High-dimensional Linear Bandits with Knapsacks

Wanteng Ma[†] Dong Xia[†] Jiashuo Jiang[‡]

† Department of Mathematics, The Hong Kong University of Science and Technology
‡ Department of Industrial Engineering & Decision Analytics, The Hong Kong University of Science and Technology

We study the contextual bandits with knapsack (CBwK) problem under the high-dimensional setting where the dimension of the feature is large. The reward of pulling each arm equals the multiplication of a sparse high-dimensional weight vector and the feature of the current arrival, with additional random noise. In this paper, we investigate how to exploit this sparsity structure to achieve improved regret for the CBwK problem. To this end, we first develop an online variant of the hard thresholding algorithm that performs the sparse estimation in an online manner. We further combine our online estimator with a primal-dual framework, where we assign a dual variable to each knapsack constraint and utilize an online learning algorithm to update the dual variable, thereby controlling the consumption of the knapsack capacity. We show that this integrated approach allows us to achieve a sublinear regret that depends logarithmically on the feature dimension, thus improving the polynomial dependency established in the previous literature. We also apply our framework to the high-dimension contextual bandit problem without the knapsack constraint and achieve optimal regret in both the data-poor regime and the data-rich regime. We finally conduct numerical experiments to show the efficient empirical performance of our algorithms under the high dimensional setting.

1. Introduction

Introduced in the seminal paper Badanidiyuru et al. (2013), the bandit with knapsacks problem (BwK) is defined by solving an online *knapsack* problem with global size constraints. This kind of problem is a special but important case of the online allocation problem, which imposes a reward-agnostic assumption on resource allocations. The bandit with knapsacks problem has been broadly applied to many scenarios, e.g., ad allocation, dynamic pricing, repeated auctions, etc. In fact, in several applications like online recommendation or online advertising, many contexts (or features, covariates) of rewards that we can observe are possibly high-dimensional, which significantly contribute to the decision-making and motivate us to consider a variant of the BwK problem, i.e., the contextual bandit with knapsacks problem (Badanidiyuru et al. 2014). However, although the contextual bandit with knapsacks problem has been extensively studied under different settings (Agrawal and Devanur 2014, 2016, Immorlica et al. 2022, Liu et al. 2022), previous studies largely neglect the inherent high dimensionality of covariates, and in turn, incur regrets that depend polynomially on the large dimension d, making these methods less feasible in the high-dimensional

setting. This motivates us to explore further approaches that can handle the BwK problem in the high-dimensional case, which is an emergent topic in online learning.

In this paper, we address this challenge by proposing efficient methods to solve the high-dimensional linear contextual bandit with knapsacks problem. Our method consists of two parts, primal estimation and dual-based allocation. We will show that our online method in primal estimation can achieve exact sparse recovery with optimal statistical error, which is comparable with the renowned LASSO method but with less computational cost. Together with dual allocation, our primal-dual method can effectively control the regret of BwK problem in the order $\widetilde{O}\left(\frac{V^{\mathrm{UB}}}{C_{\min}}\sqrt{T}+\left(\frac{V^{\mathrm{UB}}}{C_{\min}}\right)^{\frac{1}{3}}T^{\frac{2}{3}}\right)$, which is logarithmically dependent on the dimension d. Moreover, we also show that the regret can be further improved to $\widetilde{O}\left(\frac{V^{\mathrm{UB}}}{C_{\min}}\sqrt{T}\right)$ with additional diverse covariate condition.

Our method also brings new insights into the general online sparse estimation and sparse bandit problem. For the sparse bandit problem, most of the existing literature heavily relies on LASSO, which explores sparsity by regularized sample average approximation (SAA). Although LASSO guarantees good theoretical results, it is hard to perform in an online fashion. In this paper, we solve the sparse recovery problem through a novel stochastic approximation approach with hard thresholding, which is more aligned with online learning and is also statistically optimal. This estimation algorithm leads to a by-product, i.e., a unified sparse bandit algorithm framework that reaches desired optimal regrets $\widetilde{O}(s_0^{2/3}T^{2/3})$ and $\widetilde{O}(\sqrt{s_0T})$, in both data-poor and data-rich regimes respectively, which satisfies the so-called "the best of two worlds" (Hao et al. 2020).

1.1. Main Results and Contributions

Our main results and contributions can be summarized as follows.

First, we develop a new online sparse estimation algorithm, named Online HT, that performs the sparse estimation in an online manner. Note that previous methods for sparse estimation, like LASSO (e.g. Hao et al. (2020), Li et al. (2022), Ren and Zhou (2023)) and iterative hard thresholding (Blumensath and Davies 2009, Nguyen et al. 2017), perform the estimation in an offline manner and thus require us to store the entire historical data set, on the size of $O(d\cdot T)$, which can be costly when both the dimension and time epoch are large. In contrast, our algorithm is an online variant of the hard thresholding method and features a gradient-averaging technique that only requires us to store the average of the previous estimations, on the size of $O(d^2)$, instead of the entire data set. Moreover, the computation complexity of the sparse estimation step can be reduced by our approach. To be specific, the computational complexity of Online HT is $O(d^2)$ per iteration and $O(d^2T)$ in total, while the computational complexity of classical LASSO solution is $O(d^3+d^2t)$ per iteration (Efron et al. 2004), and $O(d^3T+d^2T^2)$ in total if we require constant

updates of the estimation, e.g., Kim and Paik (2019), Ren and Zhou (2023). In this way, our online estimator enjoys a greater computational benefit than the offline estimator established in the previous literature.

Second, we show that the online update of our Online HT algorithm can be naturally combined with a primal-dual framework to solve the high dimensional CBwK problem. To be specific, for each resource constraint, we introduce a dual variable. Though previous work (e.g. Badanidiyuru et al. (2013), Agrawal and Devanur (2016)) on BwK and CBwK problem has shown that a sublinear regret can be achieved by applying online learning algorithms to update the dual variables and control the resource consumption, these regret bounds depend polynomially on the feature dimension, for example, the $O(d \cdot \sqrt{T})$ regret bound in Agrawal and Devanur (2016) and the $O(\sqrt{d \cdot T})$ regret bound in Han et al. (2023b). The difference in our approach is that we use the output of the Online HT algorithm at the current step to serve as the primal estimation for the dual update. In this way, we consecutively update the primal estimation by Online HT and update the dual variable by the online mirror descent algorithm in each iteration. We show that this integrated approach can effectively exploit the sparsity structure of our problem and achieve a regret that depends logarithmically on both the dimension d and constraints number m. Thus, our approach performs the online allocation of the CBwK problem more efficiently in the high-dimensional setting when d is relatively large. We conduct numerical experiments to further illustrate the superiority of the empirical performances of our algorithm under the high-dimensional setting.

Finally, our Online HT algorithm framework can be broadly applied to many other high-dimensional problems to achieve the statistically optimal estimation rate. For example, we applied the Online HT to the high-dimensional contextual bandit problem, which can be regarded as a special case of the high-dimensional contextual CBwK problem where the resource constraints are absent. We show that our algorithm reaches the desired optimal regrets $\widetilde{O}(s_0^{2/3}T^{2/3})$ for the data-poor regime and $\widetilde{O}(\sqrt{s_0 T})$ for the general data-rich regimes under the extra diverse covariate condition. In this way, we achieve the so-called "the best of two worlds" (Hao et al. 2020) without additional phase splitting and signal requirements (Hao et al. 2020, Jang et al. 2022).

1.2. Related Literature

Bandit with knapsacks problem (Badanidiyuru et al. 2013, Agrawal and Devanur 2014) can be viewed as a special case of online allocation problem, where reward functions are unknown for decision-makers. Unlike other resource allocation problems (Jiang et al. 2020, Balseiro et al. 2023, Ma et al. 2022), BwK problem poses strong demands on balancing exploration and exploitation. In the face of uncertainty, this trade-off is mainly handled by, e.g., elimination-based algorithms (Badanidiyuru et al. 2013, 2018), or UCB (Agrawal and Devanur 2014), or primal-dual algorithms

(Badanidiyuru et al. 2013, Li et al. 2021), which are all guaranteed to be optimal for problem independent settings. In the contextual BwK problem (CBwK), some well-established methods have been proposed, including policy elimination (Badanidiyuru et al. 2014) and UCB-type algorithm (Agrawal and Devanur 2016), which both originated from contextual bandit problem. However, the currently well-known CBwK methods (Badanidiyuru et al. 2014, Agrawal and Devanur 2016) all suffer from $O(\sqrt{d})$ dependence on the dimension in the regret, which hugely confines their applicants to the low-dimensional case. The failure of classic CBwK methods for large d strongly motivates us to explore the CBwK problem with high-dimensional contexts, which is frequently encountered in the real world like user-specific recommendations and personalized treatments (Bastani and Bayati 2020).

To study high-dimensional CBwK problems, naturally, we may think of learning experiences from high-dimensional contextual bandit problems. Actually, as the origin of the CBwK problem, the contextual bandit problem has been more actively studied in high-dimensional settings. Based on the LASSO method, many sampling strategies have been devised. Noticeable force-sampling strategy in Bastani and Bayati (2020) achieves a regret $O\left(s_0^2 \cdot (\log d + \log T)^2\right)$ under the margin condition, and has been improved by Wang et al. (2018) to a sharper minimax rate $O\left(s_0^2 \cdot (\log d + s_0) \cdot \log T\right)$ with concave penalized LASSO. Kim and Paik (2019) has constructed a doubly-robust ε-greedy sampling strategy by re-solving LASSO, yielding a regret of order $\widetilde{O}(s_0\sqrt{T})$ under vanishing noise size. Hao et al. (2020) introduced an Explore-then-Commit LASSO bandit framework with the regret $\widetilde{O}(s_0^{2/3}T^{2/3})$, and this framework has been followed up by, e.g., Li et al. (2022), Jang et al. (2022). As is shown in Jang et al. (2022), the regret lower bound of sparse bandit problem is $\Omega\left(\phi_{\min}^{-2/3}s_0^{2/3}T^{2/3}\right)$ in the data-poor regime $d \geq T^{\frac{1}{3}}s_0^{\frac{4}{3}}$. However, another stream of work showed that, for the general data-rich regime, the optimal regret is of order $\Omega(\sqrt{s_0 T})$ (Chu et al. 2011, Ren and Zhou 2023) and can be obtained with additional covariate conditions, for example, diverse covariate condition (Ren and Zhou 2023), and balanced covariance condition, (Oh et al. 2021, Ariu et al. 2022), etc. The two-phase optimal regret of the sparse bandit problem leads to an open question, i.e., can we achieve "the best of two worlds" of sparse bandit problem in both data-poor and data-rich regimes with a unified framework (Hao et al. 2020)? In our paper, we will answer this question affirmatively by providing our Online HT algorithm in the sparse bandit setting.

The idea of hard thresholding is applied in our methodology for the consecutive online estimation. Hard thresholding finds its application in sparse recovery primarily for the iterative hard thresholding methods (Blumensath and Davies 2009). One of the most intriguing properties of hard thresholding is that it can return an exact sparse estimation given any sparsity level. Nonetheless, the poor smoothness behavior inhered in the hard thresholding projector (Shen and Li 2017) makes

it difficult to analyze the error for iterative methods, especially for stochastic gradient descent methods with large variances. Therefore, current applications of hard thresholding mainly focus on batch learning (Nguyen et al. 2017, Yuan and Li 2021) or hybrid learning (Zhou et al. 2018), while hard thresholding methods for online learning are still largely unexplored.

2. High-dimensional Contextual BwK

We consider the high-dimensional contextual bandit with knapsacks problem over a finite horizon of T periods. There are m resources and each resource $i \in [m]$ has an initial capacity C_i. The capacity vector is denoted by $\boldsymbol{C} \in \mathbb{R}^m$. We normalize the vector $\boldsymbol{C}$ such that $C_i/T \in [0,1]$ for each $i \in [m]$. There are K arms, together with a null arm that generate no reward and consume no resources to perform void action. At each period $t \in [T]$, one query arrives, denoted by query t, and is associated with a feature $\boldsymbol{x}_t \in \mathbb{R}^d$. We assume that the feature $\boldsymbol{x}_t$ is drawn from a distribution $F(\cdot)$ independently at each period t. For each arm $a \in [K]$, query t is associated with a reward $r_t(a, \boldsymbol{x}_t)$ and a size $\boldsymbol{b}(a, \boldsymbol{x}_t) = (b_1(a_t, \boldsymbol{x}_t), \ldots, b_m(a_t, \boldsymbol{x}_t)) \in \mathbb{R}^m_{\geq 0}$. Note that the reward $r(a_t, \boldsymbol{x}_t)$ and the size $\boldsymbol{b}(a_t, \boldsymbol{x}_t)$ depends on the feature $\boldsymbol{x}_t$ and the arm a. For each arm $a \in [K]$, we assume that the size $\boldsymbol{b}(a, \boldsymbol{x}_t)$ follows the following relationship

$$\boldsymbol{b}(a, \boldsymbol{x}_t) = \boldsymbol{W}_a^\star \boldsymbol{x}_t \tag{1}$$

where $\boldsymbol{W}_a^\star \in \mathbb{R}^{m \times d}$ is a weight matrix and is assumed to be known for simplicity, specified for each arm $a \in [K]$. Note that all our results can be directly generalized to the setting where the weight matrix $\boldsymbol{W}_a^*$ is unknown for each $a \in [K]$, as described in Section 7. The reward $r(a, \boldsymbol{x}_t)$ is stochastic and is assumed to follow the relationship

$$r(a, \boldsymbol{x}_t) = (\boldsymbol{\mu}_a^\star)^\top \boldsymbol{x}_t + \xi_t \tag{2}$$

where $\boldsymbol{\mu}_a^\star \in \mathbb{R}^d$ is an *unknown* weight vector, specified for each arm $a \in [m]$, and ξ_t is a random noise following a sub-Gaussian distribution with parameter σ independently, with expectation equals 0.

After seeing the feature $\boldsymbol{x}_t$, a decision maker must decide online which arm to pull. If arm a_t is pulled for query t, then each resource $i \in [m]$ will be consumed by $b_i(a_t, \boldsymbol{x}_t)$ units and a reward $r_t(a_t, \boldsymbol{x}_t)$ will be collected. The realized value of $r_t(a_t, \boldsymbol{x}_t)$ is also observed. Note that query t is only feasible to be served if the remaining capacities exceed $\boldsymbol{b}(a_t, \boldsymbol{x}_t)$ component-wise. The decision maker's goal is to maximize the total collected reward subject to the resource capacity constraint.

The benchmark is the offline decision maker that is aware of the value of $\boldsymbol{\mu}_a^\star$ and $\boldsymbol{x}_t$ for all $a \in [K]$, $t \in [T]$ and always makes the optimal decision in hindsight. We denote by $\{y_{a,t}^{\text{off}}, \forall a \in [K]\}_{t=1}^T$ the offline decision of the offline optimum, which is an optimal solution to the following offline problem:

$$V^{\text{Off}}(I) = \max \sum_{t=1}^{T} \sum_{a \in [K]} ((\boldsymbol{\mu}_a^\star)^\top \boldsymbol{x}_t \cdot y_{a,t} + \xi_t) \tag{3}$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{t=1}^{T} \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t \cdot y_{a,t} \le \boldsymbol{C} \\ & \sum_{a\in[K]} y_{a,t} = 1, \qquad \forall t \in [T] \\ & y_{a,t} \in \{0,1\} \qquad \forall a \in [K] \forall t \in [T]. \end{aligned}$$

For any feasible online policy π, we use *regret* to measure its performance, which is defined as follows:

$$\text{Regret}(\pi) := \mathbb{E}_{I\sim F}[V^{\text{Off}}(I)] - \mathbb{E}_{I\sim F}[V^{\pi}(I)] \tag{4}$$

where $I = \{(\boldsymbol{x}_t, \xi_t)\}_{t=1}^T \sim F$ denotes that $\boldsymbol{x}_t$ follows distribution $F(\cdot)$ independently for each $t \in [T]$, and $V^\pi(I)$ denotes the total value collected under the policy π. A common upper bound of $\mathbb{E}_{I\sim F}[V^{\text{off}}(I)]$ can be formulated as follows:

$$\begin{aligned} V^{\text{UB}} = \max \quad & \sum_{t=1}^{T} \sum_{a\in[K]} \mathbb{E}_{\boldsymbol{x}_t\sim F} \left[(\boldsymbol{\mu}_a^\star)^\top \boldsymbol{x}_t \cdot y_{a,t}(\boldsymbol{x}_t) \right] \\ \text{s.t.} \quad & \sum_{t=1}^{T} \sum_{a\in[K]} \mathbb{E}_{\boldsymbol{x}_t\sim F} \left[\boldsymbol{W}_a^\star \boldsymbol{x}_t \cdot y_{a,t}(\boldsymbol{x}_t) \right] \le \boldsymbol{C}, \quad \forall i \in [m] \\ & \sum_{a\in[K]} y_{a,t}(\boldsymbol{x}_t) = 1, \qquad \forall t \in [T], \forall \boldsymbol{x}_t \\ & y_{a,t}(\boldsymbol{x}_t) \in [0,1] \qquad \forall a \in [K], \forall t \in [T], \forall \boldsymbol{x}_t. \end{aligned} \tag{5}$$

The following result is standard in the literature, which formally establishes the fact that V^{UB} can be used to upper bound the regret of any policy π.

LEMMA 1 **(folklore)**. *It holds that* $\mathbb{E}_{I\sim F}[V^{\text{Off}}(I)] \le V^{\text{UB}}$.

Therefore, in what follows, we benchmark against V^{UB} and we exploit the structures of V^{UB} to derive our online policy and bound the regret.

2.1. High-dimensional features and sparsity structures

We consider the case where the dimension of the feature d is very large and a sparsity structure exists for the weight vector $\boldsymbol{\mu}_a^\star$. Specifically, we assume the sparsity level $\|\boldsymbol{\mu}_a^\star\|_0 \le s_0$ for each a, given $s_0 \ll d$, and a bound on the general range of arms: $\|\boldsymbol{\mu}_a^\star\|_\infty \le 1$. To establish the theory of online learning, one must ensure that the information of each $\boldsymbol{\mu}_a^\star$ can be retrieved statistically based on the observation. The following basic assumptions are necessary for such sparse learning.

ASSUMPTION 1. *We make the following assumptions throughout the paper.*

(a). There exists a constant D such that the covariate $\boldsymbol{x}_t$ is uniformly bounded: $\|\boldsymbol{x}_t\|_\infty \le D$.

(b). There exists a constant D' such that for any arm a covariate $\boldsymbol{x}$, it holds that $\|\boldsymbol{b}(a, \boldsymbol{x}_t)\|_\infty \le D'$.

(c). For any s, the covariance matrix $\mathbf{\Sigma} := \mathbb{E}\boldsymbol{x}_t\boldsymbol{x}_t^\top$ has the $2s$-sparse minimal eigenvalue $\phi_{\min}(s)$ and $2s$-sparse maximal eigenvalue $\phi_{\max}(s)$ (Meinshausen and Yu 2008):

$$\phi_{\min}(s) = \min_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0 \le \lceil 2s \rceil} \frac{\boldsymbol{\beta}^\top \mathbf{\Sigma} \boldsymbol{\beta}}{\boldsymbol{\beta}^\top \boldsymbol{\beta}}, \text{ and } \phi_{\max}(s) = \max_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0 \le \lceil 2s \rceil} \frac{\boldsymbol{\beta}^\top \mathbf{\Sigma} \boldsymbol{\beta}}{\boldsymbol{\beta}^\top \boldsymbol{\beta}}.$$

Then the condition number of our problem can be defined as $\kappa = \frac{\phi_{\max}(s)}{\phi_{\min}(s)}$.

The sparse minimal eigenvalue condition essentially shares the same idea as the restrict eigenvalue conditions that have been broadly used in the high-dimensional sparse bandit problem (Hao et al. 2020, Oh et al. 2021, Li et al. 2022). It ensures that the sparse structure can be detected from the sampling.

3. Optimal Online Sparse Estimation

The primal task for our online learning problem is to estimate the high-dimensional arms during the exploration, which serves as the foundation of our learning strategies. To this end, we focus on estimating one specific arm in this section, say, estimating $\boldsymbol{\mu}_a^\star$ for one $a \in [K]$. Since $\|\boldsymbol{\mu}_a^\star\|_0 = s_0$ for $s_0 \ll d$, for the linear problem, recovering $\boldsymbol{\mu}_a^\star$ is equivalent to the following ℓ_0 constrained optimization problem:

$$\min_{\|\boldsymbol{\mu}\|_0 \le s_0} f(\boldsymbol{\mu}) := \mathbb{E}(r_t - \boldsymbol{\mu}^\top \boldsymbol{x}_t)^2 = \|\boldsymbol{\mu} - \boldsymbol{\mu}_a^\star\|_{\mathbf{\Sigma}}^2 + \sigma^2. \tag{6}$$

To solve (6), LASSO is massively used in the literature. Despite its statistical optimality, such a method heavily relies on the accumulated data to perform the ℓ_1-regularized optimization, which can not be easily adapted to the online setting, especially sequential estimations. Thus, in high-dimensional online learning, finding an online sparse estimation algorithm that runs *fully online* and still achieves optimal statistical rate is imperative. We describe our proposed optimal online sparse estimation algorithm in Algorithm 1 in the context of ϵ-greedy sampling strategy. To ease the notation, we define the sparse projection $\mathcal{H}_s(x)$ as the hard thresholding operator that zeros out all the signals in x except the largest (in absolute value) s entries. Here we denote $\rho = s_0/s$ as the relative sparsity level.

THEOREM 1. *Define $\underline{p_j} = \inf p_{a,j}$ as the lower bound of each $p_{a,j}$ and suppose $\underline{p_j} \le O(j^{-\frac{1}{3}})$. If we take $\rho = \frac{1}{9\kappa^4}$, and $\eta_t = \frac{1}{4\kappa\phi_{\max}(s)}$, then under Assumption 1, the output of Algorithm 1 satisfies*

$$\mathbb{E}\left\|\boldsymbol{\mu}_{a,t}^{\mathsf{s}} - \boldsymbol{\mu}_a^\star\right\|_2^2 \lesssim \frac{\sigma^2 D^2 s_0}{\phi_{\min}^2(s)} \frac{\log d}{t^2} \left(\sum_{j=1}^{t} \frac{1}{\underline{p_j}}\right),$$

and the high-probability bound

$$\left\|\boldsymbol{\mu}_{a,t}^{\mathsf{s}} - \boldsymbol{\mu}_a^\star\right\|_2^2 \lesssim \frac{\sigma^2 D^2 s_0}{\phi_{\min}^2(s)} \frac{\log(dT/\varepsilon)}{t^2} \left(\sum_{j=1}^{t} \frac{1}{\underline{p_j}}\right),$$

which holds for all t with probability at least $1-\varepsilon$.

Algorithm 1 Online Hard Thresholding with Averaged Gradient (Online HT)

Input: T, step size η_t, sparsity level s, arm a, $\widetilde{\boldsymbol{\mu}}_{a,0} = \boldsymbol{\mu}_{a,0} = \mathbf{0}$

for $t = 1, \ldots, T$ **do**

Sample the reward according to the decision variable $y_{a,t} \sim \mathrm{Ber}(p_{a,t})$, where

$$p_{a,t} \in \sigma(\mathcal{H}_{t-1}, \boldsymbol{x}_t)$$

if $p_{a,t} = 0$ **then**

Treat $y_{a,t}/p_{a,t} = 0$ in the following computation

end if

Compute the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{a,t} = 1/t \cdot \left((t-1)\widehat{\boldsymbol{\Sigma}}_{a,t-1} + y_{a,t}\boldsymbol{x}_t\boldsymbol{x}_t^\top/p_{a,t}\right)$

Get averaged stochastic gradient: $\boldsymbol{g}_{a,t} = 2\widehat{\boldsymbol{\Sigma}}_{a,t}\boldsymbol{\mu}_{a,t-1} - \frac{2}{t}\sum_{j=1}^{t} y_{a,j}\boldsymbol{x}_j r_j/p_{a,j}$

Gradient descent with hard thresholding: $\boldsymbol{\mu}_{a,t} = \mathcal{H}_s(\boldsymbol{\mu}_{a,t-1} - \eta_t \boldsymbol{g}_{a,t})$

Exact s_0-sparse estimation $\boldsymbol{\mu}^{\mathrm{s}}_{a,t} = \mathcal{H}_{s_0}(\boldsymbol{\mu}_{a,t})$

end for

Output: $\{\boldsymbol{\mu}^{\mathrm{s}}_t\}$, $t \in [T]$

Algorithm 1 serves as an online counterpart of the classic LASSO method. It achieves the statistically optimal rate of sparse estimation in the sense that, if we force $p_{a,j} = 1$ for each j, then we obtain the estimation error $O\left(\frac{s_0\sigma^2\log d}{\phi^2_{\min}(s)t}\right)$, which matches the well-known optimal sparse estimation error rate (Ye and Zhang 2010, Tsybakov and Rigollet 2011). Algorithm 1 needs to continuously maintain an empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_{a,t}$, which takes up $O(d^2)$ storage space; however, all the updates of $\widehat{\boldsymbol{\Sigma}}_{a,t}$ and stochastic gradients $\boldsymbol{g}_{a,t}$ can be computed linearly, which leads to the fast $O(d^2T)$ total computational complexity. Moreover, our bound can be easily extended to the uniform bound over all arms $\mathbb{E}\max_{a\in[K]}\left\|\boldsymbol{\mu}^{\mathrm{s}}_{a,t} - \boldsymbol{\mu}^\star_a\right\|_2^2$, with only an additional $\log K$ term on the error rate. See Corollary 1 for the exact error bound.

Corollary 1. *Under the same condition as Theorem 1, we have the following uniform bound for the estimations over all arms*

$$\mathbb{E}\max_{a\in[K]}\left\|\boldsymbol{\mu}^{\mathrm{s}}_{a,t} - \boldsymbol{\mu}^\star_a\right\|_2^2 \lesssim \frac{\sigma^2 D^2 s_0}{\phi^2_{\min}(s)}\frac{\log(dK)}{t^2}\left(\sum_{j=1}^{t}\frac{1}{\underline{p_j}}\right)$$

The $\underline{p_j}$ here is used to adapt our algorithm to the ϵ-greedy exploration strategy. If for each j, the arm a can be sampled with minimum probability ϵ_j, then we have $p_{a,j} = 1-(K-1)\epsilon_j$ or $p_{a,j} = \epsilon_j$ for arm a, implying that $\underline{p_j} = \epsilon_j$. The inverse probability weight $1/p_{a,j}$ we use in Algorithm 1 serves to correct the empirical covariance matrix and the gradients of each iteration by importance sampling(Chen et al. 2021), making the gradient estimation consistent.

For the hard-thresholding type method, the major challenge that occurs in the online algorithm design is the gradient information loss caused by truncation. Specifically, in the online update, the hard thresholding operator will zero out all the small signals, which contain valuable gradient information that can be exploited for the next update (Murata and Suzuki 2018, Zhou et al. 2018). Moreover, this kind of errors caused by gradient missing will accumulate during the online iteration, rendering it difficult to recover a sparse structure. To tackle this issue, we choose a slightly larger sparsity level that allows us to preserve more information about the gradient, and use the averaged gradient in each step to obtain a more accurate characterization of the stochastic gradient. We show that a larger sparsity level (which depends on the condition number κ of the problem) allows us to keep enough information so that the truncation effect is negligible. The notion of averaging is also used in the sparse estimation in, e.g., Han et al. (2023a) but is with different objectives. (Han et al. 2023a) does averaging on estimators, which is used for online inference, and the soft thresholding in Han et al. (2023a) can not guarantee exact s_0 sparse recovery.

The fundamental cause of the gradient averaging in Algorithm 1 is actually the poor smooth property of the projection onto ℓ_0-constraint space. Unlike the convex projection or higher-order low-rank projection, the projection onto the ℓ_0-constraint space exhibits an inflating smoothness behavior. To be specific, the projection onto the convex space shares the nice property $\|\mathcal{P}(\boldsymbol{x}+\Delta)-\boldsymbol{x}\|_2 \leq \|\Delta\|_2$, with no inflation on the error. The projection onto the higher-order low-rank space (e.g., SVD or HOSVD on low-rank matrix or tensor) also satisfies $\|\mathcal{P}(\boldsymbol{x}+\Delta)-\boldsymbol{x}\|_{\mathrm{F}} \leq \|\Delta\|_{\mathrm{F}} + C\|\Delta\|_{\mathrm{F}}^2$ if Δ is in the tangent space of the manifold (Kressner et al. 2014, Cai et al. 2022), which leads to tiny inflation for small perturbations in online tensor learning (Cai et al. 2023). However, the projection onto ℓ_0-constraint space can only ensure $\|\mathcal{P}(\boldsymbol{x}+\Delta)-\boldsymbol{x}\|_2 \leq (1+\delta)\|\Delta\|_2$, where δ is a non-zero parameter depending on the relative sparsity level and is unimprovable (Shen and Li 2017), which causes trouble for online sparse recovery. To mitigate the inevitable inflation, gradient averaging is employed to decrease the variance, thereby enabling us to achieve the optimal convergence rate.

4. Online Allocation: BwK Problem

In this section, we handle the BwK problem described in Section 2. Our algorithm adopts a primal-dual framework, where we introduce a dual variable to reflect the capacity consumption of each resource. Moreover, the dual variable can be interpreted as the Lagrangian dual variable for V^{UB}, with the dual function given in the following form

$$L(\boldsymbol{\eta}) = (\boldsymbol{C})^\top \boldsymbol{\eta} + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{x}_t \sim F}\left[\max_{\boldsymbol{y}_t(\boldsymbol{x}_t)\in\Delta^K}\left\{\sum_{a\in[K]} (\boldsymbol{\mu}_a^\star)^\top \boldsymbol{x}_t \cdot y_{a,t}(\boldsymbol{x}_t) - Z\cdot(\boldsymbol{W}_a^\star \boldsymbol{x}_t)^\top \boldsymbol{\eta}\cdot y_{a,t}(\boldsymbol{x}_t)\right\}\right]$$

where Δ^K denotes the unit simplex $\Delta^K = \{\boldsymbol{y}\in\mathbb{R}^K : y_a \geq 0, \forall a\in[K], \text{ and } \sum_{a\in[K]} y_a = 1\}$ and Z is a scaling parameter that we will specify later. Note that if the weight vector $\boldsymbol{\mu}_a^\star$ is given for each arm

$a \in [K]$ and the distribution $F(\cdot)$ is known, one can directly solve the dual problem $\min_{\boldsymbol{\eta}} L(\boldsymbol{\eta})$ to obtain the optimal dual variable $\boldsymbol{\eta}^*$ and then the primal variable $y_{a,t}(\boldsymbol{x}_t)$ can be decided by solving the inner maximization problem in the definition of the dual function $L(\boldsymbol{\eta})$. However, since the weight vector $\boldsymbol{\mu}_a^\star$ for each $a \in [K]$ and the distribution $F(\cdot)$ is unknown, one cannot directly solve the dual problem. Instead, we will employ an online learning algorithm and use the information we obtained at each period as the feedback to the online learning algorithm to update the dual variable $\boldsymbol{\eta}_t$. Then, we plug in the dual variable $\boldsymbol{\eta}_t$, as well as an estimate of $\boldsymbol{\mu}_a^\star$ for each $a \in [K]$, to solve the inner maximization problem in the definition of the dual function $L(\boldsymbol{\eta})$ to obtain the primal variable $y_{a,t}(\boldsymbol{x}_t)$. Note that this primal-dual framework has been developed previously in the literature (e.g. Badanidiyuru et al. (2013), Agrawal and Devanur (2016)) of bandits with knapsacks. The innovation of our algorithm is that we obtain a new estimate of $\boldsymbol{\mu}_a^\star$ via Algorithm 1 which enables us to exploit the sparsity structure of the weight vector $\boldsymbol{\mu}_a^\star$ to obtain improved regret bound. Our formal algorithm is presented in Algorithm 2 and in the next section, where we also conduct the regret analysis of our algorithm.

4.1. Regret analysis

In this section, we conduct regret analysis of Algorithm 2. We first show how regret depends on the choice of ϵ_t, for each $t \in [T]$, as well as the estimation error of our estimator of $\boldsymbol{\mu}_a^\star$, for each $a \in [K]$. We then specify the exact value of ϵ_t and utilize the estimation error characterized in Theorem 1 to derive our final regret bound.

THEOREM 2. *Denote by π the process of our Algorithm 2, and τ the stopping time of Algorithm 2. If Z satisfies $Z \geq \frac{V^{\mathrm{UB}}}{C_{\min}}$, then, under Assumption 1, the regret of the policy π can be upper bounded as follows*

$$\begin{aligned}
&\mathrm{Regret}(\pi) \leq V^{\mathrm{UB}} - \mathbb{E}_{I\sim F}[V^{\pi}(I)] \\
&\leq Z \cdot O\left(\sqrt{TD' \cdot \log m}\right) + \mathbb{E}\left[\sum_{t=1}^{\tau} \max_{a} \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,t-1}^{\mathsf{s}} \right\rangle\right|\right] + (4R_{\max} + 2D'Z) \cdot \sum_{t=1}^{T} K\epsilon_t. \qquad (7)
\end{aligned}$$

by setting $\delta = O\left(\sqrt{\frac{\log m}{TD'}}\right)$, where $R_{\max} = \sup_{\boldsymbol{x}, a\in[K]} |\langle \boldsymbol{x}, \boldsymbol{\mu}_a^\star \rangle|$ and D' denotes an upper bound of $b_i(y_t, \boldsymbol{x}_t)$ as specified in Assumption 1.

The three terms in Theorem 2 exhibit distinct components of Algorithm 2 that contribute to the final regret bound. The first term represents the effect of the dual update using the Hedge algorithm (Freund and Schapire 1997). While the last two terms arise from online sparse estimation and ϵ-greedy exploration, both of which can be categorized as consequences of the primal update. Given that the estimation error is constrained by Corollary 1, we can establish the following regret bound:

Algorithm 2 Primal-Dual High-dimensional BwK Algorithm

Input: a parameter Z and the ϵ-greedy probability ϵ_t for each t. δ for dual update.

In the first m rounds, pull each arm once and initialize $\boldsymbol{\alpha}_1 = \mathbf{1} \in [0,1]^m$ and $\boldsymbol{\eta}_1 = \frac{1}{m} \cdot \boldsymbol{\alpha}_1$.

for $t = m+1, ..., T$ **do**

Observe the feature $\boldsymbol{x}_t$.

Compute $\text{EstCost}(a) = \boldsymbol{b}(a, \boldsymbol{x}_t)^\top \boldsymbol{\eta}_t$ for each arm $a \in [K]$.

Sample a random variable $\nu_t \sim \text{Ber}(K\epsilon_t)$, and pull the arm y_t defined as follows:

$$y_t = \begin{cases} \text{argmax}_{a\in[K]}\{(\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1})^\top \boldsymbol{x}_t - Z \cdot \text{EstCost}(a)\}, & \text{if } \nu_t = 0 \\ a, \quad \text{w.p. } 1/K \text{ for each arm } a \in [K] & \text{if } \nu_t = 1. \end{cases}$$

If $\text{argmax}_{a\in[K]}\{(\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1})^\top \boldsymbol{x}_t - Z \cdot \text{EstCost}(a)\}$ contains multiple arms, then break ties uniformly.

If one of the constraints is violated, then EXIT.

Update for each resource $i \in [m]$,

$$\alpha_{t+1}(i) = \alpha_t(i) \cdot (1+\delta)^{(b_i(y_t, \boldsymbol{x}_t) - \frac{C_i}{T}) \cdot (1-\xi_t)}$$

and we project $\boldsymbol{\alpha}_{t+1}$ into the unit simplex $\{\boldsymbol{\eta} : \|\boldsymbol{\eta}\|_1 \le 1, \boldsymbol{\eta} \ge 0\}$ to obtain $\boldsymbol{\eta}_{t+1}$ as follows.

$$\eta_{t+1}(i) = \frac{\alpha_{t+1}(i)}{\sum_{i' \in [m]} \alpha_{t+1}(i')}, \ \forall i \in [m].$$

For each arm $a \in [K]$, obtain the estimate $\boldsymbol{\mu}^{\mathsf{s}}_{a,t}$ from Algorithm 1.

end for

THEOREM 3. *Under Assumption 1, if Z satisfies $\frac{V^{\text{UB}}}{C_{\min}} \le Z \le O\left(\frac{V^{\text{UB}}}{C_{\min}} + 1\right)$, then the regret of Algorithm 2 can be upper bounded by*

$$\begin{aligned} \text{Regret}(\pi) \le & O\left(\frac{V^{\text{UB}}}{C_{\min}} + 1\right) \cdot \sqrt{D'T \cdot \log m} \\ & + O\left(\phi_{\min}^{-\frac{2}{3}}(s) \cdot \left(R_{\max} + D' \frac{V^{\text{UB}}}{C_{\min}} + 1\right)^{\frac{1}{3}} K^{\frac{1}{3}} \sigma^{\frac{2}{3}} D^{\frac{4}{3}} s_0^{\frac{2}{3}} T^{\frac{2}{3}} (\log(dK))^{\frac{1}{3}}\right) \end{aligned} \tag{8}$$

by setting $\delta = O\left(\sqrt{\frac{\log m}{TD'}}\right)$, and $\epsilon_t = O\left(\sigma^{\frac{2}{3}} D^{\frac{4}{3}} s_0^{\frac{2}{3}} (\log(dK))^{\frac{1}{3}} t^{-\frac{1}{3}} / \left((R_{\max} + D'Z)^{\frac{2}{3}} K^{\frac{2}{3}}\right) \wedge 1/K\right)$.

The result generally shows a two-phase regret of high-dimensional BwK problem, i.e., $\text{Regret}(\pi) = \widetilde{O}\left(\frac{V^{\text{UB}}}{C_{\min}}\sqrt{T} + \left(\frac{V^{\text{UB}}}{C_{\min}}\right)^{\frac{1}{3}} T^{\frac{2}{3}}\right)$, which reveals the leading effects of primal or dual updates on the regret under different situations. That is, if $\frac{V^{\text{UB}}}{C_{\min}} = O(T^{\frac{1}{4}})$, then our constraints are sufficient enough for decision-making such that learning the primal information will be the barrier of the problem, which leads to $\text{Regret}(\pi) = \widetilde{O}\left(\left(\frac{V^{\text{UB}}}{C_{\min}}\right)^{\frac{1}{3}} T^{\frac{2}{3}}\right)$; however when $\frac{V^{\text{UB}}}{C_{\min}} \ge \omega(T^{\frac{1}{4}})$, our constraints are considered scarce, positioning the dual information as the bottleneck of the problem, and thus

$\text{Regret}(\pi) = \widetilde{O}\left(\frac{V^{\text{UB}}}{C_{\min}}\sqrt{T}\right)$. Most notably, our regret only shows logarithmic dependence on the dimension d, which improves the polynomial dependency on d in previous results (Agrawal and Devanur 2016) and makes the algorithm more feasible for high-dimensional problems.

4.2. Estimating reward-constraint ratio

We now show the procedure for computing the parameter Z to serve as an input to Algorithm 2. The procedure is similar to that in Agrawal and Devanur (2016), however, we will use the estimator obtained in Algorithm 1. To be specific, we specify a parameter T_0 and we use the first T_0 periods to obtain an estimate of V^{UB}. Then, the estimate can be obtained by solving the following linear programming.

$$\hat{V} = \max \ \frac{T}{T_0} \cdot \sum_{t=1}^{T_0} \sum_{a\in[K]} (\boldsymbol{\mu}^{\mathsf{s}}_{a,T_0})^\top \boldsymbol{x}_t \cdot y_{a,t} \tag{9a}$$

$$\text{s.t.} \ \frac{T}{T_0} \cdot \sum_{t=1}^{T_0} \sum_{a\in[K]} \boldsymbol{W}^*_a \boldsymbol{x}_t \cdot y_{a,t} \le \boldsymbol{C} \tag{9b}$$

$$\sum_{a\in[K]} y_{a,t} = 1, \forall t \in [T_0] \tag{9c}$$

$$y_{a,t} \in [0,1], \forall a \in [K], \forall t \in [T_0]. \tag{9d}$$

We have the following bound regarding the gap between the value of V^{UB} and its estimate $\hat{V}$.

LEMMA 2. *With probability at least $1-\beta$, it holds that*

$$|V^{\text{UB}} - \hat{V}| \le T \cdot (R_{\max} + \frac{V^{\text{UB}}}{C_{\min}} \cdot D') \cdot \sqrt{\frac{1}{2T_0} \cdot \log\frac{4}{\beta}} + \frac{V^{\text{UB}}}{C^2_{\min}} \cdot D' \cdot \frac{T}{2T_0} \cdot \log\frac{4}{\beta} + T \cdot D \cdot \max_{a\in[K]} \|\boldsymbol{\mu}^\star_a - \boldsymbol{\mu}^{\mathsf{s}}_{a,T_0}\|_1.$$

Therefore, by uniform sampling from time 1 to T_0, we can simply set $Z = O\left(\frac{\hat{V}}{C_{\min}}\right)$, and as long as $T_0 = O\left(s_0^2\sigma^2 D^4 K \cdot \frac{T^2}{C^2_{\min}} \cdot \log\frac{1}{\beta}\right)$, we get that $Z = O(\frac{V^{\text{UB}}}{C_{\min}} + 1)$ with probability at least $1-\beta$ from the high probability bound of our sparse estimator in Theorem 1. If further the constraints grow linearly, i.e., $C_{\min} = \Omega(T)$, we only require $T_0 = O\left(\log\frac{1}{\beta}\right)$ in general.

4.3. Improved regret with diverse covariate

In Theorem 3, it is shown that the primal update may become the bottleneck of the regret. This happens because we have to compromise between exploration and exploitation. However, in some cases, when the covariates are diverse enough, our dual allocation algorithm will naturally explore sufficient arms, leading to significant improvement in the exploitation of primal updates. We now describe such a case with the notion of diverse covariate condition (Ren and Zhou 2023).

ASSUMPTION 2 **(Diverse covariate)**. *There are (possibly K-dependent) positive constants $\gamma(K)$ and $\zeta(K)$, such that for any unit vector $\boldsymbol{v} \in \mathbb{R}^d$, $\|\boldsymbol{v}\|_2 = 1$ and any $a \in [K]$, conditional on the history $\mathcal{H}_{t-1}$, there is*

$$\mathbb{P}\left(\boldsymbol{v}^\top x_t x_t^\top \boldsymbol{v} \cdot \mathbb{1}\{y_t = a\} \ge \gamma(K) \middle| \mathcal{H}_{t-1}\right) \ge \zeta(K),$$

where $y_t = \text{argmax}_{a\in[K]}\{(\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1})^\top \boldsymbol{x}_t - Z\cdot \boldsymbol{b}(a,\boldsymbol{x}_t)^\top \boldsymbol{\eta}_t\}$

Such a diverse covariate condition states that when we perform the online allocation task, our dual-based algorithm can ensure sufficient exploration. This can be viewed as a primal-dual version of the diverse covariate condition for greedy algorithms (Han et al. 2020, Ren and Zhou 2023). If our covariate is diverse enough, we can just set $\epsilon_t = 0$ in Algorithm 2 to obtain a good performance of primal exploration. We present the primal behavior of our algorithms in the following Theorem 4.

THEOREM 4. *Denote* $\kappa_1 = \frac{\phi_{\max}(s)}{\gamma(K)\zeta(K)}$ *If we take* $\rho = \frac{1}{9\kappa_1^4}$, *and* $\eta_t = \frac{1}{4\kappa_1\phi_{\max}(s)}$, *then under Assumption 1 - 2, setting* $\epsilon_t = 0$, *the output of Algorithm 1 satisfies*

$$\mathbb{E}\|\boldsymbol{\mu}_{a,t} - \boldsymbol{\mu}^\star_a\|_2^2 \lesssim \frac{\sigma^2 D^2 s_0}{\gamma^2(K)\zeta^2(K)}\cdot\frac{\log d}{t},$$

and the high-probability bound

$$\|\boldsymbol{\mu}_{a,t} - \boldsymbol{\mu}^\star_a\|_2^2 \lesssim \frac{\sigma^2 D^2 s_0}{\gamma^2(K)\zeta^2(K)}\cdot\frac{\log(dTK/\varepsilon)}{t},$$

which holds for all t *and arm* a *with probability at least* $1-\varepsilon$.

Theorem 4 suggests that under the diverse covariate condition, our algorithm can recover the sparse arms with a statistical error rate that is optimal for t. This greatly improves the primal performance of our algorithm and thus leads to a sharper regret bound for BwK problem. We describe this improved regret in Theorem 5.

THEOREM 5. *If* Z *satisfies* $\frac{V^{\text{UB}}}{C_{\min}} \le Z \le c\frac{V^{\text{UB}}}{C_{\min}} + c'$, *then the regret of the Algorithm 2 can be upper bounded by*

$$\text{Regret}(\pi) \le O\left(\left(\frac{V^{\text{UB}}}{C_{\min}}+1\right)\cdot\sqrt{TD'\cdot\log m}\right) + \frac{\sigma D^2 s_0\sqrt{T\log K\log(dK)}}{\gamma(K)\zeta(K)} \tag{10}$$

by setting $\delta = O\left(\sqrt{\frac{\log m}{T\cdot D'}}\right)$, *and* $\epsilon_t = 0$, *for each* $t\in[T]$.

The rationale behind setting $\epsilon_t = 0$ in Algorithm 2 is that, when our covariate vectors exhibit sufficient diversity, our strategy will automatically explore enough arms while simultaneously optimizing regret. This condition is typically met in the online allocation problem where the optimal strategy is often a distribution within arms, rather than a single arm (Badanidiyuru et al. 2018). This starkly contrasts with the classical multi-armed bandit problem, where the optimal solution is typically confined to a single arm. Theorem 5 significantly reduces the impact of primal update on the regret from $\widetilde{O}\left(\left(\frac{V^{\text{UB}}}{C_{\min}}\right)^{\frac{1}{3}} T^{\frac{2}{3}}\right)$ to a sharper $\widetilde{O}\left(s_0\sqrt{T}\right)$, which makes the impact of the dual update the dominating factor of regret, giving the bound $\text{Regret}(\pi) = \widetilde{O}\left(\frac{V^{\text{UB}}}{C_{\min}}\sqrt{T}\right)$.

5. Optimal High-dimensional Bandit Algorithm

An important application of our Algorithm 1 is the high-dimensional bandit problem (Carpentier and Munos 2012, Hao et al. 2020), where we do not consider the knapsacks but only focus on reward maximization (or, we can treat the bandit problem as a special BwK problem where the constraints are always met). Here we associate our algorithm with ϵ-greedy strategy and show that our high-dimensional bandit algorithm by Online HT can achieve both the $\widetilde{O}(s_0^{\frac{2}{3}}T^{\frac{2}{3}})$ optimal regret in the data-poor regime, and the $\widetilde{O}(\sqrt{s_0 T})$ optimal regret in the data-rich regime, which enjoys the so-called "the best of two worlds".

Algorithm 3 High Dimensional Bandit by Online HT

1: ϵ-greedy sampling probability ϵ_t for each t. Initialization $\boldsymbol{\mu}^{\mathsf{s}}_{a,0}$, step size η.

2: **for** $t = 1, ..., T$ **do**

3: Observe the feature $\boldsymbol{x}_t$.

4: Sample a random variable $\nu_t \sim \mathrm{Ber}(K\epsilon_t)$.

5: Pull the arm y_t with ϵ_t-greedy strategy defined as follows:

$$y_t = \begin{cases} \arg\max_{a\in[K]} \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} \right\rangle, & \text{if } \nu_t = 0 \\ a, \quad \text{w.p. } 1/K \text{ for each arm } a \in [K] & \text{if } \nu_t = 1, \end{cases}$$

and receive a reward r_t.

6: For each arm $a \in [K]$, update the sparse estimate $\boldsymbol{\mu}^{\mathsf{s}}_{a,t}$ by Algorithm 1 with each $p_{a,t} = (1 - K\epsilon_t)y_{a,t} + \epsilon_t$

7: **end for**

THEOREM 6. *Let* $R_{\max} = \sup |\langle \boldsymbol{x}_t, \boldsymbol{\mu}^a_\star \rangle|$. *Choosing* $\epsilon_t = \sigma^{\frac{2}{3}} D^{\frac{4}{3}} s_0^{\frac{2}{3}} (\log(dK))^{\frac{1}{3}} t^{-\frac{1}{3}} / \left(R_{\max} K\right)^{\frac{2}{3}} \wedge 1/K$, *our Algorithm 3 incurs the regret*

$$\mathrm{Regret}_{\mathsf{bandit}}(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{x}_t, \boldsymbol{\mu}_{\mathsf{opt}}(\boldsymbol{x}_t)\rangle - \sum_{t=1}^{T} \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^\star_{y_t} \right\rangle\right] \lesssim \frac{R_{\max}^{\frac{1}{3}} K^{\frac{1}{3}} \sigma^{\frac{2}{3}} D^{\frac{4}{3}} s_0^{\frac{2}{3}} T^{\frac{2}{3}} (\log(dK))^{\frac{1}{3}}}{\phi_{\min}(s)^{\frac{2}{3}}}$$

Theorem 6 states the optimality of our high-dimensional bandit algorithm under minimal assumptions, which matches the $\Omega\left(\phi_{\min}^{-2/3} s_0^{2/3} T^{2/3}\right)$ lower bound provided by Jang et al. (2022) in the data-poor regime $d \geq T^{\frac{1}{3}} s_0^{\frac{4}{3}}$. We further show that, we can use the same algorithm framework to achieve better regret given the diverse covariate condition, which also matches the general regret lower bound of high-dimensional bandit problems. We present our result in Theorem 7.

THEOREM 7. *Suppose* $\boldsymbol{x}_t$ *is further sparse marginal sub-Gaussian:*

$$\mathbb{E}\exp\left(\boldsymbol{u}^\top \boldsymbol{x}_t\right) \leq \exp\left(c\phi_{\max}(s_0)\|\boldsymbol{u}\|_2^2/2\right),$$

for any $2s_0$-sparse vector $\boldsymbol{u}$. Assume the following diverse covariate condition (Ren and Zhou 2023) holds: There are (possibly K-dependent) positive constants $\gamma(K)$ and $\zeta(K)$, such that for any unit vector $\boldsymbol{v} \in \mathbb{R}^d$, and any $a \in [K]$, conditional on the history $\mathcal{H}_{t-1}$, there is

$$\mathbb{P}\left(\boldsymbol{v}^\top x_t x_t^\top \boldsymbol{v} \cdot \mathbb{1}\left\{a_t^\star = a\right\} \geq \gamma(K) \middle| \mathcal{H}_{t-1}\right) \geq \zeta(K),$$

where $a_t^\star = \max_{a \in [K]} \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_{a,t-1}^{\mathsf{s}} \right\rangle$ is selected greedily. Denote $\kappa_1 = \frac{\phi_{\max}(s)}{\gamma(K)\zeta(K)}$. Setting $\epsilon_t = 0$, we have the following regret bound for Algorithm 3:

$$\mathrm{Regret}_{\mathsf{bandit}}(\pi) \lesssim \frac{\left(\kappa_1 \wedge \frac{s_0 D^2}{\gamma(K)\zeta(K)}\right)^{\frac{1}{2}} \sigma D \sqrt{s_0 T} (\log K \log(dK))^{\frac{1}{2}}}{\sqrt{\gamma(K)\zeta(K)}}.$$

The regret of our bandit algorithm indeed matches the known low bound of general high-dimensional bandit problems which is of order $\Omega(\sqrt{s_0 T})$ (Chu et al. 2011, Ren and Zhou 2023). Compared with previous LASSO-based frameworks, no additional assumption on the range of arms (e.g., ℓ_2-norm bound of $\boldsymbol{\mu}_a^\star$ (Ren and Zhou 2023)) or the minimum signal strength (Hao et al. 2020, Jang et al. 2022) is needed for our algorithm to achieve the optimal regret in the data-rich regime, as long as the diverse covariate condition holds. The sparse marginal sub-Gaussian assumption here is used to yield a more precise characterization of estimation errors associated with $\boldsymbol{x}_t$. If without sparse marginal sub-Gaussian assumption, there will be no κ_1 term in the regret bound of Theorem 7.

6. Numerical Results

6.1. Sparse recovery

We first examine the feasibility of our primal algorithm in the sparse recovery problem. To check the performance of Algorithm 1, suppose now we only consider one arm $\boldsymbol{\mu}_\star$, and we want to estimate it in an online process. To this end, we always choose $y_t = 1$ and thus $p_t = 1$. At each t, we measure the sparse estimation error $\|\boldsymbol{\mu}_t^{\mathsf{s}} - \boldsymbol{\mu}_\star\|_2^2$, and the support recovery rate $|\mathrm{supp}(\boldsymbol{\mu}_t^{\mathsf{s}}) \cap \Omega_\star|/s_0$, which indicates the ratio of the support set we have detected. The result is presented in Figure 1. Here we set $d = 1000$, $s_0 = 10$, $\sigma = 0.5$, and $\boldsymbol{\Sigma}$ to be the power decaying covariance matrix: $\boldsymbol{\Sigma}_{ij} = \alpha^{|i-j|}$, where $\alpha = 0.5$. Compared with the prevalent LASSO method used in online high dimensional bandit problem (Kim and Paik 2019, Hao et al. 2020, Ren and Zhou 2023), our method shares efficient computational cost while achieving better estimation error. See Figure 1 for the arm estimation and support set recovery of our method. To be specific, the computational cost of Online HT is $O(d^2)$ per iteration and $O(d^2 T)$ in total, while the computational cost of classical LASSO solution is $O(d^3 + d^2 t)$ per iteration (Efron et al. 2004), and $O(d^3 T + d^2 T^2)$ in total if we require constant updates of the estimation, e.g., Kim and Paik (2019), Ren and Zhou (2023). Here in the LASSO,

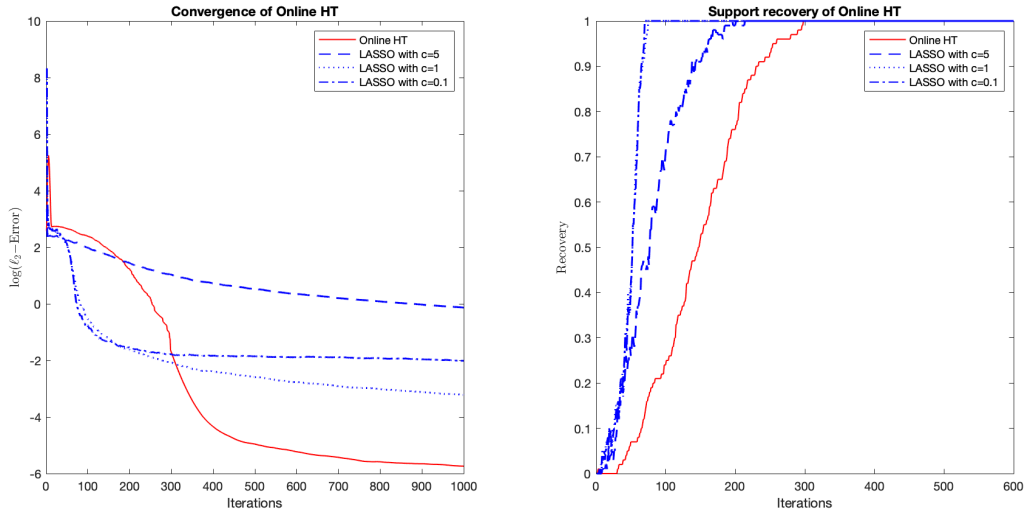

Figure 1 **Primal performance of Online HT vs LASSO.**

we select the regularization level $\lambda = c \cdot \sqrt{\frac{\log(dt)}{t}}$, where c is selected to be $\{5, 1, 0.1\}$ respectively. One huge advantage that distinguishes our method from LASSO or soft thresholding method (Han et al. 2023a) is that we can achieve a guaranteed exact s_0-sparse estimation without parameter tuning.

6.2. Online bandit problem

We then apply our Algorithm 3 to the high-dimensional linear bandit problem, and Primal-dual based Algorithm 2 to the linear BwK problem to corroborate our study on the regret.

For the bandit problem, we choose $d = 100$, $s_0 = 10$, $K = 5$. The covariates are still generated following Section 6.1. We study the regret accumulation for a fixed T and regret growth with respect to different Ts, respectively. The result is presented in Figure 2. Here, we mainly compare our ϵ-greedy Online HT method with LASSO bandit algorithm (Explore-Then-Commit method) in, e.g., Hao et al. (2020), Li et al. (2022), Jang et al. (2022). In our simulation, we try different lengths of exploration phases t_1 as $t_1 = 0.3T^{\frac{2}{3}}$ and $t_1 = 0.5T^{\frac{2}{3}}$ for LASSO bandit algorithm. The greedy Online HT means we simply treat each $\epsilon_t = 0$. It can be observed that our method outperforms the LASSO bandit algorithm in the regret growth, and the greedy Online HT shows far slower regret growth than other algorithms.

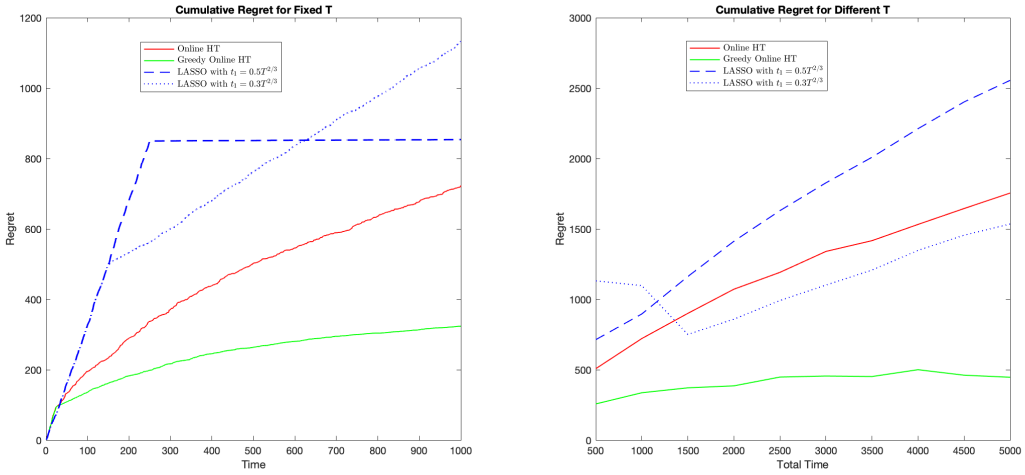

Figure 2 **Regret of Online HT vs LASSO Bandit.**

6.3. High-dimensional BwK

We now focus on the linear BwK problem with high-dimensional sparse arms. We show the performance of our algorithm, together with the classic UCB-based linear BwK algorithm, i.e., the linCBwK (Agrawal and Devanur 2016), to demonstrate the feasibility of the Online HT method. Notice that, in the original paper of Agrawal and Devanur (2016), the linCBwK algorithm is designed for Model-C bandit problem, but it can be easily generalized to our Model-P setting by computing the UCB of multiple arms at the same time. We set $d = 200$, $s_0 = 10$, $K = 5$, with generated following Section 6.1. The constraints are randomly generated following uniform distribution with $m = 5$, and each row of $W_a^\star$ is also sparse with the support set same as $\boldsymbol{\mu}_a^\star$. We present our methods' regret and relative regret control in Figure 3. The relative regret is defined by $\frac{\mathsf{Regret}}{\mathsf{OPT}}$. It can be observed that when T is small, linCBwK fails to control the cumulative regret due to the high dimensionality of the problem. As T grows larger, the impact of high dimensionality is decreased and thus two methods behave comparably. The relative regret curves also show this phenomenon. Our Online HT methods share faster convergence rates for the relative regret in the data-poor regime.

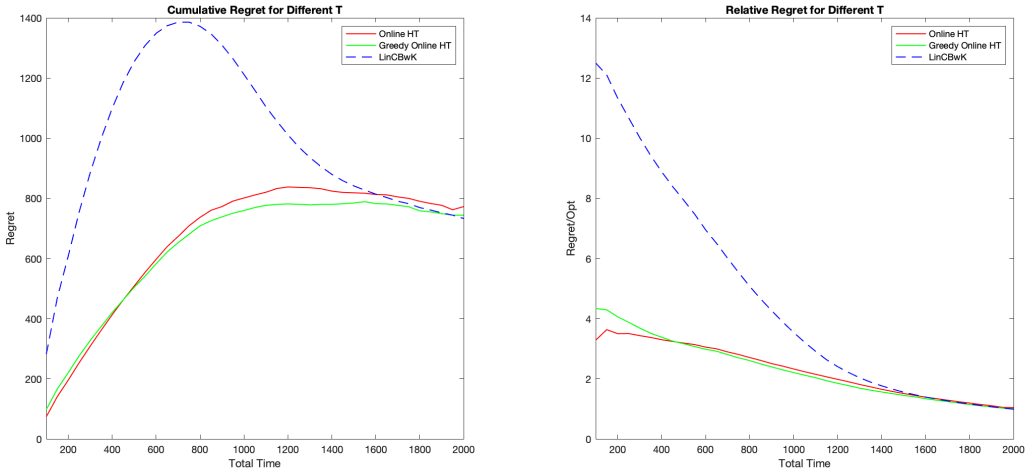

Figure 3 **Regret of Online HT vs linCBwK for CBwK problem.**

7. Discussions

Although in this paper we mainly focus on the case when the consumption $\boldsymbol{W}_a^\star$ for each arm is known, it is direct to generalize our results to the unknown $\boldsymbol{W}_a^\star$ by estimating them with Algorithm 1. Substituting $\boldsymbol{W}_a^\star$ with an estimated version $\widehat{\boldsymbol{W}}_a$ may incur additional estimation error, but this error can be generally controlled in a similar fashion to Theorem 1. As we proceed to discuss the consumption-agnostic instance, we will also posit that $\boldsymbol{W}_a^\star$ is row-wise sparse, a necessary assumption to render the problem tractable. The exploration of this particular aspect is earmarked for future work.

References

Agrawal, S. and Devanur, N. (2016). Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29.

Agrawal, S. and Devanur, N. R. (2014). Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006.

Ariu, K., Abe, K., and Proutière, A. (2022). Thresholded lasso bandit. In *International Conference on Machine Learning*, pages 878–928. PMLR.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2013). Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2018). Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55.

Badanidiyuru, A., Langford, J., and Slivkins, A. (2014). Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR.

Balseiro, S. R., Lu, H., and Mirrokni, V. (2023). The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 71(1):101–119.

Bastani, H. and Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294.

Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274.

Cai, J.-F., Li, J., and Xia, D. (2022). Generalized low-rank plus sparse tensor estimation by fast riemannian optimization. *Journal of the American Statistical Association*, pages 1–17.

Cai, J.-F., Li, J., and Xia, D. (2023). Online tensor learning: Computational and statistical trade-offs, adaptivity and optimal regret. *arXiv preprint arXiv:2306.03372*.

Carpentier, A. and Munos, R. (2012). Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pages 190–198. PMLR.

Chen, H., Lu, W., and Song, R. (2021). Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Han, R., Luo, L., Lin, Y., and Huang, J. (2023a). Online inference with debiased stochastic gradient descent. *Biometrika*, page asad046.

Han, Y., Zeng, J., Wang, Y., Xiang, Y., and Zhang, J. (2023b). Optimal contextual bandits with knapsacks under realizability via regression oracles. In *International Conference on Artificial Intelligence and Statistics*, pages 5011–5035. PMLR.

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J., Glynn, P. W., and Ye, Y. (2020). Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*.

Hao, B., Lattimore, T., and Wang, M. (2020). High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763.

Immorlica, N., Sankararaman, K., Schapire, R., and Slivkins, A. (2022). Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47.

Jang, K., Zhang, C., and Jun, K.-S. (2022). Popart: Efficient sparse regression and experimental design for optimal sparse linear bandits. *Advances in Neural Information Processing Systems*, 35:2102–2114.

Jiang, J., Li, X., and Zhang, J. (2020). Online stochastic optimization with wasserstein based non-stationarity. *arXiv preprint arXiv:2012.06961*.

Kim, G.-S. and Paik, M. C. (2019). Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32.

Kressner, D., Steinlechner, M., and Vandereycken, B. (2014). Low-rank tensor completion by riemannian optimization. *BIT Numerical Mathematics*, 54:447–468.

Li, W., Barik, A., and Honorio, J. (2022). A simple unified framework for high dimensional bandit problems. In *International Conference on Machine Learning*, pages 12619–12655. PMLR.

Li, X., Sun, C., and Ye, Y. (2021). The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *International Conference on Machine Learning*, pages 6483–6492. PMLR.

Liu, S., Jiang, J., and Li, X. (2022). Non-stationary bandits with knapsacks. *Advances in Neural Information Processing Systems*, 35:16522–16532.

Ma, W., Cao, Y., Tsang, D. H., and Xia, D. (2022). Optimal regularized online convex allocation by adaptive re-solving. *arXiv preprint arXiv:2209.00399*.

Meinshausen, N. and Yu, B. (2008). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1).

Murata, T. and Suzuki, T. (2018). Sample efficient stochastic gradient iterative hard thresholding method for stochastic sparse linear regression with limited attribute observation. *Advances in Neural Information Processing Systems*, 31.

Nguyen, N., Needell, D., and Woolf, T. (2017). Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63(11):6869–6895.

Oh, M.-h., Iyengar, G., and Zeevi, A. (2021). Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR.

Ren, Z. and Zhou, Z. (2023). Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science*.

Shen, J. and Li, P. (2017). A tight bound of hard thresholding. *The Journal of Machine Learning Research*, 18(1):7650–7691.

Tsybakov, A. and Rigollet, P. (2011). Exponential screening and optimal rates of sparse estimation. *Annals of Statistics*, 39(2):731–771.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, X., Wei, M., and Yao, T. (2018). Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pages 5200–5208. PMLR.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 11:3519–3540.

Yuan, X. and Li, P. (2021). Stability and risk bounds of iterative hard thresholding. In *International Conference on Artificial Intelligence and Statistics*, pages 1702–1710. PMLR.

Zhou, P., Yuan, X., and Feng, J. (2018). Efficient stochastic gradient hard thresholding. *Advances in Neural Information Processing Systems*, 31.

Supplement to "High-dimensional Linear Bandits with Knapsacks"

8. Proofs of Main Results

8.1. Proof of Theorem 1

Proof. We first denote $\widetilde{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_{t-1} - \eta_t \boldsymbol{g}_t$, and the support $\Omega = \Omega_{t+1} \cup \Omega_t \cup \Omega_\star$ as the union of the support set of $\boldsymbol{\mu}_{t+1}$, $\boldsymbol{\mu}_t$, and $\boldsymbol{\mu}_\star$. For the iterative method, we have

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_\star\|_2^2 = \|\mathcal{H}_s(\Omega(\widetilde{\boldsymbol{\mu}}_t)) - \boldsymbol{\mu}_\star\|_2^2 \leq \left(1 + \frac{\rho + \sqrt{\rho(4+\rho)}}{2}\right) \|\Omega(\widetilde{\boldsymbol{\mu}}_t) - \boldsymbol{\mu}_\star\|_2^2,$$

by the tight bound of hard thresholding operator (Shen and Li 2017). Here $\rho = s_0/s$ is the relative sparsity level. By selecting a small enough ρ (e.g., $\rho \leq \frac{1}{4}$), it is clear that

$$\begin{aligned}
\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_\star\|_2^2 &\leq \left(1 + \frac{3}{2}\sqrt{\rho}\right) \|\Omega(\widetilde{\boldsymbol{\mu}}_t) - \boldsymbol{\mu}_\star\|_2^2 \\
&= \left(1 + \frac{3}{2}\sqrt{\rho}\right) \left(\|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 - 2\eta_t \langle \Omega(\boldsymbol{g}_t), \boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\rangle + \eta_t^2 \|\Omega(\boldsymbol{g}_t)\|_2^2\right) \\
&\leq \left(1 + \frac{3}{2}\sqrt{\rho}\right) \Big(\|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 - 2\eta_t \langle \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\rangle + 2\eta_t^2 \|\Omega(\boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}))\|_2^2 \\
&\quad + 2\eta_t^2 \|\Omega(\nabla f(\boldsymbol{\mu}_{t-1}))\|_2^2 + 2\eta_t \|\Omega(\boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}))\|_2 \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2\Big),
\end{aligned}$$

where we use the fact that $\langle \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\rangle = \langle \Omega(\nabla f(\boldsymbol{\mu}_{t-1})), \boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\rangle$ by the definition of $\Omega(\cdot)$. Now, applying the restricted strong convexity and smoothness condition from Assumption 1:

$$\begin{aligned}
\langle \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\rangle &\geq 2\phi_{\min}(s) \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 \\
\|\Omega(\nabla f(\boldsymbol{\mu}_{t-1}))\| &\leq 2\phi_{\max}(s) \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2,
\end{aligned}$$

We can show that

$$\begin{aligned}
\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_\star\|_2^2 &\leq \left(1 + \frac{3}{2}\sqrt{\rho}\right) \left(1 - 4\phi_{\min}(s)\eta_t + 8\eta_t^2 \phi_{\max}^2(s)\right) \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 \\
&\quad + 6\eta_t^2 \|\Omega(\boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}))\|_2^2 + 6\eta_t \|\Omega(\boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}))\|_2 \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2 \\
&\leq \left(1 + \frac{3}{2}\sqrt{\rho}\right) \left(1 - 4\phi_{\min}(s)\eta_t + 8\eta_t^2 \phi_{\max}^2(s)\right) \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 \\
&\quad + 18 s \eta_t^2 \max_{i \in [d]} |\langle \boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{e}_i\rangle|^2 + 18\eta_t \sqrt{s} \max_{i \in [d]} |\langle \boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{e}_i\rangle| \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2
\end{aligned} \tag{11}$$

The following lemma quantifies the variation of the stochastic gradient:

LEMMA 3. *Define $\{\boldsymbol{e}_i\}_1^d$ as the canonical basis of $\mathbb{R}^d$. The variance of stochastic gradient $\boldsymbol{g}_t$ at the point $\boldsymbol{\mu}_{t-1}$ can be bounded by the following inequality:*

$$\mathbb{E} \max_{i \in [d]} |\langle \boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{e}_i\rangle|^2 \leq C \frac{s D^2 \log(dt)}{t^2} \left(\sum_{j=1}^t 1/\underline{p_j}\right) \mathbb{E}\|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 + C \frac{\sigma^2 D^2 (\sum_{j=1}^t 1/\underline{p_j}) \log d}{t^2}. \tag{12}$$

Moreover, the following inequality also holds with probability at least $1-\epsilon$

$$\max_{i\in[d]}|\langle \boldsymbol{g}_t-\nabla f(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle|^2\leq CsD^2\frac{\log(d/\epsilon)}{t^2}\left(\sum_{j=1}^t\frac{1}{\underline{p_j}}\right)\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2+C\frac{\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})\log(d/\epsilon)}{t^2}$$

With Lemma 3, we are able to derive the expectation bound and probability bound respectively. For the expectation bound, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2\leq&\left(1+\frac{3}{2}\sqrt{\rho}\right)\left(1-4\phi_{\min}(s)\eta_t+8\eta_t^2\phi_{\max}^2(s)\right)\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\&+18s\eta_t^2\mathbb{E}\max_{i\in[d]}|\langle \boldsymbol{g}_t-\nabla f(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle|^2\\&+18\eta_t\sqrt{s}\sqrt{\mathbb{E}\max_{i\in[d]}|\langle \boldsymbol{g}_t-\nabla f(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle|^2}\sqrt{\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2}\end{aligned}$$

We set $\rho=\frac{1}{9\kappa^4}$, and $\eta_t=\frac{1}{4\kappa\phi_{\max}(s)}$. Plugging in the expectation bound in Lemma 3, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2\leq&\left(1-\frac{1}{4\kappa^4}+C\frac{s_0D\sqrt{\log(dt)}}{\phi_{\min}(s)t}\sqrt{\sum_{j=1}^t 1/\underline{p_j}}\right)\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\&+C\frac{s_0\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})\log d}{\phi_{\min}^2(s)t^2}+C\sqrt{\frac{s_0\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})\log d}{\phi_{\min}^2(s)t^2}\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2}.\end{aligned}$$

When t is sufficiently large, essentially we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2\leq&\left(1-\frac{1}{5\kappa^4}\right)\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\&+C\frac{s_0\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})\log d}{\phi_{\min}^2(s)t^2}+C\sqrt{\frac{s_0\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})\log d}{\phi_{\min}^2(s)t^2}\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2}.\end{aligned}$$

This instantly gives us the expectation bound

$$\mathbb{E}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2\lesssim\frac{\sigma^2D^2s_0}{\phi_{\min}^2(s)}\frac{\log d}{t^2}\left(\sum_{j=1}^t\frac{1}{\underline{p_j}}\right),$$

which proves the first claim. Following a similar fashion, we can also prove the high-probability bound: with probability at least $1-\epsilon$, we have

$$\begin{aligned}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2\leq&\left(1-\frac{1}{4\kappa^4}+C\frac{s_0D\sqrt{\log(dT/\epsilon)}}{\phi_{\min}(s)t}\sqrt{\sum_{j=1}^t 1/\underline{p_j}}\right)\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\&+C\frac{s_0\sigma^2(\sum_{j=1}^t 1/\underline{p_j})\log(dT/\epsilon)}{\phi_{\min}^2(s)t^2}+C\sqrt{\frac{s_0\sigma^2(\sum_{j=1}^t 1/\underline{p_j})\log(dT/\epsilon)}{\phi_{\min}^2(s)t^2}}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2,\end{aligned}$$

for all the $t\in[T]$. When t is sufficiently large, essentially we have

$$\begin{aligned}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2\leq&\left(1-\frac{1}{5\kappa^4}\right)\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\&+C\frac{s_0\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})\log(dT/\epsilon)}{\phi_{\min}^2(s)t^2}+C\sqrt{\frac{s_0\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})\log(dT/\epsilon)}{\phi_{\min}^2(s)t^2}}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2.\end{aligned}$$

It is therefore clear that

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_\star\|_2^2 \lesssim \frac{\sigma^2 D^2 s_0}{\phi_{\min}^2(s)} \frac{\log(dT/\varepsilon)}{t^2} \left(\sum_{j=1}^t \frac{1}{\underline{p_j}} \right)$$

holds for all $t \in [T]$ with probability at least $1-\epsilon$. Thus, we finish the proof.

□

8.2. Proof of Lemma 3

Proof. Define $\{\boldsymbol{e}_i\}_1^d$ as the canonical basis of $\mathbb{R}^d$. Since

$$\begin{aligned}
\boldsymbol{g}_t &= 2\widehat{\boldsymbol{\Sigma}}_t \boldsymbol{\mu}_{t-1} - \frac{2}{t}\sum_{j=1}^t y_j \boldsymbol{x}_j r_j / p_t = \frac{2}{t}\sum_{j=1}^t \left(\frac{y_j \boldsymbol{x}_j \boldsymbol{x}_j^\top}{p_j} \right) (\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star) - \frac{2}{t}\sum_{j=1}^t y_j \boldsymbol{x}_j \xi_j / p_t, \\
&= 2\widehat{\boldsymbol{\Sigma}}_t (\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star) - \frac{2}{t}\sum_{j=1}^t y_j \boldsymbol{x}_j \xi_j / p_t
\end{aligned}$$

we have

$$\begin{aligned}
|\langle \boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{e}_i \rangle| &= \left| \left\langle 2\left(\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}\right)(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star) - \frac{2}{t}\sum_{j=1}^t y_j \boldsymbol{x}_j \xi_j / p_t, \boldsymbol{e}_i \right\rangle \right| \\
&\le \underbrace{2\left| \left\langle \left(\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}\right)(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star), \boldsymbol{e}_i \right\rangle \right|}_{\text{Part 1}} + \underbrace{2\left| \frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i} \xi_j / p_t \right|}_{\text{Part 2}}
\end{aligned}$$

We consider the two parts separately. Notice that, in the first part, $\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star$ is at most $2s$-sparse, which means that the first part can be bounded by

$$\begin{aligned}
\max_{i\in[d]} \left| \left\langle \left(\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}\right)(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star), \boldsymbol{e}_i \right\rangle \right| &\le 2 \max_{i,j\in[d]} \left| \widehat{\boldsymbol{\Sigma}}_{t,ij} - \boldsymbol{\Sigma}_{ij} \right| \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_{\ell_1} \\
&\le 2\sqrt{2s} \max_{i,k\in[d]} \left| \frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i} \boldsymbol{x}_{j,k} / p_j - \boldsymbol{\Sigma}_{ik} \right| \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2.
\end{aligned}$$

Here we use the Hölder's inequality. The concentration of $\max_{i,k\in[d]} \left| \frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i} \boldsymbol{x}_{j,k} / p_j - \boldsymbol{\Sigma}_{ik} \right|$ implies that:

$$\mathbb{P}\left(\max_{i,k\in[d]} \left| \frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i} \boldsymbol{x}_{j,k} / p_j - \boldsymbol{\Sigma}_{ik} \right| \ge z \right) \le d^2 \max_{i,k\in[d]} \mathbb{P}\left(\left| \frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i} \boldsymbol{x}_{j,k} / p_j - \boldsymbol{\Sigma}_{ik} \right| \ge z \right),$$

By the martingale structure of $\frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i} \boldsymbol{x}_{j,k} / p_j - \boldsymbol{\Sigma}_{ik}$:

$$\mathbb{E}\left[y_j \boldsymbol{x}_{j,i} \boldsymbol{x}_{j,k} / p_j - \boldsymbol{\Sigma}_{ik} | \mathcal{H}_{j-1} \right] = 0, \quad |y_j \boldsymbol{x}_{j,i} \boldsymbol{x}_{j,k} / p_j - \boldsymbol{\Sigma}_{ik}| \le 2D^2 / \underline{p_j},$$

We can use the Bernstein-type martingale concentration inequality in Lemma 4 to derive the following bound:

$$\mathbb{P}\left(\left| \frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i} \boldsymbol{x}_{j,k} / p_j - \boldsymbol{\Sigma}_{ik} \right| \ge z \right) \le 2\exp\left\{ -\frac{cz^2}{D^4 (\sum_{j=1}^t 1/\underline{p_j})/t^2 + 2D^2 z/(t\underline{p_t})} \right\},$$

where we select $v^2 = D^4(\sum_{j=1}^t 1/\underline{p_j})/t^2$, and $b = 2D^2/(t\underline{p_t})$. Thus, with the probability at least $1-\epsilon$, we can control the concentration at the level:

$$\left|\frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}/p_j - \boldsymbol{\Sigma}_{ik}\right| \leq CD^2\frac{1}{t}\sqrt{\sum_{j=1}^t \frac{1}{\underline{p_j}}}\sqrt{\log(1/\epsilon)} + CD^2\frac{1}{t\underline{p_t}}\log(1/\epsilon).$$

For simplicity, we only consider $\underline{p_j} = j^{-\alpha}$. Then, when $\alpha \leq \frac{1}{3}$, the tail can be controlled by the level

$$\left|\frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}/p_j - \boldsymbol{\Sigma}_{ik}\right| \leq CD^2\frac{1}{t}\sqrt{\sum_{j=1}^t \frac{1}{\underline{p_j}}}\sqrt{\log(1/\epsilon)} = L_\epsilon$$

For the bound on the expectation, we have

$$\mathbb{E}\max_{i\in[d]}\left|\left\langle\left(\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}\right)(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star), \boldsymbol{e}_i\right\rangle\right|^2 \leq 8s\mathbb{E}\max_{i,k\in[d]}\left|\frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}/p_j - \boldsymbol{\Sigma}_{ik}\right|^2 \|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2$$

Define $\bar{\mu}$ as an upper bound of the $\|\boldsymbol{\mu}_\star\|_2$ which can as large as $O(\mathrm{Poly}(d))$. We choose $\epsilon = \frac{\sigma^2}{s^2d^2(\sum_{j=1}^t 1/\underline{p_j})\bar{\mu}^2D^2}$. It follows that

$$\begin{aligned}
&\mathbb{E}\max_{i\in[d]}\left|\left\langle\left(\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}\right)(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star), \boldsymbol{e}_i\right\rangle\right|^2 \\
&\leq \mathbb{E}8s\mathbb{1}\left\{\max_{i,k\in[d]}\left|\frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}/p_j - \boldsymbol{\Sigma}_{ik}\right| \leq L_\epsilon\right\} L_\epsilon^2\|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 \\
&\quad + C\mathbb{E}s\mathbb{1}\left\{\max_{i,k\in[d]}\left|\frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}/p_j - \boldsymbol{\Sigma}_{ik}\right| > L_\epsilon\right\} s\bar{\mu}^2D^4\left(\frac{1}{t}\sum_{j=1}^t 1/\underline{p_j}\right)^2 \qquad (13)\\
&\leq CsL_\epsilon^2\mathbb{E}\|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 + C\frac{\sigma^2}{t^2}(\sum_{j=1}^t 1/\underline{p_j}) \\
&\leq Cs\frac{D^2}{t^2}(\sum_{j=1}^t 1/\underline{p_j})\left(\log(dt) + \log\left(\frac{\bar{\mu}D^2}{\sigma}\right)\right)\mathbb{E}\|\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\|_2^2 + C\frac{\sigma^2D^2}{t^2}(\sum_{j=1}^t 1/\underline{p_j})
\end{aligned}$$

This gives the upper bound of Part 1. We now proceed to control Part 2 analogously. Invoke Lemma 4 again, we select $v^2 = \sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})/t^2$, and $b = \sigma D/(t\underline{p_t})$. We then have the concentration bound:

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i}\xi_j/p_t\right| \geq z\right) &\leq 2\exp\left\{-\frac{cz^2}{\sigma^2(\sum_{j=1}^t 1/\underline{p_j})/t^2 + 2\sigma z/(t\underline{p_t})}\right\} \\
&\leq 4\exp\left\{-\frac{cz^2}{2\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})/t^2}\right\} + 4\exp\left\{-\frac{cz}{4\sigma D/(t\underline{p_t})}\right\}
\end{aligned}$$

and the tail on the maximum:

$$\begin{aligned}
\mathbb{P}\left(\max_{i\in[d]}\left|\frac{1}{t}\sum_{j=1}^t y_j \boldsymbol{x}_{j,i}\xi_j/p_t\right| \geq z\right) &\leq 4d\exp\left\{-\frac{cz^2}{2\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})/t^2}\right\} + 4d\exp\left\{-\frac{cz}{4\sigma D/(t\underline{p_t})}\right\} \\
&= 4\exp\left\{-\frac{cz^2}{2\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})/t^2} + \log d\right\} + 4d\exp\left\{-\frac{cz}{4\sigma D/(t\underline{p_t})} + \log d\right\}
\end{aligned}$$

According to the tail-to-expectation formula: $\mathbb{E}X^2 = 2\int z\mathbb{P}(|X|>z)dz$, we have

$$\begin{aligned}
\mathbb{E}\max_{i\in[d]}\left|\frac{1}{t}\sum_{j=1}^{t} y_j \boldsymbol{x}_{j,i}\xi_j/p_t\right|^2 \leq & 8\int_0^\infty z\exp\left\{-\frac{cz^2}{2\sigma^2 D^2(\sum_{j=1}^t 1/\underline{p_j})/t^2}+\log d\right\}dz \\
& +8\int_0^\infty z\exp\left\{-\frac{cz}{4\sigma D/(t\underline{p_t})}+\log d\right\}dz \\
\leq & 8\int_0^{z_1} z dz + 8\int_{z_1}^\infty z\exp\left\{-\frac{cz^2}{2\sigma^2 D^2(\sum_{j=1}^t 1/\underline{p_j})/t^2}+\log d\right\}dz \\
& +8\int_0^{z_2} z dz + 8\int_{z_2}^\infty z\exp\left\{-\frac{cz}{4\sigma D/(t\underline{p_t})}+\log d\right\}dz \\
\lesssim & \frac{\sigma^2 D^2(\sum_{j=1}^t 1/\underline{p_j})\log d}{t^2}+\frac{\sigma D\log d}{tp_t}+\frac{\sigma^2 D^2\log d^2}{t^2 p_t^2} \\
\leq & C\frac{\sigma^2 D^2(\sum_{j=1}^t 1/\underline{p_j})\log d}{t^2}.
\end{aligned}$$

Here in the second inequality we choose $z_1 = \sqrt{c\sigma^2 D^2(\sum_{j=1}^t 1/\underline{p_j})\log d/t^2}$, and $z_2 = c\sigma D\log d/(t\underline{p_j})$, and compute the integration by substitution. Combining Part 1 and Part 2, we have

$$\begin{aligned}
\mathbb{E}\max_{i\in[d]}|\langle \boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{e}_i\rangle|^2 \leq & 8\mathbb{E}\max_{i\in[d]}\left|\left\langle\left(\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}\right)(\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star), \boldsymbol{e}_i\right\rangle\right|^2 + 8\mathbb{E}\max_{i\in[d]}\left|\frac{1}{t}\sum_{j=1}^t y_j\boldsymbol{x}_{j,i}\xi_j/p_t\right|^2 \\
\leq & Cs\frac{D^2}{t^2}(\sum_{j=1}^t 1/\underline{p_j})\left(\log(dt)+\log\left(\frac{\bar{\mu}D^2}{\sigma}\right)\right)\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2 \\
& + C\frac{\sigma^2 D^2(\sum_{j=1}^t 1/\underline{p_j})\log d}{t^2}. \\
\leq & C\frac{sD^2\log(dt)}{t^2}(\sum_{j=1}^t 1/\underline{p_j})\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2 + C\frac{\sigma^2 D^2(\sum_{j=1}^t 1/\underline{p_j})\log d}{t^2},
\end{aligned}$$

which gives us the first claim, the expectation bound. For the second claim, the probability bound, we only need to apply the aforementioned tail bound to Part 1 and 2 again. With Lemma 4, it is clear that with probability at least $1-\epsilon$,

$$\max_{i,k\in[d]}\left|\frac{1}{t}\sum_{j=1}^t y_j\boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}/p_j - \boldsymbol{\Sigma}_{ik}\right| \leq CD^2\frac{1}{t}\sqrt{\sum_{j=1}^t\frac{1}{\underline{p_j}}}\sqrt{\log(d/\epsilon)},$$

and with probability at least $1-\epsilon$,

$$\max_{i\in[d]}\left|\frac{1}{t}\sum_{j=1}^t y_j\boldsymbol{x}_{j,i}\xi_j/p_t\right| \leq \frac{\sigma D\log(d/\epsilon)}{t\underline{p_t}} + C\frac{\sigma D}{t}\sqrt{\sum_{j=1}^t\frac{1}{\underline{p_j}}}\sqrt{\log(d/\epsilon)} \leq C\frac{\sigma D}{t}\sqrt{\sum_{j=1}^t\frac{1}{\underline{p_j}}}\sqrt{\log(d/\epsilon)}.$$

Therefore, with probability at least $1-\epsilon$, the variation can be controlled by

$$\max_{i\in[d]}|\langle \boldsymbol{g}_t - \nabla f(\boldsymbol{\mu}_{t-1}), \boldsymbol{e}_i\rangle|^2 \leq CsD^2\frac{\log(d/\epsilon)}{t^2}\left(\sum_{j=1}^t\frac{1}{\underline{p_j}}\right)\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2 + C\frac{\sigma^2 D^2(\sum_{j=1}^t 1/\underline{p_j})\log(d/\epsilon)}{t^2}$$

□

LEMMA 4 **(Bernstein-type Martingale Concentration for Heterogeneous Variables)**. *Suppose* $\{D_t\}_{t=1}^T$ *are martingale differences that are adapted to the filtration* $\{\mathcal{F}_t\}_{t=0}^{T-1}$, *i.e.,* $\mathbb{E}[D_t|\mathcal{F}_{t-1}]=0$. *If* $\{D_t\}_{t=1}^T$ *satisfies*

1. $\sum_{t=1}^T \mathrm{Var}(D_t|\mathcal{F}_{t-1}) \le v^2$,
2. $\mathbb{E}\left[|D_t|^k \middle| \mathcal{F}_{t-1}\right] \le k! b^{k-2}$, *for any* $k \ge 3$.

Then, there exists a universal constant c *such that the following probability bound holds*

$$\mathbb{P}\left(\left|\sum_{t=1}^T D_t\right| \ge z\right) \le 2\exp\left\{-\frac{cz^2}{v^2+bz}\right\}$$

This is a general version of Bernstein-type martingale concentration inequality (Freedman 1975). The Lemma 4 can be easily justified by applying the martingale argument to the classic Bernstein inequality (see, for example, Wainwright (2019)). The key idea is to prove that, conditional on the history $\mathcal{F}_{t-1}$, the moment-generating function of D_t can be bounded by $\exp\left\{-\frac{\lambda^2\sigma_t^2}{1-b|\lambda|}\right\}$ (up to some constant factor) with the individual variance σ_t^2.

8.3. Proof of Corollary 1

From the proof of Theorem 1, we can easily derive the following bound from equation (11):

$$\begin{aligned}\max_a \|\boldsymbol{\mu}_{a,t} - \boldsymbol{\mu}_a^\star\|_2^2 \le& \left(1+\frac{3}{2}\sqrt{\rho}\right)\left(1-4\phi_{\min}(s)\eta_t + 8\eta_t^2\phi_{\max}^2(s)\right)\max_a \|\boldsymbol{\mu}_{a,t}-\boldsymbol{\mu}_a^\star\|_2^2 \\ &+ 18s\eta_t^2 \max_{i\in[d],a} |\langle \boldsymbol{g}_{a,t} - \nabla f_a(\boldsymbol{\mu}_{a,t-1}), \boldsymbol{e}_i\rangle|^2 + 18\eta_t\sqrt{s}\max_{i\in[d],a}|\langle \boldsymbol{g}_{a,t} - \nabla f_a(\boldsymbol{\mu}_{a,t-1}), \boldsymbol{e}_i\rangle| \max_a \|\boldsymbol{\mu}_{a,t-1}-\boldsymbol{\mu}_a^\star\|_2. \end{aligned} \tag{14}$$

Analogous to the proof of Lemma 3, we can also prove that

LEMMA 5. *We have*

$$\begin{aligned}\mathbb{E}\max_{i\in[d],a} |\langle \boldsymbol{g}_{a,t} - \nabla f_a(\boldsymbol{\mu}_{a,t-1}), \boldsymbol{e}_i\rangle|^2 \le& C\frac{sD^2\log(dKt)}{t^2}\left(\sum_{j=1}^t 1/\underline{p_j}\right)\mathbb{E}\max_a \|\boldsymbol{\mu}_{a,t}-\boldsymbol{\mu}_a^\star\|_2^2 \\ &+ C\frac{\sigma^2D^2(\sum_{j=1}^t 1/\underline{p_j})\log(dK)}{t^2}.\end{aligned}$$

Here we have an extra $\log K$ term compared with Lemma 3 because we take the maximum overall arms. Together with (14), we can essentially show that

$$\mathbb{E}\max_a \left\|\boldsymbol{\mu}_{a,t}^{\mathsf{s}} - \boldsymbol{\mu}_a^\star\right\|_2^2 \lesssim \frac{\sigma^2D^2s_0}{\phi_{\min}^2(s)}\frac{\log(dK)}{t^2}\left(\sum_{j=1}^t \frac{1}{\underline{p_j}}\right),$$

8.4. Proof of Theorem 2

Proof . For simplicity, we just write the sparse estimations of all $\boldsymbol{\mu}_{a,t}^{\mathsf{s}}$ as $\boldsymbol{M}_t \in \mathbb{R}^{d\times K}$ collectively in the following regret analysis of the BwK problem, with the corresponding optimal value $\boldsymbol{M}^\star \in \mathbb{R}^{d\times K}$. We denote by τ the time period that one of the resources are depleted or let $\tau = T$ if there

are remaining resources at the end of the horizon. Note that by the decision rule of the algorithm, for each t, with probability $1-K\epsilon_t$, we have

$$(\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}_t(\boldsymbol{x}_t) - Z\cdot \boldsymbol{\eta}_t^\top \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) \geq (\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}^*(\boldsymbol{x}_t) - Z\cdot \boldsymbol{\eta}_t^\top \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t) \quad (15)$$

where we denote by $\boldsymbol{y}^*\in\mathbb{R}^K$ one optimal solution to V^{UB}. On the other hand, with probability $K\epsilon_t$, we pull an arm randomly in the execution of Algorithm 2, which implies that

$$\begin{aligned}&(\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}_t(\boldsymbol{x}_t) - Z\cdot \boldsymbol{\eta}_t^\top \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)\\ \geq&(\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}^*(\boldsymbol{x}_t) - Z\cdot \boldsymbol{\eta}_t^\top \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t) - 2R_{\max} - D'Z\end{aligned} \quad (16)$$

since $(\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}_t(\boldsymbol{x}_t) - Z\cdot \boldsymbol{\eta}_t^\top \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) \geq -R_{\max} - D'Z$ and $(\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}^*(\boldsymbol{x}_t) - Z\cdot \boldsymbol{\eta}_t^\top \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t) \leq R_{\max} + D'Z$. Then, we take expectations on both sides of (15) and sum up t from $t=1$ to $t=\tau$ to obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{\tau}\left((\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}_t(\boldsymbol{x}_t) - Z\cdot \boldsymbol{\eta}_t^\top \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)\right)\right]\\ \geq&\mathbb{E}\left[\sum_{t=1}^{\tau}\left((\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}^*(\boldsymbol{x}_t) - Z\cdot \boldsymbol{\eta}_t^\top \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t)\right)\right] - 2(R_{\max}+D'Z)\cdot\sum_{t=1}^{T} K\epsilon_t.\end{aligned} \quad (17)$$

We have

$$\begin{aligned}\sum_{t=1}^{\tau}\mathbb{E}\left[(\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}^*(\boldsymbol{x}_t)\right] &\geq \sum_{t=1}^{\tau}\mathbb{E}\left[((\boldsymbol{M}^\star)^\top \boldsymbol{x}_t)^\top \boldsymbol{y}^*(\boldsymbol{x}_t)\right] - \mathbb{E}\sum_{t=1}^{\tau}\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,t-1}^{\mathsf{s}}\right\rangle\right|\\ &= \frac{\tau}{T}\cdot V^{\mathrm{UB}} - \mathbb{E}\sum_{t=1}^{\tau}\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,t-1}^{\mathsf{s}}\right\rangle\right|,\end{aligned} \quad (18)$$

and

$$\mathbb{E}\left[\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t)\right] \leq \frac{\boldsymbol{C}}{T}. \quad (19)$$

Moreover, from the dual update rule, for any $\boldsymbol{\eta}$, we have the following result.

Lemma 6. *For any $\boldsymbol{\eta}$, it holds that*

$$\sum_{t=1}^{\tau}\boldsymbol{\eta}_t^\top\left(\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) - \frac{\boldsymbol{C}}{T}\right) \geq \boldsymbol{\eta}^\top \sum_{t=1}^{\tau}\left(\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) - \frac{\boldsymbol{C}}{T}\right) - R(T) - 2R_{\max}\cdot\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=1}.$$

Therefore, from Lemma 6, we know that

$$\begin{aligned}&\sum_{t=1}^{\tau}\boldsymbol{\eta}_t^\top\left(\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) - \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t)\right) \geq \sum_{t=1}^{\tau}\boldsymbol{\eta}_t^\top\left(\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) - \frac{\boldsymbol{C}}{T}\right)\\ \geq&\boldsymbol{\eta}^\top \sum_{t=1}^{\tau}\left(\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) - \frac{\boldsymbol{C}}{T}\right) - R(T) - 2R_{\max}\cdot\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=1}.\end{aligned} \quad (20)$$

Then, if $\tau < T$ which implies that $\sum_{t=1}^{\tau}\sum_{a\in[K]} \boldsymbol{W}_a^\star x_{t,i} y_{a,t}(\boldsymbol{x}_t) \geq C_i$ for some resource $i \in [m]$, we set $\boldsymbol{\eta} = \boldsymbol{e}_i$ in (20) and we have

$$\begin{aligned}\sum_{t=1}^{\tau} \boldsymbol{\eta}_t^\top \left(\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) - \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t) \right) &\geq C_i \cdot \frac{T-\tau}{T} - R(T) - 2R_{\max} \cdot \sum_{t=1}^{\tau} \mathbb{1}_{\nu_t=1} \\ &\geq C_{\min} \cdot \frac{T-\tau}{T} - R(T) - 2R_{\max} \cdot \sum_{t=1}^{\tau} \mathbb{1}_{\nu_t=1}\end{aligned} \tag{21}$$

If $\tau = T$ which implies $\frac{T-\tau}{T} = 0$, we set $\boldsymbol{\eta} = 0$ in (20) and we have

$$\begin{aligned}\sum_{t=1}^{\tau} \boldsymbol{\eta}_t^\top \left(\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) - \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t) \right) &\geq -R(T) - 2R_{\max} \cdot \sum_{t=1}^{\tau} \mathbb{1}_{\nu_t=1} \\ &= C_{\min} \cdot \frac{T-\tau}{T} - R(T) - 2R_{\max} \cdot \sum_{t=1}^{\tau} \mathbb{1}_{\nu_t=1}\end{aligned} \tag{22}$$

where $C_{\min} = \min_{i\in[m]}\{C_i\}$. Therefore, combining (21) and (22), we obtain

$$\sum_{t=1}^{\tau} \boldsymbol{\eta}_t^\top \left(\sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t) - \sum_{a\in[K]} \boldsymbol{W}_a^\star \boldsymbol{x}_t y_a^*(\boldsymbol{x}_t) \right) \geq C_{\min} \cdot \frac{T-\tau}{T} - R(T) - 2R_{\max} \cdot \sum_{t=1}^{\tau} \mathbb{1}_{\nu_t=1}. \tag{23}$$

Plugging (18) and (23) into (17), we obtain

$$\begin{aligned}&\mathbb{E} \sum_{t=1}^{\tau} \left[(\boldsymbol{M}_{t-1}^\top \boldsymbol{x}_t)^\top \boldsymbol{y}_t(\boldsymbol{x}_t) \right] \\ &\geq \frac{\tau}{T} \cdot V^{\mathrm{UB}} + Z \cdot C_{\min} \cdot \frac{T-\tau}{T} - Z \cdot \mathbb{E}[R(T)] - \mathbb{E} \sum_{t=1}^{\tau} \max_a \left| \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,t-1}^{\mathsf{s}} \right\rangle \right| - (4R_{\max} + 2D'Z) \cdot \sum_{t=1}^{T} K\epsilon_t.\end{aligned} \tag{24}$$

Note that $Z \geq \frac{V^{\mathrm{UB}}}{C_{\min}}$. We have

$$\sum_{t=1}^{\tau} \mathbb{E}\left[(\boldsymbol{\mu}_t^\top \boldsymbol{x}_t)^\top \boldsymbol{y}_t(\boldsymbol{x}_t) \right] \geq V^{\mathrm{UB}} - Z \cdot \mathbb{E}[R(T)] - \mathbb{E} \sum_{t=1}^{\tau} \max_a \left| \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,t-1}^{\mathsf{s}} \right\rangle \right| - (4R_{\max} + 2D'Z) \cdot \sum_{t=1}^{T} K\epsilon_t. \tag{25}$$

Finally, we plug in the regret bound of the Hedge algorithm (from Theorem 2 of Freund and Schapire (1997)), which is the algorithm used to update the dual variable $\boldsymbol{\eta}_t$, and we obtain that

$$\mathbb{E}[R(T)] \leq \sqrt{D' \cdot T \cdot \log m}$$

by setting $\delta = \sqrt{\frac{\log m}{T \cdot D'}}$, where D' denotes an upper bound of $b_i(y_t, \boldsymbol{x}_t)$ for each $i \in [m]$, $t \in [T]$ and every y_t, $\boldsymbol{x}_t$. Therefore, our proof is completed. □

Proof of Lemma 6. We denote by $\mathcal{T}$ the number of periods from $t=1$ to $t=\tau$ such that $\nu_t=0$. Then, from the regret bound of the embedded OCO algorithm, we know that

$$\begin{aligned}\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=0}\cdot\boldsymbol{\eta}_t^\top\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)&\geq\boldsymbol{\eta}^\top\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=0}\cdot\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)-R(\mathcal{T})\\&\geq\boldsymbol{\eta}^\top\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=0}\cdot\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)-R(T).\end{aligned}\tag{26}$$

Moreover, from the boundedness of $\boldsymbol{\eta}_t$ and $\boldsymbol{x}_t$, we know that

$$\sum_{t=1}^{\tau}\boldsymbol{\eta}_t^\top\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)\geq\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=0}\cdot\boldsymbol{\eta}_t^\top\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)-R_{\max}\cdot\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=1}\tag{27}$$

and

$$\boldsymbol{\eta}^\top\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=0}\cdot\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)\geq\boldsymbol{\eta}^\top\sum_{t=1}^{\tau}\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)-R_{\max}\cdot\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=1}.\tag{28}$$

Therefore, plugging (27) and (28) into (26), we have that

$$\sum_{t=1}^{\tau}\boldsymbol{\eta}_t^\top\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)\geq\boldsymbol{\eta}^\top\sum_{t=1}^{\tau}\left(\sum_{a\in[K]}\boldsymbol{W}_a^\star\boldsymbol{x}_t y_{a,t}(\boldsymbol{x}_t)-\frac{\boldsymbol{C}}{T}\right)-2R_{\max}\cdot\sum_{t=1}^{\tau}\mathbb{1}_{\nu_t=1}-R(T),$$

which completes our proof. □

8.5. Proof of Lemma 2

Proof. We define an intermediate benchmark as follows.

$$\bar{V}=\max\ \frac{T}{T_0}\cdot\sum_{t=1}^{T_0}\sum_{a\in[K]}(\boldsymbol{\mu}_a^*)^\top\boldsymbol{x}_t\cdot y_{a,t}\tag{29a}$$

$$\text{s.t.}\ \frac{T}{T_0}\cdot\sum_{t=1}^{T_0}\sum_{a\in[K]}\boldsymbol{W}_a^*\boldsymbol{x}_t\cdot y_{a,t}\leq\boldsymbol{C}\tag{29b}$$

$$\sum_{a\in[K]}y_{a,t}=1,\forall t\in[T_0]\tag{29c}$$

$$y_{a,t}\in[0,1],\forall a\in[K],\forall t\in[T_0].\tag{29d}$$

The only difference between $\bar{V}$ in (29) and $\hat{V}$ is that the estimation $\boldsymbol{\mu}_{a,T_0}^{\mathsf{s}}$ is replaced by the true value $\boldsymbol{\mu}_a^*$, for all $a\in[K]$. Then, we can bound the gap between $\hat{V}$ and V^{UB} by bounding the two terms $|V^{\mathrm{UB}}-\bar{V}|$ and $|\bar{V}-\hat{V}|$ separately.

Bound the term $|\bar{V} - V^{\mathrm{UB}}|$: We denote by $L(\boldsymbol{\eta})$ the dual function of V^{UB} as follows:

$$\begin{aligned} L(\boldsymbol{\eta}) &= (\boldsymbol{C})^\top \boldsymbol{\eta} + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{x}_t \sim F} \left[\max_{\sum_{a\in[K]} y_{a,t}(\boldsymbol{x}_t)=1} \left\{ \sum_{a\in[K]} \left[(\boldsymbol{\mu}_a^\star)^\top \boldsymbol{x}_t \cdot y_{a,t}(\boldsymbol{x}_t) - (\boldsymbol{\eta})^\top \boldsymbol{W}_a^\star \boldsymbol{x}_t \cdot y_{a,t}(\boldsymbol{x}_t) \right] \right\} \right] \\ &= (\boldsymbol{C})^\top \boldsymbol{\eta} + T \cdot \mathbb{E}_{\boldsymbol{x}\sim F} \left[\max_{\sum_{a\in[K]} y_a(\boldsymbol{x})=1} \left\{ \sum_{a\in[K]} \left[(\boldsymbol{\mu}_a^\star)^\top \boldsymbol{x} \cdot y_a(\boldsymbol{x}) - (\boldsymbol{\eta})^\top \boldsymbol{W}_a^\star \boldsymbol{x} \cdot y_a(\boldsymbol{x}) \right] \right\} \right]. \end{aligned} \tag{30}$$

We also denote by $\bar{L}(\boldsymbol{\eta})$ the dual function of $\bar{V}$ as follows:

$$\bar{L}(\boldsymbol{\eta}) = (\boldsymbol{C})^\top \boldsymbol{\eta} + \frac{T}{T_0} \cdot \sum_{t=1}^{T_0} \max_{\sum_{a\in[K]} y_{a,t}=1} \left\{ \sum_{a\in[K]} \left[(\boldsymbol{\mu}_a^\star)^\top \boldsymbol{x}_t \cdot y_{a,t} \right] - \sum_{a\in[K]} (\boldsymbol{\eta})^\top \left[\boldsymbol{W}_a^\star \boldsymbol{x}_t \cdot y_{a,t} \right] \right\}. \tag{31}$$

Then, the function $\bar{L}(\boldsymbol{\eta})$ can be regarded as a sample average approximation of $L(\boldsymbol{\eta})$. We then proceed to bound the range of the optimal dual variable for V^{UB} and $\hat{V}$. Denote by $\boldsymbol{\eta}^*$ an optimal dual variable for V^{UB}. Then, it holds that

$$(\boldsymbol{C})^\top \boldsymbol{\eta}^* \le V^{\mathrm{UB}}$$

which implies that

$$\boldsymbol{\eta}^* \in \Omega^* := \left\{ \boldsymbol{\eta} \ge 0 : \|\boldsymbol{\eta}\|_\infty \le \frac{V^{\mathrm{UB}}}{C_{\min}} \right\}.$$

Similarly, denote by $\bar{\boldsymbol{\eta}}^*$ an optimal dual variable for $\hat{V}$ and we can obtain that

$$\bar{\boldsymbol{\eta}}^* \in \bar{\Omega}^* := \left\{ \boldsymbol{\eta} \ge 0 : \|\boldsymbol{\eta}\|_\infty \le \frac{\bar{V}}{C_{\min}} \right\}.$$

Note that

$$V^{\mathrm{UB}} = L(\boldsymbol{\eta}^*) \ge \bar{L}(\boldsymbol{\eta}^*) - |L(\boldsymbol{\eta}^*) - \bar{L}(\boldsymbol{\eta}^*)| \ge \bar{L}(\bar{\boldsymbol{\eta}^*}) - |L(\boldsymbol{\eta}^*) - \bar{L}(\boldsymbol{\eta}^*)| = \bar{V} - |L(\boldsymbol{\eta}^*) - \bar{L}(\boldsymbol{\eta}^*)| \tag{32}$$

and

$$\bar{V} = \bar{L}(\bar{\boldsymbol{\eta}}^*) \ge L(\bar{\boldsymbol{\eta}}^*) - |\bar{L}(\bar{\boldsymbol{\eta}}^*) - L(\bar{\boldsymbol{\eta}}^*)| \ge L(\boldsymbol{\eta}^*) - |\bar{L}(\bar{\boldsymbol{\eta}}^*) - L(\bar{\boldsymbol{\eta}}^*)| = V^{\mathrm{UB}} - |\bar{L}(\bar{\boldsymbol{\eta}}^*) - L(\bar{\boldsymbol{\eta}}^*)|. \tag{33}$$

Therefore, we have

$$|\bar{V} - V^{\mathrm{UB}}| \le \max\left\{ |\bar{L}(\bar{\boldsymbol{\eta}}^*) - L(\bar{\boldsymbol{\eta}}^*)|, |\bar{L}(\boldsymbol{\eta}^*) - L(\boldsymbol{\eta}^*)| \right\}. \tag{34}$$

Define a random variable $H(\boldsymbol{x}) = \max_{\sum_{a\in[K]} y_a(\boldsymbol{x})=1} \left\{ [(\boldsymbol{\mu}_a^\star)^\top \boldsymbol{x} \cdot y_a(\boldsymbol{x}) - (\boldsymbol{\eta}^*)^\top \boldsymbol{W}_a^\star \boldsymbol{x} \cdot y_a(\boldsymbol{x})] \right\}$ where $\boldsymbol{x} \sim F$. It is clear to see that $|H(\boldsymbol{x})| \le (R_{\max} + \frac{V^{\mathrm{UB}}}{C_{\min}} \cdot D')$ where D' denotes an upper bound on $\boldsymbol{W}_a^* \boldsymbol{x}$ for every $a \in [K]$ and $\boldsymbol{x}$. Then, we have

$$|\bar{L}(\bar{\boldsymbol{\eta}}^*) - L(\bar{\boldsymbol{\eta}}^*)| = \left| \mathbb{E}_{\boldsymbol{x}\sim F}[H(\boldsymbol{x})] - \frac{T}{T_0} \cdot \sum_{t=1}^{T_0} H(\boldsymbol{x}_t) \right| \le T \cdot (R_{\max} + \frac{V^{\mathrm{UB}}}{C_{\min}} \cdot D') \cdot \sqrt{\frac{1}{2T_0} \cdot \log \frac{4}{\beta}} \tag{35}$$

holds with probability at least $1-\frac{\beta}{2}$, where the inequality follows from the standard Hoeffding's inequality. Similarly, we have

$$|\bar{L}(\boldsymbol{\eta}^*) - L(\boldsymbol{\eta}^*)| \leq T \cdot (R_{\max} + \frac{\bar{V}}{C_{\min}} \cdot D') \cdot \sqrt{\frac{1}{2T_0} \cdot \log \frac{4}{\beta}} \tag{36}$$

holds with probability at least $1-\frac{\beta}{2}$. From the union bound, we know that with probability at least $1-\beta$, both (35) and (36) hold. Therefore, we have the following inequality

$$|V^{\mathrm{UB}} - \bar{V}| \leq T \cdot (R_{\max} + \frac{V^{\mathrm{UB}}}{C_{\min}} \cdot D') \cdot \sqrt{\frac{1}{2T_0} \cdot \log \frac{4}{\beta}} + \frac{V^{\mathrm{UB}}}{C_{\min}^2} \cdot D' \cdot \frac{T}{2T_0} \cdot \log \frac{4}{\beta} \tag{37}$$

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

Bound the term $|\bar{V} - \hat{V}|$: We first denote by $\bar{\boldsymbol{y}}$ an optimal solution to $\bar{V}$. Then, it is clear to see that $\bar{\boldsymbol{y}}$ is a feasible solution to $\hat{V}$. Also, note that

$$\left| \frac{T}{T_0} \cdot \sum_{t=1}^{T_0} \sum_{a \in [K]} (\boldsymbol{\mu}_a^*)^\top \boldsymbol{x}_t \cdot \bar{y}_{a,t} - \frac{T}{T_0} \cdot \sum_{t=1}^{T_0} \sum_{a \in [K]} (\boldsymbol{\mu}_{a,T_0}^{\mathsf{s}})^\top \boldsymbol{x}_t \cdot \bar{y}_{a,t} \right| \leq T \cdot D \cdot \max_{a \in [K]} \|\boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,T_0}^{\mathsf{s}}\|_1. \tag{38}$$

Therefore, we know that

$$\bar{V} \leq \frac{T}{T_0} \cdot \sum_{t=1}^{T_0} \sum_{a \in [K]} (\boldsymbol{\mu}_{a,T_0}^{\mathsf{s}})^\top \boldsymbol{x}_t \cdot \bar{y}_{a,t} + T \cdot D \cdot \max_{a \in [K]} \|\boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,T_0}^{\mathsf{s}}\|_1 \leq \hat{V} + T \cdot D \cdot \max_{a \in [K]} \|\boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,T_0}^{\mathsf{s}}\|_1.$$

On the other hand, we denote by $\hat{\boldsymbol{y}}$ an optimal solution to $\hat{V}$. Then, $\hat{\boldsymbol{y}}$ is a feasible solution to $\bar{V}$ and again, from (38), it holds that

$$\hat{V} \leq \frac{T}{T_0} \cdot \sum_{t=1}^{T_0} \sum_{a \in [K]} (\boldsymbol{\mu}_{a,T_0}^{\mathsf{s}})^\top \boldsymbol{x}_t \cdot \hat{y}_{a,t} + T \cdot D \cdot \max_{a \in [K]} \|\boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,T_0}^{\mathsf{s}}\|_1 \leq \bar{V} + T \cdot D \cdot \max_{a \in [K]} \|\boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,T_0}^{\mathsf{s}}\|_1.$$

Therefore, we conclude that

$$|\bar{V} - \hat{V}| \leq T \cdot D \cdot \max_{a \in [K]} \|\boldsymbol{\mu}_a^\star - \boldsymbol{\mu}_{a,T_0}^{\mathsf{s}}\|_1. \tag{39}$$

Our proof is completed by combining (37) and (39). □

8.6. Proof of Theorem 6 and 7

Our proof essentially follows the basic ideas of regret analysis for ϵ-greedy algorithms, with a fine-grained process on the estimation error. For the ϵ-greedy algorithm, we have

$$\begin{aligned}
&\mathrm{Regret} = \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{x}_t, \boldsymbol{\mu}_{\mathsf{opt}}(\boldsymbol{x}_t)\rangle - \sum_{t=1}^{T} \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_{y_t}^\star \right\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_{\mathsf{opt}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_{\mathsf{opt},t-1}^{\mathsf{s}}(\boldsymbol{x}_t)\right\rangle - \sum_{t=1}^{T} \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_{y_t^*,t-1}^{\mathsf{s}} - \boldsymbol{\mu}_{\mathsf{opt},t-1}^{\mathsf{s}}(\boldsymbol{x}_t)\right\rangle \right. \\
&\quad \left. + \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_{y_t^*,t-1}^{\mathsf{s}} - \boldsymbol{\mu}_{y_t^*}^\star \right\rangle + \left\langle \boldsymbol{x}_t, \boldsymbol{\mu}_{y_t^*}^\star - \boldsymbol{\mu}_{y_t}^\star \right\rangle\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\|\boldsymbol{x}_t\|_\infty \left(\left\|\boldsymbol{\mu}_{\mathsf{opt}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_{\mathsf{opt},t-1}^{\mathsf{s}}(\boldsymbol{x}_t)\right\|_1 + \left\|\boldsymbol{\mu}_{y_t^*}^{\mathsf{s}} - \boldsymbol{\mu}_{y_t^*,t-1}^\star\right\|_1 \right) + 2\sum_{t=1}^{T} K\epsilon_t R_{max}
\end{aligned}$$

where y_t^* means the greedy action $y_t^* = \arg\max_{a\in[K]} \langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1}\rangle$, and $\boldsymbol{\mu}^{\mathsf{s}}_{\mathsf{opt},t-1}(\boldsymbol{x}_t)$ indicates the estimation of the optimal arm $\boldsymbol{\mu}_{\mathsf{opt}}(\boldsymbol{x}_t)$. The inequality uses the fact of greedy action, and the uniform risk bound. This leads to the regret-bound

$$\begin{aligned}\text{Regret} &\leq 2D\sum_{t=1}^{T}\mathbb{E}\sqrt{s_0}\max_a \left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t} - \boldsymbol{\mu}^{\star}_a\right\|_2 + 2\sum_{t=1}^{T} K\epsilon_t R_{max} \\ &\lesssim \frac{\sigma D^2 s_0\sqrt{\log(dK)}}{\phi_{\min}(s)}\sum_{t=1}^{T}\left(\frac{1}{t}\sqrt{\sum_{j=1}^{t}\frac{1}{\epsilon_j}}\right) + \sum_{t=1}^{T} K\epsilon_t R_{max}.\end{aligned}$$

Choosing $\epsilon_t = \sigma^{\frac{2}{3}} D^{\frac{4}{3}} s_0^{\frac{2}{3}} (\log(dK))^{\frac{1}{3}} t^{-\frac{1}{3}} / (KR_{\max})^{\frac{2}{3}} \wedge 1/K$, the statement in Theorem 6 can be justified. For the Theorem 7, since it can be viewed as a special case of ϵ-greedy strategy (with $\epsilon = 0$), we have

$$\text{Regret} \leq 2D\sum_{t=1}^{T}\mathbb{E}\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right|,$$

where the estimation error can be guaranteed by

$$\mathbb{E}\max_a \left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t} - \boldsymbol{\mu}^{\star}_a\right\|_2^2 \lesssim \frac{\sigma^2 D^2 s_0}{\gamma^2(K)\zeta^2(K)}\frac{\log(dK)}{t}. \tag{40}$$

This error bound can be easily derived from the proof of Theorem 4. Here each term $\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right|$ in the regret can be controlled by two ways:

$$\mathbb{E}\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right| \leq D\mathbb{E}\max_a \left\|\boldsymbol{\mu}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\|_1, \tag{41}$$

and

$$\begin{aligned}&\mathbb{E}\left[\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right| - \mathbb{E}\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right|\right] \\ &\leq \int_0^{\infty}\mathbb{P}\left(\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right| - \mathbb{E}\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right| \geq z\right) dz\end{aligned} \tag{42}$$

Combining (40) with (41), it is easy to show that the regret bound:

$$\text{Regret} \leq 2D\sum_{t=1}^{T}\mathbb{E}\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right| \lesssim \frac{\sigma D^2 s_0\sqrt{\log(dK)T}}{\gamma(K)\zeta(K)}.$$

We use (42) to give another bound. Notice that $\boldsymbol{x}_t$ is independent of the history $\mathcal{H}_{t-1}$, which implies that, conditional on the history $\mathcal{H}_{t-1}$,

$$\begin{aligned}\mathbb{E}\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right| &\leq \sqrt{\mathbb{E}\left(\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right)^{\top}\boldsymbol{x}_t\boldsymbol{x}_t^{\top}\left(\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right)} \leq \sqrt{\left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\|_{\boldsymbol{\Sigma}}^2}. \\ &\leq \sqrt{\phi_{\max}(s_0)}\left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\|_2.\end{aligned}$$

Since $\boldsymbol{x}_t$ is marginal sub-Gaussian, the $\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right|$ has a tail behavior by Chernoff bound:

$$\mathbb{P}\left(\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right| - \mathbb{E}\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\rangle\right| \geq z\right) \leq \exp\left\{-\frac{cz^2}{\phi_{\max}(s_0)\left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^{\star}_a\right\|_2^2}\right\},$$

and also

$$\begin{aligned}&\mathbb{P}\left(\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right| - \mathbb{E}\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right| \geq z\right)\\ &\leq 1 \wedge \exp\left\{\log K - \frac{cz^2}{\phi_{\max}(s_0)\max_a \left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\|_2^2}\right\}.\end{aligned}$$

This instantly gives rise to the maxima inequality by (42)

$$\begin{aligned}&\mathbb{E}\left[\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right| - \mathbb{E}\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right|\right]\\ &\leq \int_0^\infty 1 \wedge \exp\left\{\log K - \frac{cz^2}{\phi_{\max}(s_0)\max_a \left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\|_2^2}\right\} dz\\ &\lesssim \sqrt{\log K \phi_{\max}(s_0)} \max_a \left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\|_2\end{aligned}$$

We thus have

$$\begin{aligned}&\mathbb{E}\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right|\\ &\leq \mathbb{E}\left[\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right| - \mathbb{E}\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right|\right] + \max_a \mathbb{E}\left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right|\\ &\lesssim \sqrt{\log K \phi_{\max}(s_0)} \max_a \left\|\boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\|_2,\end{aligned}$$

conditional on the history $\mathcal{H}_{t-1}$. Together with the estimation error (40), we can derive another regret bound:

$$\begin{aligned}\text{Regret} &\leq 2D\sum_{t=1}^T \mathbb{E}\max_a \left|\left\langle \boldsymbol{x}_t, \boldsymbol{\mu}^{\mathsf{s}}_{a,t-1} - \boldsymbol{\mu}^\star_a\right\rangle\right| \lesssim \sqrt{\log K \phi_{\max}(s_0)}\frac{\sigma D\sqrt{s_0 \log(dK)T}}{\gamma(K)\zeta(K)}\\ &\lesssim \frac{\sqrt{\kappa_1}\sigma D\sqrt{s_0 \log K \log(dK)T}}{\sqrt{\gamma(K)\zeta(K)}}\end{aligned}$$

Associate these two regret bounds, we finish the proof.

8.7. Proof of Theorem 4

Proof. The proof shares a similar fashion with the proof of Theorem 1. The key difference is that, instead of focusing on the concentration of the gradient $\boldsymbol{g}_{a,t}$ to the population version $\nabla f^a(\boldsymbol{\mu}_{a,t-1})$, we consider a series of new objective functions $\{f^a_t\}$ that is changing over time, and derive the concentration of $\boldsymbol{g}_{a,t}$ to $\nabla f^a_t(\boldsymbol{\mu}_{t-1})$. To this end, we defined the history-dependent covariance matrices $\mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top \cdot \mathbb{1}\left\{y_t = a\right\}\middle|\mathcal{H}_{t-1}\right]$, and their average: $\bar{\boldsymbol{\Sigma}}_{a,t} = \sum_{j=1}^t \mathbb{E}\left[\boldsymbol{x}_j\boldsymbol{x}_j^\top \cdot \mathbb{1}\left\{y_j = a\right\}\middle|\mathcal{H}_{j-1}\right]/t$. We write the corresponding objective function that $\bar{\boldsymbol{\Sigma}}_{a,t}$ represents as $f^a_t(\boldsymbol{\mu}) = \|\boldsymbol{\mu} - \boldsymbol{\mu}_{\star,a}\|^2_{\bar{\boldsymbol{\Sigma}}_{a,t}}$. In the following proof, since we will mainly focus on one arm, we will write $\boldsymbol{\mu}_t$, $\boldsymbol{\mu}_\star$, $\boldsymbol{g}_t$, f_t, $\widehat{\boldsymbol{\Sigma}}_t$, $\bar{\boldsymbol{\Sigma}}_t$ etc instead of $\boldsymbol{\mu}_{a,t}$, $\boldsymbol{\mu}^\star_a$, $\boldsymbol{g}_{a,t}$, f^a_t, $\widehat{\boldsymbol{\Sigma}}_{a,t}$ and $\bar{\boldsymbol{\Sigma}}_{a,t}$, etc to easy the notation. An argument analog to the proof of Theorem 1 gives that:

$$\begin{aligned}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2 &\le \left(1+\frac{3}{2}\sqrt{\rho}\right)\left(\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2-2\eta_t\langle\Omega(\boldsymbol{g}_t),\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\rangle+\eta_t^2\|\Omega(\boldsymbol{g}_t)\|_2^2\right)\\ &\le \left(1+\frac{3}{2}\sqrt{\rho}\right)\Big(\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2-2\eta_t\langle\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\rangle+2\eta_t^2\|\Omega(\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}))\|_2^2\\ &\quad+2\eta_t^2\|\Omega(\nabla f_t(\boldsymbol{\mu}_{t-1}))\|_2^2+2\eta_t\|\Omega(\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}))\|_2\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2\Big),\end{aligned}$$

where we use the fact that $\langle\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\rangle=\langle\Omega(\nabla f_t(\boldsymbol{\mu}_{t-1})),\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\rangle$ by the definition of $\Omega(\cdot)$. Because we are interested in the new objective function $f_t(\boldsymbol{\mu})=\|\boldsymbol{\mu}-\boldsymbol{\mu}_\star\|_{\bar{\boldsymbol{\Sigma}}_t}^2$, we need to check the sparse eigenvalue of $\bar{\boldsymbol{\Sigma}}_t$. Since for any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\|_0\le\lceil 2s\rceil$, we have $\boldsymbol{\beta}^\top\mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top\cdot\mathbb{1}\left\{y_t=a\right\}\middle|\mathcal{H}_{t-1}\right]\boldsymbol{\beta}\le\boldsymbol{\beta}^\top\mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top\middle|\mathcal{H}_{t-1}\right]\boldsymbol{\beta}\le\phi_{\max}(s)\|\boldsymbol{\beta}\|_2^2$, then it is clear that the $2s$-sparse maximal eigenvalue of $\bar{\boldsymbol{\Sigma}}_t=\sum_{j=1}^t\mathbb{E}\left[\boldsymbol{x}_j\boldsymbol{x}_j^\top\cdot\mathbb{1}\left\{y_j=a\right\}\middle|\mathcal{H}_{j-1}\right]/t$ is bounded by $\phi_{\max}(s)$. For the minimum eigenvalue, it follows by Assumption 2 that given any unit vector $\boldsymbol{v}$,

$$\begin{aligned}\boldsymbol{v}^\top\mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top\mathbb{1}\left\{y_t=a\right\}\middle|\mathcal{H}_{t-1}\right]\boldsymbol{v}&\ge\mathbb{E}\left[\boldsymbol{v}^\top\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{v}\mathbb{1}\left\{y_t=a\right\}\mathbb{1}\left\{\boldsymbol{v}^\top\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{v}\mathbb{1}\left\{y_t=a\right\}\ge\gamma(K)\right\}\middle|\mathcal{H}_{t-1}\right]\\ &\ge\mathbb{E}\left[\gamma(K)\mathbb{1}\left\{\boldsymbol{v}^\top\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{v}\mathbb{1}\left\{y_t=a\right\}\ge\gamma(K)\right\}\middle|\mathcal{H}_{t-1}\right]\\ &\ge\gamma(K)\zeta(K).\end{aligned}\tag{43}$$

It is clear that the $2s$-sparse minimum eigenvalue of $\bar{\boldsymbol{\Sigma}}_t$ can be lower bounded by $\gamma(K)\zeta(K)$. We therefore take the condition number of $\bar{\boldsymbol{\Sigma}}_t$ as $\kappa_1=\frac{\phi_{\max}(s)}{\gamma(K)\zeta(K)}$. The eigenvalues of $\bar{\boldsymbol{\Sigma}}_t$ also imply:

$$\begin{aligned}\langle\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\rangle&\ge2\gamma(K)\zeta(K)\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2,\\ \|\Omega(\nabla f_t(\boldsymbol{\mu}_{t-1}))\|&\le2\phi_{\max}(s)\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2.\end{aligned}$$

We can show that

$$\begin{aligned}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2&\le\left(1+\frac{3}{2}\sqrt{\rho}\right)\left(1-4\gamma(K)\zeta(K)\eta_t+8\eta_t^2\phi_{\max}^2(s)\right)\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\ &\quad+6\eta_t^2\|\Omega(\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}))\|_2^2+6\eta_t\|\Omega(\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}))\|_2\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2\\ &\le\left(1+\frac{3}{2}\sqrt{\rho}\right)\left(1-4\gamma(K)\zeta(K)\eta_t+8\eta_t^2\phi_{\max}^2(s)\right)\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\ &\quad+18s\eta_t^2\max_{i\in[d]}|\langle\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle|^2+18\eta_t\sqrt{s}\max_{i\in[d]}|\langle\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle|\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2\end{aligned}\tag{44}$$

The following lemma, which echoes with aforementioned Lemma 3, quantifies the variation of the averaged stochastic gradient under the diverse covariate condition without ε-greedy strategy:

LEMMA 7. *Define $\{\boldsymbol{e}_i\}_1^d$ as the canonical basis of $\mathbb{R}^d$. Under Assumption 1, 1 and 2, the variance of stochastic gradient $\boldsymbol{g}_t$ at the point $\boldsymbol{\mu}_{t-1}$ given in Algorithm 1 can be bounded by the following inequality:*

$$\mathbb{E}\max_{i\in[d]}|\langle\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle|^2\le C\frac{sD^2\log(dt)}{t}\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2+C\frac{\sigma^2D^2\log d}{t}.\tag{45}$$

Moreover, the following inequality also holds with probability at least $1-\epsilon$

$$\max_{i\in[d]} |\langle \boldsymbol{g}_t - \nabla f_t(\boldsymbol{\mu}_{t-1}), \boldsymbol{e}_i\rangle|^2 \leq CsD^2\frac{\log(d/\epsilon)}{t}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2 + C\frac{\sigma^2D^2\log(d/\epsilon)}{t}.$$

We defer the proof of Lemma 7 to the next section.

We set $\rho = \frac{1}{9\kappa_1^4}$, and $\eta_t = \frac{1}{4\kappa_1\phi_{\max}(s)}$. Plugging in the expectation bound in Lemma 7, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2 \leq &\left(1-\frac{1}{4\kappa_1^4}+C\frac{s_0D\sqrt{\log(dt)}}{\gamma(K)\zeta(K)\sqrt{t}}\right)\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\ &+C\frac{s_0\sigma^2D^2\log d}{\gamma^2(K)\zeta^2(K)t}+C\sqrt{\frac{s_0\sigma^2D^2\log d}{\gamma^2(K)\zeta^2(K)t}\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2}.\end{aligned}$$

When t is sufficiently large, essentially we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2 \leq &\left(1-\frac{1}{5\kappa_1^4}\right)\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2\\ &+C\frac{s_0\sigma^2D^2\log d}{\gamma^2(K)\zeta^2(K)t}+C\sqrt{\frac{s_0\sigma^2D^2\log d}{\gamma^2(K)\zeta^2(K)t}\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2}.\end{aligned}$$

This instantly gives us the expectation bound

$$\mathbb{E}\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2 \lesssim \frac{\sigma^2D^2s_0}{\gamma^2(K)\zeta^2(K)}\frac{\log d}{t},$$

which proves the first claim. Apply Lemma 7 again to the recursive relationship in (44), we also have the second claim:

$$\|\boldsymbol{\mu}_t-\boldsymbol{\mu}_\star\|_2^2 \lesssim \frac{\sigma^2D^2s_0}{\gamma^2(K)\zeta^2(K)}\frac{\log(dT/\varepsilon)}{t}$$

holds for all $t\in[T]$ with probability at least $1-\epsilon$

□

8.8. Proof of Lemma 7

Proof. The idea essentially follows the proof of Lemma 3, with some modifications in the martingale concentration argument. Notice that, in Algorithm 1, for any arm $a\in[K]$, we have

$$\begin{aligned}\boldsymbol{g}_t &= 2\widehat{\boldsymbol{\Sigma}}_t\boldsymbol{\mu}_{t-1} - \frac{2}{t}\sum_{j=1}^t \mathbb{1}\{y_t=a\}\boldsymbol{x}_j r_j = \frac{2}{t}\sum_{j=1}^t\left(\mathbb{1}\{y_j=a\}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)(\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star) - \frac{2}{t}\sum_{j=1}^t\mathbb{1}\{y_t=a\}\boldsymbol{x}_j\xi_j,\\ &= 2\widehat{\boldsymbol{\Sigma}}_t(\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star) - \frac{2}{t}\sum_{j=1}^t\mathbb{1}\{y_j=a\}\boldsymbol{x}_j\xi_j.\end{aligned}$$

Still, we can write

$$\begin{aligned}|\langle \boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle| &= \left|\left\langle 2\left(\widehat{\boldsymbol{\Sigma}}_t-\bar{\boldsymbol{\Sigma}}_t\right)(\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star) - \frac{2}{t}\sum_{j=1}^t y_j\boldsymbol{x}_j\xi_j/p_t, \boldsymbol{e}_i\right\rangle\right|\\ &\leq \underbrace{2\left|\left\langle\left(\widehat{\boldsymbol{\Sigma}}_t-\bar{\boldsymbol{\Sigma}}_t\right)(\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star),\boldsymbol{e}_i\right\rangle\right|}_{\text{Part 1}} + \underbrace{2\left|\frac{1}{t}\sum_{j=1}^t\mathbb{1}\{y_j=a\}\boldsymbol{x}_{j,i}\xi_j\right|}_{\text{Part 2}}\end{aligned}$$

We consider the two parts separately.

In Part 1, for any $i,k\in[d]$, by the martingale structure of $\frac{1}{t}\sum_{j=1}^{t}\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}-\bar{\boldsymbol{\Sigma}}_{t,ik}$:

$$\mathbb{E}\sum_{j=1}^{t}\left[\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}|\mathcal{H}_{j-1}\right]-t\bar{\boldsymbol{\Sigma}}_{t,ik}=0,\ \ \left|\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}-\mathbb{E}\left[\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}|\mathcal{H}_{t-1}\right]\right|\leq 2D^2,$$

We can use the Bernstein-type martingale concentration inequality in Lemma 4 to derive the following bound:

$$\mathbb{P}\left(\left|\frac{1}{t}\sum_{j=1}^{t}\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}-\bar{\boldsymbol{\Sigma}}_{t,ik}\right|\geq z\right)\leq 2\exp\left\{-\frac{cz^2}{D^4/t+2D^2z/t}\right\},$$

where we select $v^2=D^4/t$, and $b=2D^2/t$. This leads to the concentration that with probability at least $1-\epsilon$,

$$\max_{i,k\in[d]}\left|\frac{1}{t}\sum_{j=1}^{t}\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\boldsymbol{x}_{j,k}-\bar{\boldsymbol{\Sigma}}_{t,ik}\right|\leq CD^2\sqrt{\frac{\log(d/\epsilon)}{t}}.$$

It follows from the process in (13) that

$$\begin{aligned}&\mathbb{E}\max_{i\in[d]}\left|\left\langle\left(\widehat{\boldsymbol{\Sigma}}_t-\bar{\boldsymbol{\Sigma}}_t\right)(\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star),\boldsymbol{e}_i\right\rangle\right|^2\\&\leq Cs\frac{D^2}{t}\left(\log(dt)+\log\left(\frac{\bar{\mu}D^2}{\sigma}\right)\right)\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2+C\frac{\sigma^2D^2}{t}\end{aligned}$$

We now proceed to control Part 2 analogously. Invoke Lemma 4 again by selecting $v^2=\sigma^2D^2/t$, and $b=\sigma D/t$. We then have the concentration bound:

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{t}\sum_{j=1}^{t}\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\xi_j\right|\geq z\right)&\leq 2\exp\left\{-\frac{cz^2}{\sigma^2D^2/t+2\sigma Dz/t}\right\}\\&\leq 4\exp\left\{-\frac{ctz^2}{2\sigma^2D^2}\right\}+4\exp\left\{-\frac{ctz}{4\sigma D}\right\},\end{aligned}$$

which gives the tail bound with probability at least $1-\epsilon$:

$$\max_{i\in[d]}\left|\frac{1}{t}\sum_{j=1}^{t}\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\xi_j\right|^2\leq C\sigma D\sqrt{\frac{\log(d/\epsilon)}{t}}.$$

and also the expectation bound for the maxima:

$$\mathbb{E}\max_{i\in[d]}\left|\frac{1}{t}\sum_{j=1}^{t}\mathbb{1}\left\{y_j=a\right\}\boldsymbol{x}_{j,i}\xi_j\right|^2\leq C\frac{\sigma^2D^2\log d}{t}.$$

Combining Part 1 and Part 2 gives us the first claim on the expectation bound:

$$\mathbb{E}\max_{i\in[d]}|\langle\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle|^2\leq C\frac{sD^2\log(dt)}{t}\mathbb{E}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2+C\frac{\sigma^2D^2\log d}{t}.$$

The high probability bound in Part 1 and Part 2 directly leads to the probability bound: with a probability at least $1-\epsilon$, the variation can be controlled by

$$\max_{i\in[d]}|\langle\boldsymbol{g}_t-\nabla f_t(\boldsymbol{\mu}_{t-1}),\boldsymbol{e}_i\rangle|^2\leq CsD^2\frac{\log(d/\epsilon)}{t}\|\boldsymbol{\mu}_{t-1}-\boldsymbol{\mu}_\star\|_2^2+C\frac{\sigma^2D^2\log(d/\epsilon)}{t}$$

□