t}\|_{\hat{\Sigma}_{t, \hat{S}_t, \hat{S}_t}^{-1}} \left\{ \|\tilde{\beta}_{\hat{S}_t} - \hat{\beta}_{t, \hat{S}_t}^{ols}\|_{\hat{\Sigma}_{t, \hat{S}_t, \hat{S}_t}} + \|\hat{\beta}_{t, \hat{S}_t}^{ols} - \beta_{\hat{S}_t}^*\|_{\hat{\Sigma}_{t, \hat{S}_t, \hat{S}_t}} \right\} + x_{\max} \|\beta_{S \setminus \hat{S}_t}^*\|_1 \\ &\leq \frac{2x_{\max} s_0 \tau_0}{\phi_*} \sqrt{\frac{\log s_0 + \log t}{t}} + \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right) \|\beta_{S \setminus \hat{S}_t}^*\|_1. \end{aligned}$$

Here $\hat{\Sigma}_{t, \hat{S}_t, \hat{S}_t}$ is the submatrix of $\hat{\Sigma}_t$, which restricted on the columns of X_t belonging to $\hat{S}_t$. Thus, we have inequality $\|x_{\hat{S}_t}\|_{\hat{\Sigma}_{t, \hat{S}_t, \hat{S}_t}^{-1}} \leq \frac{1}{\phi_*} \|x_{\hat{S}_t}\|_2 \leq \frac{1}{\phi_*} \sqrt{s_0} \|x_{\hat{S}_t}\|_{\infty} \leq \frac{x_{\max} \sqrt{s_0}}{\phi_*}$. Now we set

$$\xi_t = 2 \left[\frac{2x_{\max} s_0 \tau_0}{\phi_*} \sqrt{\frac{\log s_0 + \log t}{t}} + \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right) \|\beta_{S \setminus \hat{S}_t}^*\|_1 \right].$$

When both Γ_{ξ_t} and A_{t-1} hold, we have for any $k \neq a_t^*$ that

$$\max_{\beta \in \mathcal{C}_{t-1}} \langle X_{a_t^*, t}, \beta \rangle - \max_{\beta \in \mathcal{C}_{t-1}} \langle X_{k, t}, \beta \rangle \geq \langle X_{a_t^*, t}, \beta^* \rangle - \langle X_{k, t}, \beta^* \rangle - \xi_t \geq 0.$$

Thus, policy π will always select the optimal arm a_t^* when both events A_t and Γ_{ξ_t} hold, and the first part of

the cumulative regret in I_3 will be zero under such events.

We consider the time when both A_{t-1} and $\Gamma_{\xi_t}^c$ hold. If a sub-optimal arm $k \neq a_t^*$ is selected, the regret incurred will be

$$\langle X_{a_t^*,t} - X_{k,t}, \beta_* \rangle \leq \max_{\beta \in \mathcal{C}_{t-1}} \langle X_{a_t^*,t}, \beta \rangle - \max_{\beta \in \mathcal{C}_{t-1}} \langle X_{k,t}, \beta \rangle + \xi_t \leq \xi_t.$$

Moreover, according to Assumption 3(b), we have

$$P \left(\langle X_{a_t^*,t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b,t}, \beta_* \rangle \leq \xi_t \right) \leq \frac{\xi_t}{2\Delta_*}.$$

Therefore, the second term in I_3 can be bounded as

$$\begin{aligned} & \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_{t-1} \cap \Gamma_{\xi_t}^c\}}] \leq \sum_{t=T''}^T \mathbb{E}[\xi_t \mathbb{1}_{\{\Gamma_{\xi_t}^c\}}] \\ &= \sum_{t=T''}^T \xi_t P \left(\langle X_{a_t^*,t}, \beta_* \rangle - \max_{b \neq a_t^*} \langle X_{b,t}, \beta_* \rangle \leq \xi_t \right) \\ &\leq \sum_{t=T''}^T \frac{1}{2\Delta_*} \xi_t^2 \\ &\leq \frac{2}{\Delta_*} \sum_{t=T''}^T \left(\frac{2x_{\max}s_0\tau_0}{\phi_*} \sqrt{\frac{\log s_0 + \log t}{t}} + \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right) \|\beta_{S \setminus \hat{S}_t}^*\|_1 \right)^2 \\ &\leq \frac{4}{\Delta_*} \sum_{t=T''}^T \left\{ \left(\frac{2x_{\max}s_0\tau_0}{\phi_*} \sqrt{\frac{\log s_0 + \log t}{t}} \right)^2 + \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right)^2 \|\beta_{S \setminus \hat{S}_t}^*\|_1^2 \right\} \\ &= \frac{4}{\Delta_*} \sum_{t=T''}^T \left\{ \frac{4x_{\max}^2 s_0^2 \tau_0^2 (\log s_0 + \log t)}{\phi_*^2 t} + \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right)^2 \|\beta_{S \setminus \hat{S}_t}^*\|_1^2 \right\}. \end{aligned}$$

The selection-policy pair $(\mathcal{V}, \pi)$ satisfies (M_0, t_0, M_1, M_2) -sequentially selection consistency property. According to the proof of Theorem 3, for $t > T'$ and $\varepsilon_t = C_g \sqrt{(\log d + \log t)/t}$, where $C_g = \sqrt{M_1(d)/\log d}$, the event $\{I_{\varepsilon_t} \subseteq \hat{S}_t \subseteq S\}$ holds with high probability. This implies that for $j \in S \setminus \hat{S}_t$,

$$|\beta_j^*| < \varepsilon_t = C_g \sqrt{\frac{\log d + \log t}{t}} \iff t \lesssim \frac{C_g^2 \log d}{|\beta_j^*|^2},$$

Therefore, we have the following bound.

$$\begin{aligned}
I_3 &= \sum_{t=T''}^T \mathbb{E}[r_t \mathbb{1}_{\{A_t \cap \Gamma_{\xi_t}^c\}}] \\
&\leq \sum_{t=T''}^T \mathbb{E}[\xi_t \mathbb{1}_{\{\Gamma_{\xi_t}^c\}}] \\
&\leq \frac{4}{\Delta_*} \sum_{t=T''}^T \left[\frac{4x_{\max}^2 s_0^2 \tau_0^2}{\phi_*^2} \frac{\log s_0 + \log t}{t} + \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right)^2 \|\beta_{S \setminus \hat{S}_t}^*\|_1^2 \right] \\
&\leq \frac{16x_{\max}^2 s_0^2 \tau_0^2}{\Delta_* \phi_*^2} [\log s_0 + \log T] \log T + \frac{4}{\Delta_*} \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right)^2 \sum_{t=T''}^T \sum_{j \in S \setminus \hat{S}_t} |\beta_j^*|^2 \\
&\leq \frac{16x_{\max}^2 s_0^2 \tau_0^2}{\Delta_* \phi_*^2} [\log s_0 + \log T] \log T + \frac{4}{\Delta_*} \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right)^2 \sum_{j \in S} \frac{C_g^2 \log d}{|\beta_j^*|^2} |\beta_j^*|^2 \\
&\leq \frac{512x_{\max}^2 s_0^2 \tau_0^2}{\Delta_* \phi_0^2} [\log s_0 + \log T] \log T + \frac{4s_0 M_1(d)}{\Delta_*} \left(\frac{4\sqrt{2}}{\phi_0} s_0 x_{\max}^2 + x_{\max} \right)^2 \\
&= \mathcal{O}(s_0^2 (\log s_0 + \log T) \log T + s_0^3 M_1(d)).
\end{aligned}$$

Combining with the above result, we can prove the regret upper bound in Theorem 4.

3.7.3 Proof of Theorem 5

We will apply the Bayesian decision theory to show the lower bound of the cumulative regret up to horizon time T . To prove the $\log T$ term in Theorem 5, we can follow the procedure in [19]. For simplicity, we will focus on the term in lower bound involving the dimension d .

Suppose we have two arms, the first of which is $X^{(1)} = (X_0, X_1, \dots, X_d)$ and the second is $X^{(2)} = (0, -X_1, \dots, -X_d)$. Here the two arms are not independent from each other. Entry X_0 in the arm vector follows a discrete distribution which will be specified later. The other entries $\{X_1, \dots, X_d\}$ are i.i.d. truncated normal distribution of $N(0, 1)$. For simplicity, we first consider the standard normal case, later the proof can be easily applied to the truncated normal case. The common parameter vector is β is from a set $\mathcal{M} = \{\beta \in \mathbb{R}^{d+1} : \beta_0 = 1, \beta_u = \theta, \beta_j = 0 \text{ for } j \neq u, u \in [d]\}$. So the sparsity of parameters in set $\mathcal{M}$ is $s_0 = 2$. Within the set $\mathcal{M}$, the first entry of the parameter vector is assumed to be known as 1, so we do not need to estimate it. The remaining d entries of the parameter vector has exactly one nonzero entry with value θ . We assume the parameter vector β is uniformly distributed within the set $\mathcal{M}$. Here to make sure the assumptions in section 3.4.1 hold for the configuration of the parameters in $\mathcal{M}$ and data, we let $\beta_{\min} = \sqrt{\frac{\log d}{T}}$, $\theta = c\beta_{\min}$ where c is sufficiently small, and

$$\begin{cases} P(X_0 = 0) = C\beta_{\min} \\ P(X_0 = a) = \frac{1}{2}(1 - C\beta_{\min}), \quad \text{when } a = \pm 1 \end{cases} \quad (3.33)$$

Here constant C can be chosen such that Assumption (b) can be satisfied.

$$\begin{aligned} P(|\beta^T(X^{(1)} - X^{(2)})| < h) &= P(|X_0 + 2\theta X_u| < h) \\ &\leq P(|X_0 + 2\theta X_u| < h, X_0 = 0) + 2P(|X_0 + 2\theta X_u| < h, X_0 = 1) \\ &\leq C\beta_{\min}P(|2\theta X_u| < h) + (1 - C\beta_{\min})P(|1 + 2\theta X_u| < h). \end{aligned}$$

Since X_u is truncated normal, the density of X_u is bounded. Then, we can choose C such that there exists Δ_* and Assumption (b) is satisfied.

Define $\mathcal{P}$ as the bandit environment where all assumptions hold, and $\mathcal{P}_{\mathcal{M}}$ be the set of environment defined above with β from $\mathcal{M}$. We have $\mathcal{P}_{\mathcal{M}} \subseteq \mathcal{P}$ according to the configuration of $\mathcal{P}_{\mathcal{M}}$. Then for any policy π , the supreme of the cumulative regret of π at horizon T within $\mathcal{P}$ can be lower bounded by:

$$\begin{aligned} &\sup \{R_T(\pi, \pi^*) : P_{X,Y} \in \mathcal{P}\} \\ &\geq \sup_{P_{X,Y} \in \mathcal{P}_{\mathcal{M}}} \mathbb{E} \sum_{t=1}^n |\beta^T \Delta_{x,t}| [I \{ \beta^T \Delta_{x,t} \geq 0, \pi_t = 2 \} + I \{ \beta^T \Delta_{x,t} < 0, \pi_t = 1 \}] \\ &\geq \mathbb{E} \left\{ \sum_{t=1}^n (\mathbb{E}_{\beta} [\beta^T \Delta_{x,t} I \{ \beta^T \Delta_{x,t} \geq 0 \} | \mathcal{F}_{t-1}^+] I \{ \pi_t = 2 \} \right. \\ &\quad \left. - \mathbb{E}_{\beta} [\beta^T \Delta_{x,t} I \{ \beta^T \Delta_{x,t} < 0 \} | \mathcal{F}_{t-1}^+] I \{ \pi_t = 1 \}) \right\}, \end{aligned} \quad (3.34)$$

where $P_{X,Y} = (P_{X^{(1)}, Y^{(1)}}, P_{X^{(2)}, Y^{(2)}})$ as the joint distribution of feature vectors and response variable, and $\Delta_{x,t} = X_t^{(1)} - X_t^{(2)} = (\Delta_{0,t}, \dots, \Delta_{d,t})$. Here $\mathbb{E}_{\beta}[\cdot | \mathcal{F}_{t-1}^+]$ denotes the expectation of β in $\mathcal{M}$ conditioned on $\mathcal{F}_{t-1}^+$, which is the σ -algebra generated by feature vectors up to time t and observed response variable Y up to time $t-1$. More precisely, $\mathcal{F}_{t-1}^+ = \sigma(X_1^{(1)}, X_1^{(2)}, Y_1^{(\pi_1)}, \dots, X_{t-1}^{(1)}, X_{t-1}^{(2)}, Y_{t-1}^{(\pi_t)}, X_t^{(1)}, X_t^{(2)})$. Then according to the Bayesian decision rule in [19], the optimal policy is $\hat{\pi}_t = 1$ if

$$\begin{aligned} \mathbb{E}_{\beta} [\beta^T \Delta_{x,t} I \{ \beta^T \Delta_{x,t} \geq 0 \} | \mathcal{F}_{t-1}^+] &\geq -\mathbb{E}_{\beta} [\beta^T \Delta_{x,t} I \{ \beta^T \Delta_{x,t} < 0 \} | \mathcal{F}_{t-1}^+] \\ \iff \mathbb{E}_{\beta} [\beta^T \Delta_{x,t} | \mathcal{F}_{t-1}^+] &\geq 0 \end{aligned} \quad (3.35)$$

and $\hat{\pi}_t = 2$ otherwise. Set $\hat{\beta}_{t-1} = \mathbb{E} [\beta | \mathcal{F}_{t-1}^+]$, since $\Delta_{x,t}$ is independent from $\mathcal{F}_{t-1}$ and β , then $\mathbb{E} [\beta | \mathcal{F}_{t-1}^+] =$

$\mathbb{E}[\beta|\mathcal{F}_{t-1}]$, where $\mathcal{F}_{t-1} = \sigma(\sigma(X_1^{(1)}, X_1^{(2)}, Y_1^{(\pi_1)}, \dots, X_{t-1}^{(1)}, X_{t-1}^{(2)}, Y_{t-1}^{(\pi_t)}))$ is the σ -algebra generated by feature vectors and observed responses up to time $t-1$. Then the Bayesian policy is $\hat{\pi} = \{\hat{\pi}_t : t \in [T]\}$, where $\hat{\pi}_t = I \left\{ \hat{\beta}_{t-1}^T \Delta_{x,t} \geq 0 \right\} + 2I \left\{ \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right\}$. If we define $u \in [d]$ as the location of the nonzero entry θ in parameter vector β , then according to the distribution of β within $\mathcal{M}$, u is uniform distributed in $[d]$. Then we have,

$$\begin{aligned}
& \sup \{R_T(\pi, \pi^*) : P_{X,Y} \in \mathcal{P}\} \\
& \geq \sup \{R_T(\hat{\pi}, \pi^*) : P_{X,Y} \in \mathcal{P}_{\mathcal{M}}\} \\
& \geq \sum_{t=1}^T \mathbb{E}_{\beta} \mathbb{E}_{X,Y|\beta} \left[|\beta^T \Delta_x| I \left\{ \text{sign}(\beta^T \Delta_{x,t}) \neq \text{sign}(\hat{\beta}_{t-1}^T \Delta_{x,t}) \right\} \right] \\
& = \sum_{t=1}^T \mathbb{E}_{\beta,u} \mathbb{E} \left[|X_{0,t} + 2\theta X_{u,t}| I \left\{ \text{sign}(\beta^T \Delta_{x,t}) \neq \text{sign}(\hat{\beta}_{t-1}^T \Delta_{x,t}) \right\} \right] \\
& \geq \sum_{t=1}^T \mathbb{E}_{\beta,u} P(X_0 = 0) \mathbb{E}_{X,Y|\beta} \left[|2\theta X_{u,t}| I \left\{ \text{sign}(\beta^T \Delta_{x,t}) \neq \text{sign}(\hat{\beta}_{t-1}^T \Delta_{x,t}) \right\} | X_0 = 0 \right] \\
& \geq \sum_{t=1}^T \mathbb{E}_{\beta,u} P(X_0 = 0) \frac{1}{4} \mathbb{E}_{X,Y|\beta} \left[|2\theta X_{u,t}| I \left\{ \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right\} | X_0 = 0, \beta^T \Delta_{x,t} \geq 0 \right] \\
& \geq \sum_{t=1}^T \mathbb{E}_{\beta,u} \frac{1}{2} C c \beta_{\min}^2 \mathbb{E}_{X,Y|\beta} \left[I \left\{ 1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right\} | X_0 = 0, \beta^T \Delta_{x,t} \geq 0 \right] \\
& = \sum_{t=1}^T \frac{1}{2} C c \beta_{\min}^2 \mathbb{E}_{\beta,u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right]. \tag{3.36}
\end{aligned}$$

The third inequality above is by the conditional expectation with respect to the event $\{X_0 = 0\}$, and the fourth inequality is by the symmetric distribution of $\beta^T \Delta_{x,t}$ when $X_0 = 0$. The last inequality is by the fact that when $X_{u,t} \in (1, 2)$, then $|\theta X_{u,t}| \geq c\beta_{\min}$.

Now it suffices to lower bound the expectation term in the last equation, which has

$$\begin{aligned}
& \mathbb{E}_{\beta,u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right] \\
& = \mathbb{E}_{\beta,u} \left[P(1 < X_{u,t} < 2) P_{X,Y|\beta} \left(\hat{\beta}_{u,t-1} X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 | 1 < X_{u,t} < 2 \right) \right] \\
& \geq \tilde{C} P \left(\hat{\beta}_{u,t-1} X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 | 1 < X_{u,t} < 2 \right). \tag{3.37}
\end{aligned}$$

Here $\hat{\beta}_{-u,t-1}, X_{-u,t} \in \mathbb{R}^{d-1}$ is the corresponding vector without the 0^{th} and u^{th} entries. And $\tilde{C} = P(1 < X_{u,t} < 2)$ is positive, since $X_{u,t}$ is standard normal by definition.

In order to lower bound the probability in (3.37), we use the Fano's inequality to reduce the probability into the false discovery rate in multiple hypothesis testing problem. Firstly, we define a testing function

$\hat{u}_{t-1} = \arg \max_{j \in [d]} |\hat{\beta}_{j,t-1}|$. Then we will show that the probability in (3.37) is lower bounded by $P(\hat{u}_{t-1} \neq u)$. In fact, if event $\{\hat{u}_{t-1} \neq u\}$ holds, then we have $|\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2$ by the definition of $\hat{u}_{t-1}$.

$$\begin{aligned}
& P\left(\hat{\beta}_{u,t-1}X_{u,t} + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2, |\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2\right) \\
& \geq P\left(|2\hat{\beta}_{u,t-1}| + \hat{\beta}_{-u,t-1}^T X_{-u,t} < 0 \mid 1 < X_{u,t} < 2, |\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2\right) \\
& = P\left(2 + N(0, \|\hat{\beta}_{-u,t-1}\|_2^2 / |\hat{\beta}_{u,t-1}|^2) < 0 \mid 1 < X_{u,t} < 2, |\hat{\beta}_{u,t-1}|^2 \leq \|\hat{\beta}_{-u,t-1}\|_2^2\right) \\
& \geq P(N(2, 1) < 0) = C'
\end{aligned} \tag{3.38}$$

The last equation is since X_t is independent from $\hat{\beta}_{t-1}$, and conditioned on $\hat{\beta}_{t-1}$, we have $\hat{\beta}_{-u,t-1}^T X_{-u,t}$ is also normal distribution with mean 0 and variance $\|\hat{\beta}_{-u,t-1}\|_2^2$. Then applying the above inequalities into (3.37), we have

$$\mathbb{E}_{\beta,u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right] \geq \tilde{C}C'P(\hat{u}_{t-1} \neq u). \tag{3.39}$$

Now, it suffices to lower bound $P(\hat{u}_{t-1} \neq u)$ for each time t . We first state a variant of Fano lower bound in multiple hypothesis testing, the details of which can be found in Chapter 15 of [22].

Proposition 6. *Assume that U is uniform on $\mathcal{U}$. For any Markov chain $U \rightarrow (X, Y) \rightarrow \hat{U}$,*

$$P(\hat{U} \neq U) \geq 1 - \frac{I(X, Y; U) + \log 2}{\log(|\mathcal{U}|)}. \tag{3.40}$$

Here $I(X, Y; U)$ is the mutual information between (X, Y) and U , and $|\mathcal{U}|$ is the cardinality of set $\mathcal{U}$.

In our case, since u is uniform in $[d]$, we have $\log(|\mathcal{U}|) = \log d$. Then we only need to upper bound the mutual information $I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u)$ to make sure the probability $P(\hat{u}_{t-1} \neq u)$ in (3.39) is bounded away from 0. Here $X_{\{1:t-1\}}$ and $Y_{\{1:t-1\}}$ represent the feature set and response set up to time $t-1$.

By the chain rule of mutual information, we have

$$I(X, Y; u) = I(X; u) + I(Y; u|X) = I(Y; u|X). \tag{3.41}$$

The last equality is by the independence between X and the uniformly distributed u .

Moreover, according to [22], we can bound the conditional mutual information of Y and u conditional on X by the Kullback-Leibler divergence, which is

$$I(Y; u|X) \leq \frac{1}{|\mathcal{U}|^2} \sum_{j,k=1}^{|\mathcal{U}|} D(P_{\beta^j|X} \| P_{\beta^k|X}), \tag{3.42}$$

where $P_{\beta^j|X}$ is the distribution of Y conditioned on X and parameter vector $\beta^j \in \mathcal{M}$ corresponding to $u^j \in \mathcal{U}$.

Then we have for the KL-divergence of normal location models $P_{\beta^j|X} = N(X\beta^j, \sigma^2)$ and $P_{\beta^k|X} = N(X\beta^k, \sigma^2)$ that

$$D(P_{\beta^j|X} \| P_{\beta^k|X}) = \frac{1}{2\sigma^2} \|X(\beta^j - \beta^k)\|_2^2 = \frac{1}{2\sigma^2} \|(X_j - X_k)\theta\|_2^2 \leq \frac{1}{2\sigma^2} \theta^2 4x_{\max}^2 = \frac{2x_{\min}^2}{\sigma^2} \cdot \frac{c^2 \log d}{T}. \quad (3.43)$$

Summing up the mutual information up to time $t-1$, and according to (3.41), we have

$$I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u) = I(Y_{\{1:t-1\}}; u) \leq \sum_{s=1}^{t-1} \frac{2x_{\min}^2}{\sigma^2} \cdot \frac{c^2 \log d}{T} \leq \frac{2x_{\max}^2 c^2}{\sigma^2} \log d. \quad (3.44)$$

Then we have for c small enough that $\frac{2x_{\max}^2 c^2}{\sigma^2} \leq \frac{1}{2}$ and $\log d \geq 4 \log 2$, we have

$$\begin{aligned} P(\hat{u}_{t-1} \neq u) &\geq 1 - \frac{I(X_{\{1:t-1\}}, Y_{\{1:t-1\}}; u) + \log 2}{\log |\mathcal{U}|} \\ &\geq 1 - \frac{\frac{1}{2} \log d + \log 2}{\log d} \geq \frac{1}{4}. \end{aligned} \quad (3.45)$$

Now applying equations (3.45) and (3.39) back to (3.36), we have

$$\begin{aligned} &\sup \{R_T(\pi, \pi^*) : P_{X,Y} \in \mathcal{P}\} \\ &\geq \sum_{t=1}^T \frac{1}{2} C c \beta_{\min}^2 \mathbb{E}_{\beta, u} \left[P_{X,Y|\beta} \left(1 < X_{u,t} < 2, \hat{\beta}_{t-1}^T \Delta_{x,t} < 0 \right) \right] \\ &\geq \sum_{t=1}^T \frac{1}{2} C c \beta_{\min}^2 \tilde{C} C' P(\hat{u}_{t-1} \neq u) \\ &\geq \frac{\tilde{C} C' C c}{8} \log d. \end{aligned}$$

Now we prove the lower bound for any policy π .

$$\sup \{R_T(\pi, \pi^*) : P_{X,Y} \in \mathcal{P}\} \geq \tilde{C}_L \log d, \quad (3.46)$$

where $\tilde{C}_L$ is some positive constant.

Combined with a variant result of [19] for the $\log T$ lower bound, we prove the lower bound in Theorem 5.

3.7.4 Proof of Theorem 6

In this section, we provide the proof for Theorem 6. Let $\nabla^2 \mathcal{L}_t(\beta)$ denote the Hessian matrix of loss function $\mathcal{L}_t(\beta)$, then

$$\Gamma = \nabla^2 \mathcal{L}_t(\beta^*), \quad Q(\beta) = \int_0^1 \nabla^2 \mathcal{L}_t(\beta^* + u(\beta - \beta^*)) du.$$

In order to prove the theorem, we will apply the results (Lemma 19) from [34] on support recovery properties of the regularized estimators. In the following, we will prove Lemma 19 for the bandit settings, which implies the selection consistency of the regularized estimators as described in Lemma 20.

We first display the assumptions on the loss function $\mathcal{L}_t(\beta)$ and regularization function $\rho(\beta)$ which are required for the analysis of Lemma 19. In Lemma 19, it is assumed that the loss function $\mathcal{L}_t(\beta)$ possesses three properties: *restricted strong convexity* (RSC), *locally bounded gradient* (LBG), and *locally bounded second-order gradient* (LBSG). We introduce the definitions of these properties, where set S_1 and S_2 are defined as in Theorem 6.

Assumption 6 (Restricted strong convexity). *We say $\mathcal{L}_t$ is (α, τ) -RSC if it satisfies the following inequalities for any pair of $\beta, \Delta \in \mathbb{R}^d$:*

$$\langle \nabla \mathcal{L}_t(\beta + \Delta) - \nabla \mathcal{L}_t(\beta), \Delta \rangle \geq \begin{cases} \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log d}{t} \|\Delta\|_1^2, & \forall \|\Delta\|_2 \leq 1 \\ \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log d}{t}} \|\Delta\|_1, & \forall \|\Delta\|_2 \geq 1 \end{cases} \quad (3.47)$$

Assumption 7 (Locally bounded gradient). *We say $\mathcal{L}_t$ is (ν, λ) -LBG if for any vector $\beta \in \mathbb{R}$ with $\text{supp}(\beta) = S_2$ and $\|\beta - \beta^*\|_\infty \leq \nu$, the following inequalities are satisfied:*

$$\|(\nabla \mathcal{L}_t(\beta) - \nabla \mathcal{L}_t(\beta^*))_{S_1}\|_\infty \leq \tilde{K} \|\nabla \mathcal{L}_t(\beta^*)\|_\infty \leq \lambda - \|\nabla \mathcal{L}_t(\beta^*)\|_\infty, \quad (3.48)$$

where $\tilde{K}$ is some positive constant. Here β and β^* only differ at most s_0 entries since both are zero on S^c .

Assumption 8 (Locally bounded second-order gradient). *If S_2 is not a null set, we say $\mathcal{L}_t$ is ν -LBSG if for any vector $\beta \in \mathbb{R}^d$ with $\text{supp}(\beta) = S_2$ and $\|\beta - \beta^*\|_\infty \leq \nu$, the following inequality holds:*

$$2\|(\Gamma_{S_2 S_2})^{-1}\|_\infty \|\Gamma_{S_2 S_2} - Q(\beta)_{S_2 S_2}\|_\infty \leq 1. \quad (3.49)$$

If S_2 is a null set, we say locally bounded second-order gradient is satisfied.

Assumptions 6–8 are all in deterministic sense. In order to ensure that these assumptions are satisfied

with high probability in bandit settings, we need to carefully choose the scale of the parameters λ , r and r_1 , which is highly dependent on the gradient/subgradient of the loss function $\mathcal{L}_t(\beta)$ and the matrix norm of Γ and $Q(\beta)$.

In the following, we describe the main results from [34] on the estimation accuracy of the regularized estimator $\hat{\beta}_t$ solving optimization problem (3.5). The results are stated deterministically.

Lemma 19 ([34], Theorem 3.1). *Suppose $\mathcal{L}_t$ is a twice-differentiable, RSC with parameters $(\alpha_1, \tau_1, \alpha_2, \tau_2)$, (r_1, λ) -LBG and r -LBSG function, and regularization function $\rho(\cdot)$ satisfies Assumption 5 with parameters (λ, ϵ_0, u) . Further assume:*

(a) *The parameter λ satisfies*

$$\max \left\{ 4\|\nabla \mathcal{L}_t(\beta^*)\|_\infty, 4\alpha_2 \sqrt{\frac{\log s_0}{t}} \right\} \leq \lambda \leq \sqrt{\frac{(4\alpha_1 - 3u)\alpha_2}{384s_0}}.$$

(b) *The parameters (R, u) satisfy*

$$\begin{cases} u < \alpha_1, \\ \max \left\{ 2\|\beta^*\|_1, \|\beta^*\|_1 + s_0 r, \frac{48s_0\lambda}{4\alpha_1 - 3u} \right\} \leq R \leq \min \left\{ \frac{\alpha_2}{8\lambda}, \frac{\alpha_2}{\tau_2} \sqrt{\frac{t}{\log d}} \right\}. \end{cases}$$

(c) *For any vector $\tilde{\beta} \in \mathbb{R}^d$ with $\text{supp}(\tilde{\beta}) = S$ and $\|\tilde{\beta}_S - \beta_S^*\|_\infty \leq r$, the following condition holds for some $\delta \in [\frac{4R\tau_1 \log d}{t\lambda}, 1]$:*

$$\| -\nabla \mathcal{L}_t(\beta^*)_{S^c} + \tilde{Q}_{S^c S}(\tilde{Q}_{SS})^{-1} \nabla \mathcal{L}_t(\beta^*)_S \|_\infty \leq (1 - \delta - \epsilon)\lambda, \quad (3.50)$$

where $\epsilon\lambda = \|\tilde{Q}_{S^c S}(\tilde{Q}_{SS})^{-1}\|_\infty \max(\epsilon_0, \tilde{K}\|\nabla \mathcal{L}_t(\beta^*)\|_\infty)$, and $\tilde{Q} = Q(\tilde{\beta})$. Then (3.5) has a unique stationary point $\hat{\beta}_t$ such that

$$\|\hat{\beta}_t - \beta^*\|_\infty \leq r_1,$$

if sample size satisfies $t \geq \frac{2\tau_1}{2\alpha_1 - u} s_0 \log d$. Moreover, $\text{supp}(\hat{\beta}_t) \subseteq S$.

Before we provide the proof for Lemma 19 in bandit settings, we show that the regularized estimator $\hat{\beta}_t$ possesses the selection consistency property for significant parameters that belong to set S_2 .

Lemma 20. *Suppose conditions in Lemma 19 hold and $t \geq \frac{2\tau_1}{2\alpha_1 - u} s_0 \log d$, then*

$$S_2 \subseteq \hat{S}_t = \text{supp}(\hat{\beta}_t) \subseteq S.$$

Proof. The second inclusion is proved in Lemma 19, so we only need to prove the first inclusion. From Lemma 19, the estimator $\hat{\beta}_t$ satisfies $\|\hat{\beta}_t - \beta^*\|_\infty \leq r_1$. Then, according to the definition of $S_2 = \{i \in [d] : |\beta_i^*| > r_1\}$

in Theorem 6, we have for all $i \in S_2$ that

$$|\hat{\beta}_i| \geq |\beta_i^*| - \|\hat{\beta}_t - \beta^*\|_\infty > r_1 - r_1 = 0.$$

This implies that $i \in \hat{S}_t$ for all $i \in S_2$. Thus, it can be derived that $S_2 \subseteq \hat{S}_t$. $\square$

Lemma 20 demonstrates the selection consistency of the regularized estimator $\hat{\beta}_t$ for significant parameters in set S_2 when t exceeds some threshold, i.e., when $t \geq \frac{2\tau_1}{2\alpha_1 - u} s_0 \log d$.

The inequality (3.14) in Theorem 6 can be derived from Lemma 19 and Lemma 20. The crucial part of the proof is to show that the assumptions in Lemma 19 are satisfied in bandit settings with high probability, for which we will provide details in Section 3.7.4.

Proof Sketch for Lemma 19

In this section, for the completeness of the proof, we provide an outline of the proof for Lemma 19 from [34] and display some critical lemmas for the analysis. [34] proves Theorem 19 in deterministic sense, therefore the results in Lemma 19 can be applied to bandit settings when the assumptions in the lemma are satisfied. The outline of the proof for Theorem 19 is as below:

- (i) Prove the existence and uniqueness of the optimization problem constrained on subset S_2 , which is defined in Theorem 6.

$$\arg \min_{\beta_{(S_2)^c} = 0, \|\beta\|_1 \leq R} \left\{ \mathcal{L}_t(\beta) + \sum_{i=1}^d \rho(\beta_i) \right\}. \quad (3.51)$$

Construct a solution set $\mathcal{A}$ for this restricted optimization problem,

$$\mathcal{A} = \{\beta \in \mathbb{R}^d : \beta_{(S_2)^c} = 0, \nabla \mathcal{L}_t(\beta) + \nabla \rho(\beta) = 0, \|\beta\|_1 \leq R\} \quad (3.52)$$

- (ii) Use Lemma 21 to show that the restricted problem (3.51) has a unique solution $\tilde{\beta}$.
- (iii) Use Lemma 22 and Lemma 23 to show that the solution $\tilde{\beta}$ satisfies $\|\tilde{\beta}_{S_2} - \beta_{S_2}^*\|_\infty \leq r$ and is equal to

$$\arg \min_{\beta_{S^c} = 0, \|\beta\|_1 \leq R} \left\{ \mathcal{L}_t(\beta) + \sum_{j=1}^d \rho(\beta_j) \right\}. \quad (3.53)$$

- (iv) Establish that $\tilde{\beta}$ is the unique solution of the original optimization problem (3.5) by showing that $\tilde{\beta}$

satisfies the following zero-subgradient condition,

$$\nabla \mathcal{L}_t(\tilde{\beta}) + \nabla \rho(\tilde{\beta}) = 0. \quad (3.54)$$

To establish (3.54), define $\tilde{z}_S \in \partial \|\tilde{\beta}_S\|_1$ and choose $\hat{z}_{S^c}$ to satisfy the following:

$$\nabla \mathcal{L}_t(\tilde{\beta}) + \left(\nabla \rho(\tilde{\beta}) - \lambda \nabla \|\tilde{\beta}\|_1 \right) + \lambda \tilde{z} = 0,$$

where $\tilde{z} = (\tilde{z}_S, \tilde{z}_{S^c})$. Then use Lemma 24 to check the *strict dual feasibility* of $\tilde{z}_{S^c}$, i.e., $\|\tilde{z}_{S^c}\|_\infty < 1 - \delta$ for some $\delta > 0$.

Lemma 21. *Consider any (α, τ) -RSC function $\mathcal{L}_t$ and regularization function (3.10) with $u < \alpha_1$. If the sample size satisfies $t \geq \frac{2\tau_1}{2\alpha_1 - u} s_0 \log d$ and β^* is s_0 -sparse, then the restricted program (3.51) is strictly convex and has a unique stationary point.*

The proof can be found in Lemma 6.1 of [34].

Lemma 22. *Assume conditions in Lemma 21 hold. Further assume that the regularization function satisfies $\rho'(r_1 - r) = \epsilon_0$ and $\mathcal{L}_t$ satisfies r -LBSG condition, i.e.,*

$$2\|(\Gamma_{SS})^{-1}\|_\infty \|\Gamma_{SS} - \tilde{Q}_{SS}\|_\infty \leq 1, \quad (3.55)$$

where $\tilde{Q} = Q(\tilde{\beta}) = \int_0^1 \nabla^2 \mathcal{L}_t(\beta^* + t(\tilde{\beta} - \beta^*)) dt$, then $\|\tilde{\beta}_{S_2} - \beta_{S_2}^*\|_\infty \leq r$.

The proof can be found in Lemma 6.2 of [34]. Lemma 21 and Lemma 22 together prove the uniqueness of the solution to problem (3.51) and property that $\|\tilde{\beta}_{S_2} - \beta_{S_2}^*\|_\infty \leq r$.

Lemma 23. *Suppose $\mathcal{L}_t$ is strictly convex in the subspace of $\mathbb{R}^d$ corresponding to S and $\rho(\cdot)$ is a regularization function (3.10), if $\mathcal{L}_t$ is (r_1, λ) -LBG, then $\tilde{\beta}$, the solution to the restricted optimization problem (3.51), is the unique solution to (3.53).*

Lemma 23 shows that the solution $\tilde{\beta}$ for problem (3.51) is actually the unique solution for problem (3.53). Thus, we have that the restricted optimization problem solution $\tilde{\beta}$ is in fact a solution for problem (3.53). The following Lemma shows that it is the solution for the original problem (3.5).

Lemma 24. *Suppose conditions in Lemma 21–23 hold and $\mathcal{L}_t$ is (r_1, λ) -LBG. Then $\tilde{\beta}$ is the unique solution to the original problem (3.5), provided that*

(a) The parameters (R, u, λ) satisfy the bounds

$$\begin{aligned} \max \left\{ 2\|\beta^*\|_1, \frac{48s_0\lambda}{4\alpha_1-3u} \right\} &\leq R \leq \min \left\{ \frac{\alpha_2}{8\lambda}, \frac{\alpha_2}{\tau_2} \sqrt{\frac{t}{\log d}} \right\}, u < \frac{4}{3}\alpha_1, \text{ and} \\ 4 \max \left\{ \|\nabla \mathcal{L}_t(\beta^*)\|_\infty, \alpha_2 \sqrt{\frac{\log s_0}{t}} \right\} &\leq \lambda \leq \sqrt{\frac{(4\alpha_1-3u)\alpha_2}{384s_0}}. \end{aligned}$$

(b) For some $\delta \in [\frac{4R\tau_1 \log d}{t\lambda}, 1]$, the following condition holds

$$\| -\nabla \mathcal{L}_t(\beta^*)_{S^c} + \tilde{Q}_{S^c S}(\tilde{Q}_{SS})^{-1} \nabla \mathcal{L}_t(\beta^*)_S \|_\infty \leq (1 - \delta - \epsilon)\lambda, \quad (3.56)$$

where $\epsilon\lambda = \|\tilde{Q}_{S^c S}(\tilde{Q}_{SS})^{-1}\|_\infty \max(\epsilon_0, \tilde{K}r_1)$.

The proof of Theorem 19 is deterministic and can be found in [34].

Based on the above Lemma 21–24, Lemma 19 can be proved deterministically. In order to apply the results from Lemma 19 to prove Theorem 6 in bandit settings, we need to prove that all assumptions in Lemma 19 are satisfied with high probability according to the construction of parameters in Theorem 6. These assumptions are challenging to prove since the data are dependent due to the sequential decision-making process involved, while the standard statistical analysis techniques are for i.i.d. data. In the following sections, we will justify the assumptions in Lemma 19 in bandit settings.

Proof of Assumptions in Lemma 19

In this section, we prove that the assumptions required for Lemma 19 are satisfied in high probability with the choices of parameters in Lemma 6. Since we are considering the assumptions for all time t , in order to be coherent with Lemma 19, we denote parameter $\lambda = \lambda_t = C_L \sqrt{\frac{\log d + \log t}{t}}$.

Proof of Assumption 8: In the assumption, $\mathcal{L}_t$ needs to be r -LBSG in the linear bandit settings. Since the loss function $\mathcal{L}_t(\beta) = \frac{1}{2t} \|Y_t - X_t \beta\|_2^2$ is a quadratic function with respect to β , the second-order derivative of $\mathcal{L}_t(\beta)$ is

$$\nabla^2 \mathcal{L}_t(\beta) = \frac{1}{t} X_t^T X_t,$$

which is independent of β . Therefore, $\Gamma = Q(\beta) = \frac{1}{t} X_t^T X_t$ for all β and $\mathcal{L}_t(\beta)$ satisfies the r -LBSG assumption in linear bandit settings with probability one, i.e.

$$P(\{\mathcal{L}_t(\beta) \text{ is } r\text{-LBSG}\}) = 1.$$

Proof of Assumption 7: We show that the loss function $\mathcal{L}_t(\beta)$ is (r_1, λ) -LBG with high probability. For the left-hand side of (3.48), we can upper bound it as following:

$$\begin{aligned} \|(\nabla \mathcal{L}_t(\beta) - \nabla \mathcal{L}_t(\beta^*))_{S_1}\|_\infty &\leq \left\| \frac{1}{t} (X_t^T X_t (\beta - \beta^*))_S \right\|_\infty \\ &\leq \left\| \frac{1}{t} X_{t,S}^T X_{t,S} \right\|_\infty \|\beta - \beta^*\|_\infty \\ &\leq r_1 \left\| \frac{1}{t} X_{t,S}^T X_{t,S} \right\|_\infty, \end{aligned} \quad (3.57)$$

where $X_{t,S}$ is the submatrix of X_t corresponding to the columns of X_t within subset S . The second inequality is because that both β and β^* are supported within S . In order to prove Assumption 7 for Lemma 19, it suffices to show that

$$r_1 \left\| \frac{1}{t} X_{t,S}^T X_{t,S} \right\|_\infty + \|\nabla \mathcal{L}_t(\beta^*)\|_\infty \leq \lambda.$$

In Lemma 17, it is shown that $\|\nabla \mathcal{L}_t(\beta^*)\|_\infty \leq \tilde{\lambda} = \sigma x_{\max} \sqrt{(2 \log t + 2 \log d)/t}$ with probability greater than $1 - 2/t$. In addition, we recall Proposition 4, which indicates that the maximum eigenvalue of matrix $\frac{1}{t} X_{t,S}^T X_{t,S}$ is upper bounded with high probability. In particular, we have

$$P\left(\rho_{\max}^2(\hat{\Sigma}_{t,SS}) \leq 2\tilde{c} \log(K) \Lambda_0^2\right) \geq 1 - \mathcal{O}\left(e^{-\tilde{C} \log(K)t + s_0 \log(\Lambda_0^2 \zeta K)}\right).$$

This implies that $\left\| \frac{1}{t} X_{t,S}^T X_{t,S} \right\|_\infty \leq \sqrt{s_0} \left\| \frac{1}{t} X_{t,S}^T X_{t,S} \right\|_2 \leq 2\tilde{c} \sqrt{s_0} \log(K) \Lambda_0^2$ (Lemma 12) with high probability. Consequently, based on the definition of $r_1 = C_g \sqrt{(\log d + \log t)/t}$ in Theorem 6 and condition (a)–(c) in Lemma 19, it suffices to define $\lambda_t = C_L \sqrt{(\log d + \log t)/t}$ with a sufficiently large constant C_L . For example, we can define $C_L = 2\tilde{c} C_g \Lambda_0^2 \sqrt{s_0} \log K + 4\sqrt{2} \sigma x_{\max}$. Thus, Assumption 7 with high probability, i.e.,

$$P(\{\mathcal{L}_t(\beta) \text{ is } (r_1, \lambda)\text{-LBG}\}) \geq 1 - 2/t - \mathcal{O}\left(e^{-\tilde{C} \log(K)t + s_0 \log(\Lambda_0^2 \zeta K)}\right).$$

Proof of Assumption 6: In the remaining of this section, we prove that Assumption 6 on $\mathcal{L}_t$ holds with high probability. In the original problem (3.5), β is restricted such that $\|\beta\|_1 \leq R$, and $\|\beta^*\|_1 \leq \frac{R}{2}$ according to Lemma 19. Define $\Delta = \beta - \beta^*$, then we have $\|\Delta\|_1 \leq 2R$. In the below, we introduce Lemma 25 from [32], which demonstrates that it suffices to prove inequality (3.47) only for Δ such that $\|\Delta\|_2 \leq 1$.

Lemma 25 ([32], Lemma 8). *Suppose $\mathcal{L}_t$ is convex. If condition (3.47) holds for $\|\Delta\|_2 \leq 1$ and $t \geq 4R^2 \tau_1^2 \log d$,*

then

$$\langle \nabla \mathcal{L}_t(\beta + \Delta) - \nabla \mathcal{L}_t(\beta), \Delta \rangle \geq \alpha_1 \|\Delta\|_2 - \sqrt{\frac{\log d}{t}} \|\Delta\|_1, \quad \text{for all } \|\Delta\|_2 \geq 1.$$

The proof of Lemma 25 can be found in [32].

In the following, we prove Assumption 6 for $\|\Delta\|_2 \leq 1$. Based on the definition of $\mathcal{L}_t$, the first inequality in 6 can be written as

$$\begin{aligned} \langle \nabla \mathcal{L}_t(\beta + \Delta) - \nabla \mathcal{L}_t(\beta), \Delta \rangle &\geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log d}{t} \|\Delta\|_1^2 \\ \iff \Delta^T \left(\frac{X_t^T X_t}{t} \right) \Delta &\geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log d}{t} \|\Delta\|_1^2. \end{aligned}$$

Since the above inequality is scale invariant, it suffices to prove that the inequality holds for vector Δ with ℓ_2 -norm $\|\Delta\|_2 = 1$, for which the above inequality is equivalent to

$$\Delta^T \left(\frac{X_t^T X_t}{t} \right) \Delta \geq \alpha_1 - \tau_1 \frac{\log d}{t} \|\Delta\|_1^2. \quad (3.58)$$

When vector Δ has ℓ_1 -norm $\|\Delta\|_1 > c\sqrt{\frac{t}{\log d}}$, where c is a constant such that $c > \sqrt{\frac{\alpha_1}{\tau_1}}$, the right-hand side of the above inequality will be negative. As a result, the above inequality holds automatically since the left-hand side of the inequality is non-negative. This indicates that it is sufficient to prove the strong convexity for vector Δ with ℓ_1 -norm $\|\Delta\|_1 \leq c\sqrt{\frac{t}{\log d}}$. The proof is similar to that of Proposition 4. The main idea is to use a properly constructed ϵ -net $\mathcal{N}_1$ to approximate the set $\Omega = \{v \in S^{d-1} : \|v\|_1 \leq c\sqrt{\frac{t}{\log d}}\}$. We first prove the strong convexity of matrix $\hat{\Sigma}_t$ on $\mathcal{N}_1$, then transfer the same property to set Ω through approximation. Before the proof, we describe a result from [16].

Lemma 26 (Transfer Principle). *Suppose $\hat{\Sigma}_t$ and Σ are matrix with non-negative diagonal entries, and assume $\eta \in (0, 1)$, $m \in \{1, \dots, d\}$ are such that*

$$\forall v \in \mathbb{R}^d \text{ with } \|v\|_0 \leq m, v^T \hat{\Sigma}_t v \geq (1 - \eta) v^T \Sigma v. \quad (3.59)$$

Assume D is a diagonal matrix whose elements $D_{j,j}$ are non-negative and satisfy $D_{j,j} \geq [\hat{\Sigma}_t]_{j,j} - (1 - \eta)\Sigma_{j,j}$. Then

$$\forall x \in \mathbb{R}^d, x^T \hat{\Sigma}_t x \geq (1 - \eta) x^T \Sigma x - \frac{\|D^{1/2} x\|_1^2}{m - 1}. \quad (3.60)$$

The proof of Lemma 26 can be found in Lemma 5.1 of [16].

In the following, we prove the inequality (3.58) for vectors within the set Ω , which is defined as $\Omega =$

$\{v \in S^{d-1} : \|v\|_1 \leq c\sqrt{t/\log d}\}$ where the constant c will be defined later. According to Lemma 26, we first need to prove the eigenvalue condition (3.59) for m -sparse vectors with sufficiently large $m = Cs_0$, where constant C will be specified momentarily. To prove inequality (3.59), we introduce an ϵ -net $\mathcal{N}_1$ on unit sphere for approximation.

Denote $E_J = \text{span}\{e_j : j \in J\}$, then we define $\mathcal{N}_1$ as the ϵ -net for $S^{d-1} \cap (\cup_{|J|=Cs_0} E_J)$. For each subset $J \subseteq [d]$ with $|J| = Cs_0$, the set $S^{d-1} \cap E_J$ can be viewed as a unit sphere in $\mathbb{R}^{Cs_0}$. Similar to what we derive for $\mathcal{N}_0$ in the proof for Proposition 4, we have a bound for the covering number N_1 of $\mathcal{N}_1$ when $\epsilon \leq 1$.

$$\begin{aligned} \log N_1 &= \log \left(\binom{d}{Cs_0} \left(1 + \frac{2}{\epsilon}\right)^{Cs_0} \right) \\ &\leq \log \left(\left(\frac{ed}{Cs_0}\right)^{Cs_0} \left(\frac{3}{\epsilon}\right)^{Cs_0} \right) \\ &\lesssim Cs_0 \left[\log d + \log \left(\frac{1}{\epsilon}\right) \right]. \end{aligned} \quad (3.61)$$

In order to prove (3.58), we also need to bound the ℓ_1 -norm of vectors. Define events $\mathcal{H}_{1,t}$ and $\mathcal{H}_{2,t}$:

$$\begin{aligned} \mathcal{H}_{1,t} &= \{\forall u \in \mathcal{N}_1 : \frac{1}{t} \sum_{s=1}^t \langle X_{\hat{a}_s, s}, u \rangle \geq \frac{1}{2}h(1 - \zeta Kh)\}, \\ \mathcal{H}_{2,t} &= \{\forall u \in \mathcal{N}_1 : \langle u, \hat{\Sigma}_t u \rangle \leq \tilde{c} \log(K) \Lambda_0^2\}. \end{aligned}$$

Similar to the proof for Proposition 4, we have that events $\mathcal{H}_{1,t}$ and $\mathcal{H}_{2,t}$ hold with high probability. According to the statement after inequality (3.24), we set $h = \frac{1}{2\zeta K} = \mathcal{O}(\frac{1}{K})$ and $\epsilon = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\Lambda_0^2\zeta K \log K}\}$, then for $\forall v \in S^{d-1}$ and $\|v\|_0 \leq Cs_0$, we have

$$\langle v, \hat{\Sigma}_t v \rangle \geq \frac{1}{4}h(1 - \zeta Kh) = \frac{1}{16\zeta K}.$$

Thus, the inequality (3.59) is satisfied with $\Sigma = \frac{1}{8\zeta K} I_{d \times d}$ as a $d \times d$ diagonal matrix, $m = Cs_0$ and $\eta = \frac{1}{2}$. Moreover, we choose $D_{j,j} = [\hat{\Sigma}_t]_{j,j}$, which is the j th diagonal element of matrix $\hat{\Sigma}_t$, to satisfy the condition on diagonal matrix D . By greedy method for constructing $\mathcal{N}_1$, it can be assumed that $e_j \in \mathcal{N}_1$ for all $j \in [d]$, where e_j 's are unit vectors in $\mathbb{R}^d$ having one entry at j th location equal to 1 and 0 otherwise. Then, if event $\mathcal{H}_{2,t}$ holds, we have

$$D_{j,j} = [\hat{\Sigma}_t]_{j,j} = \langle e_j, \hat{\Sigma}_t e_j \rangle \leq \tilde{c} \log(K) \Lambda_0^2.$$

According to Lemma 26, when both events $\mathcal{H}_{1,t}$ and $\mathcal{H}_{2,t}$ hold, we have for any point x belonging to

$\Omega = \left\{ v \in S^{d-1} : \|v\|_1 \leq c\sqrt{t/\log d} \right\}$ that

$$\begin{aligned}
x^T \hat{\Sigma}_t x &\geq \frac{1}{2} x^T \Sigma x - \frac{\|D^{1/2} x\|_1^2}{Cs_0 - 1} \\
&\geq \frac{1}{16\zeta K} - \frac{\tilde{c} \log(K) \Lambda_0^2 \|x\|_1^2}{Cs_0 - 1} \\
&\geq \frac{1}{16\zeta K} - \frac{\tilde{c} \log(K) \Lambda_0^2}{Cs_0 - 1} \cdot \frac{c^2 t}{\log d}.
\end{aligned} \tag{3.62}$$

Define $C = \frac{1}{s_0} \left(\frac{32\zeta \tilde{c} c^2 K \log K \Lambda_0^2 t}{\log d} + 1 \right)$, then for any $x \in \Omega$, we have

$$x^T \hat{\Sigma}_t x \geq \frac{1}{32\zeta K} \geq \frac{1}{32\zeta K} - \tau_1 \frac{\log d}{t} \|x\|_1^2,$$

where τ_1 can be any positive numbers. Therefore, we prove the inequality (3.58) for vectors in Ω with $\alpha_1 = \frac{1}{32\zeta K}$ and any positive τ_1 . In order to prove inequality (3.58) for vectors with large ℓ_1 -norm, i.e., vectors in set $\Omega' = \{v \in S^{d-1} : \|v\|_1 > c\sqrt{t/\log d}\}$, we choose $\tau_1 = \frac{1}{32c^2\zeta K}$. Through calculation, the right-hand side of the inequality (3.58) is non-positive for all vectors in Ω' , thus the inequality (3.58) holds automatically.

From the above derivation, we prove the inequality (3.58) with parameters $\alpha_1 = \frac{1}{32\zeta K}$ and $\tau_1 = \frac{1}{32c^2\zeta K}$. In the next, we bound the cardinality N_1 of ϵ -net $\mathcal{N}_1$. Based on the choice of $C = \frac{1}{s_0} \left(\frac{32\zeta \tilde{c} c^2 K \log K \Lambda_0^2 t}{\log d} + 1 \right)$ and $\epsilon = \min\{\frac{1}{4}, \frac{1}{64\tilde{c}\Lambda_0^2\zeta K \log K}\}$, we can derive from above and inequality (3.61) that

$$\begin{aligned}
\log N_1 &\lesssim Cs_0 \left[\log d + \log\left(\frac{1}{\epsilon}\right) \right] \\
&= \left(\frac{32\zeta \tilde{c} c^2 K \log K \Lambda_0^2 t}{\log d} + 1 \right) [\log d + \log(64\tilde{c}\Lambda_0^2\zeta K \log K)] \\
&= \mathcal{O}(32\zeta \tilde{c} c^2 K \log K \Lambda_0^2 t + \log d).
\end{aligned}$$

Similar to the proof for Proposition 4, we have that $P(\mathcal{H}_{1,t}^c \cup \mathcal{H}_{2,t}^c) \leq N_1 \left[e^{-c_0 t} + e^{-\tilde{C} \log(K)t} \right]$, where c_0 and $\tilde{C}$ are positive constants. In order to bound this probability, we choose constant c small enough such that

$$\begin{aligned}
32\zeta \tilde{c} c^2 \Lambda_0^2 K \log K &\lesssim \frac{1}{2} \min\{c_0, \tilde{C} \log(K)\} \\
\implies c &\lesssim \frac{1}{8} \sqrt{\frac{\min\{c_0, \tilde{C} \log K\}}{\zeta \tilde{c} \Lambda_0^2 K \log K}}.
\end{aligned}$$

Here we choose $c \sim \frac{1}{8} \sqrt{\frac{\min\{c_0, \tilde{C} \log(K)\}}{\zeta \tilde{c} \Lambda_0^2 K \log K}}$, then we have

$$\tau_1 = \frac{1}{32c^2 \zeta K} \sim \frac{2\tilde{c} \Lambda_0^2 \log K}{\min(c_0, \tilde{C} \log K)}.$$

Thus, we show that Assumption 6 with $\alpha_1 = \frac{1}{32\zeta K}$ and $\tau_1 \sim \frac{2\tilde{c} \Lambda_0^2 \log K}{\min(c_0, \tilde{C} \log K)}$ is satisfied with high probability. In particular, we have when $t \geq 4R^2 \tau_1^2 \log d$ (Lemma 25) that

$$P(\{\mathcal{L}_t \text{ is } (\alpha, \tau)\text{-RSC}\}) \geq 1 - \mathcal{O}\left(e^{-\frac{1}{2}c_0 t} + e^{-\frac{1}{2}\tilde{C} \log(K)t}\right).$$

Proof of Theorem 6

In this section, we combine the statements in previous sections to prove Theorem 6. The theorem can be derived from Lemma 20. Firstly, we recall Assumption 5 on the regularization function ρ_{λ_t} . At time t , we define the parameter r_1 as $r_1 = r + \lambda_t \gamma$ for the MCP and SCAD regularizers. Then, we have $\epsilon_0 = 0$ and $u = \frac{1}{\gamma}$ for MCP regularizer; and $\epsilon_0 = 0$ and $u = \frac{1}{\gamma-1}$ for SCAD regularizer. The parameters $\lambda_t = C_L \sqrt{(\log d + \log t)/t}$ and $r_1 = C_g \sqrt{(\log d + \log t)/t}$ are defined as in Theorem 6. According to Lemma 20 can the proofs for Assumptions 6–8, when $t \geq \max\left\{\frac{2\tau_1}{2\alpha_1 - u} s_0 \log d, 4R^2 \tau_1^2 \log d\right\} = \mathcal{O}(s_0 \Lambda_0^4 \zeta K \log^2 K \log d)$, we have

$$\begin{aligned} P\left(\left\{S_2 \subseteq \hat{S}_t \subseteq S\right\}\right) &\geq P(\{\text{Assumption 6–8 hold}\}) \\ &\geq 1 - 2/t - \mathcal{O}\left(e^{-\frac{1}{2}c_0 t} + e^{-\frac{1}{2}\tilde{C} \log(K)t + s_0 \log(\Lambda_0^2 \zeta K)}\right), \end{aligned}$$

where c_0 and $\tilde{C}$ are absolute constants as defined in Proposition 4.

3.7.5 Proof of Corollary 2

According to inequality (3.14) in Theorem 6, condition (a) in Definition 8 is satisfied with t_0 defined as $t_0 = C_0 s_0 \Lambda_0^4 \zeta K \log^2 K \log d$ and M_0 as some large constant. In order to justify condition (b) in Definition 8, for any $\varepsilon > 0$, we define $t_\varepsilon = \max\left(\frac{C'_g \log d}{\varepsilon^2}, t_0\right)$, where the constant C'_g is chosen such that

$$r_1 = C_g \sqrt{\frac{\log d + \log t_\varepsilon}{t_\varepsilon}} \simeq C_g \sqrt{\frac{\log d}{C'_g \log d / \varepsilon^2}} = \frac{C_g \varepsilon}{\sqrt{C'_g}} < \varepsilon.$$

Therefore, set $I_\varepsilon = \{i \in [d] : |\beta_i^*| > \varepsilon\} \subseteq S_2 = \{i \in [d] : |\beta_i^*| > r_1\}$. As a result, condition (b) is satisfied with $M_1(d) = C'_g \log d$ and $M_2(d, T) = t_0$.

3.7.6 Proof of Theorem 7

The proof for Theorem 7 is an application of the result in Theorem 3. Since in Corollary 2, we show that the variable selection procedure $\mathcal{V}$ induced by the regularized estimation procedure satisfies Definition 8, we can plug-in the values of t_0 , M_0 , $M_1(d)$ and $M_2(d, T)$ from Corollary 2 to Theorem 3. The parameters are defined as below:

$$t_0 = M_2(d, T) = C_0 s_0 \Lambda_0^4 \zeta K \log^2 K \log d \quad \text{and} \quad M_1(d) = C'_g \log d,$$

where C_0 , C'_g and M_0 are absolute constants.

Chapter 4

Efficient Bayesian Sampling Using Generative Models

4.1 Introduction

A central problem in modern Bayesian statistics is how to conduct valid statistical inference on model parameters in a computationally efficient manner. In a standard Bayesian model, all information about model parameter θ in the parameter space $\Theta \subset \mathbb{R}^p$ is summarized in the posterior distribution (density function)

$$\pi_n(\theta) := p(\theta | Z^{(n)}) = \frac{\pi(\theta) p(Z^{(n)} | \theta)}{\int_{\Theta} \pi(\theta) p(Z^{(n)} | \theta) d\theta}, \quad \forall \theta \in \Theta \quad (4.1)$$

obtained by applying the Bayes rule, where $\pi(\theta)$ denotes the prior distribution (density function) over Θ , $Z^{(n)} = (Z_1, Z_2, \dots, Z_n)$ is a sample of n observations, and $p(Z^{(n)} | \theta)$ represents the likelihood function specified by the statistical model.

The main hurdle in implementing Bayesian inference is the multi-dimensional integral as the normalization constant appearing in equation (4.1). This integral is often mathematically intractable unless a conjugate prior is employed, which however significantly restricts the applicability of the Bayesian procedure to complicated models where a conjugate prior is either difficult to identify or not flexible enough to use. Two main approaches, which only requires knowing the target density up to a constant, are commonly adopted in the literature for implementing Bayesian inference, thus avoiding explicitly computing the normalization constant. One is based on (approximately) sampling from the posterior distribution [8] via, for example, Markov Chain Monte Carlo (MCMC) algorithms; and another is through variational approximation [9] by

turning the integration problem into an optimizing problem. Yet these two approaches have their respective limitations. MCMC may suffer from slow-mixing due to complicated model structures, which greatly reduces its sampling efficiency due to high sample auto-correlation induced by the Markovian structure. In contrast, although variational approximation can easily scale to massive data, it is not guaranteed to produce (almost) exact samples from the target density and may lead to estimation inconsistency and misleading uncertainty quantification.

In this chapter, we consider a new class of sampling approach, which embraces the idea of generative model learning with techniques from optimal transport theory, for implementing Bayesian inference. The new approach can achieve the best of two worlds — it produces *independent and exact* samples from the target distribution while maintaining similar computational efficiency as variation approximation since the sampler is constructed by solving an optimization problem. More specifically, we construct our sampler for Bayesian posterior via generative modelling, where a generative model takes draws from a simple reference distribution μ such as the standard normal or the prior distribution π as inputs and produces outputs by passing the inputs through a non-linear and invertible transport map T . Mathematically, the output distribution can be described as the push-forward distribution $T_{\#}\mu$ of μ through map T .

This idea of using transport maps for posterior sampling is not invented by us, and first appears in [39]. However, there is a major difference. [39] and other recent works such as [40]–[42] do not impose explicit structural constraint on the class of transport maps T for identifiability, which may lead to non-uniqueness (in fact, infinitely many) of the map and incur algorithmic issues such as instability and inefficiency. In comparison, we impose the extra constraint by requiring T to be the optimal transport map sending μ to the target posterior π_n in the Monge problem [10], [11] of minimizing the ℓ_2 transportation cost $\mathbb{E}_{\mu}[\|T(\theta) - \theta\|_2^2]$ subject to the constraint that $\pi_n = T_{\#}\mu$. Statistically, this extra constraint based on transportation cost minimizing leads to a *most parsimonious* model characterizing the posterior. Practically, our estimated transport map is *more interpretable* since $\{(1-t)id + tT^* : t \in [0, 1]\}$, where id denotes the identity map, describes the shortest path (i.e. geodesic path) between prior π and posterior π_n in the space of all probability distributions. Therefore, T^* can be used for explaining and visualizing the transition of the belief (information) on θ from prior to posterior. As another example, T^* can be readily used for quantifying the closeness between prior π and posterior π_n as the 2-Wasserstein distance $W_2(\pi, \pi_n) = \{\mathbb{E}_{\mu}[\|T^*(\theta) - \theta\|_2^2]\}^{1/2}$, which can be useful for comparing different priors through the criterion in defining the reference prior [43], [44] that a good non-information prior should maximize some measure of distance or divergence between the posterior and prior, as data observations are made.

We borrow results from optimal transport theory and provide explicit characterizations of the optimal transport map T^* in the context of Bayesian inference. It is worthwhile mentioning that in the setting where

θ contains both continuous and discrete components, such as in the context of latent variable models, our characterization of T^* is interesting by its own right since it generalizes the semi-discrete optimal transport where one of the distribution is fully discrete. Based on this theoretical guidance, we propose a parsimonious parametric family for approximating the optimal transport map T^* with diminishing approximation error as the family capacity increases, where the best parameter can be efficiently obtained by solving an optimization problem that does not require the knowledge of the unknown normalization constant in the posterior.

The rest of the chapter is organized as follows. In Section 4.2, we introduce the background for generative model learning and optimal transport and describe the problem formulation. In Section 4.3, we introduce our optimal generative model learning method for Bayesian posterior sampling and provide a finer characterization (Theorem 9) of optimal transport maps for general setting where the target distribution contains both continuous and discrete components. In Section 4.4, we discuss the computation details for estimating the optimal transport map and its application to Bayesian latent variable models. Numerical results for our proposed sampling method are presented in Section 4.5.

4.2 Background and Problem Setup

In this section, we begin with an introduction on generative model learning and its specialization to the context of Bayesian posterior approximation. After that, we briefly review the optimal transport theory. Finally, we describe our problem formulation whose goal is combining optimal transport techniques to aid the generative model learning for Bayesian posterior sampling.

4.2.1 Generative model learning

In the machine learning literature, generative models, such as Generative Adversarial Network (GAN, [45]–[47]), Wasserstein GAN (WGAN, [48]) and Wasserstein Auto-Encoder (WAE, [49], [50]), have received great success in generating synthetic realistic-looking images and texts [51], [52] by implicit distribution estimation over complex data space bypassing the need of evaluating probability densities. Generative model based approaches for statistical inference [53], [54] avoid explicitly specifying the likelihood function while utilizing the expressive power of neural networks to capture underlying probabilistic structures. In our problem of Bayesian posterior sampling, the data space of the generative model is the parameter space $\Theta \subset \mathbb{R}^p$, which may also contain latent variables.

A *generative model* is mathematically defined as a pair (μ, T) , where μ is a distribution on a latent space $\mathcal{X} \subset \mathbb{R}^p$, known as reference distribution or generative distribution, that is easy to sample from; and $T : \mathcal{X} \rightarrow \Theta$ is a map from $\mathcal{X}$ to the data space Θ , termed transport map or generative map. Ideally, if

$X \sim \mu$, then $T(X) \sim \nu$, where ν denotes the target distribution to be estimated. In other words, the target distribution ν can be expressed via the generative model (μ, T) via $\nu = T_{\#}\mu$, the ¹pushforward measure of ν using map T . The set $\mathcal{D}_T = \{T_{\#}\mu : \mu \in \Upsilon, T \in \mathcal{T}\}$ of all generative models (μ, T) with $\mu \in \Upsilon$ and $T \in \mathcal{T}$ for some distribution family Υ on $\mathcal{X}$ and function class $\mathcal{T}$ (consists of maps from $\mathcal{X}$ to Θ) is a corresponding *generator class*. In practice, Υ is often set as a fixed simple distribution that is easy to sample from, such as the standard Gaussian or uniform distribution, so that each generative model in the generator class $\mathcal{D}_T$ specifies a generating process, from which samples following the underlying distribution $T_{\#}\mu$ can be easily and efficiently produced. In the Bayesian framework where the posterior distribution π_n is the target, we may choose the reference distribution as the prior π . More generally, if the prior is characterized by a hyperparameter, then we may choose Υ as the parametric family indexed by the hyperparameter.

Generative modeling learning refers to the problem of finding a distribution $\hat{\nu}$ in the generator class $\mathcal{D}_T$ that is closest to the target distribution ν^* over the data space Θ with respect to certain discrepancy measurement d_T , or $\hat{\nu} = \arg \min_{\nu \in \mathcal{D}_T} d_G(\nu, \nu^*)$. In the standard context of learning a distribution over images or texts, one only has access to a finite number n of i.i.d. samples $\{Z_1, Z_2, \dots, Z_n\}$ from the unknown distribution ν^* . For these problems, it is convenient to set the discrepancy measurement as an adversarial risk $d_G(\nu_1, \nu_2) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{\nu_1}[f(Z)] - \mathbb{E}_{\nu_2}[f(Z)]|$ defined through a discriminator class $\mathcal{F}$, since the adversarial risk admits a simple approximation (surrogate) by replacing the expectation $\mathbb{E}_{\nu^*}[f(Z)]$ with the empirical average $n^{-1} \sum_{i=1}^n f(Z_i)$, which facilitates the computation.

In the context of Bayesian inference, the goal is to generate (nearly) independent samples from the target posterior π_n , which is known up to a multiplicative constant. As a consequence, it is no longer convenient to choose the discrepancy measurement d_G as an adversarial risk. Instead, it is natural to resort to the idea from variational approximation by using the Kullback-Leibler (KL) divergence as the discrepancy measurement, since the evaluation (up to a fixed constant) of KL divergence from any distribution to π_n does not require knowledge of the unknown normalization constant. More specifically, we consider the following optimization problem for Bayesian transport map learning with a fixed reference distribution μ ,

$$\min_{T \in \mathcal{T}} d_{\text{KL}}(T_{\#}\mu \parallel \pi_n), \quad (4.2)$$

where the transport map class $\mathcal{T}$ is a flexible function space free of the likelihood (c.f. Section 4.3.1 for its choice). However, unlike variational approximation, the optimization problem as stated above is in general computationally infeasible since the evaluation of the KL divergence requires an analytical expression of the density function of $T_{\#}\mu$, which further requires knowledge of the inverse transformation T^{-1} . Fortunately,

¹For any measure μ on $\mathcal{X}$ and map $T : \mathcal{X} \rightarrow \Theta$, the pushforward measure $\nu = T_{\#}\mu$ is defined as the unique measure on Θ such that $\nu(A) = \mu(T^{-1}(A))$ holds for any measurable set A on Θ .

based on the property that KL divergence is invariant under invertible transformations (applying to both μ and π_n), we have the following lemma showing that optimization problem (4.2) can be cast into an equivalent problem that does not require explicit evaluation of T^{-1} . Let $J_T \in \mathbb{R}^{p \times p}$ denote the Jacobian matrix associated with map T from $\mathbb{R}^p$ to $\mathbb{R}^p$, and $\det(A)$ denote the determinant of a square matrix A .

Lemma 27. *Assume that each element $T \in \mathcal{T}$ is differentiable and invertible on the support of μ . Then optimization problem (4.2) is equivalent to*

$$\max_{T \in \mathcal{T}} \mathbb{E}_\mu [\log(\pi_n \circ T) + \log |\det(J_T)|]. \quad (4.3)$$

In particular, the above lemma enables us to find the best transport map $\hat{T}$ without computing the intractable normalization constant as omitting this constant does not change the optimum. Besides, it provides an alternative view for the objective function. The optimal value of problem (4.3) is the negative differential entropy $-h(\mu) = \mathbb{E}_\mu[\log \mu]$ of the reference distribution μ . In addition, the first term $\log(\pi_n \circ T)$ in objective function (4.3) can be interpreted as a goodness-of-fit to target distribution π_n and the second term is a regularization term penalizing the “roughness” of the transport map T since for each z in the latent space, $|\det(J_T(z))|$ measures the degree of local distortions (of local volumes) produced by the differential map $J_T(z) : \mathbb{R}^p \rightarrow \mathbb{R}^p$ around z .

4.2.2 Review on optimal transport

Optimal transport (OT), or transportation theory, originates from the study by French mathematician Gaspard Monge in 1781, and has recently been found useful in solving a number of machine learning problems, perhaps most prominently in generative adversarial networks [WGAN, 48]. We refer the reader to two books, one by [11] and the other by [10], for a comprehensive and thorough discussion of theoretical developments of OT. In this subsection, we briefly review some theoretical background and fundamental results of OT that are necessary before introducing our development.

Let $\Pi(\nu_1, \nu_2)$ be the set of all couplings between two probability distributions, ν_1 over $\mathcal{X} \subset \mathbb{R}^p$ and ν_2 over $\mathcal{Y} \subset \mathbb{R}^p$. In other words, $\gamma \in \Pi(\nu_1, \nu_2)$ if and only if the marginal distributions of γ over $\mathcal{X}$ and $\mathcal{Y}$ are ν_1 and ν_2 , respectively. In this work, we focus on the OT problem with quadratic cost

$$(KP) \quad \inf_{\gamma \in \Pi(\nu_1, \nu_2)} \int \|x - y\|^2 d\gamma(x, y),$$

which is called the Kantorovich formulation of OT. Problem (KP) always admits a solution [Theorem 1.7, 10], called an optimal transport plan, where the corresponding optimal objective value defines the squared

2-Wasserstein distance $W_2^2(\nu_1, \nu_2)$ between ν_1 and ν_2 . In particular, if one of the distributions, say ν_1 , is absolutely continuous with respect to the Lebesgue measure of $\mathbb{R}^p$, Theorem 1.22 in [10] shows that under some mild conditions the optimal transport plan is unique and takes the form of $(id, T)_\# \nu_1$, implying $\nu_2 = T_\# \nu_1$ [also see, e.g., 55], [56]. The unique map T is called the optimal transport map from ν_1 to ν_2 and takes the form of $T = \nabla u$, where u is a convex function. In particular, the map T is the unique solution to the following Monge formulation of OT with quadratic cost,

$$(MP) \quad \inf_T \int \|x - T(x)\|^2 d\nu_1(x), \quad \text{s.t.} \quad T_\# \nu_1 = \nu_2.$$

The form of the unique map T can be extended to a more general case where ν_2 consists of convex regions that are strongly separated. Theorem 5.1 in [57] concludes that under quadratic cost when $\text{spt}(\nu_2) = \overline{\Omega}_1 \cup \overline{\Omega}_2$, u takes form of $\max\{u_1, u_2\}$, where ∇u_i maps to $\overline{\Omega}_i$ for $i \in \{1, 2\}$. Moreover, according to (5.5) in [57], the same holds when $\text{spt}(\nu_2) = \cup_{i \in I} \overline{\Omega}_i$ for a finite set I .

4.2.3 Problem formulation

Based on the notion of generative model and theory of OT, we are now ready to formulate our problem of learning an optimal generative model for efficient Bayesian sampling as follows:

For a given reference distribution μ on parameter space Θ , how to find (approximate) the optimal (minimal) generative model (μ, T_B) for sampling from posterior π_n ? Here, the optimality refers to T_B as the solution of the OT problem (MP) with $\nu_1 = \mu$ and $\nu_2 = \pi_n$.

Since posterior π_n is only known up to a constant, identifying the optimal generative mode (μ, T_B) by directly solving (MP) with $\nu_1 = \mu$ and $\nu_2 = \pi_n$ is computationally infeasible, and we have to find alternative ways to (approximately) compute T . One common practice in the literature to find the solution of an under-determined system (in our case is $T_\# \mu = \pi_n$) with additional identifiability conditions (in our case is the optimality) is by imposing regularization. In our case, based on the formulation of (4.2) for matching $T_\# \mu$ with π_n for example, it is natural to consider the following regularized KL optimization,

$$\min_{T \in \mathcal{T}} \left\{ d_{\text{KL}}(T_\# \mu \parallel \pi_n) + \lambda \mathbb{E}_\mu [\|T(X) - X\|^2] \right\}, \quad (4.4)$$

where $\lambda > 0$ is the regularization parameter and $\mathcal{T}$ is a pre-specified class of transport maps. However, a regularized based approach as above tends to introduce a bias term proportional to λ ; moreover, in practice a data driven method of optimally tuning λ for balancing the first goodness-of-fit term and the second

penalty term in the objective function can be computationally expensive. Moreover, in order for Lemma 27 to be applicable for facilitating tractable computations, one need to add an extra constraint of each $T \in \mathcal{T}$ being invertible, which further complicates the implementation. In this work, we propose a more direct and non-regularization based approach for computing the optimal generative model by utilizing more underlying structures of the optimal transport map T_B . Moreover, we will also generalize the characterization of T_B from continuous parameter θ to cases where θ itself may contain discrete components, which naturally appears in Bayesian latent variable models, so that the assumption of reference distribution μ being absolutely continuous therein violates.

4.3 Generative Model Learning for Bayesian Inference

In this section, we introduce our optimal generative model learning method for Bayesian posterior sampling. We separately treat two common settings in Bayesian inference: 1. parameter θ is a continuous random variable in $\mathbb{R}^p$; 2. parameter $\theta = (\tau, \zeta)$ with τ collecting all discrete components such as unknown latent variables in the model and $\zeta \in \mathbb{R}^p$ collecting all continuous components such as model parameters.

4.3.1 Optimal transport map with continuous variables

As we mentioned in the previous section, the regularized approach based on solving (4.4) is a universal black-box approach not utilizing any specialized structure of the optimal transport map T_B . Instead, our method will utilize characterizations of T_B to design a transport class $\mathcal{T}$ so that T_B is (nearly) the unique member within $\mathcal{T}$ such that $T_{\#}\mu = \pi_n$. As a consequence, optimization problem (4.2) admits a unique solution as (an approximation to) T_B . In this subsection, we will start from the relatively easy case when θ is a continuous parameter, or continuous random variable in the Bayesian model, so that T_B is differentiable and invertible and the equivalently (4.3) can be applied to facilitate the computation. In Section 4.3.2, we will steer to the general case where θ is a mixed parameter (vector) that contains both continuous and discrete components. In this general case, Lemma 4.3 no longer applies and we need to utilize a finer characterization of optimal transport map T_B to design a computation-friendly optimization formulation for computing T_B .

According to discussions in Section 4.2.2, if the reference distribution μ admits a density function with respect to the Lebesgue measure of $\mathbb{R}^p$, then optimal transport map T_B has the characterization as the gradient of a convex function u . Conversely, if we can find a convex function u such that $T_B = \nabla u$ pushforwards μ to π_n , then this T_B is the unique optimal transport map. Moreover, if both reference distribution μ and posterior π_n admit continuous densities functions, then T_B which pushforwards μ to π_n is differentiable and invertible. This makes the convex function u defining $T = \nabla u$ to be at least twice differentiable. Therefore, we can use

the standard characterization of the twice differentiable convex function u as its Hessian $\nabla^2 u = J_T$ being a positive semidefinite (psd) matrix on its domain. These properties motivate us to specify the transport map class in (4.3) for identifying T_B as $\mathcal{T} = \{\nabla u \mid u : \text{supp}(\mu) \rightarrow \Theta \mid u \in C^2(\text{supp}(\mu)), \nabla^2 u \in \mathbb{R}^{p \times p} \text{ is pd}\}$, which also makes Lemma 4.3 applicable. In practice, we can choose $\mathcal{T}$ as any parametric family approximating the space of all twice differentiable and strongly convex functions over $\text{supp}(\mu)$. Our choice of such an approximating family to be described below is a personal preference.

4.3.2 Optimal transport map with mixed variables

In the more general setting where θ contains both continuous and discrete components (i.e. mixed variables), we may write $\theta = (\tau, \zeta)$, where $\tau \in [K] := \{1, 2, \dots, K\}$ denotes the discrete component with K levels and $\zeta \in \mathbb{R}^p$ denotes the continuous component of dimension p . Since τ is discrete, we assume the knowledge of a prespecified embedding of the K categories to K vectors $\{b_k\}_{k=1}^K$ in $\mathbb{R}^r$ such that for each distinct pair $(k, \ell) \in [K]^2$, the ℓ_2 distance $\|b_k - b_\ell\|_2$ corresponds to the discrepancy measure between category k and ℓ . For example, if τ is a nominal variable, then order does not matter and we may choose $\{b_k\}_{k=1}^K$ as the K vertices of a regular polyhedron in $\mathbb{R}^{K-1}$ ($r = K - 1$) such that $\|b_k - b_\ell\|_2 = 1$ if $k \neq \ell$; and if τ is an ordinal variable where order matters, then we may choose $\{b_k\}_{k=1}^K$ as an increasing sequence in $\mathbb{R}$ ($r = 1$) so that $b_{k+1} - b_k$ quantifies the gap between two consecutive levels k and $k + 1$ for $k \in [K - 1]$. Let us correspondingly decompose the random variable $X \sim \mu$ from reference distribution μ into $X = (X^{(1)}, X^{(2)}) \in \mathbb{R}^r \times \mathbb{R}^p$, where r is the embedding space dimension induced by the discrete component τ , and decompose any transport map as $T = (T^{(1)}, T^{(2)}) : \mathbb{R}^r \times \mathbb{R}^p \rightarrow [K] \times \mathbb{R}^p$, where the ranges of $T^{(1)}$ and $T^{(2)}$ are $[K]$ and $\mathbb{R}^p$, respectively.

With the specified embedding of discrete components, the Monge formulation of OT with mixed random variables then takes the form as

$$\begin{aligned}
(\text{MPm}) \quad & \inf_{T=(T^{(1)}, T^{(2)})} \int_{\mathbb{R}^r \times \mathbb{R}^p} \left[\underbrace{\kappa \|x^{(1)} - b_{T^{(1)}(x)}\|_2^2}_{\text{discrete part cost}} + \underbrace{\|x^{(2)} - T^{(2)}(x)\|_2^2}_{\text{continuous part cost}} \right] d\mu(x), \\
\text{s.t.} \quad & T_{\#}\mu = \pi_n \quad \text{and} \quad T^{(1)}(x) \in [K], \quad \text{for all } x = (x^{(1)}, x^{(2)}) \in \mathbb{R}^r \times \mathbb{R}^p,
\end{aligned}$$

where $\kappa > 0$ is a relative weight parameter controlling the impact of the discrete component in the overall cost. By default, we choose $\kappa = 1$ throughout the paper. In this case, since posterior π_n contains singularity due to the discrete component, the optimal transport map $T = \nabla u$ with a convex function u is non-invertible. As a consequence, although we can still model the transport map class $\mathcal{T}$ as the space of all convex function gradients, Lemma 4.3 is no long applicable to make Bayesian transport map learning computationally tractable. This observation demands us to find a finer characterization of T_B and its associated convex

function u_B and a counterpart of Lemma 4.3 to facilitate the computation. The following theorem provides such a characterization, suggesting that we should treat the continuous and discrete components separately when modelling T or u .

Theorem 9. *Suppose the reference distribution μ is absolutely continuous with respect to the Lebesgue measure of $\mathbb{R}^r \times \mathbb{R}^p$ and the target distribution π_n is over $\theta = (\tau, \zeta)$, where $\tau \in [K]$ and $\zeta \in \mathbb{R}^p$. Then the convex function u_B associated with the unique optimal transport map $T_B = (T_B^{(1)}, T_B^{(2)})$ of problem (MPm) that pushforwards μ to π_n takes the form of*

$$u_B(x^{(1)}, x^{(2)}) = \kappa \max_{k \in [K]} \{ \langle x^{(1)}, b_k \rangle + \phi_k(x^{(2)}) \},$$

where $b_k \in \mathbb{R}^r$ is the pre-specified embedding vector associated with label k and ϕ_k is a convex function over $\mathbb{R}^p$ for $k \in [K]$, and T_B maps $(x^{(1)}, x^{(2)})$ to (τ, ζ) with

$$\begin{cases} \tau &= T_B^{(1)}(x^{(1)}, x^{(2)}) = \arg \max_{k \in [K]} \{ \langle x^{(1)}, b_k \rangle + \phi_k(x^{(2)}) \}, \\ \zeta &= T_B^{(2)}(x^{(1)}, x^{(2)}) = \kappa \nabla \phi_\tau(x^{(2)}). \end{cases} \quad (4.5)$$

Moreover, if for $k \in [K]$, the conditional posterior distribution $\pi_n(\zeta | \tau = k)$ is absolutely continuous with respect to the Lebesgue measure of $\mathbb{R}^p$, then function ϕ_k is at least twice differentiable and strongly convex, so that $\nabla \phi_k$ is differentiable and invertible.

In the case when μ is a product measure over $\mathbb{R}^r \times \mathbb{R}^p$, the k -th map $T_k := \nabla \phi_k$ over $\mathbb{R}^p$ in the lemma can be interpreted as the optimal transport map that pushforwards the marginal of μ over $\mathbb{R}^p$ to the condition posterior $\pi_n(\zeta | \tau = k)$ for $k \in [K]$. The piecewise linear component of u_B over $\mathbb{R}^r$ induced by the linear function $\langle x^{(1)}, b_k \rangle$ causes the probability mass accumulation in discrete component τ since the gradient of a linear function is a constant vector, which is also the reason of the non-invertibility of the global map T_B as a location-dependent “mixture” of local maps $\{T_k\}_{k=1}^K$. For reference, [41] also proposed a random transport map as a mixture of transport maps $\{T_i\}_{i=1}^N$, which is motivated by Bayesian non-parametric approximation of the conditional density [58]. In their method, each point is transported by map T_i with a local-dependent probability ω_i from $\{\omega_i\}_{i=1}^N$. However, since they only consider location-scale changes for T_i ’s, it is costly to represent the target distributions. Furthermore, the total number of the maps $\{T_k\}_{k=1}^N$ would increase exponentially if the target distribution contains discrete components. [39] proposed a non-unique lower-triangular map. In order to guarantee uniqueness, extra constraints must be enforced and will complicate the optimization. In addition, since polynomial functions are used to express their transport map, the number of polynomials grows rapidly with dimension r and p . In particular, when μ is the standard normal distribution

over $\mathbb{R}^{p+r}$, the first equation of (4.5) that determines the level of discrete component τ under a given $x^{(2)}$ resembles a multi-class extension of the Probit model, where $c_k = \langle x^{(1)}, b_k \rangle + \phi_k(x^{(2)}) \sim N(\phi_k(x^{(2)}), \|b_k\|^2)$ can be viewed as the latent score associated with class k and the final output is the one with the highest latent score; then the continuous component ζ is generated from the conditional posterior $\pi_n(\zeta | \tau)$ by pushforwarding the marginal reference distribution $N(0, I_p)$ through transport map T_τ . This factorization perspective of interpreting the transport map motivates the Lemma 28 in Section 4.4.1 below as the building block for facilitating the computation.

4.4 Computation

In this section, we first provide a general optimization formulation for identifying the optimal transport map from a prespecified transport map class $\mathcal{T}$. After that, we describe our specific construction of the transport map class $\mathcal{T}$ in both cases (i.e. continuous variables and mixed variables), and describe the corresponding iterative algorithms for numerical computations.

4.4.1 Optimization formulation of optimal transport map

In the continuous case, we can apply the equivalent optimization formulation (4.3) to the original KL minimization problem (4.2) for finding the (approximated) optimal transport map T_B within the candidate class $\mathcal{T}$. In this section, we describe an equivalent optimization formulation for finding the optimal transport map in the form of (4.18) when θ contains discrete components. Recall that in the original formulation (4.2), we minimize the KL-divergence between push-forward distribution $T_\# \mu$ and the target distribution ν . As an alternative to Lemma 27, we propose a different objective function to tackle the issues from discrete random variables, as summarized in the following lemma.

Lemma 28. *Let $T : \mathcal{X} \rightarrow \mathcal{Y}$ be the optimal transport map defined in (4.18), where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 := \mathbb{R}^r \times \mathbb{R}^p$ and $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 := [K] \times \mathbb{R}^p$. We define the intermediate transport map $\tilde{T} : \mathcal{Y}_1 \times \mathcal{X}_2 \rightarrow \mathcal{Y}_1 \times \mathcal{Y}_2$ induced by T as*

$$\begin{pmatrix} \tau \\ x^{(2)} \end{pmatrix} \mapsto \begin{pmatrix} \tau \\ \zeta = \nabla \phi_\tau(x^{(2)}) \end{pmatrix}.$$

Then the optimization problem (4.2) is equivalent to

$$\min_{T \in \mathcal{T}} \mathbb{E} \left[\log \mathbb{P}(\tau | x^{(2)}) - \log \pi_n \circ \tilde{T}(\tau, x^{(2)}) - \log |\det(J_{\tilde{T}}(\tau, x^{(2)}))| \right], \quad (4.6)$$

where the expectation is taken with respect to random variables $\tau = \arg \max_{k \in [K]} \{ \langle x^{(1)}, b_k \rangle + \phi_k(x^{(2)}) \}$ and

$x^{(2)}$ with $(x^{(1)}, x^{(2)}) \sim \mu$.

To evaluate the objective function in optimization problem (4.6) under a given T , we use the Monte Carlo method by using an empirical average for approximating the expectation. To generate the Monte Carlo samples, we first generate N samples $\{x_1, x_2, \dots, x_N\}$ i.i.d. from μ , and then compute the corresponding transformed variables $\{\theta_1, \theta_2, \dots, \theta_N\}$ through transport map T in (4.18), where $\theta_i = (\tau_i, \zeta_i)$ with $\zeta_i = \nabla \phi_{\tau_i}(x_i^{(2)})$ and

$$\tau_i = \arg \max_{k \in [K]} \left\{ \langle x_i^{(1)}, b_k \rangle + \phi_k(x_i^{(2)}) \right\}, \quad \text{for } i = 1, 2, \dots, N.$$

To numerically approximate the (generally) analytically intractable conditional probability $\mathbb{P}(\tau_i \mid x_i^{(2)})$ in objective (4.6) for each sample $(\tau_i, x_i^{(2)})$, we apply the Monte Carlo method again by generating M samples $\{z_{i,1}^{(1)}, z_{i,2}^{(1)}, \dots, z_{i,M}^{(1)}\}$ i.i.d. from μ_1 for each sample $x_i^{(2)}$, and computing the transported variables $\{w_{i,1}^{(1)}, w_{i,2}^{(1)}, \dots, w_{i,M}^{(1)}\}$ as

$$w_{i,j}^{(1)} = \arg \max_{k \in [K]} \left\{ \langle z_{i,j}^{(1)}, b_k \rangle + \phi_k(x_i^{(2)}) \right\}, \quad \text{for } j = 1, 2, \dots, M.$$

Then probability $\mathbb{P}(\tau_i \mid x_i^{(2)})$ can be approximated by the empirical distribution,

$$\widehat{\mathbb{P}}(\tau_i = \ell \mid x_i^{(2)}) := \frac{1}{M} \sum_{j=1}^M \mathbb{1}_{\{w_{i,j}^{(1)} = \ell\}} \approx \frac{1}{M} \sum_{j=1}^M W_{i,j}(\ell), \quad \ell \in [K],$$

where $W_{i,j}$ denotes the softmax approximation of the indicator function, i.e.,

$$W_{i,j}(\ell) = \frac{\exp \left(\gamma (\langle z_{i,j}^{(1)}, b_\ell \rangle + \phi_\ell(x_i^{(2)})) \right)}{\sum_{k \in [K]} \exp \left(\gamma (\langle z_{i,j}^{(1)}, b_k \rangle + \phi_k(x_i^{(2)})) \right)}, \quad \ell \in [K],$$

with γ being a positive tuning parameter. Here we use the softmax approximation to make the objective function differentiable, so that first-order optimization algorithms such as (stochastic) gradient descent can be applied for numerical optimization. In practice, we may reuse the samples $\{x_1^{(1)}, x_2^{(1)}, \dots, x_N^{(1)}\}$ when generating the τ_i 's as $\{z_{i,1}^{(1)}, z_{i,2}^{(1)}, \dots, z_{i,M}^{(1)}\}$ to facilitate computational efficiency and reduce space complexity.

The calculation of the middle term in the objective (4.6) is straightforward, and only requires the knowledge of the target posterior π_n up to a multiplicative constant (without changing the optimum). To calculate the (last) determinant of the Jacobian matrix of the transport map $\tilde{T}$ in the objective (4.6) for each sample each

sample $(\tau_i, x_i^{(2)})$, we notice that $\tilde{T}$ does not change the first component $\tau \in \mathcal{Y}_1$ of the input, so that

$$J_{\tilde{T}}(\tau, x^{(2)}) = \begin{pmatrix} I & 0 \\ 0 & \nabla^2 \phi_\tau(x^{(2)}) \end{pmatrix}.$$

Therefore, the determinant of the Jacobian matrix at $(\tau_i, x_i^{(2)})$ can be simplified into

$$\det(J_{\tilde{T}}(\tau_i, x_i^{(2)})) = \det(\nabla^2 \phi_{\tau_i}(x_i^{(2)})).$$

4.4.2 Construction of transport map class $\mathcal{T}$

In this section, we define an approximation family to the transport class $\mathcal{T}$. In the continuous case, the class $\mathcal{T}$ consists of the gradients of parameterized convex functions as in Section 4.3.1. To start with, we introduce a class of convex units as our building blocks, which is validated in the following lemma:

Lemma 29. *For constant $a \in \mathbb{R}^p$ and $b \in \mathbb{R}$, let*

$$n(x; a, b) := \langle a, x \rangle + b$$

be a linear function of x , and φ an increasing and bounded function defined on $\mathbb{R}$. Then a unit defined by

$$f(x) := \int_{-\infty}^{n(x; a, b)} \varphi(u) du \tag{4.7}$$

is a convex function of x . As a result,

$$g(x) := \alpha f(x) + \langle \beta, x \rangle + w$$

is convex as well for $\alpha > 0$, $w \in \mathbb{R}$, and $\beta \in \mathbb{R}^p$.

In our formulation, we choose φ as nonlinear functions which we will provide some examples below. Therefore, the constant $\alpha > 0$ can be viewed as the parameter controlling the degree of non-linearity: the smaller α is, the closer the convex unit to an affine transform. Adding an affine function $\langle \beta, x \rangle + w$ does not impact the convexity of the activation function, while adding its flexibility. The coefficient β works both as an adjustment to the separation of boundary and a translation after taking gradient with respect to x .

We employ f of form in Lemma 29 as our local approximation to the convex function u , where φ can be selected from functions that admit the constraints while introducing certain amount of non-linearity for

model flexibility. It can be easily verified that the following activation functions meet our criteria:

- The Tanh function: $\varphi(x) = \tanh x$.
- The Softsign function: $\varphi(x) = \frac{x}{1+|x|}$.
- The Square Nonlinearity (SQNL) function:

$$\varphi(x) = \begin{cases} 1 & \text{if } x > 2 \\ x - \frac{x^2}{4} & \text{if } 0 \leq x \leq 2 \\ x + \frac{x^2}{4} & \text{if } -2 \leq x < 0 \\ -1 & \text{if } x < -2 \end{cases}$$

With the convex units defined in Lemma 29, our formulation for the convex functions is defined as summation of convex units:

$$u(x) := \sum_{k=1}^K g_k(x) = \sum_{k=1}^K \alpha_k f_k(x) + \langle \beta_k, x \rangle + \omega_k,$$

where

$$f_k(x) := \int_{-\infty}^{n(x; a_k, b_k)} \varphi(u) du = \int_{-\infty}^{\langle a_k, x \rangle + b_k} \varphi(u) du.$$

Then, the transport map in $\mathcal{T}$ is defined as the gradient of the convex function u , i.e.,

$$T(x) := \nabla u(x) = \sum_{k=1}^K \nabla g_k(x) = \sum_{k=1}^K \alpha_k \varphi_k(\langle a_k, x \rangle + b_k) a_k + \beta_k. \quad (4.8)$$

The form of the transport map T motivates our choice of φ in Lemma 29. In particular, each convex unit g_k only has a local effect on the transport map T due to the boundedness of φ , since $\nabla g_k(x)$ will become constant when x is far from the hyperplane $\langle a_k, x \rangle + b_k = 0$. As a result, the transport map T will gain much flexibility.

In order to solve problem (4.3), we still need to compute the Jacobian matrix of T , which we also provide in the following.

$$J_T(x) := \nabla T(x) = \nabla^2 u(x) = \sum_{k=1}^K \alpha_k \varphi'_k(\langle a_k, x \rangle + b_k) a_k a_k^T. \quad (4.9)$$

Inspired by the discussions in Section 4.2.2, when the support of the target distribution ν consists of separated

regions, we can construct the convex function u as the maximum of a set of convex functions $\{u_i\}_{i=1}^M$:

$$u(x) := \max_{i \in [M]} u_i(x), \quad (4.10)$$

where

$$u_i(x) := \sum_{j=1}^{K_i} g_{ij}(x) = \sum_{j=1}^{K_i} \alpha_{ij} f_{ij}(x) + \langle \beta_{ij}, x \rangle + \omega_{ij}.$$

Then, the derived transport map T and the Jacobian matrix have the following form.

Lemma 30. *Let $K := \arg \max_{k \in [M]} u_k(\theta)$. The transform derived from u in equation (4.10) takes the following form:*

$$T(\theta) = \nabla u_K(\theta),$$

where ∇u_i , $i = 1, \dots, M$ is given as equation (4.8). The Jacobian matrix of T is given by

$$J_T(\theta) = \nabla^2 u_K(\theta),$$

where $\nabla^2 u_i$, $i = 1, \dots, M$ is given as equation (4.9).

In the simplest case when $M = 1$, the transport map T will be reduced to the gradient of a single convex function. In the optimal case, the number M of the convex functions equals to that of the separate regions in the target distribution. If the exact number M is not known, we can start the computation with a small number of M and increase M if needed.

In addition, we can extend the construction of the transport map to the more general setting where θ contains discrete components as in Theorem 9. According to the definition of the optimal transport map T_B , we can construct the convex function ϕ_k as the maximum of multiple convex functions according to equation (4.10).

4.4.3 Algorithm description

In this section, we describe the details of the optimization procedure for the transport map. Since the optimization problem (4.3) can be highly non-convex, e.g., when the density of the target distribution is not log-concave, a good initialization of the parameters in the transport map is critical to achieve good optimization performance. Therefore, our optimization of the transport map can be decomposed into two

steps: (1) initialization of the parameters; (2) optimization of the parameters by minimizing the KL-divergence. We employ the stochastic gradient-based optimization methods to optimize the parameters in the transport map.

Initialization: In the implementation, it is helpful to initialize the parameters close to their optimal values. A transport map with good initial parameters will map points from the reference distribution close to those from the target distribution. Therefore, we first generate N samples $\{y_1, y_2, \dots, y_N\}$ i.i.d. from the target distribution. If points from the target distribution are not accessible, e.g. posterior distribution in Bayesian inference, we can apply approximation methods such as MCMC to obtain approximate points of the target distribution. Since our goal is only to find good initialization of parameters, it is not required that the approximation points have high accuracy characterizing the target distribution. After obtaining points from the target distribution, we will minimize the distance between the transported points $\{T(x_1), T(x_2), \dots, T(x_N)\}$ from the reference distribution through map T and $\{y_1, y_2, \dots, y_N\}$ to optimize the parameters in the transport map T , where we can employ Sinkhorn distance [59] or Wasserstein GAN distance [48]. We apply stochastic gradient-based methods to optimize parameters in T , where the objective function is defined as $d(\{T(x_i)\}_{i=1}^N, \{y_i\}_{i=1}^N)$ with d being Sinkhorn distance or Wasserstein GAN distance. In the implementation, we use the Python Optimal Transport [60] framework to compute the Sinkhorn distance and ADAM optimizer [61] for gradient descent.

Optimization: Since the optimization methods for both continuous and mixed variables cases are similar except for the approximation of the conditional probability in objective (4.6) which is discussed in Section 4.4.1, we will focus on the continuous variables case in the following for conciseness.

In the implement, in order to fit into the more general case where the number of the discontinuous regions in the target distribution is not known, we construct the approximate family of transport map through the maximum of convex functions as in (4.10).

$$\mathcal{T}_0 := \{T = \nabla u \mid u = \max_{i \in [M]} u_i \mid u_i(x) = \sum_{j=1}^{K_i} \alpha_{ij} f_{ij}(x) + \langle \beta_{ij}, x \rangle + \omega_{ij}\}. \quad (4.11)$$

Starting from the parameters obtained in the initialization step, we apply stochastic gradient-based methods to solve the optimization problem (4.3). In each iteration, the estimated objective function and its gradient with respect to the parameters $\gamma := (a_{ij}, b_{ij}, \alpha_{ij}, \beta_{ij}, \omega_{ij})$ are obtained through Monte Carlo method. We first generate N samples $\{x_1, x_2, \dots, x_N\}$ i.i.d. from the reference distribution μ . The estimated objective

function and its gradient are defined as

$$\begin{aligned}\widehat{d}_{\text{KL}} &:= \frac{1}{N} \sum_{i=1}^N [\log(\pi_n \circ T(x_i)) + \log |\det(J_T(x_i))|] \\ \widehat{\Gamma}_t(\gamma) &:= \frac{1}{N} \sum_{i=1}^N [\nabla_{\gamma} \log(\pi_n \circ T(x_i)) + \nabla_{\gamma} \log |\det(J_T(x_i))|],\end{aligned}$$

and the parameters are updated through

$$\gamma_{t+1} \leftarrow \gamma_t + \eta \widehat{\Gamma}_t(\gamma_t),$$

where η is the learning rate and can be updated based on ADAM optimizer [61].

In the implementation, we use the PyTorch auto-differentiation framework to obtain estimated gradient $\Gamma_t(\gamma)$, which requires the objective function to be differential with respect to the parameters. In order to make the objective function $\widehat{d}_{\text{KL}}$ differential, we modify the transport map T and its gradient through softmax approximation, i.e.,

$$\begin{aligned}T(x) &\approx \sum_{i=1}^M W_i(x) \nabla u_i(x) \\ J_T(x) &\approx \sum_{i=1}^M W_i(x) \nabla^2 u_i(x)\end{aligned}$$

with

$$W_i(x) = \frac{\exp(\gamma u_i(x))}{\sum_{j \in [M]} \exp(\gamma u_j(x))}, \quad i \in [M],$$

where $\gamma > 0$ is the concentration parameter. When γ goes to infinity, only one W_i are close to one while all other W_i 's will be close to zero. As a result, the softmax approximation will be close to the maximum function.

In the optimization, a stopping criterion is also needed. However, since the optimal value of the problem (4.3) is often not known in practice, we will stop the optimization when the empirical objective $\widehat{d}_{\text{KL}}$ does not improve within a number of iterations. Alternatively, we can use a quantitative method considered in [39], where they use the empirical variance of the objective function (4.3) over reference samples $\{x_1, x_2, \dots, x_N\}$ to monitor the convergence. When the push-forward distribution $T_{\#}\mu$ is close to π_n , the empirical variance is close to zero. Therefore, we can stop the optimization when the empirical variance falls below some positive threshold.

4.4.4 Application to Bayesian latent variable models

Many commonly encountered Bayesian latent variable models involve multiple latent variables $\{c_j\}_{j=1}^J$. Using the notation in Section 4.3.2, the mixed parameter $\theta = (\tau, \zeta)$ consists of a discrete variable $\tau = (c_1, \dots, c_J) \in [K]^J$ collecting all discrete latent variables and a continuous variable $\zeta \in \mathbb{R}^p$ collecting all model parameters (and other continuous latent variables). For example, in Gaussian mixture models and hidden Markov models [62], we have observation specific latent variables where $J = n$ and c_j is associated with the j -th observation Z_j for $j \in [n]$; in spike-and-slab regression [63], J equals to the number of covariates and c_j corresponds to the binary variable indicating the inclusion or exclusion of the j -th covariate.

In this subsection, we focus on the Bayesian Gaussian mixture model (GMM) as a representative example of Bayesian latent variable models, and illustrate how to formulate and compute the optimal transport map for approximating its posterior. We consider the following standard Bayesian hierarchical formulation of GMM in $\mathbb{R}^d$ with K isotropic mixture components, n observations $\mathbf{Z}^n = \{\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n\}$ and n observation specific latent variables $\{c_1, c_2, \dots, c_n\}$,

$$\begin{aligned} \mathbf{Z}_i \mid c_i, \{\boldsymbol{\mu}_k\}_{k=1}^K &\sim \mathcal{N}(\boldsymbol{\mu}_{c_i}, \sigma^2 \mathbf{I}_d), & i = 1, \dots, N, \\ c_i &\sim \text{Categorical}\left(\frac{1}{K}, \dots, \frac{1}{K}\right), & i = 1, \dots, N, \\ \boldsymbol{\mu}_k &\sim \mathcal{N}(\boldsymbol{\alpha}, \lambda^2 \mathbf{I}_d), & k = 1, \dots, K, \end{aligned}$$

where $\{\boldsymbol{\mu}_k\}_{k=1}^K$ are the means of K clusters and c_i is the i -th latent variables recording the cluster label of observation $\mathbf{Z}_i$. The first two lines together specify the complete data likelihood and the third line specifies the prior of model parameter $\boldsymbol{\zeta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K)$, where we assume the common variance σ^2 of clusters to be known to simplify the presentation. Based on this hierarchical formulation and the Bayes formula, we can express the joint posterior distribution π_n of $\theta = (\tau, \zeta) := (\{c_i\}_{i=1}^n, \{\boldsymbol{\mu}_k\}_{k=1}^K)$ as

$$\pi_n(\theta) = p(\tau, \zeta \mid \mathbf{Z}^n) = C \prod_{i=1}^n [p(\mathbf{Z}_i \mid c_i, \zeta) \cdot p(c_i)] \cdot \prod_{k=1}^K p(\boldsymbol{\mu}_k),$$

for all $c_i \in [K]$ with $i \in [n]$ and $\boldsymbol{\mu}_k \in \mathbb{R}^d$ with $k \in [K]$, where C is a normalization constant. In the following, we will construct a transport map class $\mathcal{T}$ consisting of transport maps in form of (4.5) to approximate the posterior π_n .

To construct a transport map as in (4.5), we first need to specify the embedding of the concatenated discrete variable $\tau \in [K]^n$ composed of n latent variables $\{c_i\}_{i=1}^n$, one per observation. Since the latent variables take the same set $[K]$ of unordered labels, it is natural to assign a vector $b_k \in \mathbb{R}^{K-1}$ to each label $k \in [K]$ such that $\|b_k - b_\ell\| = 1$ for each distinct pair $(k, \ell) \in [K]^2$, and to set the embedding of $\tau = \{c_i\}_{i=1}^n$

into $\mathcal{X}_1 = \mathbb{R}^{(K-1)n}$ as $b_\tau = (b_{c_1}, b_{c_2}, \dots, b_{c_n})$, so that the support of reference distribution μ of input $X = (X_1^{(1)}, X_2^{(1)}, \dots, X_n^{(1)}, X^{(2)})$ is $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 = \mathbb{R}^r \times \mathbb{R}^p$ with $r = (K-1)n$ and $p = Kd$. For simplicity, we use the default choice for μ as the standard multivariate normal distribution in $\mathbb{R}^{r+p}$. Unfortunately, in this setup the K^n cardinality of discrete variable τ makes the evaluation of the transport map specified by (4.5) computationally infeasible with a large sample size n . Therefore, we propose to further impose the following “mean-field” approximation to the K^n convex functions $\{\phi_{k_1, k_2, \dots, k_n} : k_i \in [K] \text{ and } i \in [n]\}$ in the formulation of transport map T_B as

$$\phi_{k_1, k_2, \dots, k_n}(x^{(2)}) = \sum_{i=1}^n \phi_{k_i}^{(i)}(x^{(2)}), \quad \forall x^{(2)} \in \mathbb{R}^p, \quad (4.12)$$

which reduces the number of convex functions to be learned from K^n to Kn . Correspondingly, the transport map from (4.5) can be decomposed into $T_B = (T_{B,1}^{(1)}, T_{B,2}^{(1)}, \dots, T_{B,n}^{(1)}, T_B^{(2)})$, where

$$\begin{cases} c_i = T_{B,i}^{(1)}(x_i^{(1)}, x^{(2)}) = \arg \max_{k \in [K]} \left\{ \langle x_i^{(1)}, b_k \rangle + \phi_k^{(i)}(x^{(2)}) \right\}, & \text{for } i \in [n], \\ \zeta = T_B^{(2)}(\{x_i^{(1)}\}_{i=1}^n, x^{(2)}) = \kappa \sum_{i=1}^n \nabla \phi_{c_i}^{(i)}(x^{(2)}). \end{cases} \quad (4.13)$$

Here $x_i^{(1)} \in \mathbb{R}^{K-1}$, $x^{(2)} \in \mathcal{X}_2$ and $x := (x_1^{(1)}, \dots, x_n^{(1)}, x^{(2)}) \in \mathcal{X}$. According to Theorem 9, any transport map T in form of (4.5) is the optimal transport map between μ and $T_\#(\mu)$. Therefore, we consider the push-forward distribution $T_\#(\mu)$ in the approximated form of (4.13) and apply numerical methods to minimize objective function (4.6), which according to lemma 28 is equivalent to minimizing the KL-divergence between the posterior distribution π_n and $T_\#(\mu)$.

The approximation in (4.13) borrows the idea from mean-field variational inference [64] to reduce of the complexity of the transport map T . However, our approximation lies between the fully flexible model and the conventional mean-field approximation that requires the approximating distribution to fully factorize over its components. In particular, under approximation scheme (4.12), the i -th latent variable c_i only depends on the i -th input variable $x_i^{(1)}$ associated with the discrete part given the input variable $x^{(2)}$ associated with the continuous part, as opposed to relying on all $x_i^{(1)}$'s, which can be interpreted as imposing some partial conditional independence structure among latent variables $\{c_i\}_{i=1}^n$ given model parameter ζ (determined by $x^{(2)}$) and significantly relaxes the full independence structure imposed by the conventional mean-field. Since it is proved in [65] that the conventional mean-field approximation already leads to consistent parameter estimation in Bayesian latent variable models, our more flexible approximation scheme will also lead to estimation consistency. Our numerical study in Section XX also illustrates that the simplifying structure (4.12) leads to a good approximation to the target posterior distribution.

To further facilitate the computation by making the scale of convex functions $\phi_k^{(i)}$ less affected by the sample size n , we may choose the relative weight parameter κ in (MPm) to be n^{-1} , so that the second equation in (4.13) for generating ζ becomes an average,

$$\zeta = T_B^{(2)}(\{x_i^{(1)}\}_{i=1}^n, x^{(2)}) = \frac{1}{n} \sum_{i=1}^n \nabla \phi_{c_i}^{(i)}(x^{(2)}),$$

where $\nabla \phi_{c_i}^{(i)}(x^{(2)})$ can be interpreted as the message about cluster centers carried by observation $\mathbf{Z}_i$. Another way to justify this choice of κ is as follows. Since the concatenated discrete variable τ is composed of n latent variables, the corresponding discrete part cost in (MPm) becomes

$$\kappa \sum_{i=1}^n \|x_i^{(1)} - T_{B,i}^{(1)}(x_i^{(1)}, x^{(2)})\|^2.$$

To avoid this discrete part cost, which is of order $\mathcal{O}(n\kappa)$, to dominate the continuous part cost of order $\mathcal{O}(1)$, we should take $\kappa = \mathcal{O}(n^{-1})$. To make the computational even simpler, we may further approximate each additive component $\phi_{k_i}^{(i)}(x^{(2)})$ in (4.12) via an additive form as $\sum_{k=1}^K \phi_{k_i,k}^{(i)}(x_k^{(2)})$, where $x_k^{(2)} \in \mathbb{R}^d$ denotes the k -th component of $x^{(2)}$ associated with the k -th cluster mean vector μ_k . This corresponds to a block mean-field approximation to model parameter $\zeta = (\mu_1, \mu_2, \dots, \mu_K)$ over the K (blocks of) parameters $\{\mu_k\}_{k=1}^K$, and preserves (ignores) the within (between) block dependence. Similarly, we may consider the full mean-field approximation on model parameter ζ by further decomposing each convex function $\phi_{k_i,k}^{(i)}(x_k^{(2)})$ with respect to the coordinates $x_k^{(2)} = (x_{k,1}^{(2)}, x_{k,2}^{(2)}, \dots, x_{k,d}^{(2)})$ in $\mathbb{R}^d$.

4.5 Numerical Results

4.5.1 Bayesian Logistic Regression

In this section, we apply our transport map to the Bayesian logistic regression model to investigate the performance of our proposed transport map when the target distribution is continuous and uni-modal. We consider the following Bayesian logistic regression model in $\mathbb{R}^p$ with N binary observations $\{y_1, y_2, \dots, y_N\}$ taking values in $\{0, 1\}$ and covariate vectors $\{x_1, x_2, \dots, x_N\} \subseteq \mathbb{R}^p$,

$$P(y_i = 1 \mid x_i, \beta) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}, \quad i = 1, \dots, N,$$

$$\beta \sim N(0, \sigma^2 I_p).$$

The first line specifies the likelihood for the binary observation y_i and the second line specifies the prior distribution of the parameter vector $\beta \in \mathbb{R}^p$, where we choose $\sigma = 10$ to mimic a non-informative prior.

We generate the data in the simulation as follows. The number of observations is $N = 1000$ and the dimension $p = 20$. The entries of the ground truth parameter β^* are generated from normal distribution $N(0, 2^2)$. We generate $x_i \in \mathbb{R}^p$ for $i = 1, \dots, N$ from normal distribution $N(0, \Sigma)$ with Toeplitz covariance matrix $\Sigma_{ij} = 0.5^{|i-j|}$.

In our experiment, we use the approximate family (4.11) to approximate the posterior distribution of the parameter β . In particular, since the posterior distribution of β is uni-model, it suffices to include $M = 1$ convex function in the family (4.11). In addition, we define the convex function u as the summation of $K = 32$ convex units with Softsign function for nonlinear function φ . For the optimization, since the target distribution is simple, we can initialize the parameters in transport maps with normal distributions and skip the initialization step described in Section 3.3.3. In order to characterize the quality of our transport map, we compare our transport map with the Gibbs sampler.

We draw 2000 samples from the Gibbs sampler and our optimized transport map. Figure 4.1 shows the approximated marginal posterior density of four parameters $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ from our optimized transport map and the Gibbs sampler, while Table 4.1 presents the 95% credible intervals for parameters from both sampling methods. From both the approximated posterior density plot and the credible intervals, we can observe that our transport map can achieve similar performance to the Gibbs sampler. Furthermore, in order to quantify the closeness between the samples drawn from Gibbs sampler and our transport map, we draw two sets of samples, say $\{x_i\}_{i=1}^B$ and $\{y_i\}_{i=1}^B$, from Gibbs sampler and one set of samples, say $\{z_i\}_{i=1}^B$, from our transport map with $B = 1000$. Then, the Sinkhorn distance between $\{x_i\}_{i=1}^B$ and $\{y_i\}_{i=1}^B$ is 0.416, while the distance between $\{x_i\}_{i=1}^B$ and $\{z_i\}_{i=1}^B$ is 0.450, which indicates that the approximated distributions from both sampling methods are close.

The simulation result shows the advantages of our transport map over the Gibbs sampler when both methods have similar sampling performance. In particular, the transport map generates independent samples from the posterior distribution, while the Gibbs sampler generates dependent samples. In addition, the transport map can be used repeatedly to generate new data once trained, while the Gibbs sampler needs to run the whole MCMC chains to obtain new samples.

	β_1	β_2	β_3	β_4
True value	1.586	1.045	3.492	-2.543
Transport map	(1.612, 2.646)	(0.560, 1.440)	(3.595, 5.189)	(-3.697, -2.450)
Gibbs sampler	(1.596, 2.655)	(0.541, 1.427)	(3.474, 5.212)	(-3.753, -2.339)

Table 4.1: 95% credible intervals for $\{\beta_1, \beta_2, \beta_3, \beta_4\}$.

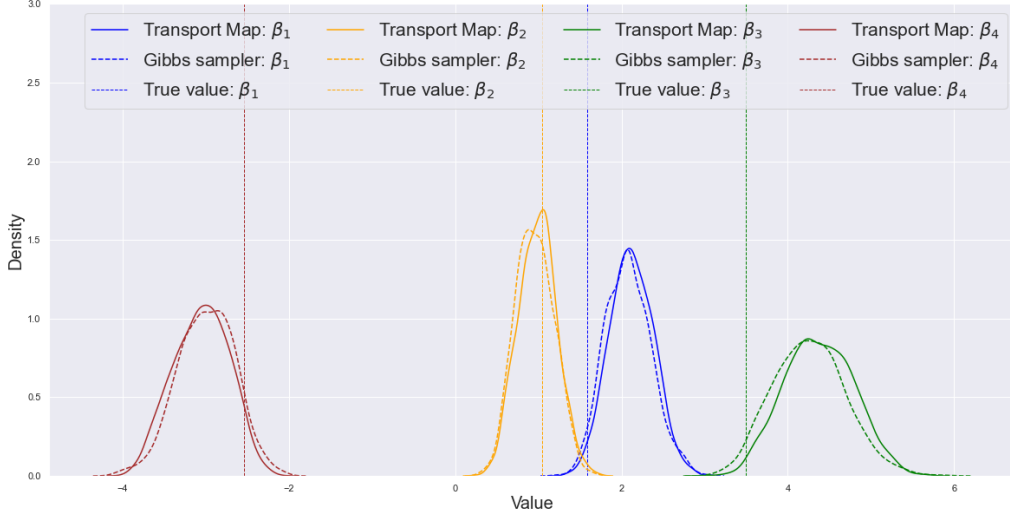


Figure 4.1: Approximated marginal posteriors of $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ from the transport map and Gibbs sampler.

4.6 Supplementary Material

4.6.1 Proof of Theorem 9

Our proof of the theorem consists of two parts whose details are provided in the following subsections. In the first part, we show that any optimal transport map must take the form of formula (4.5); and in the second part, we show that if any transport map T takes the form of (4.5) and pushforwards the reference measure μ to the target posterior π_n , or $T_{\#}(\mu) = \pi_n$, then it must be an optimal transport map from μ to π_n . Finally, these two properties together with the uniqueness and regularity properties of the optimal transport maps [55], [56] lead to the claimed results.

Formulation of optimal transport map

In this part of the proof, we derive the optimal transport map between the reference distribution μ and target distribution π_n which contains both discrete and continuous components. We define a random variable $Y = (Y^{(1)}, Y^{(2)}) := (b_{\tau}, \zeta) \in \{b_k\}_{k=1}^K \times \mathbb{R}^p$ associated with $\theta = (\tau, \zeta)$. For notation simplicity, we also assume Y to follow π_n .

The Kantorovich problem between $X \sim \mu$ and $Y \sim \pi_n$ associated with the problem (MPm) is defined as

$$(KPm) \quad \inf_{\gamma \in \Pi(\nu, \pi_n)} \int \kappa \|x^{(1)} - y^{(1)}\|^2 + \|x^{(2)} - y^{(2)}\|^2 d\gamma(x, y).$$

Then, the dual formulation for (KPm), known as Kantorovich duality, is defined as

$$\begin{aligned} & \sup_{\varphi, \psi} \int_{\mathbb{R}^r \times \mathbb{R}^p} \varphi(x) d\mu(x) + \int_{\mathbb{R}^r \times \mathbb{R}^p} \psi(y) d\pi_n(y), \\ & \text{s.t. } \varphi \in L^1(\mu), \quad \psi \in L^1(\pi_n) \\ & \text{and } \varphi(x) + \psi(y) \leq \kappa \|x^{(1)} - y^{(1)}\|^2 + \|x^{(2)} - y^{(2)}\|^2 \\ & \text{for all } x \in \mathbb{R}^r \times \mathbb{R}^p \quad \text{and} \quad y \in \{b_k\}_{k=1}^K \times \mathbb{R}^p. \end{aligned} \tag{4.14}$$

For each eligible function pair (φ, ψ) in formulation (4.14), we have

$$\begin{aligned} & \varphi(x) + \psi(y) \leq \kappa \|x^{(1)} - y^{(1)}\|^2 + \|x^{(2)} - y^{(2)}\|^2 \\ \iff & \underbrace{\kappa \|x^{(1)}\|^2 + \|x^{(2)}\|^2 - \varphi(x)}_{f(x)} + \underbrace{\kappa \|y^{(1)}\|^2 + \|y^{(2)}\|^2 - \psi(y)}_{g(y)} \\ & \geq 2\kappa \langle x^{(1)}, y^{(1)} \rangle + 2\langle x^{(2)}, y^{(2)} \rangle. \end{aligned}$$

Therefore, the dual formulation (4.14) can be equivalently formulated as

$$\begin{aligned} (DPm) \quad & C_{\mu, \pi_n} - \inf_{f, g} \left\{ \int_{\mathbb{R}^r \times \mathbb{R}^p} f(x) d\mu(x) + \int_{\mathbb{R}^r \times \mathbb{R}^p} g(y) d\pi_n(y) \right\}, \\ & \text{s.t. } f \in L^1(\mu), \quad g \in L^1(\pi_n) \\ & \text{and } f(x) + g(y) \geq 2\kappa \langle x^{(1)}, y^{(1)} \rangle + 2\langle x^{(2)}, y^{(2)} \rangle \\ & \text{for all } x \in \mathbb{R}^r \times \mathbb{R}^p \quad \text{and} \quad y \in \{b_k\}_{k=1}^K \times \mathbb{R}^p, \end{aligned}$$

where

$$C_{\mu, \pi_n} := \int_{\mathbb{R}^r \times \mathbb{R}^p} \kappa \|x^{(1)}\|^2 + \|x^{(2)}\|^2 d\mu(x) + \int_{\mathbb{R}^r \times \mathbb{R}^p} \kappa \|y^{(1)}\|^2 + \|y^{(2)}\|^2 d\pi_n(y).$$

Therefore, for any eligible pair of functions (f, g) in problem (DPm), we have

$$\begin{aligned} f(x) &\geq \sup_{z \in \{b_k\}_{k=1}^K \times \mathbb{R}^p} \left\{ 2\kappa\langle x^{(1)}, z^{(1)} \rangle + 2\langle x^{(2)}, z^{(2)} \rangle - g(z) \right\} \\ &= 2 \sup_{k \in [K]} \left\{ \kappa\langle x^{(1)}, b_k \rangle + \sup_{z^{(2)} \in \mathbb{R}^p} \left\{ \langle x^{(2)}, z^{(2)} \rangle - \frac{1}{2}g(z^{(1)}, z^{(2)}) \right\} \right\}. \end{aligned} \quad (4.15)$$

We define $h(y) := \frac{1}{2}g(y)$ over $\{b_k\}_{k=1}^K \times \mathbb{R}^p$ and $h_k(y^{(2)}) := h(b_k, y^{(2)})$ for $k \in [K]$. Then, for any function pair (f, h) eligible for problem (DPm), we have

$$f(x) + 2h(y) \geq 2\kappa\langle x^{(1)}, y^{(1)} \rangle + 2\langle x^{(2)}, y^{(2)} \rangle.$$

In the following, we show that it suffices to only consider convex functions for h_k to solve the problem (DPm).

For any function h_k in $\mathbb{R}^p$, its convex conjugate function is defined as

$$h_k^*(x^{(2)}) := \sup_{z^{(2)} \in \mathbb{R}^p} \left\{ \langle x^{(2)}, z^{(2)} \rangle - h_k(z^{(2)}) \right\}.$$

Then, we define a function $\tilde{f}$ over $\mathbb{R}^r$ as

$$\begin{aligned} \tilde{f}(x) &:= 2 \sup_{k \in [K]} \left\{ \kappa\langle x^{(1)}, b_k \rangle + h_k^*(x^{(2)}) \right\} \\ &= 2 \sup_{k \in [K]} \left\{ \kappa\langle x^{(1)}, b_k \rangle + \sup_{z^{(2)} \in \mathbb{R}^p} \left\{ \langle x^{(2)}, z^{(2)} \rangle - h_k(z^{(2)}) \right\} \right\}. \end{aligned} \quad (4.16)$$

For any function pair (f, h) eligible for problem (DPm), the pair $(\tilde{f}, h)$ improves the value of the objective function in (DPm), since $f(x) \geq \tilde{f}(x)$ for all $x \in \mathbb{R}^r$ according to the definition of $\tilde{f}$ and the inequality (4.15).

Then, we can define a function h_k^{**} over $\mathbb{R}^p$ as the convex conjugate function of h_k^* and function $h^{**}(b_k, y^{(2)}) := h_k^{**}(y^{(2)})$ over $\{b_k\}_{k=1}^K \times \mathbb{R}^p$. The function pair $(\tilde{f}, h^{**})$ will further improve $(\tilde{f}, h)$ on the value of the objective function in (DPm), since $h(y) \geq h^{**}(y)$ for all $y = (b_k, y^{(2)}) \in \{b_k\}_{k=1}^K \times \mathbb{R}^p$ by the definition of the conjugate

functions, i.e.,

$$\begin{aligned}
h^{**}(y) &= h_k^{**}(y^{(2)}) \\
&= \sup_{x^{(2)} \in \mathbb{R}^p} \left\{ \langle x^{(2)}, y^{(2)} \rangle - h_k^*(x^{(2)}) \right\} \\
&= \sup_{x^{(2)} \in \mathbb{R}^p} \left\{ \langle x^{(2)}, y^{(2)} \rangle - \sup_{z^{(2)} \in \mathbb{R}^p} \left\{ \langle x^{(2)}, z^{(2)} \rangle - h_k(z^{(2)}) \right\} \right\} \\
&\leq \sup_{x^{(2)} \in \mathbb{R}^p} \left\{ \langle x^{(2)}, y^{(2)} \rangle - \langle x^{(2)}, y^{(2)} \rangle + h_k(y^{(2)}) \right\} \\
&= h_k(y^{(2)}) = h(y).
\end{aligned}$$

Based on the above discussion, the problem (DPm) can be formulated as a problem with K convex functions $h_k(y^{(2)})$ for $k \in [K]$, i.e.,

$$\begin{aligned}
(\text{DPm}') \quad & C_{\mu, \pi_n} - 2 \cdot \inf_{f, h} \left\{ \int_{\mathbb{R}^r \times \mathbb{R}^p} f(x) d\mu(x) + \int_{\{b_k\}_{k=1}^K \times \mathbb{R}^p} h(y) d\pi_n(y) \right\}, \\
\text{s.t.} \quad & h \in L^1(\pi_n) \quad \text{and} \quad h_k(y^{(2)}) := h(b_k, y^{(2)}) \text{ is convex over } \mathbb{R}^p \text{ for } k \in [K] \\
\text{and} \quad & f(x) := \max_{k \in [K]} \left\{ \kappa \langle x^{(1)}, b_k \rangle + h_k^*(x^{(2)}) \right\}.
\end{aligned}$$

According to the duality equality between (KPm) and (DPm) and the equivalence between (DPm) and (DPm'), it is satisfied that if point (x, y) belongs to the support of the optimal transport plan γ between μ and π_n minimizing (KPm) and (f, h) is a pair of functions maximizing (DPm'), then

$$f(x) + h(y) = \kappa \langle x^{(1)}, y^{(1)} \rangle + \langle x^{(2)}, y^{(2)} \rangle.$$

This implies that for $(x, y) \in \gamma$ with $y = (b_\tau, \zeta)$, it is satisfied that

$$\begin{aligned}
& \kappa \langle x^{(1)}, b_\tau \rangle + \langle x^{(2)}, \zeta \rangle - h_\tau(\zeta) \\
&= \max_{k \in [K]} \left\{ \kappa \langle x^{(1)}, b_k \rangle + h_k^*(x^{(2)}) \right\} \\
&= \max_{k \in [K]} \left\{ \kappa \langle x^{(1)}, b_k \rangle + \sup_{z^{(2)} \in \mathbb{R}^p} \left\{ \langle x^{(2)}, z^{(2)} \rangle - h_k(z^{(2)}) \right\} \right\} \\
&\geq \left\{ \kappa \langle x^{(1)}, b_\tau \rangle + \langle x^{(2)}, \zeta \rangle - h_\tau(\zeta) \right\},
\end{aligned} \tag{4.17}$$

where the first equation is by the definition of f . Therefore, the equality is achieved in the last step. According

to equation (4.17), the discrete random variable $\tau \in [K]$ associated with b_τ is then defined as

$$\tau = \arg \max_{k \in [K]} \left\{ \kappa \langle x^{(1)}, b_k \rangle + h_k^*(x^{(2)}) \right\}.$$

Then, by plugging the definition of τ into (4.17), we have

$$h_\tau^*(x^{(2)}) = \langle x^{(2)}, \zeta \rangle - h_\tau(\zeta).$$

According to the definition of the conjugate function h_τ^* of h_τ , the following equation is satisfied.

$$\nabla_\zeta \left[\langle x^{(2)}, \zeta \rangle - h_\tau(\zeta) \right] = 0 \implies x^{(2)} - \nabla h_\tau(\zeta) = 0.$$

Therefore, we have

$$\zeta = (\nabla h_\tau)^{-1}(x^{(2)}) = \nabla h_\tau^*(x^{(2)}),$$

where the second equality is according to Lemma 31.

Based on the above relationship between the optimal transport plan γ solving (KPm) and the function pair (f, h) solving (DPm'), the optimal transport map T_B that pushforwards μ to π_n can be characterized through K convex functions $\{\phi_k\}_{k=1}^K$ over $\mathbb{R}^p$ with $\phi_k(x^{(2)}) := \frac{1}{\kappa} h_k^*(x^{(2)})$ for $k \in [K]$. Specifically, T_B maps $(x^{(1)}, x^{(2)})$ to (τ, ζ) with

$$\begin{cases} \tau &= T_B^{(1)}(x^{(1)}, x^{(2)}) = \arg \max_{k \in [K]} \left\{ \langle x^{(1)}, b_k \rangle + \phi_k(x^{(2)}) \right\}, \\ \zeta &= T_B^{(2)}(x^{(1)}, x^{(2)}) = \kappa \nabla \phi_\tau(x^{(2)}). \end{cases} \quad (4.18)$$

In addition, the convex function u_B associated with the optimal transport map T_B is in form of function f , i.e.,

$$u_B(x^{(1)}, x^{(2)}) = \kappa \max_{k \in [K]} \left\{ \langle x^{(1)}, b_k \rangle + \phi_k(x^{(2)}) \right\}.$$

Lemma 31. *Suppose function $f \in \mathbb{R}^p$ is closed, strictly convex and differentiable. Let $f^* \in \mathbb{R}^p$ denote the conjugate function of f . Then,*

$$(\nabla f^*)^{-1}(x) = \nabla f(x).$$

Proof. Since function f is closed and strictly convex, the conjugate function of f^* is still f . Therefore, for any $x \in \mathbb{R}^p$, we have

$$f(x) = \sup_{z \in \mathbb{R}^p} \{\langle x, z \rangle - f^*(z)\}.$$

Suppose the above optimization problem is maximized at $y \in \mathbb{R}^p$, then we have

$$\nabla_y [\langle x, y \rangle - f^*(y)] = 0 \implies \nabla f^*(y) = x \implies y = (\nabla f^*)^{-1}(x),$$

and

$$f(x) = \langle x, y \rangle - f^*(y). \quad (4.19)$$

By the definition of the conjugate function f^* , we have

$$f^*(y) = \sup_{z \in \mathbb{R}^p} \{\langle y, z \rangle - f(z)\}.$$

Then, according to the equation (4.19), the above optimization problem is maximized at point x . Therefore, we have

$$\nabla_x [\langle y, x \rangle - f(x)] = 0 \implies \nabla f(x) = y.$$

Combined with the above equation, it can be shown that

$$y = \nabla f(x) = (\nabla f^*)^{-1}(x).$$

□

4.6.2 Optimality of map with form (4.5)

In this part of the proof, we show the following property: suppose there exists a transport map T_0 in form of (4.18) satisfying $(T_0)_\#(\mu) = \pi_n$. The transport map T is then the optimal transport map for the problem (MPm). In addition, the induced transport plan $(id, T_0)_\#(\mu)$ is the optimal plan for the Kantorovich problem (Kpm) between μ and $\pi_n = (T_0)_\#(\mu)$.

To see this, suppose there exist convex functions $\{\phi_k\}_{k=1}^K$ such that $(T_0)_\#(\mu) = \pi_n$. We can define a pair

of (f, h) that is eligible for problem (DPm'), i.e.,

$$\begin{aligned} f(x^{(1)}, x^{(2)}) &:= \kappa \max_{k \in [K]} \left\{ \langle x^{(1)}, b_k \rangle + \phi_k(x^{(2)}) \right\} \\ h(b_\tau, \zeta) &:= (\kappa \phi_\tau)^*(\zeta) \quad \text{for } \tau \in [K], \end{aligned}$$

where $(\kappa \phi_k)^*$ is the convex conjugate function of $\kappa \phi_k$ for $k \in [K]$.

For any point (x, y) with $y = (b_\tau, \zeta)$ that belongs to the support $\gamma_0 = (id, T_0)_\#(\mu)$, we have

$$\begin{aligned} f(x) + h(y) &= \max_{k \in [K]} \kappa \left\{ \langle x^{(1)}, b_k \rangle + \phi_k(x^{(2)}) \right\} + (\kappa \phi_\tau)^*(\zeta) \\ &= \kappa \langle x^{(1)}, b_\tau \rangle + \kappa \phi_\tau(x^{(2)}) + (\kappa \phi_\tau)^*(\zeta) \\ &= \kappa \langle x^{(1)}, b_\tau \rangle + \langle x^{(2)}, \zeta \rangle, \end{aligned}$$

where the second equation is by the definition of τ in (4.18) and the third equation is by the property of the convex function $\kappa \phi_\tau$.

Therefore, we have the following inequality.

$$\begin{aligned} \min(\text{KPm}) &\leq \int \kappa \|x^{(1)} - y^{(1)}\|^2 + \|x^{(2)} - y^{(2)}\|^2 d\gamma_0(x, y) \\ &= C_{\mu, \pi_n} - 2 \int \kappa \langle x^{(1)}, y^{(1)} \rangle + \langle x^{(2)}, y^{(2)} \rangle d\gamma_0(x, y) \\ &= C_{\mu, \pi_n} - 2 \int f(x) + h(y) d\gamma_0(x, y) \\ &= C_{\mu, \pi_n} - 2 \left[\int f(x) d\mu(x) + \int h(x) d\pi_n(y) \right] \\ &\leq \max(\text{DPm}'). \end{aligned}$$

In addition, since $\min(\text{KPm}) \geq \max(\text{DPm}')$ by the inequality between the primal and dual problems, we have $\min(\text{KPm}) = \max(\text{DPm}')$. Therefore T_0 is the optimal transport map that pushforwards μ to π_n and $(id, T_0)_\#(\mu)$ is the optimal plan between μ and $\pi_n = (T_0)_\#(\mu)$.

4.6.3 Proof of Lemma 27

We first consider the case where μ and π_n are continuous, and slightly abuse the same notation for their corresponding density function

$$\begin{aligned}
D(T_{\#}\mu||\pi_n) &\equiv \int \mu(T^{-1}(\theta)) |\det JT^{-1}(\theta)| \log \frac{\mu(T^{-1}(\theta)) |\det JT^{-1}(\theta)|}{\pi_n(\theta)} d\theta \\
&= \int \mu(\xi) \log \frac{\mu(\xi) |\det JT(\xi)|^{-1}}{\pi_n(T(\xi))} d\xi \\
&= E_{\mu}[-\log \pi_n \circ T - \log |\det JT| + \log \mu].
\end{aligned} \tag{4.20}$$

Because the source distribution μ is fixed throughout the training, the third term $E_{\mu}[\log \mu]$ is a constant term. Therefore, the optimization is equivalent to the minimization of the first two terms.

If one of the distribution is not continuous, we replace the underlying Lebesgue measure by the source measure μ , and the ratio of density functions by the corresponding Radon-Nicodym derivative.

4.6.4 Proof of Lemma 29

Since n is a linear function over $x \in \mathbb{R}^p$, it is convex. Also as φ is increasing and bounded, the function $u \mapsto \int_0^u \varphi(s) ds$ is convex. As a result, the functions f and g as compositions of convex function and linear function are also convex.

References

- [1] N. Abe, A. W. Biermann, and P. M. Long, “Reinforcement learning with immediate rewards and linear hypotheses,” *Algorithmica*, vol. 37, no. 4, pp. 236–293, 2003.
- [2] P. Auer, “Using confidence bounds for exploitation-exploration trade-offs,” *Journal of Machine Learning Research*, pp. 397–422, 2003.
- [3] W. Chu, L. Li, L. Reyzin, and R. E. Schapire, “Contextual bandits with linear payoff functions,” *AISTATS*, pp. 208–214, 2011.
- [4] H. Bastani and M. Bayati, “Online decision making with high-dimensional covariates,” *Operations Research*, vol. 68, no. 1, pp. 276–294, 2019.
- [5] X. Wang, M. Wei, and T. Yao, “Minimax concave penalized multi-armed bandit model with high-dimensional covariates,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 5200–5208.
- [6] C.-H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, vol. 38, pp. 94–942, 2010.
- [7] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American statistical Association*, vol. 96, pp. 1348–1360, 2001.
- [8] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.
- [9] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.
- [10] F. Santambrogio, “Optimal transport for applied mathematicians,” *Birkäuser, NY*, vol. 55, no. 58-63, p. 94, 2015.
- [11] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.
- [12] V. Dani, T. P. Hayes, and S. M. Kakade, “Stochastic linear optimization under bandit feedback,” *Conference On Learning Theory*, pp. 355–366, 2008.

- [13] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.
- [14] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.
- [15] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [16] R. I. Oliveira, “The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties,” *Probability Theory and Related Fields*, vol. 166, no. 3, pp. 1175–1194, 2016.
- [17] M. Rudelson and S. Zhou, “Reconstruction from anisotropic random measurements,” *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3434–3447, 2013.
- [18] P. Bühlmann and S. V. D. Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- [19] A. Goldenshluger and A. Zeevi, “A linear response bandit problem,” *Stochastic Systems*, vol. 3, no. 1, pp. 230–261, 2013.
- [20] J.-Y. Audibert, A. B. Tsybakov, *et al.*, “Fast learning rates for plug-in classifiers,” *The Annals of statistics*, vol. 35, pp. 608–633, 2007.
- [21] E. Candes and T. Tao, “Decoding by linear programming,” *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [22] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.
- [23] A. Goldenshluger and A. Zeevi, “Woodroffe’s one-armed bandit problem revisited,” *The Annals of Applied Probability*, vol. 19, no. 4, pp. 1603–1633, 2009.
- [24] P. Rusmevichientong and J. N. Tsitsiklis, “Linearly parameterized bandits,” *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, 2010.
- [25] C.-H. Zhang *et al.*, “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of statistics*, vol. 38, pp. 894–942, 2010.
- [26] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American statistical Association*, vol. 96, pp. 1348–1360, 2001.

- [27] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Online-to-confidence-set conversions and application to sparse stochastic bandits,” *AISTATS*, pp. 1–9, 2012.
- [28] G.-S. Kim and M. C. Paik, “Doubly-robust lasso bandit,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.
- [29] I. W. P. Consortium, “Estimation of the warfarin dose with clinical and pharmacogenetic data,” *New England Journal of Medicine*, vol. 360, no. 8, p. 753, 2009.
- [30] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds., Cambridge University Press, 2012, pp. 210–268.
- [31] K. Li, Y. Yang, and N. N. Narisetty, “Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit,” *Electronic Journal of Statistics*, vol. 15, no. 2, pp. 5652–5695, 2021.
- [32] P.-L. Loh and M. J. Wainwright, “Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima,” 2015.
- [33] —, “Support recovery without incoherence: A case for nonconvex regularization,” *The Annals of Statistics*, vol. 45, no. 6, pp. 2455–2482, 2017.
- [34] L. Gan, N. N. Narisetty, and F. Liang, “On the optimal theoretical properties of bayesian regularization,” 2019.
- [35] P. Zhao and B. Yu, “On model selection consistency of lasso,” *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.
- [36] R. R. Hocking, “A biometrics invited paper. the analysis and selection of variables in linear regression,” *Biometrics*, vol. 32, no. 1, pp. 1–49, 1976.
- [37] A. S. I. Service, “Africa soil property prediction challenge,” *Kaggle*,
- [38] P. Bühlmann, M. Kalisch, and L. Meier, “High-dimensional statistics with a view toward applications in biology,” *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 255–278, 2014.
- [39] T. A. El Moselhy and Y. M. Marzouk, “Bayesian inference with optimal maps,” *Journal of Computational Physics*, vol. 231, no. 23, pp. 7815–7850, 2012.
- [40] M. Hoffman, P. Sountsov, J. V. Dillon, I. Langmore, D. Tran, and S. Vasudevan, “Neutra-lizing bad geometry in hamiltonian monte carlo using neural transport,” *arXiv preprint arXiv:1903.03704*, 2019.

- [41] L. L. Duan, “Transport monte carlo: High-accuracy posterior approximation via random transport,” *Journal of the American Statistical Association*, no. just-accepted, pp. 1–30, 2021.
- [42] M. Katzfuss and F. Schäfer, “Scalable bayesian transport maps for high-dimensional non-gaussian spatial fields,” *arXiv preprint arXiv:2108.04211*, 2021.
- [43] J. O. Berger and J. M. Bernardo, “On the development of the reference prior method,” *Bayesian statistics*, vol. 4, no. 4, pp. 35–60, 1992.
- [44] J. O. Berger, J. M. Bernardo, and D. Sun, “The formal definition of reference priors,” *The Annals of Statistics*, vol. 37, no. 2, pp. 905–938, 2009.
- [45] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, *Generative adversarial networks*, 2014. arXiv: [1406.2661 \[stat.ML\]](#).
- [46] Y. Li, K. Swersky, and R. Zemel, *Generative moment matching networks*, 2015. arXiv: [1502.02761 \[cs.LG\]](#).
- [47] G. Biau, B. Cadre, M. Sangnier, and U. Tanielian, “Some theoretical properties of GANS,” *The Annals of Statistics*, vol. 48, no. 3, pp. 1539–1566, 2020. DOI: [10.1214/19-AOS1858](#). [Online]. Available: <https://doi.org/10.1214/19-AOS1858>.
- [48] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*, PMLR, 2017, pp. 214–223.
- [49] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, *Wasserstein auto-encoders*, 2019. arXiv: [1711.01558 \[stat.ML\]](#).
- [50] S. Zhao, J. Song, and S. Ermon, *Infovae: Information maximizing variational autoencoders*, 2018. arXiv: [1706.02262 \[cs.LG\]](#).
- [51] A. Brock, J. Donahue, and K. Simonyan, *Large scale gan training for high fidelity natural image synthesis*, 2018. arXiv: [1809.11096 \[cs.LG\]](#).
- [52] A. van den Oord, S. Dieleman, H. Zen, *et al.*, *Wavenet: A generative model for raw audio*, 2016. arXiv: [1609.03499 \[cs.SD\]](#).
- [53] S. Mohamed and B. Lakshminarayanan, “Learning in implicit generative models,” *arXiv preprint arXiv:1610.03483*, 2016.
- [54] F. Huszár, “Variational inference using implicit distributions,” *arXiv preprint arXiv:1702.08235*, 2017.
- [55] Y. Brenier, “Polar factorization and monotone rearrangement of vector-valued functions,” *Communications on pure and applied mathematics*, vol. 44, no. 4, pp. 375–417, 1991.

- [56] R. J. McCann *et al.*, “Existence and uniqueness of monotone measure-preserving maps,” *Duke Mathematical Journal*, vol. 80, no. 2, pp. 309–324, 1995.
- [57] J. Kitagawa and R. McCann, “Free discontinuities in optimal transport,” *Archive for Rational Mechanics and Analysis*, vol. 232, no. 3, pp. 1505–1541, 2019.
- [58] D. B. Dunson, N. Pillai, and J.-H. Park, “Bayesian density regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 163–183, 2007.
- [59] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS’13, Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 2292–2300.
- [60] R. Flamary, N. Courty, A. Gramfort, *et al.*, “Pot: Python optimal transport,” *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021. [Online]. Available: <http://jmlr.org/papers/v22/20-451.html>.
- [61] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1412.6980>.
- [62] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [63] H. Ishwaran and J. S. Rao, “Spike and slab variable selection: Frequentist and bayesian strategies,” *The Annals of Statistics*, vol. 33, no. 2, pp. 730–773, 2005.
- [64] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
- [65] Y. Yang, D. Pati, and A. Bhattacharya, “ α -variational inference with statistical guarantees,” *The Annals of Statistics*, vol. 48, no. 2, pp. 886–905, 2020.