# Online Learning and Decision-Making under Generalized Linear Model with High-Dimensional Data

Xue Wang* Mike Mingcheng Wei* Tao Yao†

*DAMO Academy, Alibaba Group, xue.w@alibaba-inc.com
*School of Management, University at Buffalo, mcwei@buffalo.edu
†Antai College of Economics and Management, Shanghai Jiao Tong University, taoyao@sjtu.edu.cn

We propose a minimax concave penalized multi-armed bandit algorithm under the generalized linear model (G-MCP-Bandit) for decision-makers facing high-dimensional data in an online learning and decision-making environment. We demonstrate that in the data-rich regime, the G-MCP-Bandit algorithm attains a regret upper bound of $\tilde{\mathcal{O}}(s^2 \log d \log T)$, which attains the optimal cumulative regret in the sample size dimension $T$ and a tight bound in the covariate dimension $d$ and the significant covariate dimension $s$, where $\tilde{\mathcal{O}}(\cdot)$ suppresses the logarithmic dependence on $s$. In the data-poor regime, the G-MCP-Bandit algorithm maintains a tight regret upper bound of $\tilde{\mathcal{O}}(s^2(\log d + \log T)\log T)$. In addition, we develop a local linear approximation method, the 2-step weighted Lasso procedure, to identify the Minimax Concave Penalty (MCP) estimator for the G-MCP-Bandit algorithm under non-i.i.d. samples. Under this procedure, the MCP estimator can match the oracle estimator with high probability and converge to the true parameters at the optimal convergence rate. Finally, through experiments based on both synthetic and real datasets, we show that the G-MCP-bandit algorithm outperforms other benchmarking algorithms in terms of cumulative regret and that the benefits of the G-MCP-Bandit algorithm increase in the data's sparsity level and the size of the decision set.

*Key words*: Multi-armed bandits, minimax concave penalty, high-dimensional data, online learning and decision-making, generalized linear model.

## 1. Introduction

Individual-level data has become increasingly accessible in the internet age, allowing decision-makers in various industries, such as healthcare, retail, and advertising, to accumulate data at an extraordinary speed. User-specific data, including demographics, geographic information, medical records, and search/browsing history, are now widely available. This growing availability of data offers decision-makers unprecedented opportunities to tailor decisions to individual users. For instance, doctors can personalize treatments for patients based on their medical history, clinical tests, and biomarkers; search engines can offer personalized advertisements based on users' queries, demographics, and geographic information.

These user-specific data are often collected sequentially over time, during which decision-makers adaptively learn to predict the expected rewards based on users' responses to each available decision as a function of the user-specific data (i.e., the user's covariates) and optimally adjust decisions to maximize their rewards – an *online* learning and decision-making process, which requires a careful balance between exploration and exploitation. Consider decision-makers who select decisions for incoming users and obtains rewards based on users' responses. To maximize their expected rewards, decision-makers first need an accurate predictive model for users' responses, typically uncertain at the beginning but can be partially learned through collecting samples of users' responses. On the one hand, decision-makers could select a decision that yields the "highest", based on their best knowledge so far, expected reward (i.e., exploitation). Yet, this decision can be suboptimal, as the selection is based on a potentially wrong prediction of users' responses, misled by limited samples. Even worse, decision-makers could incorrectly estimate the expected reward of the true optimal decision to be low and never have a chance to correct such a mistake (as decision-makers will not select the true optimal decision due to the current low reward prediction, they will not generate additional samples to be able to learn and correct their incorrect estimation). On the other hand, decision-makers can improve their predictive ability and learn users' responses by collecting more response samples, which often are obtained through costly random user experiments and/or clinical trials (i.e., exploration). The exploration and exploitation dilemma has been extensively studied in the multi-armed bandit model (Robbins 1952), but the growing dimensionality and data availability have added another layer of complexity.

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

However, the search engine's ability to accurately predict CTR is often crippled by the high-dimensional data with limited samples. Counting more than three-quarters of a million distinct words and their combinations (OxfordDictionaries 2018), there are nearly infinite possible queries the user can submit to the search engine. For example, from 2003 to 2012, Google answered 450

billion unique queries, and it has been estimated that 16% to 20% of queries submitted every day have never been used before (Mitchell 2012). Hence, to accurately estimate a single ad's CTR to these queries, the search engine requires billions, if not trillions, of samples. The craving for samples will be further intensified if the search engine practices personalized advertising by taking users' individual information (such as demographics and geographic information) into consideration. However, the available samples for the search engine to learn and predict CTR are greatly limited. Consider a 45-day marketing campaign promoting a sales event or merchandise, during which time an average ad is expected to reach approximately one-third of a million users (WordStream 2017, Shewan 2017). Among these users, a very small portion can be selected to perform costly experiments to learn CTR, so the size of samples is much smaller compared to the dimension of queries and individual data.

In this paper, we propose the G-MCP-Bandit algorithm for online learning and decision-making problems under high-dimensional settings. Our algorithm follows the ideas of the bandit model and develops a $\epsilon$-decay random sampling method to balance the exploration-and-exploitation trade-off. We allow decision-makers' reward function to follow the generalized linear model (McCullagh and Nelder 1989), which is a large class of models including the linear model, the logistic model, the Poisson regression model, etc., and we develop the MCP estimator, which builds on the Minimax Concave Penalized (MCP) method (Zhang 2010) and solved by a 2-step weighted Lasso (2sWL) procedure, to improve the parameter estimations in high-dimensional settings.

**Main Contribution:**

1. We derive new oracle inequalities for the MCP estimator under non-i.i.d. samples. In particular, we show that the MCP estimator matches the oracle estimator with high probability and converges to the true parameter with the optimal convergence rate. Since the bandit model mixes the exploitation and exploration steps, samples generated under the exploitation steps may be non-i.i.d.. Therefore, we adopt a matrix perturbation technique to derive new oracle inequalities for the MCP estimator under non-i.i.d. samples. To the best of our knowledge, this work is the first one that applies MCP to handle non-i.i.d. samples.
2. We prove that the G-MCP-Bandit algorithm improves the regret upper bound to $\tilde{\mathcal{O}}\left(s^2 \log d \log T\right)$ in the data-rich regime and $\tilde{\mathcal{O}}\left(s^2(\log d + \log T)\log T\right)$ in the data-poor regime. Specifically, we show that in the data-rich regime, where $T$ exceeds a time threshold that depends on the magnitude of the signal for significant covariates, the cumulative regret of the G-MCP-Bandit algorithm over $T$ users is at most $\mathcal{O}(\log T)$, which improves the poly-logarithmic bound from the Lasso-Bandit algorithm under high-dimensional setting (i.e., $\mathcal{O}(s^2(\log d + \log T)^2)$ in Bastani and Bayati 2020) and also is the optimal/lowest theoretical bound for all possible algorithms (Goldenshluger and Zeevi 2013). Further, we show that the

G-MCP-Bandit algorithm also attains a tight bound in the covariate dimension $d$ and the significant covariate dimension $s$, $\tilde{\mathcal{O}}(s^2 \log d)$, in both data-poor and data-rich regimes.

3. Through both synthetic-data-based and real-data-based experiments, we demonstrate that the G-MCP-Bandit algorithm performs favorably to other benchmarking algorithms. Through two synthetic-data-based experiments, we benchmark the G-MCP-Bandit algorithm's performance to other state-of-the-art bandit algorithms designed both in low-dimensional settings (i.e., OLS-Bandit by Goldenshluger and Zeevi 2013 and OFUL by Abbasi-Yadkori et al. 2011) and in high-dimensional settings (i.e., Lasso-Bandit by Bastani and Bayati 2020). We observe that the G-MCP-Bandit algorithm has the lowest cumulative regret. Furthermore, the benefits of the G-MCP-Bandit algorithm over other benchmarking algorithms tend to increase with the data's sparsity level and the size of the decision set. Finally, we evaluate the G-MCP-Bandit algorithm's performance through a real-data-based experiment via the Tencent search advertising dataset, where the technical assumptions specified for the theoretical analysis of the G-MCP-Bandit algorithm's expected cumulative regret may not hold. We observe that the G-MCP-Bandit algorithm continues to perform favorably and that the choice of the underlying reward model can significantly influence the G-MCP-Bandit algorithm's performance. In particular, under the logistic model, which is a special case of the generalized linear model, the G-MCP-Bandit algorithm merely needs 20 user samples to outperform other benchmarking algorithms. This observation suggests that understanding the context of the underlying managerial problem and identifying the appropriate model for the G-MCP-Bandit algorithm can be critical and bring decision-makers substantial revenue improvement.

## 2. Literature Review

This research is closely related to the exploration-exploitation trade-off in the multi-armed bandit literature. Rigollet and Zeevi (2010), Slivkins (2014) follow the non-parametric approach and consider that the arm reward can be any smooth non-parametric function. Under this approach, the expected cumulative regret has an exponential dependence on the covariate dimension $d$, which is undesirable under high-dimensional settings where $d$ can be extremely large. Such exponential dependence can be improved by following the parametric approach. Auer (2002) proposes the UCB algorithm for a linear bandit model, where the arm reward can be approximated by linear combinations of covariates. Since Auer (2002), other UCB-type algorithms (e.g., Dani et al. 2008, Rusmevichientong and Tsitsiklis 2010, Abbasi-Yadkori and Szepesvari 2012, Deshpande and Montanari 2012) and Bayesian-type algorithms (e.g., Agrawal and Goyal 2013, Russo and Van Roy 2014) have been proposed and shown to improve on the expected cumulative regret. Yet, allowing the adversary and without regulating the sample generating process, the statistical performance

of the parameter vector estimation in the learning process may suffer. As a result, the expected cumulative regret bound typically has a sublinear dependence on the sample size dimension $T$ (e.g., $\mathcal{O}(\sqrt{T})$) and a polynomial dependence on the covariate dimension $d$. However, in high-dimensional settings, where the covariate dimension and the sample size dimension can be exceedingly large, these algorithms can perform poorly.

By introducing a forced sampling approach to the linear K-armed bandit model, Goldenshluger and Zeevi (2013) ensure that enough i.i.d. samples are generated in their algorithm and show that their proposed OLS-Bandit algorithm can achieve $\mathcal{O}(\log T)$ dependence on the sample size dimension $T$ in low-dimensional settings. Following a similar approach, Bastani and Bayati (2020) consider the high-dimensional setting and adopt the Lasso method to explore the sparsity structure in estimation. They propose the Lasso-Bandit algorithm, which attains a poly-logarithmic dependence on the sample size dimension $\mathcal{O}(\log^2 T)$ and the covariate dimension $\mathcal{O}(\log^2 d)$ in high-dimensional settings. In this paper, we allow the reward function to follow the generalized linear model, which contains a wide family of models that includes the linear K-armed bandit model in Goldenshluger and Zeevi (2013), Bastani and Bayati (2020). We develop a 2sWL procedure to identify the unbiased MCP estimator and propose a $\epsilon$-decay random sampling method to hurdle the high-dimensional data challenge. We show that in the data-rich regime, our proposed G-MCP-Bandit algorithm achieves the optimal cumulative regret bound on the sample size dimension $\mathcal{O}(\log T)$ and attains a tight bound in the covariate dimension $\mathcal{O}(\log d)$ in high-dimensional settings. Recently, there has been a growing interest in the sparse linear bandit model. By adopting sparse regularization, Bastani and Bayati (2020), Kim and Paik (2019), Ren and Zhou (2020), Wang et al. (2020), Hao et al. (2020), Ariu et al. (2020) establish ploy-logarithmic dependence bounds on $d$. Furthermore, conditioning on the minimum signal strength in the data-rich regime, Hao et al. (2020), Ariu et al. (2020) also prove nearly optimal regret. Different from these two papers, we consider the generalized linear model and prove that the G-MCP-Bandit algorithm attains $\mathcal{O}(\log T)$ bound for the data-rich regime. Instead of adopting a greedy algorithm, Wang et al. (2020) propose a UCB-type algorithm for better numerical performance and use the best subset selection, which requires a time-consuming combinatorial optimization procedure, to debias and show that their algorithm reaches $\tilde{O}(\text{poly-log}(d)\sqrt{T})$. By applying the matrix sketching techniques (e.g., random projection and frequent directions), Yu et al. (2017), Carpentier and Munos (2012), Kuzborskij et al. (2018) also break the polynomial dependence on $d$ but may lead to a linear regret on $T$ due to the sketching distortion.

Our research is also related to the regret analysis for bandit problems that go beyond the classical linear model framework. Filippi et al. (2010) analyze the K-armed bandit problem under the generalized linear model framework in low-dimensional settings. They point out that the confidence

regions under the generalized linear model pose more complicated geometry in the parameter space than simple ellipsoids and highlight the technical difficulties in generalizing the linear bandit model to the generalized linear framework. They propose a UCB-based bandit algorithm, GLM-UCB, which attains a regret upper bound of $\mathcal{O}(d\sqrt{T})$. Li et al. (2017) further propose a UCB-GLM algorithm and SupCB-GLM algorithm in low-dimensional settings, where the second algorithm improves the regret bound to $\mathcal{O}(\sqrt{dT})$. Compared to this stream of literature that uses the generalized linear model framework, our paper considers the high-dimensional settings and attains the $\mathcal{O}(\log d \log T)$ bound. In addition, some results of EXP-type algorithms can also be applied to the generalized linear model. Yet, these algorithms (e.g., Auer et al. 2002, Beygelzimer et al. 2011, Agarwal et al. 2014) typically obtain an $\mathcal{O}(\sqrt{dT})$ bound and can be expensive to run.

Our research is also connected to the statistical learning literature. In high-dimensional statistics, Lasso type methods (Tibshirani 1996) have become the golden standard for high-dimensional learning (Meinshausen et al. 2006, 2009, Zhang et al. 2008, Van de Geer et al. 2008). Yet, Lasso-type regularizations may lead to estimation bias, and strong conditions are needed for analyzing its theoretical performance guarantee (Fan et al. 2014a). Recently, Zhang (2010) proposed Minimax Concave Penalty (MCP), a folded concave penalty function, which entails better statistical properties, such as the unbiasedness and a strong oracle property for high-dimensional sparse estimation, and requires weaker conditions than Lasso (Zou 2006, Fan et al. 2014b, Meinshausen et al. 2006). Although it is statistically favorable to adopt MCP, solving the MCP estimator (an NP-complete problem) could be computationally challenging (Liu et al. 2017, 2016). Various approximation methods have been developed in the literature. For example, Fan and Li (2001) use the local quadratic approximation, Fan et al. (2014b, 2018), Zou (2006), Zhao et al. (2014) adopt the local linear approximation, Zhang (2010) choose the path following algorithm, and Liu et al. (2017) propose the second-order approximation. Liu et al. (2022) further extend the second-order approximation to neural network settings. Our proposed solution procedure (the 2sWL procedure) is analogous to the local linear approximation and guarantees that the solution has desirable statistical properties for theoretical analysis and can be efficiently solved. In the literature, the theoretical analysis of MCP's statistical properties relies on the assumption that all samples are i.i.d., which is hardly the case under bandit models. This paper also contributes to the statistical learning literature by deriving new oracle inequalities for MCP under non-i.i.d. samples.

## 3. Model Settings

We consider a sequential stochastic arrival process for $t \in \{1, 2, ..., T\}$. At each time step $t$, a single user, described by a high-dimensional feature covariate vector $\boldsymbol{X}_t \in \mathbb{R}^d$ where $d$ is the number of features, arrives, and the covariate vector is observable to decision-makers. The covariate vector

combines all available (but not necessarily valuable for decision-makers to base their decision on) user-specific data, such as demographics, geographic information, browsing/shopping history, and medical records. Users' covariate vectors $\{\boldsymbol{X}_t\}_{t\geq 1}$ are i.i.d. distributed according to an unknown distribution $\mathcal{P}_{\boldsymbol{X}}$.

Based on the user's covariate vector $\boldsymbol{X}_t$, decision-makers will select a decision from a decision set $\mathcal{K}=\{1,2,...,K\}$, where $K\geq 2$, to maximize their expected reward. The user will respond to the chosen decision $k\in\mathcal{K}$, and such a response will generate a reward for decision-makers. Take search advertising, for example. The search engine can recommend one of $K$ different ads to the user; the user can respond to the recommended ad by clicking, which generates revenue for the search engine. We denote this reward under the chosen decision $k$ at time $t$ as $R_{k,t}\in\mathbb{R}$, which follows the generalized linear model (McCullagh and Nelder 1989):

$$R_{k,t}=\mu(\boldsymbol{X}_t^\top\boldsymbol{\beta}_k^{\text{true}})+\epsilon_t, \tag{1}$$

where $\boldsymbol{X}_t$ is the user's covariate vector at time $t$, $\boldsymbol{\beta}_k^{\text{true}}\in\mathbb{R}^d$ is the unknown time-independent (i.e., $\boldsymbol{\beta}_k^{\text{true}}$ is independent on $t$) parameter vector corresponding to decision $k\in\mathcal{K}$, $\epsilon_t\in\mathbb{R}$ is an independent sub-gaussian random variable, and $\mu:\mathbb{R}\to\mathbb{R}$ is a link function.

The generalized linear model covers a large class of models, including the linear model and the logistic model. For example, by setting the link function $\mu(\boldsymbol{X}^\top\boldsymbol{\beta})=\boldsymbol{X}^\top\boldsymbol{\beta}$, we have the classic linear multi-armed bandit model, which has been extensively studied by Dani et al. (2008) and Goldenshluger and Zeevi (2013), among others, under low-dimensional settings and by Bastani and Bayati (2020) under high-dimensional settings. Furthermore, the generalized linear model facilitates us to go beyond the classic linear bandit model, as the reward may take a *nonlinear* form in practice. For instance, the search engine collects revenue only when a user has clicked the recommended ad; otherwise, the search engine earns nothing – a nonlinear logistic model by nature and belongs to the class of the generalized linear model. It is also worth noting that the generalized linear model facilitates a separation between the link function and the sub-gaussian random variable. Therefore, all our analysis and results do not require any detailed knowledge of the reward density function (similar to Bastani and Bayati 2020) and merely need to access the link function (e.g., $\mu(\boldsymbol{X}^\top\boldsymbol{\beta})=\boldsymbol{X}^\top\boldsymbol{\beta}$ for the linear bandit model and $\mu(\boldsymbol{X}^\top\boldsymbol{\beta})=(1+\exp(-\boldsymbol{X}^\top\boldsymbol{\beta}))^{-1}$ for the logistic model).

The parameter vector $\boldsymbol{\beta}_k^{\text{true}}$ is high-dimensional with latent sparse structure, and we denote $\mathcal{S}^k=\{j:\beta_{k,j}^{\text{true}}\neq 0\}$ as the index set for significant covariates of decision $k$, where $\beta_{k,j}^{\text{true}}$ denotes the $j$-th element in $\boldsymbol{\beta}_k^{\text{true}}$. Note that $\mathcal{S}^k$ contains non-zero coefficient parameters and therefore is important for decision-makers to predict the user's response. This index set is also unknown to

decision-makers. We denote $s = \max_{k\in\mathcal{K}} |\mathcal{S}^k|$, where $|\mathcal{S}^k|$ is the cardinality of $\mathcal{S}^k$ (i.e., the number of significant covariates), and is typically much smaller than the dimension of the covariate vector.

The decision-makers' objective is to maximize their expected cumulative reward. Denote decision-makers' current policy as $\boldsymbol{\pi} = \{\pi_t\}_{t\geq 1}$, where $\pi_t \in \mathcal{K}$ is the decision prescribed by policy $\boldsymbol{\pi}$ at time $t$. To benchmark the performance of policy $\boldsymbol{\pi}$, we first introduce an *oracle policy* $\boldsymbol{\pi}^* = \{\pi_t^*\}_{t\geq 1}$ under which decision-makers know the values of the true parameter vector $\boldsymbol{\beta}_k^{\text{true}}$ for all $k \in \mathcal{K}$ and chooses the best decision to maximize their expected reward for all $t \geq 1$:

$$\pi_t^* \doteq \arg\max_{k\in\mathcal{K}} \left\{\mathbb{E}_{\epsilon_t}[R_{k,t}|\boldsymbol{X}_t^\top \boldsymbol{\beta}_k^{\text{true}}]\right\}, \tag{2}$$

where $\mathbb{E}_{\epsilon_t}[R_{k,t}|\boldsymbol{X}_t^\top \boldsymbol{\beta}_k^{\text{true}}] = \mu(\boldsymbol{X}_t^\top \boldsymbol{\beta}_k^{\text{true}})$, following the definition in (1), and the $\pi_t^*$ is the optimal decision that maximizes the expected reward given the true parameter vectors $\boldsymbol{\beta}_k^{\text{true}}$ for all $k \in \mathcal{K}$ and the covariate vector for the $t$-th user $\boldsymbol{X}_t$.

Note that in practice, the parameter vector $\boldsymbol{\beta}_k^{\text{true}}$, for $k \in \mathcal{K}$, is unknown to decision-makers, and therefore the construction and definition of the oracle policy directly imply that decision-makers' reward under policy $\boldsymbol{\pi}$ is upper-bounded by that of the oracle policy. We, therefore, define decision-makers' expected cumulative regret up to time $T$ under the policy $\boldsymbol{\pi}$, $R^C(T)$, as follows:

$$R^C(T) \doteq \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{X}_t,\epsilon_t}[R_{\pi_t^*,t} - R_{\pi_t,t}],$$

where $R_{\pi_t,t}$ and $R_{\pi_t^*,t}$ are the rewards at time $t$ under policy $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$, respectively. The expected cumulative regret is defined as the expected cumulative reward difference between the optimal policy $\boldsymbol{\pi}^*$ and decision-makers' alternative policy $\boldsymbol{\pi}$. To maximize their expected cumulative reward, decision-makers are equivalent to exploring the policy $\boldsymbol{\pi}$ that minimizes the cumulative regret up to time $T$.

As the true model parameter vector $\boldsymbol{\beta}_k^{\text{true}}$, for $k \in \mathcal{K}$, is unknown to decision-makers, we will use the maximum likelihood estimation (MLE) to learn the model parameters. Given observed $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ and corresponding rewards $R_1, ..., R_n$, we define the negative log-likelihood loss function $\mathcal{L}(\boldsymbol{\beta})$ as follows:

$$\mathcal{L}(\boldsymbol{\beta}) \doteq \frac{1}{n}\sum_{i=1}^{n} f(R_i|\boldsymbol{X}_i^\top \boldsymbol{\beta}),$$

where $f(R_i|\boldsymbol{X}_i^\top \boldsymbol{\beta})$ is the sample-wise loss function, $n$ is the sample size. For example, given the observed covariate vector $\boldsymbol{x}_i$ and the corresponding reward $r_i$, $f(r_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{x}_i^\top \boldsymbol{\beta} - r_i\|^2$ for the linear bandit model and $f(r_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}) = -\mathbb{1}(r_i = 1)\log(\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))) - \mathbb{1}(r_i = 0)\log(1/(1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})))$ for the logistic model with binary $r_i \in \{0, 1\}$.

Before presenting the proposed G-MCP-Bandit algorithm, we will first state five technical assumptions necessary for the theoretical analysis of decision-makers' expected cumulative regret. The first three assumptions are adopted directly from the multi-armed bandit literature, and the last two assumptions are from the high-dimensional statistics literature. Note that both $\epsilon_t$ and $\boldsymbol{X}_t$ are i.i.d. with respect to $t$, so we omit the subscript $t$ in $\mathbb{E}_{\epsilon_t}$ and $\mathbb{E}_{\boldsymbol{X}_t}$ hereafter for brevity.

**A. 1** (Parameter set) There exist positive constants $x_{\max}$, $R_{\max}$, and $b$ such that for any $t \geq 1$ and $k \in \mathcal{K}$, we have $\|\boldsymbol{x}_t\|_\infty \leq x_{\max}$, $\|\boldsymbol{\beta}\|_\infty \leq b$ for all feasible $\boldsymbol{\beta}$, and $\mathbb{E}_\epsilon[R_{k,t}|\boldsymbol{x}_t^\top \boldsymbol{\beta}_k^{\text{true}}] = \mu(\boldsymbol{x}_t^\top \boldsymbol{\beta}_k^{\text{true}}) \in (0, R_{\max}]$ for all realization $\boldsymbol{x}_t$ of $\boldsymbol{X}_t$.

The first assumption is a standard assumption in the bandit literature (Rusmevichientong and Tsitsiklis 2010) and ensures that the covariate vector, the estimated/true coefficient vector, and the expected reward are bounded so that the maximum regret at every time step will also be bounded to avoid trivial decisions. Most real-world applications, including the real-data experiment in §6.2, satisfy this assumption.

**A. 2** (Margin condition) There exists a $C > 0$ such that $\mathbb{P}(|\mathbb{E}_\epsilon[R_{i,t}|\boldsymbol{X}_t^\top \boldsymbol{\beta}_i^{\text{true}}] - \mathbb{E}_\epsilon[R_{j,t}|\boldsymbol{X}_t^\top \boldsymbol{\beta}_j^{\text{true}}]| \leq \gamma) \leq C R_{\max} \gamma$ for all $\gamma > 0$, $t \geq 1$, $i \neq j$, and $i, j \in \mathcal{K}$.

The second assumption is first introduced in the classification literature by Tsybakov et al. (2004). Goldenshluger and Zeevi (2013) and Bastani and Bayati (2020) adopt this assumption to the linear bandit model, under which the Margin condition ensures only a fraction of covariates can be drawn near the boundary hyperplane $\boldsymbol{X}_t^\top(\boldsymbol{\beta}_i^{\text{true}} - \boldsymbol{\beta}_j^{\text{true}}) = 0$ in which rewards for both arms are nearly equal for all $i \neq j$. Clearly, if a large proportion of covariates are drawn from the vicinity of the boundary hyperplane, then for any bandit algorithm, a small estimation error in the decision parameter vectors may lead decision-makers to choose the suboptimal decision and perform poorly (Bastani and Bayati 2020). Therefore, this margin condition ensures that given a user's covariate vector, decisions can be properly separated and ordered based on rewards.

**A. 3** (Arm optimality) There exists a partition $\mathcal{K}_s$ and $\mathcal{K}_o$ for decision set $\mathcal{K}$ such that for some $h > 0$, (a) if $k_1 \in \mathcal{K}_s$, then $\mathbb{E}_\epsilon[R_{k_1,t}|\boldsymbol{x}_t^\top \boldsymbol{\beta}_{k_1}^{\text{true}}] + h < \max_{k \neq k_1, k \in \mathcal{K}} \mathbb{E}_\epsilon[R_{k,t}|\boldsymbol{x}_t^\top \boldsymbol{\beta}_k^{\text{true}}]$ for all realization $\boldsymbol{x}_t$ of $\boldsymbol{X}_t$, $t \geq 1$; and (b) if $k_2 \in \mathcal{K}_o$, then there exists a positive constant $p^*$ such that for $t \geq 1$, $\min_{k_2 \in \mathcal{K}_o} \mathbb{P}(\boldsymbol{X}_t \in U_{k_2}) \geq p^*$, where $U_{k_2} \doteq \{\boldsymbol{x} : \mathbb{E}_\epsilon[R_{k_2,t}|\boldsymbol{x}^\top \boldsymbol{\beta}_{k_2}^{\text{true}}] > \max_{k \neq k_2, k \in \mathcal{K}} \mathbb{E}_\epsilon[R_{k,t}|\boldsymbol{x}^\top \boldsymbol{\beta}_k^{\text{true}}] + h\}$.

The arm optimality condition (Goldenshluger and Zeevi 2013, Bastani and Bayati 2020) ensures that as the sample size increases, the parameter vectors for optimal decisions can eventually be learned. In particular, this condition separates decisions into an optimal decision subset $\mathcal{K}_o$ and a suboptimal decision subset $\mathcal{K}_s$. Decision $i$ in $\mathcal{K}_o$ is strictly optimal for some users' covariate vectors (denoted by set $U_i$); yet, decision $j$ in $\mathcal{K}_s$ must be strictly suboptimal for all users' covariate vectors. Therefore, even if there is a small estimation error for decision $i$ in $\mathcal{K}_o$, decision-makers will be more likely to choose decision $i$ for a user with a covariate vector drawn from the set $U_i$.

These first three assumptions are directly adopted from the multi-armed bandit literature and have been shown to be satisfied for all discrete distributions with finite support and a very large class of continuous distributions (see Bastani and Bayati 2020 for detailed examples and discussions).

**A. 4** (Sample-wise loss function) Let $r \in [0, R_{\max}]$ and $|y| \leq x_{\max} b$. We assume (i) $f(r|y)$ is convex and has smooth gradient in $y$, and (ii) there exists positive constants $\sigma$ and $\sigma_2$ such that $|f_y^{'}(r|y)| \leq \sigma$ and $f_{yy}^{''}(r|y) \leq \sigma_2$, where $f_y^{'}(r|y)$ and $f_{yy}^{''}(r|y)$ are the first and second order partial derivatives of $f(r|y)$ with respect to the second augment $y$. Moreover, $f_y^{'}(R_{k,t}|\boldsymbol{x}_t^\top \boldsymbol{\beta}_k^{\text{true}})$ is a zero mean $\sigma^2$-sub-gaussian random variable for all $k \in \mathcal{K}$, $t \geq 1$, and any realization $\boldsymbol{x}_t$ of $\boldsymbol{X}_t$.

The sample-wise loss function assumption enables us to use the estimated parameters to statistically infer the true parameters. It is a fairly weak technical assumption and shares the same spirit as the log-concavity assumption widely discussed in the literature (Bagnoli and Bergstrom 2005, Boyd et al. 2004). For example, in the linear bandit model, $f_y^{'}(r|y) = y - r$ and $f_{yy}^{''}(r|y) = 1$; in the logistic model, $f_y^{'}(r|y) = -\mathbb{1}(r = 1) + \exp(y)/(1 + \exp(y))$ and $f_{yy}^{''}(r|y) = \exp(y)/(1 + \exp(y))^2$.

**A. 5** (Restricted eigenvalue condition) There exists a $\kappa > 0$ such that for all feasible $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}\|_1 \leq b$ and $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}_{(\mathcal{S}^k)^c}\|_1 \leq 3\|\boldsymbol{u}_{\mathcal{S}^k}\|_1$, we have $\frac{\kappa}{s}\|\boldsymbol{u}_{\mathcal{S}^k}\|_1^2 \leq \boldsymbol{u}^T \mathbb{E}_{\epsilon,\boldsymbol{X}}[f_{yy}^{''}(R|\boldsymbol{X}^\top\boldsymbol{\beta})\boldsymbol{X}\boldsymbol{X}^\top]\boldsymbol{u}$ for $k \in \mathcal{K}_s$ and $\frac{\kappa}{s}\|\boldsymbol{u}_{\mathcal{S}^k}\|_1^2 \leq \boldsymbol{u}^T \mathbb{E}_{\epsilon,\boldsymbol{X}|\boldsymbol{X}\in U_k}[f_{yy}^{''}(R|\boldsymbol{X}^\top\boldsymbol{\beta})\boldsymbol{X}\boldsymbol{X}^\top]\boldsymbol{u}$ for $k \in \mathcal{K}_o$.

The restricted eigenvalue condition assumption is a standard assumption in high-dimensional statistics and is necessary for the identifiability and consistency of high-dimensional estimators (Fan et al. 2018, 2014b). This assumption considers the local geometry of the standard loss function for the generalized linear model (e.g., Negahban et al. 2009, Li et al. 2017, Oh et al. 2021) with i.i.d. samples in $U_k$. To intuit, note that under low-dimensional settings, the literature (Montgomery et al. 2012) requires that $\mathcal{L}(\boldsymbol{\beta})$ is strongly convex around the true parameter vector $\boldsymbol{\beta}^{\text{true}}$ (e.g., the Hessian matrix in OLS estimator is positive-definite and invertible) in order to achieve identifiability of the parameter vector. However, the strong convexity assumption is typically violated in high-dimensional settings, as the sample size can be much smaller than the covariate dimension. Therefore, a weaker condition is adopted: The $\mathcal{L}(\boldsymbol{\beta})$ exhibits local strongly convex behavior only in some restricted subspace of $\boldsymbol{u}$. In high-dimensional linear models, the restricted eigenvalue condition assumption is analogous to the compatibility condition (Bastani and Bayati 2020, Bühlmann and Van De Geer 2011), restrict strongly convexity condition (Negahban et al. 2009, Loh and Wainwright 2013), and sparse eigenvalue condition (Zhang et al. 2012, Fan et al. 2018).

It is worth-noting that most common sub-Gaussian distributions satisfies the restricted eigenvalue condition assumption. Specifically, in the linear model case, where $f_{yy}^{''}(R|\boldsymbol{X}^\top\boldsymbol{\beta}) = 1$, the restricted eigenvalue condition reduces to Assumption 4 in Bastani and Bayati (2020) so that common distributions (e.g., Bernoulli distribution, uniform distribution, truncated Gaussian distribution, etc.) or discrete distribution with finite support will satisfy this assumption (see the end

of §3 in Bastani and Bayati 2020 for detailed discussions). When going beyond the linear model, one sufficient condition, to ensure that the restricted eigenvalue condition continues to hold under the distributions mentioned earlier, is that the second derivative of the sample-wise loss function is positive (i.e., $f''_{yy}(R|\boldsymbol{X}^\top\boldsymbol{\beta}) > 0$). One example that satisfies this sufficient condition is the logistic regression where we have $f''_{yy}(R|\boldsymbol{X}^\top\boldsymbol{\beta}) = \frac{\exp(-\boldsymbol{X}^\top\boldsymbol{\beta})}{(1+\exp(-\boldsymbol{X}^\top\boldsymbol{\beta}))^2} \geq \frac{\exp(x_{\max}b)}{(1+\exp(x_{\max}b)^2} > 0$.

Finally, we will follow Bastani and Bayati (2020) to present specific examples that satisfy all five assumptions. In the first example, we start with a modified version of the "Discrete Covariates" example in Bastani and Bayati (2020): Let the underlying true parameter vectors for covariates to be arbitrarily set to be $\boldsymbol{\beta}_1^{\text{true}} = (1, 0, 0, 0, ...)$, $\boldsymbol{\beta}_2^{\text{true}} = (0, 1, 0, 0, ...)$, and $\boldsymbol{\beta}_3^{\text{true}}(1/4, 1/4, 0, 0, ...)$; for each incoming user, we randomly draw a covariate vector from the $d$-dimensional unit cube $[0, 1]^d$; rewards of arm $i$ are sampled from Bernoulli distributions with success probability $\frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta}_i^{\text{true}})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta}_i^{\text{true}})}$ for user $X$. As discussed by the end of §3 in Bastani and Bayati (2020), Assumptions **A.1** - **A.3** are satisfied; as the logistic regression function (associated with the MLE of Bernoulli distribution) is strongly convex in the bounded domain, Assumption **A.4** is satisfied; based on the previous analysis of the restricted eigenvalue conditions, Assumption **A.5** is also satisfied. In the second example, following the "Generic Example" in Bastani and Bayati (2020), we describe the corresponding generic example that satisfies all assumptions in our paper: The problem is in a bounded domain, and both continuous and discrete values are allowed (Assumption **A.1**) with strongly convex loss functions (Assumption **A.4**); for a given user's covariate, the rewards of different arms are likely to be properly separated (Assumption **A.2**); each arm is either optimal for some users or strictly suboptimal for all users (Assumption **A.3**). Therefore, in practical applications, the aforementioned distributions (e.g., Bernoulli, uniform, and truncated Gaussian distributions) together with commonly used loss functions (e.g., the least squares and the logistic regressions) will satisfy all five assumptions.

## 4. G-MCP-Bandit Algorithm

One of the major challenges for online learning and decision-making problems is discovering the underlying sparse data structure and estimating the parameter vector for high-dimensional data with limited samples. Lasso (Tibshirani 1996) has been proposed as an efficient statistical learning method and adopted in the multi-armed bandit literature (Bastani and Bayati 2020) to hurdle this challenge. However, the standard single-step Lasso estimator can be biased and performs inadequately, especially when the magnitude of true parameters is not too small (Fan and Li 2001). One way to address this bias issue is to use multi-step variants of the Lasso (e.g., adaptive Lasso in Zou 2006 or relaxed Lasso in Meinshausen 2007), and in this paper, we propose a new multi-step Lasso-based method that builds on the novel convex MCP penalty function (Zhang 2010) and solved by a 2-step weighted Lasso procedure.

### 4.1. Parameter Vector Estimation

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

$$\boldsymbol{\beta}^{\text{oracle}}(\bar{\boldsymbol{X}}, \boldsymbol{R}) \doteq \arg \min_{\substack{\boldsymbol{\beta}_{\mathcal{S}^c}=0 \\ \boldsymbol{\beta}_{\mathcal{S}}}} \mathcal{L}(\boldsymbol{\beta}), \tag{3}$$

where $\bar{\boldsymbol{X}} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n] \in \mathbb{R}^{d \times n}$ is the matrix contains $n$ row user covariates and $\boldsymbol{R} = [R_1, R_2, ..., R_n] \in \mathbb{R}^n$ stores the corresponding rewards.

When solving for the oracle estimator, decision-makers can ignore all insignificant covariates by forcing their corresponding coefficients to be zero and essentially reduce the high-dimensional problem to a low-dimensional counterpart. In the classical statistical analysis, the literature primarily focuses on analyzing the statistical behavior of the problem (3) when all samples are i.i.d. drawn from a given distribution. In online learning and decision-making settings, however, future samples could depend on historical data through the parameter estimation of the decision model, whose process suggests that a large portion of the samples could be non-i.i.d.. Hence, in this paper, we will use $\mathcal{A}$ to denote the sample set that contains only i.i.d. samples out of the whole sample set. Further, we use $n$ and $|\mathcal{A}|$ to represent the sample size of whole samples and the sample size of i.i.d. samples, respectively. Clearly, $n \geq |\mathcal{A}|$. Now, we present the result for the oracle estimator with partial i.i.d. samples in the following lemma.

LEMMA 1. *Let $n$ be the size of the whole samples, $\boldsymbol{\beta}_k^{\text{true}}$ be the underlying true parameters, and $|\mathcal{A}|$ be the size of i.i.d. samples with $\boldsymbol{X} \in \mathbf{R}^d$ for $k \in \mathcal{K}_s$ and $\boldsymbol{X} \in U_k$ for $k \in \mathcal{K}_o$. Under assumptions A.1, A.4, and A.5, when $|\mathcal{A}| \geq C_1^{-1} \log s$, the following inequality for the oracle estimator holds for any $\zeta > 0$*

$$\mathbb{P}\left( \|\boldsymbol{\beta}^{\text{oracle}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \frac{2ns\zeta}{|\mathcal{A}|\kappa} \right) \geq 1 - \delta_1(n, |\mathcal{A}|, \zeta), \tag{4}$$

*where*

$$\delta_1(n, |\mathcal{A}|, \zeta) \doteq 2s \exp\left( -\frac{n\zeta^2}{2\sigma^2 x_{\max}^2} \right) + \exp(-C_1|\mathcal{A}|) \tag{5}$$

*and $C_1 = \mathcal{O}(s^{-2})$, where the detailed expression of $C_1$ is given in* (EC.125) *in the Appendix.*

Lemma 1 considers the scenario in which the sample set is mixed with i.i.d. samples and non-i.i.d. samples. If we set $\zeta = \mathcal{O}(\sqrt{1/n})$, then the convergence rate for the oracle estimator is on the order

$\mathcal{O}(s\sqrt{n/|\mathcal{A}|^2})$. Hence, when the number of i.i.d. samples is on the same order as the total sample size (i.e., $|\mathcal{A}| = \mathcal{O}(n)$), the oracle estimator will match the optimal convergence rate of $\mathcal{O}(s\sqrt{1/n})$ commonly stated in the literature (Fan et al. 2018, Zhao et al. 2018).

However, the significant covariates index set $\mathcal{S}$ is typically unknown to decision-makers in practice, so we are not able to directly apply Eq. (3) to obtain the oracle estimator. In this research, we propose to use the MCP penalty (Zhang 2010) to recover this latent sparse structure and estimate the unknown parameter vector. To better understand the rationale behind the MCP penalty, we start with the following weighted Lasso estimator:

$$\boldsymbol{\beta}^{\mathrm{W}}(\bar{\boldsymbol{X}}, \boldsymbol{R}, \boldsymbol{w}) \doteq \arg\min_{\boldsymbol{\beta}} \left\{ \mathcal{L}(\boldsymbol{\beta}) + \sum_{i=1}^{d} w_i |\beta_i| \right\}, \tag{6}$$

where $\boldsymbol{w} = (w_1, w_2, ..., w_d)$ is a non-negative weights vector chosen by decision-makers. Note that when we set $w_i = \lambda$ for all $i$, $\boldsymbol{\beta}^{\mathrm{W}}(\bar{\boldsymbol{X}}, \boldsymbol{R}, \boldsymbol{w})$ reduces to the Lasso estimator, which can be biased when the magnitude of true parameters is not too small. To recover the sparse structure and provide an unbiased parameter estimator, an ideal way to select $\{w_i\}$ is to set $w_i = \lambda > 0$ for all $i \in \mathcal{S}^{\mathrm{c}}$ and $w_j = 0$ for all $j \in \mathcal{S}$. By doing so, when the weight $\lambda$ is large enough, the weighted Lasso estimator converges to the oracle estimator $\boldsymbol{\beta}^{\mathrm{oracle}}(\bar{\boldsymbol{X}}, \boldsymbol{R})$. The benefits of the weighted Lasso method have attracted considerable attention recently, and various mechanisms have been proposed in the literature aiming to improve the weight selection process (Zou 2006, Huang et al. 2008, Candes et al. 2008). The MCP method, adopted in our paper, reflects such a process.

In particular, we define the following MCP penalty function:

$$P_{\lambda,a}(x) \doteq \int_0^{|x|} \max\left(0, \lambda - \frac{1}{a}t\right) dt, \tag{7}$$

where $a$ and $\lambda$ are positive parameters selected by decision-makers. The MCP estimator can be presented as follows:

$$\boldsymbol{\beta}^{\mathrm{MCP}}(\bar{\boldsymbol{X}}, \boldsymbol{R}, \lambda, a) \doteq \arg\min_{\boldsymbol{\beta}} \left\{ \mathcal{L}(\boldsymbol{\beta}) + \sum_{i=1}^{d} P_{\lambda,a}(\beta_i) \right\}. \tag{8}$$

Denote the index set for non-zero coefficients solutions in Equation (8) as $\mathcal{J} \doteq \{j : \beta_j^{\mathrm{MCP}} \neq 0\}$. If we have $|\beta_j^{\mathrm{MCP}}| \geq a\lambda$ for all $j \in \mathcal{J}$, then based on the definition of $P_{\lambda,a}(x)$, we can verify that $P_{\lambda,a}(\beta_j^{\mathrm{MCP}}) = \frac{1}{2}a\lambda^2$ for $j \in \mathcal{J}$. Similarly, for all $j \notin \mathcal{J}$, we have $P_{\lambda,a}(\beta_j^{\mathrm{MCP}}) = 0$. In other words, the statistical performance of solving the MCP estimator is equivalent to solving the following problem: $\min_{\boldsymbol{\beta}_{\mathcal{J}^c}=0, \boldsymbol{\beta}_{\mathcal{J}}} \mathcal{L}(\boldsymbol{\beta})$. Hence, if $\mathcal{J} = \mathcal{S}$, then the MCP estimator converges to the oracle estimator.

Solving the MCP estimator can be challenging. Liu et al. (2017) have shown that it is an NP-complete problem to find the MCP estimator by globally solving Equation (8). In the next subsection, we propose a local linear approximation method, the 2-step Weighted Lasso (2sWL) procedure, to tackle this challenge and demonstrate that the estimator solved by the 2sWL procedure will match the oracle estimator $\boldsymbol{\beta}^{\mathrm{oracle}}$ with high probability.

### 4.2. 2-Step Weighted Lasso Procedure

The 2sWL procedure consists of two steps. We first solve a Lasso problem by setting all positive weights in Equation (6) to a given parameter $\lambda$. Then, we use the Lasso estimator obtained in the first step to update the weights vector $\boldsymbol{w}$ by taking the first-order derivatives of the MCP penalty function, and by applying this updated weight vector, we re-solve the weighted Lasso problem in Equation (6) to obtain the MCP estimator. Let $\mathbf{1}$ be the vector filled with 1, and the 2sWL procedure at time $t$ can be outlined as follows:

**2-Step Weighted Lasso (2sWL) Procedure**:

**Require**: input parameters $a$, $\lambda$ and dataset $\{\bar{\boldsymbol{X}}, \boldsymbol{R}\}$

**Step 1**: solve the standard Lasso problem

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}^{\mathrm{W}}(\bar{\boldsymbol{X}}, \boldsymbol{R}, \lambda \mathbf{1})$$

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

and solve the weighted Lasso Problem

$$\boldsymbol{\beta}^{\mathrm{MCP}} = \boldsymbol{\beta}^{\mathrm{W}}(\bar{\boldsymbol{X}}, \boldsymbol{R}, \boldsymbol{w})$$

Next, we will use the following proposition to show that the MCP estimator identified by the 2sWL procedure can recover the oracle estimator with high probability. We denote a new index set

$$\mathcal{S}_1 \doteq \left\{ i : |\beta_i^{\mathrm{true}}| \geq \left( \frac{24ns}{|\mathcal{A}|\kappa} + a \right) \lambda,\ i \in \mathcal{S} \right\}, \tag{9}$$

which is a subset of the index set for significant covariates $\mathcal{S}$. To simplify the notation and presentation, let's consider a special case where all samples in the whole sample set are i.i.d. (i.e., $n = |\mathcal{A}|$) and postpone the proof of the general case where $n \geq |\mathcal{A}|$ to Proposition 3 in §5.1.

Proposition 1. *Under assumptions A.1, A.4, and A.5, when $n = |\mathcal{A}|$, if $|\mathcal{A}| > C_1^{-1} \log d$ and $a > \frac{48s}{\kappa}$, then for $\zeta > 0$, the MCP estimator solved by the 2sWL procedure $\boldsymbol{\beta}^{\mathrm{MCP}}$ satisfies the following inequality*

$$\mathbb{P}\left( \|\boldsymbol{\beta}^{\mathrm{MCP}} - \boldsymbol{\beta}^{\mathrm{true}}\|_1 \leq \frac{16s\zeta}{\kappa} + \frac{16s\rho^{\mathrm{MCP}}_{\mathcal{S}/\mathcal{S}_1}}{\kappa}\lambda \right) \geq 1 - \delta_1(n, n, \zeta) - \delta_2(n, n, \lambda), \tag{10}$$

*where*

$$\delta_2(n, |\mathcal{A}|, \lambda) = 4d \exp\left( -\frac{n\lambda^2}{2\sigma^2 x_{\max}^2} \cdot \left( \frac{1}{2} - \frac{18ns}{|\mathcal{A}|\kappa a} \right)^2 \right), \tag{11}$$

$$\rho^{\mathrm{MCP}}_{\mathcal{S}/\mathcal{S}_1} = \frac{\|\boldsymbol{\beta}^{\mathrm{MCP}}_{\mathcal{S}/\mathcal{S}_1} - \boldsymbol{\beta}^{\mathrm{true}}_{\mathcal{S}/\mathcal{S}_1}\|_1}{\|\boldsymbol{\beta}^{\mathrm{MCP}}_{\mathcal{S}} - \boldsymbol{\beta}^{\mathrm{true}}_{\mathcal{S}}\|_1} \in [0, 1], \tag{12}$$

*$C_1$ is the same as defined in Lemma 1, $\mathcal{S}/\mathcal{S}_1 = \{i : i \in \mathcal{S} \text{ and } i \notin \mathcal{S}_1\}$, and $\rho^{\mathrm{MCP}}_{\mathcal{S}/\mathcal{S}_1} = 0$ if $\mathcal{S}/\mathcal{S}_1$ is a empty set.*

Note that in step 1 of the 2sWL procedure, we solve the Lasso problem, and according to the Lemma EC.2 in the E-Companion, the estimator $\boldsymbol{\beta}_1$ will converge to $\boldsymbol{\beta}^{\mathrm{true}}$ at a rate of $\frac{24s\lambda}{\kappa}$ with

high probability. Therefore, if $|\beta_i^{\text{true}}| \geq (\frac{24s}{\kappa} + a)\lambda$, then we immediately have $|\beta_{1,i}| \geq a\lambda$, leading to $w_i = P'_{a,\lambda}(|\beta_{1,i}|) = 0$, which means that the step 2 of the 2sWL procedure will not penalize the coefficient for dimension $i$ and, therefore, remove the bias issue in the $l_1$ penalty. In addition, we quantify the influence of the $\mathcal{S}_1$ set on the convergence rate by the $\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1}$ term. Because $\mathcal{S}_1$ is a subset of $\mathcal{S}$, the $\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1}$ term decreases as the $\mathcal{S}_1$ set contains more elements, which improves the convergence performance of the MCP estimator solved by the 2sWL procedure.

Compared to the oracle estimator $\boldsymbol{\beta}^{\text{oracle}}$ in Lemma 1, the probability bound on the MCP estimator under the 2sWL procedure has an extra term $\delta_2(n, n, \lambda)$, which depends on the covariate dimension $d$ and the i.i.d. sample size $n$. Note that as the sample size increases, the extra term decreases to 0 at an exponential rate. In other words, as the sample size increases, $\boldsymbol{\beta}^{\text{MCP}}$ and the oracle solution enjoy the same order of the optimal convergence rate with high probability.

Remark 1. Proposition 1 also suggests that as long as the $\mathcal{S}_1$ set is non-empty, then the MCP estimator solved by the 2sWL procedure enjoys better convergence properties than the Lasso estimator. In particular, by setting $\zeta = \frac{1}{2}\lambda$, we can show that $\boldsymbol{\beta}^{\text{MCP}}$ has $\frac{8+16\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1}}{\kappa} s\lambda$ convergence rate. In Lemma EC.2 in the E-Companion, we show that the Lasso estimator has $\frac{24}{\kappa} s\lambda$ convergence rate. In fact, since the Lasso problem is corresponding to the case with $w_i = \lambda$ in the 2sWL procedure, we can view it as a special case of the MCP estimator for $\mathcal{S}_1 = \emptyset$, under which $\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1} = 1$ and both estimators have the same convergence. Therefore, as long as not all $\beta_i^{\text{true}}$ for $i \in \{1, 2, ..., d\}$ are very small, then we will have $\mathcal{S}_1 \neq \emptyset$, which means $\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1} < 1$, so that the MCP estimator has better convergence property than the Lasso estimator. In practice, it is common to set $\lambda = \mathcal{O}\left(\sqrt{\frac{\log n + \log d}{n}}\right)$, so when the sample size $n$ is large enough, $\mathcal{S}_1$ will be nonempty (i.e., $\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1} < 1$).

It is worth mentioning that if the set $\mathcal{S}_1$ includes all significant covariates (i.e., $\mathcal{S}_1 = \mathcal{S}$), the term $\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1}$ will be 0 and the MCP estimator will attain the same order of the convergence rate as the oracle estimator. The following corollary states the MCP estimator's oracle property[1].

Corollary 1. *Let assumptions in Proposition 1 hold, if $\mathcal{S}_1 = \mathcal{S}$, then both $\boldsymbol{\beta}^{\text{MCP}}$ and $\boldsymbol{\beta}^{\text{oracle}}$ converge to $\boldsymbol{\beta}^{\text{ture}}$ on the order of $\mathcal{O}(s\zeta)$ with probability $1 - \delta_1(n, n, \zeta) - \delta_2(n, n, \lambda)$, where $\delta_1(n, n, \zeta)$ and $\delta_2(n, n, \lambda)$ are defined in* (5) *and* (11) *respectively.*

### 4.3. $\epsilon$-decay Random Sampling Method

As bandit models involve exploitation and exploration, samples generated under exploitation typically are non-i.i.d.. These non-i.i.d. samples pose analytical challenges to the existing MCP literature that relies on the assumption that all samples are i.i.d. to establish the convergence rate. Hence, to ensure desired performance of the MCP estimator, we need to secure that at least some

[1] In §5.3, we will show that under the G-MCP-Bandit Algorithm described in §4.4, the condition $\mathcal{S}_1 = \mathcal{S}$ will be satisfied for a large $T$ value.

samples generated in the online learning and decision-making process are i.i.d. (see §5.1 for detailed reasons). In this research, we propose a $\epsilon$-decay random sampling method, in which decision-makers draw random samples, with decreasing probability, by randomly selecting decisions from the decision set with equal probability. In particular, the $\epsilon$-decay random sampling method can be described as follows:

***$\epsilon$-decay Random Sampling Method***: At time $t$, with probability $\min\{1, t_0/t\}$, where $t_0$ is a predetermined positive constant, decision-makers will randomly select a decision from their decision set with equal probability. Otherwise, decision-makers will follow a bi-level decision structure, which will be specified later, to determine the optimal decision to maximize their expected reward.

The $\epsilon$-decay random sampling method can balance the exploitation and exploration trade-off by ensuring that decision-makers do not explore too much to significantly sacrifice their revenue performance (as the probability of drawing a random sample decays in time) but will secure sufficient random samples to guarantee the quality of the parameter vector estimation. In particular, we can bound the random sample size in the following proposition.

PROPOSITION 2. *Let $C_0 \geq 20$, $t_0 = 2C_0|\mathcal{K}|$, and $T > t_0$. Under the $\epsilon$-decay random sampling method, the random sample size $n_k$ for arm $k \in \mathcal{K}$ up to time $T$ is bounded by*

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

*with probability at least $1 - \delta_0(T, t_0)$, where*

$$\delta_0(T, t_0) = \frac{2(t_0+1)}{e^4(T+1)^4}. \tag{13}$$

### 4.4. G-MCP-Bandit Algorithm

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

**G-MCP-Bandit Algorithm**

**Require**: Input parameters $t_0, h, a, \lambda_1, \lambda_{2,0}$.

Initialize $\boldsymbol{\beta}_k^{\text{random}} = \boldsymbol{\beta}_k^{\text{whole}} = \mathbf{0}$, and $\mathcal{R}_k$ and $\mathcal{W}_k$ being empty sets for all $k \in \mathcal{K}$.

**For** $t = 1, 2, \ldots.$ **do**

Observe $\boldsymbol{x}_t$.

Draw a binary random variable $\mathcal{D}_t$, where $\mathcal{D}_t = 1$ with probability $\min\{1, t_0/t\}$.

**If** $\mathcal{D}_t = 1$

Assign $\pi_t$ a random decision $k \in \mathcal{K}$ with probability $\mathbb{P}(\pi_t = k) = 1/|\mathcal{K}|$.

Play decision $\pi_t$ and observe $r_t$.

Update $\mathcal{R}_{\pi_t} = \mathcal{R}_{\pi_t} \cup \{(\boldsymbol{x}_t, r_t)\}$ and $\mathcal{W}_{\pi_t} = \mathcal{W}_{\pi_t} \cup \{(\boldsymbol{x}_t, r_t)\}$.

**Else**

Construct the optimal decision set:

$\Pi_t = \left\{ i : \mathbb{E}_\epsilon[R_i | \boldsymbol{x}_t^\top \boldsymbol{\beta}_i^{\text{random}}] \geq \max_{j \in \mathcal{K}} \mathbb{E}_\epsilon[R_j | \boldsymbol{x}_t^\top \boldsymbol{\beta}_j^{\text{random}}] - \frac{1}{2}h,\ i \in \mathcal{K} \right\}$.

Set $\pi_t = \arg\max_{k \in \Pi_t} \mathbb{E}_\epsilon[R_k | \boldsymbol{x}_t^\top \boldsymbol{\beta}_k^{\text{whole}}]$.

Play decision $\pi_t$, observe $r_t$, and update $\mathcal{W}_{\pi_t} = \mathcal{W}_{\pi_k} \cup \{(\boldsymbol{x}_t, r_t)\}$.

**End If**

For $k \in \mathcal{K}$, set $\lambda_{2,t} = \lambda_{2,0}\sqrt{\frac{\log t + \log d}{t}}$, update $\boldsymbol{\beta}_k^{\text{random}}$ and $\boldsymbol{\beta}_k^{\text{whole}}$ via the 2sWL procedure with $(a, \lambda_1, \mathcal{R}_{\pi_t})$ and $(a, \lambda_{2,t}, \mathcal{W}_{\pi_t})$, respectively.

**End for**

Specifically, decision-makers will start by assigning values for system parameters ($t_0, h, a, \lambda_1$, and $\lambda_{2,0}$), which can be optimized through tuning, and initialing parameter vector estimators ($\boldsymbol{\beta}_k^{\text{random}}$ and $\boldsymbol{\beta}_k^{\text{whole}}$) and sample datasets ($\mathcal{R}_k$ and $\mathcal{W}_k$, which represent the random sample set and the whole sample set) for all arm $k \in \mathcal{K}$. Then, for an incoming user at time $t$, decision-makers will draw a binary random variable $\mathcal{D}_t$ with probability $\min\{1, t_0/t\}$. There are two possibilities:

- If $\mathcal{D}_t = 1$, then they will randomly choose a decision $k$ from their decision set $\mathcal{K}$ with equal probability of $1/|\mathcal{K}|$; then, they will implement the chosen decision (i.e., $\pi_t = k$), observe the user's response, and claim the corresponding reward; finally, decision-makers will include the user's covariate vector and the corresponding reward $\{(\boldsymbol{x}_t, r_t)\}$ in both sample datasets, $\mathcal{R}_{\pi_t}$ and $\mathcal{W}_{\pi_t}$.
- If $\mathcal{D}_t = 0$, then they will use a bi-level decision structure to determine their decision. In the upper-level decision-making process, decision-makers will first construct an optimal decision set $\Pi_t$. Specifically, all decisions in the optimal decision set $\Pi_t$ are estimated, based on the random sample MCP estimator $\boldsymbol{\beta}^{\text{random}}$, to yield expected rewards within $h/2$ of the maximum possible reward. If there is only one decision in the optimal decision set $\Pi_t$, then decision-makers will implement this decision as the optimal decision; otherwise, decision-makers will perform the lower-level decision-making process, in which decision-makers will estimate, by using the whole sample MCP estimator $\boldsymbol{\beta}^{\text{whole}}$, the rewards for all decisions in the optimal decision set $\Pi_t$ and select the decision that generates the highest expected reward. Then, observing the user's response and collecting the corresponding reward, decision-makers will only update the whole sample dataset $\mathcal{W}_{\pi_t}$ by appending the user's covariate vector and the corresponding reward $\{(\boldsymbol{x}_t, r_t)\}$.

Finally, decision-makers will update parameter $\lambda_{2,t}$, and then use the 2sWL procedure to update the random sample parameter vector estimator $\boldsymbol{\beta}^{\text{random}}$ and the whole sample parameter vector estimator $\boldsymbol{\beta}^{\text{whole}}$, based on sample data sets $\mathcal{R}_{\pi_t}$ and $\mathcal{W}_{\pi_t}$, respectively.

The expected cumulative regret upper bound for the G-MCP-Bandit algorithm can be established in the following theorem.

THEOREM 1. *Under assumptions A.1-A.5, let* $t_0 = 2C_0|\mathcal{K}|$, $a > \frac{1152s}{p^*\kappa}$, $\lambda_1 = \mathcal{O}(s^{-1})$, *and* $\lambda_{2,0} = \mathcal{O}(1)$, *where detailed expressions of* $\lambda_1$ *and* $\lambda_{2,0}$ *are given in* (EC.36) *and* (EC.62), *respectively. The cumulative regret upper bounds for the G-MCP-Bandit algorithm up to time* $T$ *are given as follows:*

$$R^C(T) \leq \begin{cases} R_{\max}|\mathcal{K}|\left[(3C_0+C_3)\log T + (7+2C_4)T_0 + C_5\log^2 T\right] = \tilde{\mathcal{O}}(s^2(\log d + \log T)\log T), & T < T_1 \\ R_{\max}|\mathcal{K}|\left[(3C_0+C_3+C_5)\log T + (7+2C_4)T_0 + C_5\log^2 T_1\right] = \tilde{\mathcal{O}}(s^2\log d\log T), & T \geq T_1 \end{cases}$$

*where* $T_0 = \tilde{\mathcal{O}}(s^2\log d)$ *by* (EC.60) *and* (EC.78), $T_1 = \tilde{\mathcal{O}}(\beta_{\min}^{-2}\cdot s^2\log d)$, $C_0 = \mathcal{O}(s^2\log d)$, $C_3 \leq \tilde{\mathcal{O}}((1+\rho_{\max})^2 s^2\log d)$, $C_4 = \mathcal{O}(1)$, *and* $C_5 = \tilde{\mathcal{O}}(s^2)$ *are defined in* (EC.61), (EC.35), (EC.112), (EC.114), *and* (EC.115), *respectively,*

$$\rho_{\max} = \max_{T_0 \leq t \leq T_1, k\in\mathcal{K}_o} \rho^{\text{whole}}_{\mathcal{S}^k/\mathcal{S}^k_{1,t}} \tag{14}$$

$$\beta_{\min} = \min_{i\in\mathcal{S}^k, k\in\mathcal{K}} |\beta^{\text{true}}_{k,i}|, \tag{15}$$

$\mathcal{S}^k_{1,t}$ *is the index set* $\mathcal{S}_1$ *of arm* $k$ *at time* $t$, $\rho^{\text{whole}}_{\mathcal{S}^k/\mathcal{S}^k_{1,t}}$ *defined in* (12), *and we use* $\tilde{\mathcal{O}}(\cdot)$ *to suppress the logarithmic dependence on* $s$.

Theorem 1 shows that the expected cumulative regret of the G-MCP-Bandit algorithm over $T$ users is upper-bounded by $\mathcal{O}(\log T)$ in the data-rich regime (i.e., $T \geq T_1$). Goldenshluger and Zeevi (2013) have shown that under low-dimensional settings, the expected cumulative regret for a linear bandit model is lower-bounded by $\mathcal{O}(\log T)$, which is directly applicable to high-dimensional settings. Further, note that the linear model is a special case of the generalized linear model. Therefore, the expected cumulative regret of the G-MCP-Bandit algorithm is also lower-bounded by $\mathcal{O}(\log T)$. In other words, the G-MCP-Bandit algorithm achieves the optimal expected cumulative regret in the sample size dimension.

The optimal $\mathcal{O}(\log T)$ order represents an improvement from Lasso-Bandit's $\mathcal{O}(\log^2 T)$ regret upper bound. To explain this improvement, first note that by design, at each time step, the G-MCP-Bandit algorithm will update the penalty term for the MCP estimator $\lambda_{2,t}$ to be $\lambda_{2,0}\sqrt{\frac{\log t + \log d}{t}}$. Such a decrease of the penalty term helps round down significant covariates with small coefficients to infer the true support of $\boldsymbol{\beta}^{\text{true}}$. As time increases, $\lambda_{2,t}$ will eventually decay below a threshold that is proportional to the $\beta_{\min}$ value so that the $\mathcal{S}_1$ set in Proposition 1 becomes $\{i : |\beta^{\text{true}}_i| \geq \beta_{\min}\} = \mathcal{S}$

for all arms, which happens in the data-rich regime (i.e., $T \geq T_1$). Therefore, in the data-rich regime, the MCP estimator will enjoy the same order of the convergence rate as the oracle solution (see Corollary 1) and will be independent on $\lambda_{2,t}$.

In addition, in the data-rich regime, Theorem 1 also demonstrates that the cumulative regret of the G-MCP-Bandit algorithm in the high-dimensional covariate vector $d$ is upper-bounded by $\mathcal{O}(\log d)$. This bound presents a significant improvement over other classic bandit algorithms (Goldenshluger and Zeevi 2013, Abbasi-Yadkori and Szepesvari 2012, Dani et al. 2008), which yield polynomial dependence on $d$, and is also a tighter bound than the Lasso-type algorithm (i.e., $\mathcal{O}(\log^2 d)$ in Bastani and Bayati 2020). It is worth noting that compared to Lasso-Bandit, the G-MCP-Bandit algorithm improves the regret bound from $\mathcal{O}(s^2(\log d + \log T)^2)$ to $\mathcal{O}(s^2 \log d \log T)$. This improvement is of particular importance in high-dimensional settings, where the covariate dimension can be extremely large, and it suggests that the G-MCP-Bandit algorithm can bring substantial regret reduction compared to existing bandit algorithms (e.g., see §6).

In the data-poor regime (i.e., $T < T_1$), the regret upper bound of the G-MCP-Bandit algorithm on $T$ will worsen to $\mathcal{O}(s^2(\log d + \log T)\log T)$, which has better dependence on $d$ but shares the same order on $T$ as Lasso-Bandit. Yet, note that $C_3$ depends on $\rho_{\max} = \max_{T_0 \leq t \leq T_1, k \in \mathcal{K}_o} \rho^{\text{whole}}_{\mathcal{S}^k/\mathcal{S}^k_{1,t}}$, and recall that Remark 1 demonstrates that for non-empty index $\mathcal{S}^k_{1,t}$, we have $\rho^{\text{MCP}}_{\mathcal{S}^k/\mathcal{S}^k_{1,t}} < 1$, which suggests that the G-MCP-Bandit algorithm has a tighter bound on $T$ than Lasso-bandit in the constant $C_3$ and performs better under the data-poor regime as well.

Remark 2. The value of $T_1$ depends on the magnitude of the signal for significant covariates $\beta_{\min}$. The $\beta_{\min}$ is often referred to as the minimal signal for the non-zero component of $\boldsymbol{\beta}^{\text{true}}$, and in the high-dimensional bandit literature (e.g., Hao et al. 2020, Ariu et al. 2020), the $\beta_{\min}$ value are often used to generate a time-threshold such that once the sample size $T$ passes such a time-threshold, the regret upper bound can be improved in the data-rich regime. This is because for a very small $\beta_{\min}$ value, it will be prohibitively difficult to distinguish all significant covariates away from 0, so we will need more samples to correct the bias from the penalized estimation, which means that the G-MCP-Bandit algorithm has to stay longer in the data-poor regime (i.e., $T < T_1$) with suboptimal regret of $\tilde{\mathcal{O}}(s^2(\log d + \log T)\log T)$. Yet, it is worth noting that the G-MCP-Bandit algorithm doesn't require the knowledge of the $\beta_{\min}$ value as an input parameter, and the regret upper bound will eventually switch to $\tilde{\mathcal{O}}(s^2 \log d \log T)$ automatically in the data-rich regime (i.e., $T \geq T_1$) as more samples are collected.

### 4.5. Computational Complexity

The average computational cost for the G-MCP-Bandit algorithm can be shown in the following theorem.

THEOREM 2. *Let $\epsilon > 0$ be an optimization tolerance constant. Under assumptions A.1 and A.5, the average computation cost of the G-MCP-Bandit Algorithm by time $T$ will be upper bounded by $\mathcal{O}(x_{\max} b^3 \epsilon^{1/2} \cdot |\mathcal{K}| d^4 T)$. Moreover, for a large $T$, the average computation cost can be improved to $\mathcal{O}(x_{\max} b \cdot |\mathcal{K}| d^2 T)$ with high probability.*

The G-MCP-Bandit algorithm's worst-case average computational cost is on the order of $\mathcal{O}(x_{\max} b^3 \epsilon^{1/2} \cdot |\mathcal{K}| d^4 T)$, when a basic accelerated gradient descent method (e.g., the FISTA method in Beck and Teboulle 2009) is used as the optimization scheme. The primary computational cost of the G-MCP-Bandit algorithm is from updating/solving model parameter $\boldsymbol{\beta}$ via the 2sWL procedure for each arm at every time step: the $|\mathcal{K}|$ dependence is because we require to update every arm at every time step, $x_{\max} bd$ part comes from the Lipschitz constant of the loss function $\mathcal{L}(\boldsymbol{\beta})$, the remaining $b^2 d^2$ stems from the distance between the initial solution and the optimal solution, and the $Td$ part describes the cost of evaluating the full gradient of $\mathcal{L}(\boldsymbol{\beta})$. The improvement in the long-run regime is mainly from the warm start in the 2sWL procedure. From Proposition 6 and Lemma EC.2 in E-Companion, if time $T$ is large enough (e.g., $T \geq \max\{T_1, \tilde{\mathcal{O}}(s^2 \epsilon^{-1/2})\}$), then with high probability, $\boldsymbol{\beta}_k^{\text{MCP}}$ differs from $\boldsymbol{\beta}_k^{\text{true}}$ at a rate lower than $\epsilon^{1/4}$. Therefore, if we use $\boldsymbol{\beta}_k^{\text{MCP}}$ as the initial solution in the 2sWL procedure, then it will be very efficient to identify the optimal solution, which suggests that the average computational cost of the G-MCP-Bandit algorithm can be improved to $\mathcal{O}(x_{\max} b \cdot |\mathcal{K}| d^2 T)$.

## 5. Key Steps of Regret Analysis for the G-MCP-Bandit Algorithm

In this section, we provide abridged technical proofs for Theorem 1, the main theorem in this paper. Specifically, we briefly lay out four key steps in establishing the expected cumulative regret upper bound for the G-MCP-Bandit algorithm. In the first step, we highlight the influence of non-i.i.d. data, inherited from the multi-armed bandit model, and provide the statistical convergence property for the MCP estimator under partially i.i.d. samples. Applying these results to the G-MCP-Bandit algorithm, in the second and third steps, we establish the convergence properties for both the random sample estimator, which is based on only samples that were generated through the $\epsilon$-decay random sampling method, and the whole sample estimator, which uses all available samples. Finally, in the last step, we establish the total expected cumulative regret by separating the regret up to time $T$ into three segments and providing a bound for each segment. The main structure and sequence of our proving steps described above are first introduced by Bastani and Bayati (2020), which presents their expected regret analysis for a linear bandit model (i.e., LASSO-Bandit algorithm) in a similar sequence. We will largely follow their presentation structure, but with different steps, proving techniques, and convergence properties, to illustrate the key steps in analyzing the G-MCP-Bandit algorithm.

### 5.1. General Non-i.i.d. Sample Estimator

Note that the restricted eigenvalue condition (i.e., assumption A.4) for high-dimensional statistics is for i.i.d. samples in the literature. Yet, in this research, we consider the G-MCP-Bandit algorithm, under which only part of the samples are i.i.d., so we first need to show that the restricted eigenvalue condition continues to hold for partially i.i.d. samples (see Lemma EC.1 in E-Companion). Then, we can establish general results for the MCP estimator under non-i.i.d. data.

We denote $\mathcal{W}_k$ as the whole sample set and $\boldsymbol{\beta}_k^{\text{MCP}}$ as the MCP estimator for the parameter vector corresponding to decision $k \in \mathcal{K}$. Similarly, in the following presentation, we will omit parameters' subscripts corresponding to the choice of arms and time index for brevity, as long as doing so will not cause any misinterpretation.

Note that as samples in $\mathcal{W}$ may be non-i.i.d., standard MCP convergence results (Fan et al. 2014b, 2018) cannot be directly applied. Recall that we proposed the $\epsilon$-decay random sampling method, in which these samples generated under randomly selected decisions are i.i.d.. Therefore, there exists a subset $\mathcal{A} \subseteq \mathcal{W}$ such that all samples in this subset are i.i.d. from the distribution $\mathcal{P}_{\boldsymbol{X}}$. The next step is to show that when the cardinality of $\mathcal{A}$ (i.e., $|\mathcal{A}|$) is large enough, $\boldsymbol{\beta}^{\text{MCP}}$ will converge to the true parameters $\boldsymbol{\beta}^{\text{true}}$.

Proposition 3. *Under assumptions A.1, A.4, and A.5, if* $|\mathcal{A}| \geq C_1^{-1} \log d$ *and* $a > \frac{48ns}{|\mathcal{A}|\kappa}$, *then for* $\zeta > 0$, *the following inequality holds for the MCP estimator under the 2sWL procedure* $\boldsymbol{\beta}^{\text{MCP}}$

$$\mathbb{P}\left( \|\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \frac{16ns\zeta}{|\mathcal{A}|\kappa} + \frac{16ns\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1}}{|\mathcal{A}|\kappa}\lambda \right) \geq 1 - \delta_1(n, |\mathcal{A}|, \zeta) - \delta_2(n, |\mathcal{A}|, \lambda), \tag{16}$$

*where* $C_1 = \mathcal{O}(s^{-2})$, $\delta_1(n, |\mathcal{A}|, \zeta)$, *and* $\delta_2(n, |\mathcal{A}|, \lambda)$ *are defined in* (EC.125), (5), *and* (11), *respectively.*

Proposition 3 describes the statistical properties of the non-i.i.d. MCP estimators under the 2sWL procedure. If we set $\zeta$ to be on the order of $\mathcal{O}(\sqrt{1/n})$, then $\|\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}\|_1$ is on the order of $\mathcal{O}(s\sqrt{n/|\mathcal{A}|^2})$. In addition, when the i.i.d. sample size $|\mathcal{A}|$ matches the whole sample size $n$, Equation (16) suggests that the MCP estimator guarantees the optimal statistical convergence at $\mathcal{O}(s\sqrt{1/n})$.

Moreover, Proposition 3 shows the necessity of generating i.i.d. random samples in high-dimension bandit settings. Non-i.i.d. samples are inevitable in online learning and decision-making process, so ensuring the desired performance of the parameter vector estimation in high-dimensional settings can only be achieved through generating a sufficient number of i.i.d. samples, as shown in Proposition 3. We will show in the next two subsections that at time $t$, the size of i.i.d. samples generated under the $\epsilon$-decay random sampling method is on the order of $\mathcal{O}(\log t)$, which can be further improved to the order of $\mathcal{O}(t)$ under the bi-level decision structure.

### 5.2. Estimator from Random Samples up to Time $t$

In Proposition 3, we show that the MCP estimator will converge to the oracle parameter as long as the sample set contains a sufficient number of i.i.d. samples. Recall that in our proposed G-MCP-Bandit algorithm, samples generated by the $\epsilon$-decay random sampling method are i.i.d., and the size of these i.i.d. samples is on the order of $\mathcal{O}(\log T)$ (i.e., see Proposition 2). Combining these observations, the following proposition establishes the statistical performance of the MCP estimator based on only random samples generated by the $\epsilon$-decay random sampling method.

PROPOSITION 4. *Let $t_0 = 2C_0|\mathcal{K}|$, $t > t_0$ and $a > 1152s/(p^*\kappa)$. If assumptions A.1, A.3, A.4, and A.5 hold, then the random sample MCP estimator for any arm in $\mathcal{K}$ under the G-MCP-Bandit algorithm $\boldsymbol{\beta}^{\text{random}}$ will satisfy the following inequality*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{random}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \min\left\{\frac{1}{\sigma x_{\max}}, \frac{h}{4e\sigma R_{\max} x_{\max}}\right\}\right) \geq 1 - 5\delta_0(t, t_0),$$

*where $\delta_0(t, t_0)$, $C_0 = \mathcal{O}(s^2 \log d)$, and $\lambda_1 = \mathcal{O}(s^{-1})$, whose detailed expressions are given in* (13), (EC.35)*, and* (EC.36)*, respectively.*

### 5.3. Estimator from Whole Samples up to Time $t$

In addition to i.i.d. samples generated by the $\epsilon$-decay random sampling method, other samples can also be used to improve the statistical performance of the MCP estimator. To intuit, recall that in the G-MCP-Bandit algorithm, when the user is not selected to perform a random sampling, decision-makers will use the bi-level structure to determine the optimal decision to maximize their expected reward. In the upper-level decision-making process, only i.i.d. samples will be used (as $\boldsymbol{\beta}^{\text{random}}$ is the MCP estimator based on samples generated only by the $\epsilon$-decay random sampling method) to determine the candidate(s) for the optimal decision set. From Proposition 4, we know that this random sample MCP estimator will not be far away from its true parameter values. In other words, if we define the event that the random sample MCP estimator at time $t$ is within a given distance from its true parameter as event $\mathcal{E}_2$:

$$\mathcal{E}_2 \doteq \left\{\|\boldsymbol{\beta}_k^{\text{random}} - \boldsymbol{\beta}_k^{\text{true}}\|_1 \leq \min\left\{\frac{1}{\sigma x_{\max}}, \frac{h}{4e\sigma R_{\max} x_{\max}}\right\},\ k \in \mathcal{K}\right\}, \tag{17}$$

then event $\mathcal{E}_2$ will happen with high probability. Further, conditioning on event $\mathcal{E}_2$ and assumption A.3, we can verify that for any $\boldsymbol{x} \in U_k$, $k \in \mathcal{K}$, the following inequality holds:

$$\mathbb{E}_\epsilon(R_k|\boldsymbol{x}^\top \boldsymbol{\beta}_k^{\text{random}}) \geq \max_{j \neq k} \mathbb{E}_\epsilon(R_j|\boldsymbol{x}^\top \boldsymbol{\beta}_j^{\text{random}}) + \frac{h}{2}. \tag{18}$$

Therefore, if using Equation (18) as the selecting criterion, decision-makers will be able to choose the optimal decision $k$ for any $\boldsymbol{x} \in U_k, k \in \mathcal{K}$ with high probability. We defer the detailed analysis to Lemma EC.3 in the E-Companion.

Formally, we can bound the total number of times under which event $\boldsymbol{X}_j \in U_k$ and event $\mathcal{E}_2$ happen simultaneously. In particular, we define

$$M_k(i) \doteq \mathbb{E}_{\epsilon,\boldsymbol{X}}\left[\sum_{j=1}^{t} \mathbb{1}(\boldsymbol{X}_j \in U_k, \mathcal{E}_2, \boldsymbol{X}_j \notin \mathcal{R}_{\boldsymbol{x},k})|\mathcal{F}_i\right] \tag{19}$$

for $i \in \{0,1,2,..,t\}$, where $\mathcal{F}_i = \{(\boldsymbol{X}_j, R_j) \text{ for } j \le i\}$ and $\mathcal{R}_{\boldsymbol{x},k}$ being the set containing the user covariate $\boldsymbol{X}$ with decision arm $k$ assigned by the $\epsilon$-decay random sampling method. Then, $\{M_k(i)\}$ is a martingale with bounded difference $|M_k(i) - M_k(i+1)| \le 1$ for $i = 0,1,2,...,t$, and we can bound the value of $M_k(t)$ in the following proposition:

PROPOSITION 5. *Let $t_0 = 2C_0|\mathcal{K}|$ for some $C_0$, $t \ge t_0$, and $a > 1152s/(p^*\kappa)$. If assumptions A.1, A.3, A.4, and A.5 hold, then $\mathbb{P}\left(M_k(t) \le \frac{p^*t}{8}\right) \le \exp\left(-\frac{(p^*)^2t}{256}\right)$ holds for all $k \in \mathcal{K}$, where $C_0 = \mathcal{O}(s^2\log d)$ and $\lambda_1 = \mathcal{O}(s^{-1})$ are defined in* (EC.35) *and* (EC.36)*, respectively.*

Intuitively, Proposition 5 suggests that with high probability, the actual i.i.d. sample size in $U_k$ for decision $k$ will be on the order of $\mathcal{O}(t)$ instead of $\mathcal{O}(\log t)$. This improvement is the reason why the whole sample MCP estimator $\boldsymbol{\beta}^{\text{whole}}$ used in the lower-level decision-making process has better statistical performance, compared to the random sample MCP estimator $\boldsymbol{\beta}^{\text{random}}$ used in the upper-level decision-making process. Specifically, we can establish the convergence property for the whole sample MCP estimator in the following proposition.

PROPOSITION 6. *Let $t_0 = 2C_0|\mathcal{K}|$, $t > T_0$, and $a > \frac{1152s}{p^*\kappa}$. If assumptions A.1, A.3, A.4, and A.5 hold, then the whole sample MCP estimator for the arm in the optimal arm set $\mathcal{K}_o$, under the G-MCP-Bandit algorithm, $\boldsymbol{\beta}^{\text{whole}}$ will satisfy the following inequality:*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{whole}} - \boldsymbol{\beta}^{\text{true}}\|_1 \le \frac{128s\zeta}{p^*\kappa} + \frac{128s\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1}}{p^*\kappa}\lambda\right) \ge 1 - 5\delta_0(t,t_0) - \frac{10}{(t+1)^2} - 2s\exp\left(-\frac{t\zeta^2}{2\sigma^2x^2_{\max}}\right),$$

*where $\zeta \ge 0$, and $\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1}$, $\delta_0(t,t_0)$, $C_0 = \mathcal{O}(s^2\log d)$, $T_0 = \mathcal{O}(s^2\log d)$, $\lambda_1 = \mathcal{O}(s^{-1})$, and $\lambda_{2,0} = \mathcal{O}(1)$ are defined in* (12), (13), (EC.35), (EC.60), (EC.36), *and* (EC.62)*, respectively. Moreover, let $T_1 = \mathcal{O}(\beta^{-2}_{\min}s^2\log d)$ be set as in* (EC.61)*. Then, when $t \ge T_1$, the above result can be improved to*

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{whole}} - \boldsymbol{\beta}^{\text{true}}\|_1 \le \frac{128s\zeta}{p^*\kappa}\right) \ge 1 - 5\delta_0(t,t_0) - \frac{10}{(t+1)^2} - 2s\exp\left(-\frac{t\zeta^2}{2\sigma^2x^2_{\max}}\right).$$

### 5.4. Cumulative Regret Up To Time $T$

Finally, to bound the cumulative regret for the G-MCP-Bandit algorithm, we need to divide the time, up to time $T$, into three groups and provide an upper bound for each group.

The first group of time contains the time before time $T_0$ and the time up to time T when i.i.d. samples are generated through the $\epsilon$-decay random sampling method. Note that before time $T_0$ (the explicit requirement for $T_0$ is given in the proof of Proposition 6 in E-Companion and on

the order of $\tilde{\mathcal{O}}(s^2 \log d)$), decision-makers does not have sufficient samples to accurately estimate covariate parameter vectors. Hence, the reward under the G-MCP-Bandit algorithm will suffer and be sub-optimal compared to that of the oracle case. We can bound the cumulative regret by the worst-case performance by $3C_0 R_{\max}|\mathcal{K}| \log T + 5R_{\max}|\mathcal{K}|T_0 = \tilde{\mathcal{O}}(s^2 \log d \log T)$, where the first part of this cumulative regret is for all samples before time $T_0$ and the second part is for all random samples up to time $T$.

Next, we will segment the remaining scenarios into two groups, depending on whether we can accurately estimate covariate parameter vectors by using only random samples collected by the $\epsilon$-decay method. In particular, the second group includes the remaining scenarios where the random-sample-based estimators are not accurate (i.e., event $\mathcal{E}_2$ doesn't hold). Under those scenarios, inevitably, decision-makers' decisions will be suboptimal with high probability. However, note that as the size of i.i.d. samples increases in $t$, the probability of event $\mathcal{E}_2$ not occurring decreases. We can bound the cumulative regret for the second group by $2R_{\max}|\mathcal{K}|T_0 = \tilde{\mathcal{O}}(s^2 \log d)$.

The last group includes the remaining scenarios where the random-sample-estimators are accurate enough (i.e., event $\mathcal{E}_2$ holds.). Benefiting from the improved estimation accuracy (Proposition 6), we can bound the cumulative regret for the last group by $R_{\max}|\mathcal{K}|(2C_4 T_0 + C_3 \log T + C_5 \log^2 T) = \tilde{\mathcal{O}}(s^2(\log d + \log T) \log T)$. Further, when $t$ goes beyond $T_1 \geq T_0$, we can prove that the expected cumulative regret for the last group will be bounded by $R_{\max}|\mathcal{K}|(2C_4 T_0 + (C_3 + C_5) \log T + C_5 \log^2 T_1) = \tilde{\mathcal{O}}(s^2 \log d \log T)$. Combining the cumulative regret for all three groups, Theorem 1 directly follows.

## 6. Empirical Experiments

In this section, we will benchmark the G-MCP-Bandit algorithm to OFUL (Abbasi-Yadkori et al. 2011), OLS-Bandit (Goldenshluger and Zeevi 2013), and Lasso-Bandit (Bastani and Bayati 2020). In particular, we seek answers to the following two questions: How does the performance of the G-MCP-Bandit algorithm compare to other bandit algorithms? And how is the performance of the G-MCP-Bandit algorithm influenced by the data availability ($T$), the data dimensions ($s$ and $d$), and the size of the decision set ($K$)?

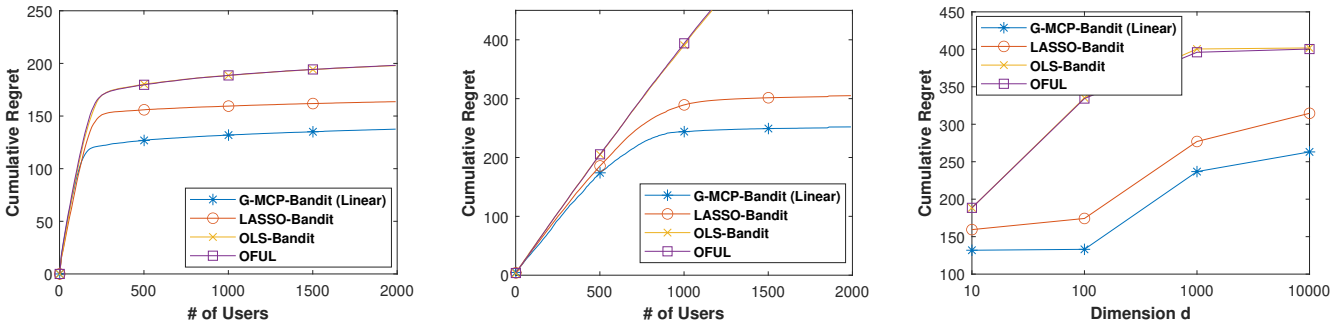
To this end, we start with two synthetic-data-based experiments in §6.1 and then conduct one real-dataset-based experiment, the Tencent search advertising data, in §6.2. Note that the algorithms and theoretical bounds of OFUL, OLS-Bandit, and Lasso-Bandit are developed under the assumption that the reward function follows the linear model, which is a special case in the G-MCP-Bandit algorithm. Therefore, for fair comparisons, we specify the underlying reward function for the G-MCP-Bandit algorithm to follow the same linear model (i.e., the reward under decision $k$ for a user with covariate vector $\boldsymbol{X}_t$ takes the form of $R_{k,t} = \boldsymbol{X}_t^\top \boldsymbol{\beta}_k^{\text{true}} + \epsilon$, where $\epsilon$ is a $\sigma$-gaussian

random variable) in both synthetic experiments. In the Tencent search advertising data experiment, besides benchmarking the G-MCP-Bandit algorithm to other bandit algorithms, we also explore the performance of the G-MCP-Bandit algorithm under both the linear model and the logistic model to examine the impacts of the model choice on decision-makers' revenue performance.

### 6.1. Synthetic Data (Linear Model)

In the first synthetic data experiment, we fix the size of the decision set $K$ and focus on the impacts of the data dimensions, $s$ and $d$, and the data availability, $T$, on the algorithms' cumulative regret performance. In particular, we consider a two-arm bandit setting (i.e., $K = 2$). To simulate different sparsity levels, we vary the covariate dimension $d = \{10, 10^2, 10^3, 10^4\}$ and keep the dimension for significant covariates unchanged at $s = 5$. Therefore, as the covariate dimension $d$ increases, the data become sparser. The underlying true parameter vectors for covariates are arbitrarily set to be $\boldsymbol{\beta}_1 = (1, 2, 3, 4, 5, 0, 0, ...)$ for the first arm and $\boldsymbol{\beta}_2 = 1.1 \cdot \boldsymbol{\beta}_1$ for the second arm. For each incoming user, we randomly draw her covariate vector from $N(0, I_{d\times d})$ and the error term in the linear model $\epsilon$ from $N(0, 1)$. We truncated the covariate vector and reward between $[-10, 10]$. Finally, we use the same parameter values for $t_0$, $h$, $\lambda_1$, and $\lambda_{2,0}$ in both the Lasso-Bandit algorithm and the G-MCP-Bandit algorithm and select the unique parameter for the G-MCP-Bandit algorithm $a$ at 2. For each algorithm, we perform 100 trials and report the average cumulative regret for OFUL, OLS-Bandit, Lasso-Bandit, and G-MCP-Bandit (under the linear model) in Figure 1.

**Figure 1** **Synthetic study 1: The impact of $T$ and $d$ on the cumulative regret, where $K = 2$ and $s = 5$.**




(a) d=10 (b) d=1000 (c) T=1000

Overall, we observe that the G-MCP-Bandit algorithm significantly outperforms OFUL, OLS-Bandit, and Lasso-Bandit and achieves the lowest cumulative regret. Facing only two decisions/arms, decision-makers can easily identify the optimal arm, and therefore OFUL and OLS-Bandit, both of which are not specifically designed for high-dimensional settings, perform nearly identically. Lasso-Bandit and G-MCP-Bandit could benefit from their abilities to recover the sparse

structure and identify the significant covariates. Therefore, compared to OFUL and OLS-Bandit, Lasso-Bandit and G-MCP-Bandit can improve their parameter estimations, especially under high-dimensional settings, and perform substantially better. Further, the improvement of the cumulative regret performance of G-MCP-Bandit over Lasso-Bandit follows from the facts that the MCP estimator is unbiased and could improve the sparse structure discovery. Next, we will discuss the influence of sample size $T$ and the covariate dimension $d$ on these algorithms' cumulative regret performance.

Figure 1(a) and 1(b) illustrate the influence of the sample size $T$ on the cumulative regret for the cases where $d = 10$ and $d = 1000$ (other cases exhibit a similar pattern and are therefore omitted)[2]. As we have proven that G-MCP-Bandit provides the optimal time dependence under both low-dimensional and high-dimensional settings (Theorem 1), G-MCP-bandit strictly improves on the cumulative regret performance from Lasso-Bandit, especially when T is not too small. Note that facing insufficient samples, all algorithms fail to accurately learn parameter vectors and therefore perform poorly. As the sample size increases, the G-MCP-bandit algorithm is able to, in an expeditious fashion, unveil the underlying sparse data structure, accurately estimate parameter vectors, and outperform all other benchmarks. For example, in Figure 1(b), we observe that the regret reduction of G-MCP-Bandit over all other algorithms is at least larger than 5% when the sample size $T$ is larger than 70. This observation echoes our theoretical findings that the G-MCP-Bandit algorithm attains the optimal regret bound in sample size dimension $\mathcal{O}(\log T)$.

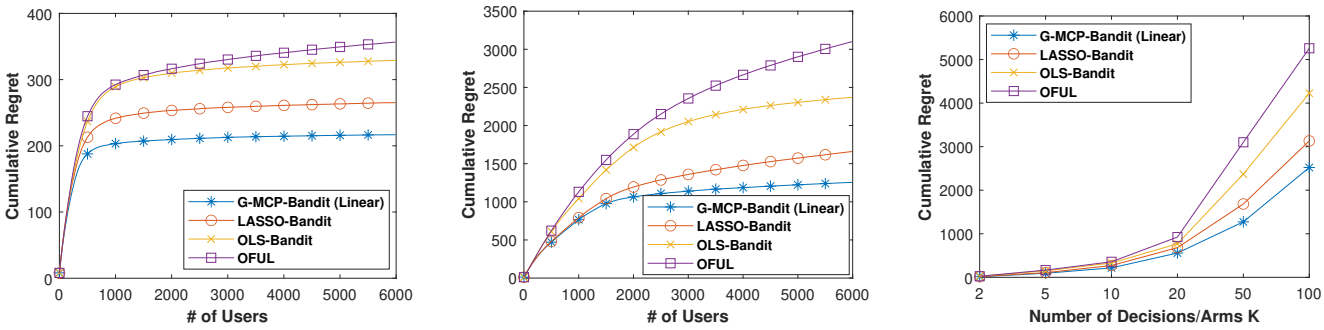
We also observe that the benefits of G-MCP-Bandit over the other three algorithms appear to increase in the data sparsity level. Figure 1(c) presents the influence of the covariate dimension $d$ on the cumulative regret for the case where $T = 1000$. Recall that we fixed the dimension for significant covariates $s = 5$. Therefore, as the covariate dimension $d$ increases, the data become sparser (i.e., $d/s$ increases). As expected, the cumulative regret for all four algorithms increases in the covariate dimension $d$, but at different rates. On the one hand, both OLS-Bandit and OFUL lack the ability to recover the sparse data structure and are ill-suited for high-dimensional problems. On the other hand, Lasso-Bandit and G-MCP-Bandit, which adopt different statistical learning methods for sparse structure discovery and are designed for high-dimensional problems, have lower cumulative regret that increases in $d$ at a slower rate. Further, we notice that the G-MCP-Bandit algorithm has the least increase in cumulative regret among all four algorithms, which confirms our theoretical finding in Theorem 1: The G-MCP-Bandit algorithm has a better dependence on

[2] In all four experiments where $d \in \{10, 10^2, 10^3, 10^4\}$, we simulated the sample size up to $10,000$ and observe that the G-MCP-Bandit algorithm's cumulative regret seems to be stabilized before $T = 2000$. Therefore, we only plot for the first 2000 samples to avoid duplication.

the covariate dimension $\mathcal{O}(\log d)$ than Lasso-Bandit $\mathcal{O}(\log^2 d)$, OFUL, and OLS-Bandit (the last two algorithms have polynomial bounds in $d$).

In the second synthetic data experiment, we study the influence of the size of the decision set by varying $K = \{2, 5, 10, 20, 50, 100\}$ and keeping the data dimensions unchanged ($s = 5$ and $d = 100$). For each decision, we randomly draw the parameter vector for the significant covariates from a uniform distribution, $U(0, 1)$. Finally, we keep other parameters the same as in the first synthetic data experiment. Figure 2 plots the average cumulative regret for OFUL, OLS-Bandit, Lasso-Bandit, and G-MCP-Bandit (under the linear model).

**Figure 2** **Synthetic study 2: The impact of $T$ and $K$ on the cumulative regret, where $d = 100$ and $s = 5$.**




(a) K=10 (b) K=50 (c) T=6000

We observe that the benefits of adopting G-MCP-Bandit over the other three algorithms increase in the size of the decision set. In particular, as $K$ increases, the cumulative regret gap between G-MCP-Bandit and any other algorithm grows; see Figure 2(c). This observation is as expected. To intuit, note that as we add more possible decisions into the decision set, the complexity and difficulty for decision-makers to select the optimal decision grow for two main reasons. First, decision-makers will need more samples to identify the significant covariates and estimate the parameter vectors. Second, as the number of decisions increases, the process of comparing the expected rewards among all decisions and selecting the optimal decision becomes more vulnerable to estimation errors. Therefore, we should expect that as the number of arms increases, the number of samples required for these algorithms to accurately learn the parameter vectors and select the optimal decision will increase as well.

Figure 2(a) and Figure 2(b) plot the cumulative regret for the case of ten arms and fifty arms, respectively. Clearly, decision-makers need far more samples before their cumulative regret can be stabilized in the case of fifty arms than in the case of ten arms. Therefore, the cumulative regret performance under all algorithms suffers from the increasing size of the decision set. As discussed earlier, the G-MCP-Bandit algorithm attains the optimal bound in the sample size dimension and

is able to learn the sparse data structure and provide accurate unbiased estimators for parameter vectors. Hence, we observe that the benefits of adopting the G-MCP-Bandit algorithm over other algorithms are amplified as the number of arms increases, as illustrated in Figure 2(c).

Finally, before moving to the real-data-based experiment, it is worth mentioning that we also conduct a systematic sensitivity analysis to test the robustness of the G-MCP-Bandit algorithm under different values of input parameters in §EC.3 of the E-Companion. In particular, we find that when we vary the G-MCP-Bandit algorithm's input parameters (i.e., $a$, $\lambda_1$, $\lambda_{2,0}$, $h$, $t_0$), the cumulative regret remains largely unchanged, which suggests that the G-MCP-Bandit algorithm is robust with respect to the choices of its input parameters.

### 6.2. Tencent Search Advertising Data (Linear & Logistic Models)

Now, we scale up the dataset's dimensionality by considering a search advertising problem at Tencent. The Tencent search advertising dataset is collected by Tencent's proprietary search engine, soso.com, and it documents the interaction sessions between users and the search engine (Tencent 2012). In the dataset, each session contains a user's demographic information (age and gender), the query generated by the user (combinations of keywords), ads information (title, URL address, and advertiser ID), the user's response (click or not), etc. This dataset is high-dimensional with a sparse data structure and contains millions of observations and covariates. To put the size of the dataset into perspective, it contains $149,639,105$ session entries, more than half a million ads, more than one million unique keywords, and more than 26 million unique queries.

For illustration purposes, we focus on a three-ad experiment[3] (with ad IDs 21162526, 3065545, and 3827183). Each of these three ads has an average CTR higher than 2% and more than $100,000$ session entries, which provide a basis for reasonably accurate estimations for parameter vectors. In total, there are $849,338$ session entries with $169,744$ unique queries and 8 covariates for users' demographic information. As the search engine receives payment from advertisers only when the user has clicked the sponsored ad, we arbitrarily set the awards for clicked ads to be \$1, \$5, and \$10, respectively.

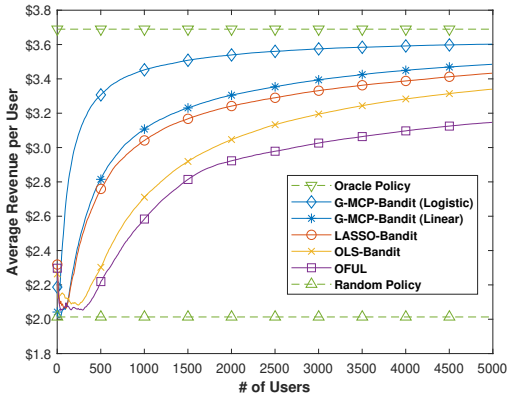
When the true underlying reward function follows the logistic model[4], Figure 3 plots the average revenue performance under OFUL (under linear model), OLS-Bandit (under linear model), Lasso-Bandit (under linear model), a random policy, the oracle policy (under the logistic model), and G-MCP-Bandit (under both linear and logistic models). It is worth noting that the "true" oracle

[3] Experiments with more ads are provided in Appendix EC.4.1. With a larger number of ads, our observations and insights remain qualitatively unchanged. Furthermore, similar to synthetic studies, the benefits of G-MCP-Bandit over other benchmarks increase in the number of ads.

[4] In Appendix EC.4.2, we further consider the case where the true underlying model follows a two-component Gaussian mixture model, which does not belong to the GLMs family, and the benefits of the G-MCP-Bandit algorithm over other benchmarks remain qualitatively unchanged.

policy is impossible to implement, as the true parameter vectors are unknown, or at least have considerable variance even when all session entries in the dataset are used for estimation. Therefore, the oracle policy in the experiment represents the scenario when the search engine has access to all data to estimate these parameter vectors and make ad selection decisions. In addition, we introduce the random policy as another benchmark to simulate the scenario in which the search engine will randomly recommend an ad with equal probability to an incoming user. Finally, note that the CTR prediction is binary in nature (i.e., click or not). We, therefore, include the G-MCP-Bandit algorithm under the logistic model and compare it to the G-MCP-Bandit algorithm under the linear model to study the influence of the underlying model choice. In the experiment, we simulate incoming users by permuting their covariate vectors randomly. For each algorithm, we perform 100 trials and report the average revenue with 5000 users, which seems to be sufficient for the G-MCP-Bandit algorithm to converge.

**Figure 3** **Tencent search advertising experiment: The average revenue under different algorithms.**




We can show that all learning algorithms generate higher average revenue than the random policy for any number of users and that the G-MCP-Bandit algorithm outperforms other algorithms under most scenarios. Specifically, when comparing all algorithms under the same linear model, we observe that the G-MCP-Bandit algorithm (under the linear model) has better average revenue performance than OFUL, OLS-Bandit, and Lasso-Bandit as soon as there are more than 140 users. This observation is consistent with that previous synthetic-data-based experiments and suggests that compared to other benchmarking algorithms, the G-MCP-Bandit algorithm can benefit from improved parameter vector estimation under high-dimensional data with limited samples and achieve better revenue performance.

Further, we find that the choice of underlying models can significantly influence the G-MCP-Bandit algorithm's average revenue performance. Note that the advertisers award the search engine

only when users have clicked the recommended ads. Therefore, the search engine's reward function is binary in nature. When comparing the G-MCP-Bandit algorithm under the logistic model to that under the linear model, both of which are special cases of the G-MCP-Bandit algorithm, we observe that the former always dominates the latter for any number of users. In addition, the G-MCP-Bandit algorithm under the logistic model merely needs 20 users to outperform the other three algorithms. This observation suggests that understanding the underlying managerial problem and identifying the appropriate model for the G-MCP-Bandit algorithm can be critical and bring substantial revenue improvement for decision-makers.

## 7. Conclusion

In this research, we develop the G-MCP-Bandit algorithm for online learning and decision-making processes in high-dimensional settings under limited samples. We adopt the matrix perturbation technique to derive new oracle inequalities for the MCP estimator under non-i.i.d. samples and further propose a linear approximation method, the 2sWL procedure, to overcome the computational and statistical challenges associated with solving the MCP estimator (an NP-complete problem) under the bandit setting. We demonstrate that the MCP estimator solved by the 2sWL procedure matches the oracle estimator with high probability and converges to the true parameters with the optimal convergence rate. Further, we show that in the data-rich regime, the cumulative regret of the G-MCP-Bandit algorithm over the sample size $T$ is bounded by $\mathcal{O}(\log T)$, which matches the theoretical lower bound for all possible algorithms under both low-dimensional and high-dimensional settings. In the covariate dimension $d$, the cumulative regret of the G-MCP-Bandit algorithm is upper bounded by $\mathcal{O}(\log d)$, which is also a tighter bound than existing bandit algorithms. Finally, we illustrate that compared to other benchmarking algorithms, the G-MCP-Bandit algorithm performs favorably in both synthetic-data-based and real-data-based experiments.

**Limitations and future research:** One limitation of this paper is that the analysis relies on the sub-Gaussian assumption, but some other models, such as the Poisson regression, merely satisfy a weaker sub-exponential assumption. Extending the current paper beyond the sub-Gaussian assumption to the sub-exponential assumption could future generalize this paper. Additionally, note that implementing the G-MCP-Bandit algorithm in an online setting could be computationally challenging in practice, especially when the covariate dimension and the decision set are extremely large. In particular, for every incoming user, the G-MCP-Bandit algorithm needs to update all arms' parameter vectors. Hence, when there are millions or billions of ads and users, implementing the G-MCP-Bandit algorithm becomes a highly time-consuming process on a single server. Hence, another future research direction could combine designing an online-offline hybrid structure, adopting parallel computing techniques, and using stochastic learning algorithms to improve the computational performance of the G-MCP-Bandit Algorithm.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 71929101 and 72371172, and the National Science Foundation (NSF) under Grant 1820702.

## References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 2312–2320.

Abbasi-Yadkori Y, Szepesvari C (2012) *Online learning for linearly parametrized control problems* (University of Alberta).

Agarwal A, Hsu D, Kale S, Langford J, Li L, Schapire R (2014) Taming the monster: A fast and simple algorithm for contextual bandits. *International Conference on Machine Learning*, 1638–1646.

Agrawal S, Goyal N (2013) Thompson sampling for contextual bandits with linear payoffs. *International Conference on Machine Learning*, 127–135.

Ariu K, Abe K, Proutière A (2020) Thresholded lasso bandit. *arXiv preprint arXiv:2010.11994* .

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3(Nov):397–422.

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002) The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32(1):48–77.

Bagnoli M, Bergstrom T (2005) Log-concave probability and its applications. *Economic theory* 26(2):445–469.

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

Bastani H, Bayati M, Khosravi K (2021) Mostly exploration-free algorithms for contextual bandits. *Management Science* 67(3):1329–1349.

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2(1):183–202.

Beygelzimer A, Langford J, Li L, Reyzin L, Schapire R (2011) Contextual bandit algorithms with supervised learning guarantees. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 19–26.

Boyd S, Boyd SP, Vandenberghe L (2004) *Convex optimization* (Cambridge university press).

Bühlmann P, Van De Geer S (2011) *Statistics for high-dimensional data: methods, theory and applications* (Springer Science & Business Media).

Candes EJ, Wakin MB, Boyd SP (2008) Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications* 14(5-6):877–905.

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

Chatzigeorgiou I (2013) Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters* 17(8):1505–1508.

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback .

Deshpande Y, Montanari A (2012) Linear bandits in high dimension and recommendation systems. *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1750–1754 (IEEE).

Fan J, Han F, Liu H (2014a) Challenges of big data analysis. *National science review* 1(2):293–314.

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 96(456):1348–1360.

Fan J, Liu H, Sun Q, Zhang T (2018) I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of statistics* 46(2):814.

Fan J, Xue L, Zou H (2014b) Strong oracle optimality of folded concave penalized estimation. *Annals of statistics* 42(3):819.

Filippi S, Cappe O, Garivier A, Szepesvári C (2010) Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 586–594.

Goemans M (2015) Chernoff bounds, and some applications.

Goldenshluger A, Zeevi A (2013) A linear response bandit problem. *Stochastic Systems* 3(1):230–261.

Hao B, Lattimore T, Wang M (2020) High-dimensional sparse linear bandits. *arXiv preprint arXiv:2011.04020* .

Huang J, Ma S, Zhang CH (2008) Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* 1603–1618.

Kim GS, Paik MC (2019) Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 5869–5879.

Kuzborskij I, Cella L, Cesa-Bianchi N (2018) Efficient linear bandits through matrix sketching. *arXiv preprint arXiv:1809.11033* .

Li L, Lu Y, Zhou D (2017) Provably optimal algorithms for generalized linear contextual bandits. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2071–2080 (JMLR. org).

Liu H, Yao T, Li R, Ye Y (2017) Folded concave penalized sparse linear regression: Sparsity, statistical performance, and algorithmic theory for local solutions. *Mathematical Programming* 1–34, URL `http://dx.doi.org/10.1007/s10107-017-1114-y`.

Liu H, Yao T, Li R, et al. (2016) Global solutions to folded concave penalized nonconvex learning. *The Annals of Statistics* 44(2):629–659.

Liu H, Ye Y, Lee HY (2022) High-dimensional learning under approximate sparsity with applications to nonsmooth estimation and regularized neural networks. *Operations Research* .

Loh PL, Wainwright MJ (2013) Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Advances in Neural Information Processing Systems*, 476–484.

McCullagh P, Nelder J (1989) *Generalized linear models* (Chapman and Hall/CRC).

Meinshausen N (2007) Relaxed lasso. *Computational Statistics & Data Analysis* 52(1):374–393.

Meinshausen N, Bühlmann P, et al. (2006) High-dimensional graphs and variable selection with the lasso. *The annals of statistics* 34(3):1436–1462.

Meinshausen N, Yu B, et al. (2009) Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* 37(1):246–270.

Mitchell J (2012) How google search really works. `https://readwrite.com/2012/02/29/interview_changing_engines_mid-flight_qa_with_goog/#awesm=~oiNkM4tAX3xhbP`, accessed: Oct 22nd, 2018.

Montgomery DC, Peck EA, Vining GG (2012) *Introduction to linear regression analysis*, volume 821 (John Wiley & Sons).

Negahban S, Yu B, Wainwright MJ, Ravikumar PK (2009) A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. *Advances in Neural Information Processing Systems*, 1348–1356.

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

OxfordDictionaries (2018) How many words are there in the english language? `https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language/`, accessed: Oct 22nd, 2018.

Ren Z, Zhou Z (2020) Dynamic batch learning in high-dimensional sparse linear contextual bandits. *arXiv preprint arXiv:2008.11918* .

Rigollet P, Zeevi A (2010) Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630* .

Robbins H (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.

Rusmevichientong P, Tsitsiklis JN (2010) Linearly parameterized bandits. *Mathematics of Operations Research* 35(2):395–411.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39(4):1221–1243.

Shewan D (2017) The comprehensive guide to online advertising costs. https://www.wordstream.com/blog/ws/2017/07/05/online-advertising-costs, accessed: Oct 22nd, 2018.

Slivkins A (2014) Contextual bandits with similarity information. *The Journal of Machine Learning Research* 15(1):2533–2568.

Tencent (2012) Predict the click-through rate of ads given the query and user information. https://www.kaggle.com/c/kddcup2012-track2, accessed: Oct 22nd, 2018.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tsybakov AB, et al. (2004) Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* 32(1):135–166.

Van de Geer SA, et al. (2008) High-dimensional generalized linear models and the lasso. *The Annals of Statistics* 36(2):614–645.

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

Wang Y, Chen Y, Fang EX, Wang Z, Li R (2020) Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection. *arXiv preprint arXiv:2009.02003* .

Wang Z, Liu H, Zhang T (2014) Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics* 42(6):2164.

WordStream (2017) Average ctr (click-through rate): Learn how your ctr compares. https://www.wordstream.com/average-ctr, accessed: Oct 22nd, 2018.

Yu X, Lyu MR, King I (2017) Cbrap: Contextual bandits with random projection. *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* 38(2):894–942.

Zhang CH, Huang J, et al. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* 36(4):1567–1594.

Zhang CH, Zhang T, et al. (2012) A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* 27(4):576–593.

Zhao T, Liu H, Zhang T (2014) Pathwise coordinate optimization for sparse learning: Algorithm and theory. *arXiv preprint arXiv:1412.7477* .

Zhao T, Liu H, Zhang T, et al. (2018) Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics* 46(1):180–218.

Zou H (2006) The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476):1418–1429.

# Electronic Companion to "Online Learning and Decision-Making under Generalized Linear Model with High-Dimensional Data" by Xue Wang, Mike Mingcheng Wei, Tao Yao

| Parameters | Explanation |
|---|---|
| $t$ and $T$ | Time indexes. |
| $\mathcal{K}$ | The decision set: $\mathcal{K}=\{1,2,...,K\}$. |
| $R_{k,t}$ and $R_i$ | The reward, where $k\in\mathcal{K}$, $t=1,2,...,T$, and $i=1,2,...,n$. |
| $\boldsymbol{X}_t$, $\boldsymbol{x}$, $\boldsymbol{x}_t$ | The covariates vectors, where $\boldsymbol{X}_t,\boldsymbol{x},\boldsymbol{x}_t\in\mathbb{R}^d$, and $t=1,2,...,T$. |
| $d$, $s$ | The dimension of total covariates and the dimension of significant covariates. |
| $\boldsymbol{\beta}_k^{\text{true}}$ | User's true parameter vector corresponding to arm/decision $k$. |
| $f(\cdot)$, $\mathcal{L}(\cdot)$ | The sample-wise loss function and the negative log-likelihood loss function. |
| $f_y^{'}(\cdot\|y)$, $f_{yy}^{''}(\cdot\|y)$ | The first and second order partial derivatives of $f(\cdot\|y)$ with respect to $y$. |
| $x_{\max}$, $R_{\max}$, $b$ | Positive constants that bound parameters defined in assumption A.1. |
| $C$ | A positive constant defined in assumption A.2. |
| $\mathcal{K}_o$, $\mathcal{K}_s$ | The optimal and suboptimal decision sets defined in assumption A.3. |
| $U_k$ | A subset of users' covariates defined in assumption A.3, where $k\in\mathcal{K}$. |
| $h$, $p^*$ | Positive constants defined in assumption A.3. |
| $\sigma$, $\sigma_2$ | Positive constants defined in assumption A.4. |
| $\kappa$ | The restricted eigenvalue constant defined in assumption A.5. |
| $\boldsymbol{\beta}^{\text{oracle}}$, $\boldsymbol{\beta}^{\text{lasso}}$, $\boldsymbol{\beta}^{\text{W}}$ | The oracle, Lasso, and weighted Lasso estimators. |
| $\boldsymbol{\pi}$ | The decision-makers' policy: $\boldsymbol{\pi}=\{\pi_t\}_{t\geq 1}$, where $\pi_t\in\mathcal{K}$ is the decision prescribed by policy $\boldsymbol{\pi}$ at time $t$. |
| $R^C(T)$ | The cumulative regret up to time $T$. |
| $\mathcal{A}$ | The sample set that contains only i.i.d. samples out of the whole sample set. |
| $\boldsymbol{w}$ | Non-negative weights vector for weighted Lasso in Eq. (6), $\boldsymbol{w}=(w_1,w_2,...,w_d)$. |
| $P_{\lambda,a}(x)$ | The MCP penalty function with positive parameters $a$ and $\lambda$. |
| $\boldsymbol{\beta}^{\text{MCP}},\boldsymbol{\beta}^{\text{random}},\boldsymbol{\beta}^{\text{whole}}$ | The MCP estimator, the MCP estimator under the random sample set $\mathcal{R}$, and the MCP estimator under the whole sample set $\mathcal{W}$. |
| $a,\lambda_1,\lambda_{2,0},t_0$ | Input parameters for the G-MCP-Bandit algorithm. |
| $\delta_0(t,t_0)$ | $\delta_0(t,t_0)\doteq 2((t_0+1)/(e(t+1)))^4$. |
| $\delta_1(n,\|\mathcal{A}\|,\zeta)$ | $\delta_1(n,\|\mathcal{A}\|,\zeta)\doteq 2s\exp(-\frac{n\zeta^2}{2\sigma^2x_{\max}^2})+\exp(-C_1\|\mathcal{A}\|)$. |
| $\delta_2(n,\|\mathcal{A}\|,\lambda)$ | $\delta_2(n,\|\mathcal{A}\|,\lambda)\doteq 4d\exp(-\frac{n\lambda^2}{2\sigma^2x_{\max}^2}\cdot(\frac{1}{2}-\frac{18ns}{\|\mathcal{A}\|\kappa a})^2)$. |
| $\mathcal{S}_1$, $\rho_{\mathcal{S}/\mathcal{S}_1}^{\text{MCP}}$ | Terms defined for Proposition 1: $\mathcal{S}_1\doteq\{i:\|\beta_i^{\text{true}}\|\geq(\frac{24ns}{\|\mathcal{A}\|\kappa}+a)\lambda\}$; $\rho_{\mathcal{S}/\mathcal{S}_1}^{\text{MCP}}\doteq\|\boldsymbol{\beta}_{\mathcal{S}/\mathcal{S}_1}^{\text{MCP}}-\boldsymbol{\beta}_{\mathcal{S}/\mathcal{S}_1}^{\text{true}}\|_1/\|\boldsymbol{\beta}_{\mathcal{S}}^{\text{MCP}}-\boldsymbol{\beta}_{\mathcal{S}}^{\text{true}}\|_1$ if $\mathcal{S}_1\neq\mathcal{S}$ and 0 otherwise. |
| $\mathcal{R}_{\boldsymbol{x},k}$ | The set contains the user covariate $\boldsymbol{X}$ generated by the $\epsilon$-decay random sampling method for arm $k$. |
| $\mathcal{F}_i$ | A filtration defined as $\mathcal{F}_i=\{(\boldsymbol{X}_j,R_j)\text{ for }j\leq i\}$. |
| $C_0$ | Defined in the proof of Proposition 4 and used in Proposition 2, 4-6, and Theorem 1; its dependence on $T$, $d$, and $s$ is $C_0=\mathcal{O}(s^2\log d)$. |
| $C_1$ | Defined in the statements of Lemma EC.1 and EC.2; $C_1=\mathcal{O}(s^{-2})$. |
| $T_0,T_1$ | Defined in the proof of Proposition 6 and used in Proposition 6 and Theorem 1; $T_0=\tilde{\mathcal{O}}(s^2\log d)$ and $T_1=\tilde{\mathcal{O}}(\beta_{\min}^{-2}\cdot s^2\log d)$, where $\beta_{\min}=\min_{i\in\mathcal{S}^k,k\in\mathcal{K}}\|\beta_{k,i}^{\text{true}}\|$. |
| $C_\rho,C_3,C_4,C_5$ | Defined in the proof of Theorem 1; $C_\rho=\mathcal{O}(1)$, $C_3\leq\tilde{\mathcal{O}}((1+\rho_{\max})^2s^2\log d)$, $C_4=\mathcal{O}(1)$, and $C_5=\mathcal{O}(s^2)$. |
| $\mathcal{E}_0,\mathcal{E}_1,...,\mathcal{E}_{5,(i,j),t}(w)$ | Series of events. |
| $\{M_k(i)\}$ | Martingale sequences used in the proof of Proposition 5. |

## EC.1. Appendix: Main Proofs

To simplify the notation in the E-companion, we denote $\nabla_{\mathcal{B}} F(\boldsymbol{x})$ as the vector with elements $(\nabla F(\boldsymbol{x}))_i,\ i \in \mathcal{B}$, where $(\cdot)_i$ is the $i$-th element in the vector. Similarly, we denote $\nabla^2_{\mathcal{B},\mathcal{C}} F(\boldsymbol{x})$ as the matrix with elements $(\nabla^2 F(\boldsymbol{x}))_{ij},\ i \in \mathcal{B}, j \in \mathcal{C}$, where $(\cdot)_{ij}$ is the element in $i$-th column and $j$-th row. To prove the main lemma, propositions, and theorems in this section, we need four additional technical lemmas (i.e., Lemma EC.1 to Lemma EC.4), whose statements and proofs are given in §EC.2 of this E-Companion. For notational convenience, we will omit parameters' subscripts corresponding to the choice of arms, as long as doing so will not cause any misinterpretation.

**Proof of Lemma 1** From the optimality condition of Eq. (3) and $\boldsymbol{\beta}^{\text{oracle}}$ being the optimal solution, we know that

$$\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{oracle}}) = \mathbf{0}. \tag{EC.1}$$

Expanding $\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta})$ in (EC.1) at $\boldsymbol{\beta}^{\text{true}}$, we can show that via the mean value theorem, for some $\boldsymbol{\xi} \in \{\tau \boldsymbol{\beta}^{\text{oracle}} + (1-\tau)\boldsymbol{\beta}^{\text{true}}, \tau \in [0,1]\}$, the following result holds:

$$\begin{aligned}
&\nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\xi})(\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}) = \mathbf{0} - \nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) \\
\Rightarrow\ &(\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}})^\top \nabla^2_{\mathcal{S},\mathcal{S}} \mathcal{L}(\boldsymbol{\xi})(\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}) = -(\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}})^\top \nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) \\
\Rightarrow\ &\boldsymbol{u}^\top \nabla^2 \mathcal{L}(\boldsymbol{\xi}) \boldsymbol{u} = -(\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}})^\top \nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}}),
\end{aligned} \tag{EC.2}$$

where in (EC.2) we denote $\boldsymbol{u} = \boldsymbol{\beta}^{\text{oracle}} - \boldsymbol{\beta}^{\text{true}}$ and use the fact $\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}^c} = \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^c} = \mathbf{0}$ to expend the left-hand side to $d$ dimensional space. By the definition of $\boldsymbol{\beta}^{\text{oracle}}$ and $\boldsymbol{\beta}^{\text{true}}$, it is direct to show that $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 = 0 \le 3\|\boldsymbol{u}_{\mathcal{S}}\|_1$. From Lemma EC.1, we know that when $|\mathcal{A}| \ge C_1^{-1} \log s$, the following inequality holds with probability at least $1 - \exp(-C_1|\mathcal{A}|)$:

$$\frac{|\mathcal{A}|\kappa}{2ns} \|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \le \boldsymbol{u}^\top \nabla^2 \mathcal{L}(\boldsymbol{\xi}) \boldsymbol{u}. \tag{EC.3}$$

Combining (EC.2) and (EC.3), we have:

$$\begin{aligned}
&\frac{|\mathcal{A}|\kappa}{2ns} \|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \le -(\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}})^\top \nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) \\
\Rightarrow\ &\frac{|\mathcal{A}|\kappa}{2ns} \|\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1^2 \le \|\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1 \|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty \\
\Rightarrow\ &\|\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1 \le \frac{2ns}{|\mathcal{A}|\kappa} \|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty.
\end{aligned} \tag{EC.4}$$

To obtain an upper bound for $\|\boldsymbol{\beta}^{\text{oracle}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1$, we need to show that $\|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty$ is also upper bounded.

- **Upper bound for $\|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty$:**

From the definition of $\mathcal{L}(\cdot)$, we have

$$\|\nabla_{\mathcal{S}} \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty = \left\| \frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{X}_{j,\mathcal{S}})^\top f_y'(R_j | \boldsymbol{X}_j^\top \boldsymbol{\beta}^{\text{true}}) \right\|_\infty, \tag{EC.5}$$

where we replace $r$ and $y$ in $f_y'(r|y)$ by $R_j$ and $\boldsymbol{X}_j^\top \boldsymbol{\beta}^{\text{true}}$ respectively, and $\boldsymbol{X}_{j,\mathcal{S}}$ is the subvector of $\boldsymbol{X}_j$ with elements in $\mathcal{S}$. Under assumption A.4, $f_y'(R_j|\boldsymbol{X}_j^\top \boldsymbol{\beta}^{\text{true}})$ is a $\sigma^2$-sub-gaussian random variable. From the Hoeffding inequality (see Proposition 2.5 in Wainwright 2019), for $\zeta > 0$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}(X_{j,i})f_y'(R_j|\boldsymbol{X}_j^\top \boldsymbol{\beta}^{\text{true}})\right| \geq \zeta\right) \leq 2\exp\left(-\frac{n\zeta^2}{2\sigma^2 x_{\max}^2}\right) \quad \forall i \in \mathcal{S}, \tag{EC.6}$$

where the right-hand side uses the fact that all realization $\|\boldsymbol{x}_j\|_\infty \leq x_{\max}$ in assumption A.1. Hence, via union bound, we can show that

$$\begin{aligned}
\mathbb{P}\left(\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty \geq \zeta\right) &= \mathbb{P}\left(\left\|\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{X}_{j,\mathcal{S}})^\top f_y'(R_j|\boldsymbol{X}_j^\top \boldsymbol{\beta}^{\text{true}})\right\|_\infty \geq \zeta\right) \\
&\leq \sum_{i\in\mathcal{S}}\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}X_{j,i}f_y'(R_j|\boldsymbol{X}_j^\top \boldsymbol{\beta}^{\text{true}})\right| \geq \zeta\right) \\
&\leq 2s\exp\left(-\frac{n\zeta^2}{2\sigma^2 x_{\max}^2}\right),
\end{aligned} \tag{EC.7}$$

where the last inequality in (EC.7) follows from $|\mathcal{S}| \leq s$. At last, the lemma follows directly by combining (EC.4) and (EC.7).

**Proof of Proposition 1** Proposition 1 directly follows Proposition 3 by setting $|\mathcal{A}| = n$.

**Proof of Proposition 2** Under the $\epsilon$-decay random sampling method, the probability of randomly drawn arm $k$ at time $t$ is $\min\{1, t_0/t\}/|\mathcal{K}|$, where $|\mathcal{K}|$ is the number of arms. Hence, at time $T$, the expected total number of times at which arm $k$ was randomly drawn is

$$\mathbb{E}[n_k] = \frac{1}{|\mathcal{K}|}\sum_{t=1}^{T}\min\left\{1, \frac{t_0}{t}\right\},$$

where the expectation is taken with respect to $n_k$, the total number of random samples.

When $T > t_0$,

$$\mathbb{E}[n_k] = \frac{1}{|\mathcal{K}|}\left(t_0 + \sum_{t=t_0+1}^{T}\frac{t_0}{t}\right) = \frac{t_0}{|\mathcal{K}|}\left(1 + \sum_{t=t_0+1}^{T}\frac{1}{t}\right). \tag{EC.8}$$

Since the function $f(t) = 1/t$ is decreasing in $t$, it can be bounded as follows

$$\int_t^{t+1}\frac{1}{x}dx < \frac{1}{t} < \int_{t-1}^{t}\frac{1}{x}dx, \ t \geq 2.$$

As $C_0 \geq 20$, we can verify that $t_0 = 2C_0|\mathcal{K}| \geq 2$. Hence, for any $t$ from $t_0+1$ to $T$, we have

$$\log(T+1) - \log(t_0+1) < \sum_{t=t_0+1}^{T}\frac{1}{t} < \log(T) - \log(t_0). \tag{EC.9}$$

Combining (EC.8) and (EC.9), we can bound $\mathbb{E}[n_k]$ as follows:

$$\frac{1}{|\mathcal{K}|}t_0(1 + \log(T+1) - \log(t_0+1)) < \mathbb{E}[n_k] < \frac{1}{|\mathcal{K}|}t_0(1 + \log(T) - \log(t_0)). \tag{EC.10}$$

Since $n_k = \sum_{t=1}^{T}\mathbb{1}\{\text{random sampling for arm } k \text{ at time } t\}$, we can view $n_k$ as the summation of bounded i.i.d. random variables. By Chernoff bound (see Theorem 4 in Goemans 2015 by setting $\delta = 0.5$), we can

have the following inequality:

$$\mathbb{P}\left(\frac{1}{2}\mathbb{E}[n_k] \le n_k \le \frac{3}{2}\mathbb{E}[n_k]\right) \ge 1 - 2\exp\left(-\frac{1}{10}\mathbb{E}[n_k]\right). \tag{EC.11}$$

We then relax the $\mathbb{E}[n_k]$ in (EC.11) by using the upper and lower bounds provided in (EC.10) to attain the following result:

$$\mathbb{P}\left(\frac{t_0(1+\log(T+1)-\log(t_0+1))}{2|\mathcal{K}|} \le n_k \le \frac{3t_0(1+\log(T)-\log(t_0))}{2|\mathcal{K}|}\right) \quad \ge 1 - 2\left(\frac{t_0+1}{e(T+1)}\right)^{\frac{t_0}{10|\mathcal{K}|}}. \tag{EC.12}$$

When $t_0 = 2C_0|\mathcal{K}|$ and $C_0 \ge 20$, we have $\frac{t_0}{10|\mathcal{K}|} = C_0/5 \ge 4$. Then, this proposition follows directly by plugging $t_0 = 2C_0|\mathcal{K}|$ back into (EC.12) and using the definition of $\delta_0(T, t_0)$ in the proposition statement.

**Proof of Proposition 3** In the first step of the 2sWL procedure, we solve a Lasso problem. From Lemma EC.2, we know that if $|\mathcal{A}| \ge C_1^{-1}\log d$, then the inequality $\|\boldsymbol{\beta}^{\text{lasso}} - \boldsymbol{\beta}^{\text{true}}\|_1 \le \frac{24ns\lambda}{|\mathcal{A}|\kappa}$ holds with high probability. Beside set $\mathcal{S}_1$ defined in (9), let's consider the following index set:

$$\mathcal{S}_2 \doteq \left\{i : |\beta_i^{\text{true}}| < \left(\frac{24ns}{|\mathcal{A}|\kappa} + a\right)\lambda, i \in \mathcal{S}\right\}. \tag{EC.13}$$

Directly, we can show that

$$\begin{aligned} &i \in \mathcal{S}_1 \Rightarrow |\beta_i^{\text{lasso}}| \ge a\lambda \text{ so that } w_i = P'_{\lambda,a}(|\beta_i^{\text{lasso}}|) = 0; \\ &i \in \mathcal{S}_2 \Rightarrow |\beta_i^{\text{lasso}}| \le \left(\frac{48ns}{|\mathcal{A}|\kappa} + a\right)\lambda \text{ and } w_i = P'_{\lambda,a}(|\beta_i^{\text{lasso}}|) \le \lambda, \end{aligned} \tag{EC.14}$$

where we use the fact that for all $x \ge 0$

$$P'_{\lambda,a}(x) = \max\left(0, \lambda - \frac{x}{a}\right) \tag{EC.15}$$

per definition of MCP penalty in (7). Similarly, for $i \in \mathcal{S}^{\text{c}} = \{i : |\beta_i^{\text{true}}| = 0, i \in \{1, 2, ..., d\}\}$, we can show that

$$i \in \mathcal{S}^{\text{c}} \Rightarrow |\beta_i^{\text{lasso}}| \le \frac{24ns}{|\mathcal{A}|\kappa}\lambda \text{ and } w_i = P'_{\lambda,a}(|\beta_i^{\text{lasso}}|) \ge \left(1 - \frac{24ns}{|\mathcal{A}|\kappa a}\right)\lambda, \tag{EC.16}$$

where the last inequality uses $1 - \frac{24ns}{|\mathcal{A}|\kappa a} > 0$ for $a > \frac{48ns}{|\mathcal{A}|\kappa}$.

Let $\boldsymbol{\beta}^{\text{MCP}}$ be the optimal solution to the second step of the 2sWL procedure. Using the fact that $\mathcal{L}(\boldsymbol{\beta}) + \sum_{j=1}^d w_j|\beta_j|$ is minimized at $\boldsymbol{\beta}^{\text{MCP}}$ and the fact that $\mathcal{L}(\boldsymbol{\beta})$ is convex, we have

$$\mathcal{L}(\boldsymbol{\beta}^{\text{MCP}}) + \sum_{j=1}^d w_j|\beta_j^{\text{MCP}}| \le \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) + \sum_{j=1}^d w_j|\beta_j^{\text{true}}| \tag{EC.17}$$

$$\Rightarrow \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) + \nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top(\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}) + \sum_{j=1}^d w_j|\beta_j^{\text{MCP}}| \le \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) + \sum_{j=1}^d w_j|\beta_j^{\text{true}}|$$

$$\Rightarrow \nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top(\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}) + \sum_{j=1}^d w_j|\beta_j^{\text{MCP}}| \le \sum_{j=1}^d w_j|\beta_j^{\text{true}}|$$

$$\Rightarrow \nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top(\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}) + \sum_{j\in\mathcal{S}_2} w_j|\beta_j^{\text{MCP}}| + \sum_{j\in\mathcal{S}^c} w_j|\beta_j^{\text{MCP}}| \le \sum_{j\in\mathcal{S}_2} w_j|\beta_j^{\text{true}}| \tag{EC.18}$$

$$\Rightarrow \nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top(\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}) + \sum_{j\in\mathcal{S}^c} w_j|\beta_j^{\text{MCP}} - \beta_j^{\text{true}}| \le \sum_{j\in\mathcal{S}_2} w_j|\beta_j^{\text{MCP}} - \beta_j^{\text{true}}|. \tag{EC.19}$$

where (EC.18) uses the observations that $w_i = 0$ for $i \in \mathcal{S}_1$ and $\beta_i^{\text{true}} = 0$ for $i \in \mathcal{S}^{\text{c}}$, and (EC.19) uses the observation that $\beta_i^{\text{true}} = 0$ for $i \in \mathcal{S}^{\text{c}}$.

Let $\boldsymbol{u} = \boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}$. Then, inequality (EC.19) can be further simplified as follows:

$$\nabla \mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top \boldsymbol{u} + \sum_{j \in \mathcal{S}^c} w_j |u_j| \leq \sum_{j \in \mathcal{S}_2} w_j |u_j| \tag{EC.20}$$

$$\Rightarrow \sum_{j \in \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}^c} \nabla_j \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) u_j + \sum_{j \in \mathcal{S}^c} w_j |u_j| \leq \sum_{j \in \mathcal{S}_2} w_j |u_j|$$

$$\Rightarrow \sum_{j \in \mathcal{S}^c} (w_j - |\nabla_j \mathcal{L}(\boldsymbol{\beta}^{\text{true}})|) |u_j| \leq \sum_{j \in \mathcal{S}_2} (w_j + |\nabla_j \mathcal{L}(\boldsymbol{\beta}^{\text{true}})|) |u_j| + \sum_{j \in \mathcal{S}_1} |\nabla_j \mathcal{L}(\boldsymbol{\beta}^{\text{true}})| |u_j|$$

$$\Rightarrow (\tilde{w}_c - \|\nabla_{\mathcal{S}^c} \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty) \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq (\tilde{w}_2 + \|\nabla_{\mathcal{S}_2} \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty) \|\boldsymbol{u}_{\mathcal{S}_2}\|_1 + \|\nabla_{\mathcal{S}_1} \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty \|\boldsymbol{u}_{\mathcal{S}_1}\|_1, \tag{EC.21}$$

where we define two positive constants, $\tilde{w}_c$ and $\tilde{w}_2$, as follows:

$$\tilde{w}_c \doteq \left(1 - \frac{24ns}{|\mathcal{A}|\kappa a}\right) \lambda \leq \min_{j \in \mathcal{S}^c} \{w_j\} \tag{EC.22}$$

and

$$\tilde{w}_2 \doteq \lambda \geq \max_{j \in \mathcal{S}_2} \{w_j\}, \tag{EC.23}$$

where the inequalities in (EC.22) and (EC.23) are from (EC.16) and (EC.14), respectively.

Now, we define the following event:

$$\mathcal{E}_{\text{sub},1} \doteq \left\{ \|\nabla \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty < \frac{3}{4}\tilde{w}_c - \frac{1}{4}\tilde{w}_2 \right\}. \tag{EC.24}$$

Then, under event $\mathcal{E}_{\text{sub},1}$, inequality (EC.21) implies

$$\left(\tilde{w}_c - \frac{3}{4}\tilde{w}_c + \frac{1}{4}\tilde{w}_2\right) \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq \left(\tilde{w}_2 + \frac{3}{4}\tilde{w}_c - \frac{1}{4}\tilde{w}_2\right) \|\boldsymbol{u}_{\mathcal{S}_2}\|_1 + \left(\frac{3}{4}\tilde{w}_c - \frac{1}{4}\tilde{w}_2\right) \|\boldsymbol{u}_{\mathcal{S}_1}\|_1$$

$$\Rightarrow \frac{1}{4}(\tilde{w}_c + \tilde{w}_2) \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq \frac{3}{4}(\tilde{w}_2 + \tilde{w}_c) \|\boldsymbol{u}_{\mathcal{S}_2}\|_1 + \frac{3}{4}(\tilde{w}_2 + \tilde{w}_c) \|\boldsymbol{u}_{\mathcal{S}_1}\|_1 - \tilde{w}_2 \|\boldsymbol{u}_{\mathcal{S}_1}\|_1$$

$$\Rightarrow (\tilde{w}_c + \tilde{w}_2) \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq 3(\tilde{w}_2 + \tilde{w}_c) \|\boldsymbol{u}_{\mathcal{S}}\|_1 - 4\tilde{w}_2 \|\boldsymbol{u}_{\mathcal{S}_1}\|_1$$

$$\Rightarrow \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq 3\|\boldsymbol{u}_{\mathcal{S}}\|_1 - \frac{4\tilde{w}_2}{\tilde{w}_c + \tilde{w}_2} \|\boldsymbol{u}_{\mathcal{S}_1}\|_1$$

$$\Rightarrow \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq 3\|\boldsymbol{u}_{\mathcal{S}}\|_1. \tag{EC.25}$$

Combining (EC.25) and Lemma EC.1, we can show that for all feasible $\boldsymbol{\xi}$, the following inequality holds:

$$\mathbb{P}\left( \frac{|\mathcal{A}|\kappa}{2ns} \|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq u^\top \nabla^2 \mathcal{L}(\boldsymbol{\xi}) \boldsymbol{u} \right) \geq 1 - \exp(-C_1 |\mathcal{A}|). \tag{EC.26}$$

Now, we go back to (EC.17) and expand the $\mathcal{L}(\boldsymbol{\beta})$ term in the left-hand side at $\boldsymbol{\beta}^{\text{true}}$. Denoting $\boldsymbol{u} = \boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}$, we can show that there exists a feasible $\boldsymbol{\xi}$ between $\boldsymbol{\beta}^{MCP}$ and $\boldsymbol{\beta}^{true}$ such that

$$\mathcal{L}(\boldsymbol{\beta}^{\text{true}}) + \nabla \mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top \boldsymbol{u} + \frac{1}{2} \boldsymbol{u}^\top \nabla^2 \mathcal{L}(\boldsymbol{\xi}) \boldsymbol{u} + \sum_{i=1}^d w_i |\beta_i^{\text{MCP}}| \leq \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) + \sum_{i=1}^d w_i |\beta_i^{\text{true}}|$$

$$\Rightarrow \nabla \mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top \boldsymbol{u} + \frac{1}{2} \boldsymbol{u}^\top \nabla^2 \mathcal{L}(\boldsymbol{\xi}) \boldsymbol{u} + \sum_{i=1}^d w_i |\beta_i^{\text{MCP}}| \leq \sum_{i=1}^d w_i |\beta_i^{\text{true}}|$$

$$\Rightarrow \nabla \mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top \boldsymbol{u} + \frac{|\mathcal{A}|\kappa}{4ns} \|\boldsymbol{u}_{\mathcal{S}}\|_1^2 + \sum_{i=1}^d w_i |\beta_i^{\text{MCP}}| \leq \sum_{i=1}^d w_i |\beta_i^{\text{true}}| \tag{EC.27}$$

$$\Rightarrow \frac{|\mathcal{A}|\kappa}{4ns} \|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \sum_{i=1}^d \left( -\nabla_i \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) u_i + w_i (|\beta_i^{\text{true}}| - |\beta_i^{\text{MCP}}|) \right)$$

$$\Rightarrow \frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \sum_{i\in\mathcal{S}_1\cup\mathcal{S}_2\cup\mathcal{S}^c}\left(-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})u_i + w_i(|\beta_i^{\text{true}}| - |\beta_i^{\text{MCP}}|)\right), \tag{EC.28}$$

where inequality (EC.27) uses (EC.26) and we defer the consideration of the probability part via union bound to (EC.34). Then, we can bound the right hand side of (EC.28) by considering $i\in\mathcal{S}_1,\mathcal{S}_2$ and $\mathcal{S}^c$ separately.

- $i\in\mathcal{S}_1$**:**

$$\begin{aligned}
&\sum_{i\in\mathcal{S}_1}\left(-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})u_i + w_i(|\beta_i^{\text{true}}| - |\beta_i^{\text{MCP}}|)\right)\\
\leq&\sum_{i\in\mathcal{S}_1}\left(\left|\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right| + w_i\right)|u_i|\\
=&\sum_{i\in\mathcal{S}_1}\left(\left|\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right|\right)|u_i|\\
\leq&\|\boldsymbol{u}_{\mathcal{S}_1}\|_1\|\nabla_{\mathcal{S}_1}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty,
\end{aligned} \tag{EC.29}$$

where the equality uses $w_i = 0$ for all $i\in\mathcal{S}_1$.

- $i\in\mathcal{S}_2$**:**

$$\begin{aligned}
&\sum_{i\in\mathcal{S}_2}\left(-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})u_i + w_i(|\beta_i^{\text{true}}| - |\beta_i^{\text{MCP}}|)\right)\\
\leq&\sum_{i\in\mathcal{S}_2}\left(\left|\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right| + w_i\right)|u_i|\\
\leq&\|\boldsymbol{u}_{\mathcal{S}_2}\|_1\left(\|\nabla_{\mathcal{S}_2}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty + \lambda\right),
\end{aligned} \tag{EC.30}$$

where the last inequality uses $w_i\leq\lambda$ for $i\in\mathcal{S}_2$.

- $i\in\mathcal{S}^c$**:**

$$\begin{aligned}
&\sum_{i\in\mathcal{S}^c}\left(-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})u_i + w_i(|\beta_i^{\text{true}}| - |\beta_i^{\text{MCP}}|)\right)\\
=&\sum_{i\in\mathcal{S}^c}\left(-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\beta_i^{\text{MCP}} - w_i|\beta_j^{\text{MCP}}|\right)\\
\leq&\sum_{i\in\mathcal{S}^c}\left(|\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})||\beta_i^{\text{MCP}}| - w_i|\beta_i^{\text{MCP}}|\right)\\
\leq&\sum_{i\in\mathcal{S}^c}\left(\frac{3}{4}\tilde{w}_c - \frac{1}{4}\tilde{w}_2 - w_i\right)|\beta_i^{\text{MCP}}|\\
\leq&0,
\end{aligned} \tag{EC.31}$$

where in the second-to-last inequality, we use the event $\mathcal{E}_{\text{sub},1}$ in (EC.24), and in the last inequality, we adopt the fact that $\tilde{w}_c\leq w_i$, $\tilde{w}_2>0$, and $w_i>0$ by definitions.

Then, we combine (EC.28), (EC.29), (EC.30) and (EC.31):

$$\begin{aligned}
&\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \|\boldsymbol{u}_{\mathcal{S}_1}\|_1\|\nabla_{\mathcal{S}_1}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty + \|\boldsymbol{u}_{\mathcal{S}_2}\|_1\left(\|\nabla_{\mathcal{S}_2}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty + \lambda\right)\\
\Rightarrow&\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \|\boldsymbol{u}_{\mathcal{S}}\|_1\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty + \lambda\|\boldsymbol{u}_{\mathcal{S}_2}\|_1\\
\Rightarrow&\|\boldsymbol{u}_{\mathcal{S}}\|_1 \leq \frac{4ns}{|\mathcal{A}|\kappa}\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty + \frac{4ns}{|\mathcal{A}|\kappa}\cdot\lambda\cdot\frac{\|\boldsymbol{u}_{\mathcal{S}_2}\|_1}{\|\boldsymbol{u}_{\mathcal{S}}\|_1}\\
\Rightarrow&\|\boldsymbol{u}\|_1 = \|\boldsymbol{u}_{\mathcal{S}}\|_1 + \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq 4\|\boldsymbol{u}_{\mathcal{S}}\|_1 \leq \frac{16ns}{|\mathcal{A}|\kappa}\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty + \frac{16ns}{|\mathcal{A}|\kappa}\cdot\lambda\cdot\frac{\|\boldsymbol{u}_{\mathcal{S}_2}\|_1}{\|\boldsymbol{u}_{\mathcal{S}}\|_1}
\end{aligned} \tag{EC.32}$$

$$\Rightarrow \|\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \frac{16ns}{|\mathcal{A}|\kappa}\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty + \frac{16ns}{|\mathcal{A}|\kappa} \cdot \frac{\|\boldsymbol{\beta}^{\text{MCP}}_{\mathcal{S}_2} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}_2}\|_1}{\|\boldsymbol{\beta}^{\text{MCP}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1} \cdot \lambda$$
$$\Rightarrow \|\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \frac{16ns}{|\mathcal{A}|\kappa}\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty + \frac{16ns}{|\mathcal{A}|\kappa} \cdot \rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1} \cdot \lambda,$$

where first inequality in (EC.32) applies (EC.25), and we use the definition of $\rho^{\text{MCP}}_{\mathcal{S}/\mathcal{S}_1}$ (i.e., Equation (12)) in the last inequality. Then, via (EC.7), we can bound $\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty$ as follows:

$$\mathbb{P}(\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty \leq \zeta) \geq 1 - 2s\exp\left(-\frac{n\zeta^2}{2\sigma^2 x^2_{\max}}\right). \tag{EC.33}$$

Next, we build the probability bound for event $\mathcal{E}_{\text{sub},1}$ in (EC.24) by the Hoeffding's inequality. For $t > 0$, from (EC.7), we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right\|_\infty \geq t\right) \leq 2d\exp\left(-\frac{nt^2}{2\sigma^2 x^2_{\max}}\right) \\
\Rightarrow &\mathbb{P}\left(\left\|\nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right\|_\infty \geq \frac{3}{4}\tilde{w}_c - \frac{1}{4}\tilde{w}_2\right) \leq 2d\exp\left(-\frac{n(\frac{3}{4}\tilde{w}_c - \frac{1}{4}\tilde{w}_2)^2}{2\sigma^2 x^2_{\max}}\right) \\
\Rightarrow &\mathbb{P}\left(\left\|\nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right\|_\infty \geq \left(\frac{1}{2} - \frac{18ns}{|\mathcal{A}|\kappa a}\right)\lambda\right) \leq 2d\exp\left(-\frac{n\lambda^2}{2\sigma^2 x^2_{\max}} \cdot \left(\frac{1}{2} - \frac{18ns}{|\mathcal{A}|\kappa a}\right)^2\right).
\end{aligned}$$

Combining this result with Lemma EC.2, we can show that if $|\mathcal{A}| \geq C_1^{-1}\log d$ and $a > \frac{48ns}{|\mathcal{A}|\kappa}$ (which also implies that $\frac{1}{2} > \frac{18ns}{|\mathcal{A}|\kappa a}$), the proposition statement holds with probability

$$\begin{aligned}
&1 - \exp(-C_1|\mathcal{A}|) - 2d\exp\left(-\frac{n\lambda^2}{8\sigma^2 x_{\max}}\right) - 2s\exp\left(-\frac{n\zeta^2}{2\sigma^2 x_{\max}}\right) - 2d\exp\left(-\frac{n\lambda^2}{2\sigma^2 x_{\max}} \cdot \left(\frac{1}{2} - \frac{18ns}{|\mathcal{A}|\kappa a}\right)^2\right) \\
\geq &1 - \exp(-C_1|\mathcal{A}|) - 4d\exp\left(-\frac{n\lambda^2}{2\sigma^2 x_{\max}} \cdot \left(\frac{1}{2} - \frac{18ns}{|\mathcal{A}|\kappa a}\right)^2\right) - 2s\exp\left(-\frac{n\zeta^2}{2\sigma^2 x^2_{\max}}\right),
\end{aligned} \tag{EC.34}$$

where the last inequality uses the fact that $\left(\frac{1}{2} - \frac{18ns}{|\mathcal{A}|\kappa a}\right)^2 \leq \frac{1}{4}$.

**Proof of Proposition 4** For clear expositions, we first state two constants that we will use in this proof:

$$C_0 \geq \max\left\{20, \frac{64}{p^*}, \frac{24\log d}{p^* C_1}, \frac{96}{p^* C_1}, \frac{3072\sigma^2 x^2_{\max}(1+\log d)}{\lambda^2}\right\} \tag{EC.35}$$

and

$$\lambda \leq \min\left\{\frac{h\kappa p^*}{3072 e\sigma s R_{\max} x_{\max}}, \frac{p^*\kappa}{768\sigma s x_{\max}}\right\}. \tag{EC.36}$$

As $C_1 = \mathcal{O}(s^{-2})$ per equation (EC.125), it is direct to verify that $C_0 = \mathcal{O}(s^2\log d)$. Further, we denote event $\mathcal{E}_3$ as follows:

$$\mathcal{E}_3 = \left\{\frac{|\mathcal{A}|}{n} \geq \frac{1}{24}p^*\right\}. \tag{EC.37}$$

Note that for the suboptimal arm set (i.e., $k \in \mathcal{K}_s$), event $\mathcal{E}_3$ holds automatically, as $|\mathcal{A}| = n$ for $k \in \mathcal{K}_s$ so that $|\mathcal{A}|/n > p^*/24$ always holds true; for the optimal arm set (i.e., $k \in \mathcal{K}_o$), $|\mathcal{A}|/n$ represents the proportion of covariate vectors $\boldsymbol{X}$ that are in the set $U_k$ (i.e., $\boldsymbol{X} \in U_k$) to all $n$ i.i.d. samples that are generated via the $\epsilon$-decay sampling scheme, and we will bound the probability for event $\mathcal{E}_3$ later in (EC.45).

Combining Proposition 2 and event $\mathcal{E}_3$, we can show that when $C_0 \geq \max\left\{20, \frac{24\log d}{p^* C_1}\right\}$, the following inequalities hold simultaneously with probability at least $1-\delta_0(t,t_0)$:

$$|\mathcal{A}| \geq \frac{p^*}{24} n \geq \frac{p^*}{24} C_0 (1+\log(t+1)-\log(t_0+1)) \geq \frac{p^*}{24} C_0 \geq C_1^{-1} \log d. \tag{EC.38}$$

In addition, given $\mathcal{E}_3$, we can show that under the condition $a > \frac{1152s}{p^*\kappa}$, the following inequality holds:

$$a > \frac{1152s}{p^*\kappa} \geq \frac{48ns}{\kappa|\mathcal{A}|}. \tag{EC.39}$$

Hence, with (EC.38) and (EC.39), it is direct to show that the inequality (16) in Proposition 3 holds: for $\zeta > 0$, we have the following inequality:

$$\begin{aligned} &\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{random}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \frac{16ns\zeta}{|\mathcal{A}|\kappa} + \frac{16ns\rho^{\text{random}}_{\mathcal{S}/\mathcal{S}_1}}{|\mathcal{A}|\kappa}\lambda\right) \geq 1-\delta_1(n,|\mathcal{A}|,\zeta)-\delta_2(n,|\mathcal{A}|,\lambda) \\ \Rightarrow &\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{random}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \frac{32ns}{|\mathcal{A}|\kappa}\lambda\right) \geq 1-\delta_1(n,|\mathcal{A}|,\lambda)-\delta_2(n,|\mathcal{A}|,\lambda), \end{aligned} \tag{EC.40}$$

where in (EC.40), we set $\zeta = \lambda$ and use $\rho^{\text{random}}_{\mathcal{S}/\mathcal{S}_1} \leq 1$.

Combining event $\mathcal{E}_3$ and Proposition 2, we can show that with probability at least $1-\delta_0(t,t_0)$, the following results hold

$$n \geq C_0(1+\log(t+1)-\log(t_0+1)) \text{ and } |\mathcal{A}| \geq \frac{p^*}{24} n \geq \frac{p^*}{24} C_0(1+\log(t+1)-\log(t_0+1)). \tag{EC.41}$$

Then, we can further simplify (EC.41) as follows:

$$n \geq C_0 \log\left(\frac{e(t+1)}{t_0+1}\right) \text{ and } |\mathcal{A}| \geq \frac{p^*}{24} C_0 \log\left(\frac{e(t+1)}{t_0+1}\right). \tag{EC.42}$$

Now, if we set $C_0 \geq \max\left\{\frac{96}{p^* C_1}, \frac{3072\sigma^2 x^2_{\max}(1+\log d)}{\lambda^2}\right\}$, then we can directly verify that $\delta_1(n,|\mathcal{A}|,\lambda) = 2s\exp\left(-\frac{n\lambda^2}{2\sigma^2 x^2_{\max}}\right) + \exp(-C_1|\mathcal{A}|) \leq \delta_0(t,t_0) + \frac{1}{2}\delta_0(t,t_0)$ and $\delta_2(n,|\mathcal{A}|,\lambda) \leq 2\delta_0(t,t_0)$, combining which we have the following result:

$$\delta_1(n,|\mathcal{A}|,\lambda) + \delta_2(n,|\mathcal{A}|,\lambda) \leq \frac{7}{2}\delta_0(t,t_0). \tag{EC.43}$$

Next, we need to bound the probability for event $\mathcal{E}_3$ for $k \in \mathcal{K}_o$. First, we can show that

$$\begin{aligned} \left\{\frac{|\mathcal{A}|}{n} \geq \frac{1}{24}p^*\right\} &\supseteq \left\{|\mathcal{A}| \geq \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\} \cap \left\{n \leq 6C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\} \\ &= \left(\left\{|\mathcal{A}| < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\} \cup \left\{n > 6C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\}\right)^{\text{c}}, \end{aligned} \tag{EC.44}$$

which infers that for $k \in \mathcal{K}_o$,

$$\begin{aligned} \mathbb{P}\left\{\frac{|\mathcal{A}|}{n} \geq \frac{1}{24}p^*\right\} &\geq \mathbb{P}\left\{\left(\left\{|\mathcal{A}| < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\} \cup \left\{n > 6C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\}\right)^{\text{c}}\right\} \\ &= 1 - \mathbb{P}\left\{\left\{|\mathcal{A}| < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\} \cup \left\{n > 6C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\}\right\} \\ &\geq 1 - \mathbb{P}\left\{|\mathcal{A}| < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\} - \mathbb{P}\left\{n > 6C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\}. \end{aligned} \tag{EC.45}$$

Now, we will separately consider bounds for $\mathbb{P}\left\{|\mathcal{A}| < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\}$ and $\mathbb{P}\left\{n > 6C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right\}$.

• **The probability bound for $n > 6C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)$:**

From Proposition 2, when $t \geq t_0$, the following result holds with probability $1-\delta_0(t,t_0)$:

$$\begin{aligned} n \leq 3C_0(1+\log(t)-\log(t_0)) &= 3C_0 \log\left(\frac{et}{t_0}\right) \\ &< 3C_0 \log\left(\frac{2e(t+1)}{2t_0}\right) \\ &< 3C_0 \log\left(\frac{2e(t+1)}{t_0+1}\right) \\ &= 3C_0 \log\left(\frac{e(t+1)}{t_0+1}\right) + 3C_0 \log(2) \\ &< 6C_0 \log\left(\frac{e(t+1)}{t_0+1}\right), \end{aligned} \tag{EC.46}$$

where the last inequality uses $2 < e < \frac{e(t+1)}{t_0+1}$.

• **The probability bound for $|\mathcal{A}| < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)$:**

By Proposition 2 and assumption A.3, we can show that the expected number of i.i.d. samples belong to $U_k$ for $k \in \mathcal{K}$ is lower bounded with high probability by

$$\begin{aligned} \mathbb{E}_{\boldsymbol{X}}\left[\sum_{i=1}^{t} \mathbb{1}(\boldsymbol{X}_i \in U_k)\right] &\geq p^* C_0(1+\log(t+1)-\log(t_0+1)) \\ &> \frac{1}{2}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right). \end{aligned} \tag{EC.47}$$

Then, we apply the Chernoff inequality (similar to the analysis for (EC.11)) on $\sum_{i=1}^{n} \mathbb{1}(x_i \in U_k)$:

$$\begin{aligned} &\mathbb{P}\left(\sum_{i=1}^{t} \mathbb{1}(\boldsymbol{X}_i \in U_k) < \frac{1}{2}\mathbb{E}_{\boldsymbol{X}}\left[\sum_{i=1}^{t} \mathbb{1}(\boldsymbol{X}_i \in U_k)\right]\right) \leq \exp\left(-\frac{1}{8}\mathbb{E}_{\boldsymbol{X}}\left[\sum_{i=1}^{t} \mathbb{1}(\boldsymbol{X}_i \in U_k)\right]\right) \\ \Rightarrow &\mathbb{P}\left(\sum_{i=1}^{t} \mathbb{1}(\boldsymbol{X}_i \in U_k) < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right) \leq \exp\left(-\frac{1}{16}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right). \end{aligned} \tag{EC.48}$$

When $C_0 \geq 64/p^*$, (EC.48) can be further simplified as follows:

$$\begin{aligned} &\mathbb{P}\left(\sum_{i=1}^{t} \mathbb{1}(\boldsymbol{X}_i \in U_k) < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right) \leq \frac{(t_0+1)^4}{e^4(t+1)^4} \\ \Rightarrow &\mathbb{P}\left(|\mathcal{A}| < \frac{1}{4}p^* C_0 \log\left(\frac{e(t+1)}{t_0+1}\right)\right) \leq \frac{(t_0+1)^4}{e^4(t+1)^4} = \frac{1}{2}\delta_0(t,t_0). \end{aligned} \tag{EC.49}$$

Having proved these two probability bounds, we can combine (EC.45), (EC.46), and (EC.49) to show that

$$\mathbb{P}\{\mathcal{E}_3 | k \in \mathcal{K}_o\} \geq 1 - \frac{3}{2}\delta_0(t,t_0),$$

which implies that

$$\begin{aligned} \mathbb{P}\{\mathcal{E}_3\} &= \mathbb{P}\{\mathcal{E}_3 | k \in \mathcal{K}_s\}\mathbb{P}\{\mathcal{K}_s\} + \mathbb{P}\{\mathcal{E}_3 | k \in \mathcal{K}_o\}\mathbb{P}\{\mathcal{K}_o\} \\ &\geq \mathbb{P}\{\mathcal{E}_3 | k \in \mathcal{K}_o\}\mathbb{P}\{\mathcal{K}_s\} + \mathbb{P}\{\mathcal{E}_3 | k \in \mathcal{K}_o\}\mathbb{P}\{\mathcal{K}_o\} \\ &= \mathbb{P}\{\mathcal{E}_3 | k \in \mathcal{K}_o\} \geq 1 - \frac{3}{2}\delta_0(t,t_0), \end{aligned} \tag{EC.50}$$

where the first inequality uses the fact that $\mathbb{P}\{\mathcal{E}_3|k\in\mathcal{K}_s\}=1\geq\mathbb{P}\{\mathcal{E}_3|k\in\mathcal{K}_o\}$. Finally, combining (EC.40), (EC.43), and (EC.50), via union bound, we have

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{random}}-\boldsymbol{\beta}^{\text{true}}\|_1\leq\frac{32ns}{|\mathcal{A}|\kappa}\lambda\right)\geq 1-5\delta_0(t,t_0). \tag{EC.51}$$

Moreover, if we pick $\lambda$ to be small enough (e.g., $\lambda\leq\min\left\{\frac{h\kappa p^*}{3072e\sigma s R_{\max}x_{\max}},\frac{p^*\kappa}{768\sigma s x_{\max}}\right\}$), then when event $\mathcal{E}_3$ holds, we have the following two results:

$$\frac{32ns\lambda}{|\mathcal{A}|\kappa}\leq\frac{32ns\cdot hp^*\kappa}{3072e\sigma s R_{\max}x_{\max}|\mathcal{A}|\kappa}=\frac{h}{4e\sigma R_{\max}x_{\max}}\cdot\frac{n}{|\mathcal{A}|}\cdot\frac{p^*}{24}\leq\frac{h}{4e\sigma R_{\max}x_{\max}} \tag{EC.52}$$

$$\frac{32ns\lambda}{|\mathcal{A}|\kappa}\leq\frac{32nsp^*\kappa}{768\sigma s x_{\max}|\mathcal{A}|\kappa}=\frac{1}{\sigma x_{\max}}\cdot\frac{n}{|\mathcal{A}|}\cdot\frac{p^*}{24}\leq\frac{1}{\sigma x_{\max}}, \tag{EC.53}$$

from which the proposition follows immediately.

**Proof of Proposition 5** Here, we will continue using the same requirement of $C_0$ stated in (EC.35). Because $\{M_k(i)\}$ for $k\in\mathcal{K}$ is a martingale with a bounded difference of 1 per the definition in (19), we can use $M_k(0)$ to bound the value of $M_k(t)$ via Azuma's inequality as follows:

$$\begin{aligned}&\mathbb{P}\left(M_k(t)-M_k(0)\leq-\frac{1}{2}M_k(0)\right)\leq\exp\left(-\frac{M_k(0)^2}{8(t+1)}\right)\\ \Rightarrow&\mathbb{P}\left(M_k(t)\leq\frac{1}{2}M_k(0)\right)\leq\exp\left(-\frac{M_k(0)^2}{8(t+1)}\right).\end{aligned} \tag{EC.54}$$

The $M_k(0)$ term can be stated as follows

$$\begin{aligned}M_k(0)&=\mathbb{E}_{\boldsymbol{X}}\left[\sum_{i=1}^{t}\mathbb{1}(\boldsymbol{X}_i\in U_k,\mathcal{E}_2,\boldsymbol{X}_i\notin\mathcal{R}_{\boldsymbol{x},k}))\right]\\&=\sum_{i=1}^{t}\mathbb{P}(\boldsymbol{X}_i\in U_k,\mathcal{E}_2,\boldsymbol{X}_i\notin\mathcal{R}_{\boldsymbol{x},k}).\end{aligned} \tag{EC.55}$$

As $\{\boldsymbol{X}_i\in U_k\}$ is independent of $\{\mathcal{E}_2,\boldsymbol{X}_i\notin\mathcal{R}_{\boldsymbol{x},k}\}$, and $\{\boldsymbol{X}_i\notin\mathcal{R}_{\boldsymbol{x},k}\}$ is independent of $\{\mathcal{E}_2\}$, (EC.55) implies the following inequality

$$\begin{aligned}M_k(0)&=\sum_{i=1}^{t}\mathbb{P}(\boldsymbol{X}_i\in U_k)\mathbb{P}(\mathcal{E}_2)\mathbb{P}(\boldsymbol{X}_i\notin\mathcal{R}_{\boldsymbol{x},k})\\&\geq\sum_{i=1}^{t}p^*\left(1-5\delta_0(t,t_0)\right)\left(1-\frac{2C_0}{t}\right),\end{aligned} \tag{EC.56}$$

where (EC.56) uses assumption A.3, Proposition 4, and the definition of $\epsilon$-decay random sampling scheme with $t_0=2C_0|\mathcal{K}|$.

When $t\geq t_0$, we have

$$5\delta_0(t,t_0)=\frac{10(t_0+1)^4}{e^4(t+1)^4}\leq\frac{1}{2}\text{ and }\frac{2C_0}{t}\leq\frac{1}{2}, \tag{EC.57}$$

where the second inequality uses $t\geq t_0=2C_0|\mathcal{K}|\geq 4C_0$. Inequalities in (EC.57) imply that

$$M_k(0)\geq\sum_{i=1}^{t}\frac{p^*}{4}=\frac{p^*t}{4}. \tag{EC.58}$$

Finally, combining (EC.54) and (EC.58), we can show that the following inequalities hold:

$$\begin{aligned}
&\mathbb{P}\left(M_k(t) \le \frac{p^* t}{8}\right) \le \exp\left(-\frac{(p^* t)^2/16}{8(t+1)}\right) \\
\Rightarrow &\mathbb{P}\left(M_k(t) \le \frac{p^* t}{8}\right) \le \exp\left(-\frac{(p^*)^2 t/16}{16}\right) \\
\Rightarrow &\mathbb{P}\left(M_k(t) \le \frac{p^* t}{8}\right) \le \exp\left(-\frac{(p^*)^2 t}{256}\right),
\end{aligned} \tag{EC.59}$$

where the second inequality uses $t/(t+1) \ge \frac{1}{2}$.

**Proof of Proposition 6** For clear expositions, we first state the following constants:

$$T_0 \ge \max\left\{\frac{48}{C_1 p^*}\log\left(\frac{16}{C_1 p^*}\right), \frac{8}{p^* C_1}\log d, 2|\mathcal{K}|C_0\right\}, \tag{EC.60}$$

$$T_1 \ge \max\left\{\frac{6(192s+\kappa p^* a)^2\lambda_{2,0}^2}{(\kappa p^* \min_{k,i\in\mathcal{S}^k}|\beta_{k,i}^{\text{true}}|)^2}\log\left(\frac{2(192s+\kappa p^* a)^2\lambda_{2,0}^2}{(\kappa p^* \min_{k,i\in\mathcal{S}^k}|\beta_{k,i}^{\text{true}}|)^2}\right), \frac{2(192s+\kappa p^* a)^2\lambda_{2,0}^2 \log d}{(\kappa p^* \min_{k,i\in\mathcal{S}^k}|\beta_{k,i}^{\text{true}}|)^2}\right\}, \tag{EC.61}$$

$$\lambda_{2,0} = \frac{4\sigma x_{\max} p^* \kappa a}{p^* \kappa a - 288 s}, \tag{EC.62}$$

where $C_0, C_1$ are defined in (EC.35) and (EC.125) respectively. We further require $a > \frac{1152 s}{p^* k} = \mathcal{O}(s)$ in the statement of this proposition, and then we can verify $T_0 = \tilde{\mathcal{O}}(s^2 \log d)$, $T_1 = \tilde{\mathcal{O}}(\beta_{\min}^{-2} s^2 \log d)$, and $\lambda_{2,0} = \mathcal{O}(1)$. Note that if the estimator $\boldsymbol{\beta}_j^{\text{random}}$ is close to $\boldsymbol{\beta}_j^{\text{true}}$ for all $j \in \mathcal{K}$, then assumption A.3 implies that for $\boldsymbol{x}_t \in U_j$, we can clearly separate $\mathbb{E}_\epsilon[R_t|\boldsymbol{x}_t\boldsymbol{\beta}_j^{\text{random}}]$ and $\max_{i\neq j}\mathbb{E}_\epsilon[R_t|\boldsymbol{x}_t\boldsymbol{\beta}_i^{\text{random}}]$. Specifically, part 2 of Lemma EC.3 shows that under event $\mathcal{E}_2$, the following inequality holds for any $\boldsymbol{x} \in U_k$ and $k \in \mathcal{K}_o$:

$$\mathbb{E}_\epsilon[R_k|\boldsymbol{x}^\top\boldsymbol{\beta}_k^{\text{random}}] > \max_{j\neq k}\mathbb{E}_\epsilon[R_j|\boldsymbol{x}^\top\boldsymbol{\beta}_j^{\text{random}}] + \frac{h}{2}, \tag{EC.63}$$

which implies

$$\mathbb{E}_\epsilon[R_k|\boldsymbol{x}_t^\top\boldsymbol{\beta}_k^{\text{random}}] = \max_{j\in\mathcal{K}}\mathbb{E}_\epsilon[R_j|\boldsymbol{x}_t^\top\boldsymbol{\beta}_j^{\text{random}}]$$

and for any $j \neq k$,

$$\mathbb{E}_\epsilon[R_j|\boldsymbol{x}_t^\top\boldsymbol{\beta}_j^{\text{random}}] < \mathbb{E}_\epsilon[R_k|\boldsymbol{x}_t^\top\boldsymbol{\beta}_k^{\text{random}}] - \frac{1}{2}h.$$

Further note that the G-MCP-Bandit algorithm constructs the optimal decision set as follows:

$$\Pi_t = \left\{i : \mathbb{E}_\epsilon[R_i|\boldsymbol{x}_t^\top\boldsymbol{\beta}_i^{\text{random}}] \ge \max_{j\in\mathcal{K}}\mathbb{E}_\epsilon[R_i|\boldsymbol{x}_t^\top\boldsymbol{\beta}_j^{\text{random}}] - \frac{1}{2}h\right\},$$

and therefore, for $\boldsymbol{x}_t \in U_k$, the optimal decision set will be a singleton, i.e., $\Pi_t = \{k\}$, which suggests that decision-makers will assign $k$ as the final decision by merely using the random-sample based estimator $\boldsymbol{\beta}^{\text{random}}$. As the event $\mathcal{E}_2$ is associated with the random estimator using randomly collected samples up to $t-1$ period, the set $\{\boldsymbol{x}_t : \boldsymbol{x}_t \in U_k, \mathcal{E}_2, \boldsymbol{x}_t \notin \mathcal{R}_{\boldsymbol{x},k}\}$ can be viewed as i.i.d. sample from the condition distribution $\mathcal{P}_{\boldsymbol{X}|\boldsymbol{X}\in U_k}$. Then, from Proposition 5, we have

$$\mathbb{P}\left(M_k(t) \le \frac{p^* t}{8}\right) \le \exp\left(-\frac{(p^*)^2 t}{256}\right), \tag{EC.64}$$

where $\{M_k(i)\}$ is defined in (19). As $M_k(t) = \mathbb{E}_{\epsilon,\boldsymbol{X}}\left[\sum_{i=1}^t \mathbb{1}(\boldsymbol{X}_i \in U_k, \mathcal{E}_2, \boldsymbol{X}_i \notin \mathcal{R}_{\boldsymbol{x},k})|\mathcal{F}_t\right] = \sum_{i=1}^t \mathbb{1}(\boldsymbol{x}_i \in U_k, \mathcal{E}_2, \boldsymbol{x}_i \notin \mathcal{R}_{\boldsymbol{x},k})$, the amount of i.i.d. samples in $U_k$ among the whole sample set for arm $k$ up to time $t$ will

be lower bounded by $M_k(t)$. Denote $\mathcal{A}$ and $n$ as the set of i.i.d. samples belonging to $U_k$ in the whole sample set and the size of the whole sample, respectively. The following two inequalities hold:

$$\mathbb{P}\left(|\mathcal{A}| \geq \frac{p^* t}{8}\right) \geq 1 - \exp\left(-\frac{(p^*)^2 t}{256}\right) \text{ and } n \leq t. \tag{EC.65}$$

If $|\mathcal{A}| \geq \frac{p^* t}{8}$ and $n \leq t$, then we can obtain the following result:

$$a > \frac{1152s}{p^*\kappa} \geq \frac{144st}{|\mathcal{A}|\kappa} > \frac{48sn}{\kappa|\mathcal{A}|}. \tag{EC.66}$$

Moreover, as $t > T_0 \geq 8(p^* C_1)^{-1} \log d$, then, by (EC.65), we have $|\mathcal{A}| \geq C_1^{-1} \log d$ with high probability. Combining this result with (EC.66) (i.e., two conditions required in Proposition 3), we have the following result via Proposition 3:

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{whole}} - \boldsymbol{\beta}^{\text{true}}\|_1 \geq \frac{16ns\zeta}{|\mathcal{A}|\kappa} + \frac{16ns\rho^{\text{whole}}_{\mathcal{S}/\mathcal{S}_1}}{|\mathcal{A}|\kappa}\lambda\right) \leq \delta_1(n, |\mathcal{A}|, \zeta) + \delta_2(n, |\mathcal{A}|, \lambda) \\
\Rightarrow &\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{whole}} - \boldsymbol{\beta}^{\text{true}}\|_1 \geq \frac{128s\zeta}{p^*\kappa} + \frac{128s\rho^{\text{whole}}_{\mathcal{S}/\mathcal{S}_1}}{p^*\kappa}\lambda\right) \leq \delta_1\left(t, \frac{p^* t}{8}, \zeta\right) + \delta_2\left(t, \frac{p^* t}{8}, \lambda\right),
\end{aligned} \tag{EC.67}$$

where (EC.67) uses (EC.65) and the fact that $n \leq t$ in the left-hand side and the facts that $\delta_1(\cdot)$ and $\delta_2(\cdot)$ are monotonically decreasing in $|\mathcal{A}|$ in the right-hand side.

When $t \geq T_1$, (EC.67) can be further simplified. We use Lemma EC.4 in E-Companion with $\alpha = \frac{(\kappa p^* \min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}|)^2}{2(192s+\kappa p^* a)^2 \lambda_{2,0}^2}$. When $t \geq T_1$, we have $t \geq 3\alpha^{-1} \log \alpha^{-1}$, combining with the nonnegativity of $t$, we can show that

$$\begin{aligned}
&\alpha t \geq \log t \\
\Rightarrow &\frac{(\kappa p^* \min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}|)^2}{2(192s + \kappa p^* a)^2 \lambda_{2,0}^2} t \geq \log t \\
\Rightarrow &\frac{t}{2} \geq \frac{(192s + \kappa p^* a)^2 \lambda_{2,0}^2}{(\kappa p^* \min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}|)^2} \log t.
\end{aligned} \tag{EC.68}$$

Moreover, as $T_1 \geq \frac{2(192s+\kappa p^* a)^2 \lambda_{2,0}^2 \log d}{(\kappa p^* \min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}|)^2}$, we can show that when $t > T_1$, the following inequality holds

$$\frac{t}{2} \geq \frac{(192s + \kappa p^* a)^2 \lambda_{2,0}^2 \log d}{(\kappa p^* \min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}|)^2}. \tag{EC.69}$$

Combining (EC.69) and (EC.68), we can verify that

$$\begin{aligned}
&t \geq \frac{(192s + \kappa p^* a)^2 \lambda_{2,0}^2 (\log t + \log d)}{(\kappa p^* \min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}|)^2} \\
\Rightarrow &\min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}| \geq \frac{192s + \kappa p^* a}{\kappa p^*} \cdot \lambda_{2,0} \sqrt{\frac{\log t + \log d}{t}} \\
\Rightarrow &\min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}| \geq \left(\frac{192s}{\kappa p^*} + a\right) \cdot \lambda_{2,0} \sqrt{\frac{\log t + \log d}{t}} \\
\Rightarrow &\min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}| \geq \left(\frac{24ns}{\kappa|\mathcal{A}|} + a\right) \cdot \lambda_{2,0} \sqrt{\frac{\log t + \log d}{t}} \qquad \text{(EC.70)} \\
\Rightarrow &\min_{i:|\beta_i^{\text{true}}|>0} |\beta_i^{\text{true}}| \geq \left(\frac{24ns}{\kappa|\mathcal{A}|} + a\right) \cdot \lambda_{2,t}. \qquad \text{(EC.71)}
\end{aligned}$$

where (EC.70) uses $|\mathcal{A}| \geq \frac{p^* t}{8}$ and $t \geq n$ and (EC.71) uses the definition of $\lambda_{2,t} = \lambda_{2,0} \cdot \sqrt{(\log t + \log d)/t}$. Now, we use the set $\mathcal{S}_1$ in the Proposition 3 by setting $\lambda = \lambda_{2,t}$, which directly implies that $\mathcal{S}_1 = \mathcal{S}$ so that we have $\mathcal{S}/\mathcal{S}_1$ being the empty set and

$$\rho^{\text{whole}}_{\mathcal{S}/\mathcal{S}_1} = 0. \tag{EC.72}$$

Hence, when $t > T_1$, we can use (EC.72) to simplify (EC.67) into

$$\mathbb{P}\left( \|\boldsymbol{\beta}^{\text{whole}} - \boldsymbol{\beta}^{\text{true}}\|_1 \geq \frac{128 s\zeta}{p^*\kappa} \right) \leq \delta_1\left(t, \frac{p^* t}{8}, \zeta\right) + \delta_2\left(t, \frac{p^* t}{8}, \lambda_{2,t}\right). \tag{EC.73}$$

Finally, we will show that when $t > T_0$, the following two inequalities hold:

$$\delta_1\left(t, \frac{p^* t}{8}, \zeta\right) \leq \frac{2}{(t+1)^2} + 2s \exp\left(-\frac{t\zeta^2}{2\sigma^2 x_{\max}^2}\right), \tag{EC.74}$$

$$\delta_2\left(t, \frac{p^* t}{8}, \lambda_{2,t}\right) \leq \frac{8}{(t+1)^2}. \tag{EC.75}$$

Let's first establish the first inequality (EC.74). Via Lemma EC.4 in E-Companion with $\alpha = \frac{C_1 p^*}{16}$, we can show that because $t > T_0 \geq \max\left\{ \frac{48}{C_1 p^*} \log\left(\frac{16}{C_1 p^*}\right), 0 \right\}$, we have $\frac{C_1 p^*}{16} t \geq \log t \Rightarrow \frac{C_1 p^*}{8} t \geq 2 \log t$, which implies that

$$\exp\left(-C_1 |\mathcal{A}|\right) \leq \exp\left(-C_1 \frac{p^*}{8} t\right) \leq \frac{1}{t^2} \leq \frac{2}{(t+1)^2}, \tag{EC.76}$$

where the last inequality uses the fact that $t^{-2} \leq 2(t+1)^{-2}$ holds for all $t \geq T_0 \geq 2|\mathcal{K}| C_0 > 3$. Combining (EC.76) with the definition of $\delta_1(t, |\mathcal{A}|, \zeta)$, we will reach (EC.74). Next, we will show the second inequality (EC.75). When $\lambda_{2,0} = \frac{4\sigma x_{\max} p^* \kappa a}{p^* \kappa a - 288 s}$, we can show that

$$\begin{aligned}
\delta_2(t, \frac{p^* t}{8}, \lambda_{2,t}) &= 4d \exp\left( -\frac{t\lambda_{2,0}^2}{2\sigma^2 x_{\max}^2} \cdot \frac{\log t + \log d}{t} \cdot \left(\frac{1}{2} - \frac{144 s}{p^* \kappa a}\right)^2 \right) \\
&= 4d \exp\left( -\frac{16\sigma^2 x_{\max}^2 (p^* \kappa a)^2}{(p^* \kappa a - 288 s)^2} \cdot \frac{t}{2\sigma^2 x_{\max}^2} \cdot \frac{\log t + \log d}{t} \cdot \left(\frac{1}{2} - \frac{144 s}{p^* \kappa a}\right)^2 \right) \\
&= 4 \exp\left(-2(\log t + \log d) + \log d\right) \leq 4 \exp\left(-2 \log t\right) \leq \frac{4}{t^2} \leq \frac{8}{(t+1)^2},
\end{aligned}$$

where the last inequality still uses the fact that $t^{-2} \leq 2(t+1)^{-2}$ for $t > 3$. Combining (EC.74) and (EC.75), we have

$$\delta_1\left(t, \frac{p^* t}{8}, \zeta\right) + \delta_2\left(t, \frac{p^* t}{8}, \lambda\right) \leq \frac{10}{(t+1)^2} + 2s \exp\left(-\frac{t\zeta^2}{2\sigma^2 x_{\max}^2}\right). \tag{EC.77}$$

Proposition 6 directly follows by combining (EC.67), (EC.73), (EC.77), and $\mathbb{P}(\mathcal{E}_2^c) \leq 5\delta_0(T, t_0)$ from Proposition 4.

**Proof of Theorem 1** We divide the time, up to time $T$, into three groups and derive the cumulative regret bound for each group separately. Consider the following three groups:

1. $t \in \{\{t : (\boldsymbol{X}_t, R_t) \in \mathcal{R}_k, k \in \mathcal{K}\} \cup \{t \leq T_0\}\}$.
2. $t \in \{\{t : (\boldsymbol{X}_t, R_t) \notin \mathcal{R}_k, k \in \mathcal{K}, t > T_0\} \cap \{\mathcal{E}_2 \text{ doesn't hold}\}\}$.
3. $t \in \{\{t : (\boldsymbol{X}_t, R_t) \notin \mathcal{R}_k, k \in \mathcal{K}, t > T_0\} \cap \{\mathcal{E}_2 \text{ holds}\}\}$.

In this proof, we follow the same choices of $C_0 = \mathcal{O}(s^2 \log d)$ in (EC.35), $T_1 = \tilde{\mathcal{O}}(\beta_{\min}^{-2} s^2 \log d)$ in (EC.61), $\lambda_1 = \mathcal{O}(s^{-1})$ in (EC.36), $\lambda_{2,0} = \mathcal{O}(1)$ in (EC.62), $T_0 = \tilde{\mathcal{O}}(s^2 \log d)$, and $a > \frac{1152s}{p^*\kappa} = \mathcal{O}(s)$. Beside the requirements for $T_0$ in (EC.60), we also require that

$$T_0 \geq \max\Big\{t_0, \Big(\frac{512s\lambda_{2,0}}{p^*\kappa}\Big)^2 \log d, 3\Big(\frac{512s\lambda_{2,0}}{p^*\kappa}\Big)^2 \log\Big(\frac{512s\lambda_{2,0}}{p^*\kappa}\Big)^2, \Big(\frac{1024\sigma e^{\sigma x_{\max}} x_{\max} s\lambda_{2,0}}{p^*\kappa}\Big)^2 \log d, \\ 3\Big(\frac{1024\sigma e^{\sigma x_{\max}} x_{\max} s\lambda_{2,0}}{p^*\kappa}\Big)^2 \log\Big(\frac{1024\sigma e^{\sigma x_{\max}} x_{\max} s\lambda_{2,0}}{p^*\kappa}\Big)^2\Big\}, \tag{EC.78}$$

and $T_0$ remains on the order of $\tilde{\mathcal{O}}(s^2 \log d)$.

• **Regret in part 1:**

Denote the regret for the first part as $R_1(T)$, and we have

$$R_1(T) \leq R_{\max}\left(\sum_{t=T_0}^{T} \mathbb{1}\left((\boldsymbol{X}_t, R_t) \in \mathcal{R}_k, k \in \mathcal{K}\right) + T_0\right) \leq R_{\max}\left(\sum_{k\in\mathcal{K}} |\mathcal{R}_k| + T_0\right), \tag{EC.79}$$

where $|\mathcal{R}_k|$ is the cardinality of $\mathcal{R}_k$. From Proposition 2, when $t_0 = 2C_0|\mathcal{K}|$ and $C_0 \geq 20$ , we know that

$$\begin{aligned} &\mathbb{P}\left(|\mathcal{R}_k| \leq 3C_0(1 + \log(T) - \log(t_0))\right) \geq 1 - \delta_0(T, t_0) \\ \Rightarrow &\mathbb{P}\left(|\mathcal{R}_k| \leq 3C_0 \log(T)\right) \geq 1 - \delta_0(T, t_0), \end{aligned} \tag{EC.80}$$

which implies

$$\mathbb{P}\left(\sum_{k\in\mathcal{K}} |\mathcal{R}_k| > |\mathcal{K}| \cdot 3C_0 \log(T)\right) \leq \sum_{k\in\mathcal{K}} \mathbb{P}\left(|\mathcal{R}_k| > 3C_0 \log(T)\right) \leq |\mathcal{K}|\delta_0(T, t_0). \tag{EC.81}$$

We then combine (EC.79) and (EC.81) to bound the regret in part 1:

$$\begin{aligned} R_1(T) \leq R_{\max}\left(\sum_{k\in\mathcal{K}} |\mathcal{R}_k| + T_0\right) &\leq R_{\max}\left(\sum_{k\in\mathcal{K}} |\mathcal{R}_k|\right) \mathbb{P}\left(\sum_{k\in\mathcal{K}} |\mathcal{R}_k| > 3C_0|\mathcal{K}| \log(T)\right) \\ &\quad + R_{\max}\left(3C_0|\mathcal{K}| \log(T)\right) \mathbb{P}\left(\sum_{k\in\mathcal{K}} |\mathcal{R}_k| \leq 3C_0|\mathcal{K}| \log(T)\right) \\ &\quad + R_{\max} T_0 \\ &\leq R_{\max} T|\mathcal{K}|\delta_0(T, t_0) + R_{\max} 3C_0|\mathcal{K}| \log(T) + R_{\max} T_0 \\ &\leq 2R_{\max}|\mathcal{K}|(t_0 + 1) + 3R_{\max} C_0|\mathcal{K}| \log T + R_{\max} T_0 \qquad \text{(EC.82)} \\ &\leq 3C_0 R_{\max}|\mathcal{K}| \log T + 5R_{\max}|\mathcal{K}|T_0, \qquad \text{(EC.83)} \end{aligned}$$

where (EC.82) uses $T\delta_0(T, t_0) \leq (T+1)\frac{2(t_0+1)^4}{e^4(T+1)^4} \leq 2(t_0+1)$ for $T \geq t_0$ and (EC.83) uses $(t_0+1) < 2t_0 \leq 2T_0$ and $|\mathcal{K}| > 1$.

• **Regret in part 2:**

Denote the cumulative regret for the second part as $R_2(T)$. From Proposition 4, at time $t$, we know that

$$\begin{aligned} &\mathbb{P}\left(\|\boldsymbol{\beta}_k^{\text{random}} - \boldsymbol{\beta}_k^{\text{true}}\|_1 \leq \min\left\{\frac{1}{\sigma x_{\max}}, \frac{h}{4e\sigma R_{\max} x_{\max}}\right\}\right) \geq 1 - 5\delta_0(t, t_0),\ k \in \mathcal{K} \\ \Rightarrow &\mathbb{P}(\mathcal{E}_2(t)^{\text{c}}) \leq 5|\mathcal{K}|\delta_0(t, t_0), \end{aligned} \tag{EC.84}$$

where $\mathcal{E}_2(t)$ denotes the event $\mathcal{E}_2$ at time $t$.

Therefore, $R_2(T)$ can be bounded as follows:

$$
\begin{aligned}
R_2(T) &\le \mathbb{E}_{\boldsymbol{X},\epsilon}\left[\sum_{i=T_0+1}^{T}\mathbb{1}(\mathcal{E}_2(i)^{\mathrm{c}})R_{\max}\right] = R_{\max}\sum_{i=T_0+1}^{T}\mathbb{P}(\mathcal{E}_2(i)^{\mathrm{c}}) \\
\Rightarrow R_2(T) &\le R_{\max}\sum_{i=T_0+1}^{T} 5|\mathcal{K}|\delta_0(i,t_0) \\
&\le 5R_{\max}|\mathcal{K}|\int_{i=T_0}^{T-1}\delta_0(i,t_0)di \\
&= 10R_{\max}|\mathcal{K}|\int_{i=T_0}^{T-1}\frac{(t_0+1)^4}{e^4(i+1)^4}di \\
&= -\frac{10}{3}R_{\max}|\mathcal{K}|\cdot\frac{(t_0+1)^4}{e^4(i+1)^3}\Big|_{T_0}^{T-1} \\
&= \frac{10}{3}e^{-4}R_{\max}|\mathcal{K}|(t_0+1)^4(T_0+1)^{-3} - \frac{10}{3}e^{-4}R_{\max}|\mathcal{K}|(t_0+1)^4(T)^{-3} \\
&\le 2R_{\max}|\mathcal{K}|T_0,
\end{aligned}
$$

where last inequality we use $\frac{10}{3}e^{-4} < \frac{1}{8}$ and $(t_0+1) < 2t_0 \le 2T_0$.

• **Regret in part 3:**

Denote the cumulative regret for the third part as $R_3(T)$. We first consider the case where $T \le T_1$. By the second part of Lemma EC.3, it is direct to show that the optimal decision set $\Pi_t$ constructed in the G-MCP-Bandit Algorithm only contains arms in the optimal decision subset $\mathcal{K}_o$. Without loss of generality, we assume that arm $i$ is the true optimal arm at time $t$. Then, the regret at time $t$ can be bounded as follows

$$
\begin{aligned}
\text{regret}_t \le& \mathbb{E}_{\boldsymbol{X}}\left[\sum_{j\in\Pi_t}\mathbb{1}\left(j=\arg\max_{k\in\Pi_t}\mathbb{E}_\epsilon[R_k|\boldsymbol{X}_t^\top\boldsymbol{\beta}_k^{\text{whole}}]\right)\left(\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{true}}]-\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{true}}]\right)\right] \\
\le& \mathbb{E}_{\boldsymbol{X}}\left(\sum_{j\neq i}\mathbb{1}\left(\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{whole}}]\ge\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{whole}}]\right)\left(\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{true}}]-\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{true}}]\right)\right).
\end{aligned}
\tag{EC.85}
$$

We then denote

$$
\mathcal{E}(t,w,\delta_t)_{4,k} = \left\{\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{true}}]-\mathbb{E}_\epsilon[R_k|\boldsymbol{X}_t^\top\boldsymbol{\beta}_k^{\text{true}}]\in[w\delta_t,(w+1)\delta_t)\right\},
\tag{EC.86}
$$

where $k\neq i, k\in\mathcal{K}_o, w=0,1,...,$ and $\delta_t>0$. Then, we have the following bound:

$$
\begin{aligned}
\text{regret}_t \le& \mathbb{E}_{\boldsymbol{X}}\left(\sum_{w=0}^{w_{1,t}}\sum_{j\neq i}\mathbb{1}\left(\left\{\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{whole}}]\ge\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{whole}}]\right\}\cap\mathcal{E}(t,w,\delta_t)_{4,j}\right)(w+1)\delta_t\right) \\
=& \sum_{w=0}^{w_{1,t}}(w+1)\delta_t\overbrace{\sum_{j\neq i}\mathbb{P}\left(\left\{\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{whole}}]\ge\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{whole}}]\right\}\cap\mathcal{E}(t,w,\delta_t)_{4,j}\right)}^{(*)},
\end{aligned}
\tag{EC.87}
$$

where $w_{1,t}=\lceil R_{\max}/\delta_t\rceil$. Now we consider the (*) term in (EC.87), which can be bounded as follows:

$$
\begin{aligned}
(*) &\le \sum_{j\neq i}\mathbb{P}\left(\left\{\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{whole}}]-\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{true}}]\ge\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{whole}}]-\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{true}}]+w\delta_t\right\}\cap\mathcal{E}(t,w,\delta_t)_{4,j}\right) \\
&\le \sum_{j\neq i}\mathbb{P}\left(\left\{\left|\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{whole}}]-\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{true}}]\right|+\left|\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{whole}}]-\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{true}}]\right|\ge w\delta_t\right\}\cap\mathcal{E}(t,w,\delta_t)_{4,j}\right),
\end{aligned}
\tag{EC.88}
$$

where the first inequality uses the fact that $\mathbb{E}_\epsilon[R_i|\boldsymbol{X}_t^\top\boldsymbol{\beta}_i^{\text{true}}]-\mathbb{E}_\epsilon[R_j|\boldsymbol{X}_t^\top\boldsymbol{\beta}_j^{\text{true}}]\in[w\delta_t,(w+1)\delta_t)$ when event $\mathcal{E}(t,w,\delta_t)_{4,j}$ holds. To simplify the notation, we denote $\Delta_k=\boldsymbol{\beta}_k^{\text{whole}}-\boldsymbol{\beta}_k^{\text{true}}$ for $k\in\Pi_t$. Combining (EC.88) with the first part of Lemma EC.3, we can show

$$\begin{aligned}(*)\leq&\sum_{j\neq i}\mathbb{P}\left(\left\{\|\Delta_j\|_1+\|\Delta_i\|_1\geq\frac{w\delta_t}{R_{\max}\sigma e^{\sigma x_{\max}\max\{\|\Delta_j\|_1,\|\Delta_i\|_1\}}x_{\max}}\right\}\cap\mathcal{E}(t,w,\delta_t)_{4,j}\right)\\=&\sum_{j\neq i}\mathbb{P}\left(\|\Delta_j\|_1+\|\Delta_i\|_1\geq\frac{w\delta_t}{R_{\max}\sigma e^{\sigma x_{\max}\max\{\|\Delta_j\|_1,\|\Delta_i\|_1\}}x_{\max}}\right)\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j}),\end{aligned}\tag{EC.89}$$

where the last equality uses the fact that in $\Delta_i$ and $\Delta_j$, the terms $\boldsymbol{\beta}_t^{\text{whole}}$ only depend on historical samples upto $t-1$ (independent on $t$ step's information), which implies their independence on $\mathcal{E}(t,w,\delta_t)_{4,j}$.

Denote event $\mathcal{E}_{5,(i,j),t}(w)$ as follows

$$\mathcal{E}_{5,(i,j),t}(w)=\left\{\left\{\|\Delta_i\|_1\geq\min\left\{\frac{w\delta_t}{2R_{\max}\sigma e^{\sigma x_{\max}}x_{\max}},1\right\}\right\}\cup\left\{\|\Delta_j\|_1\geq\min\left\{\frac{w\delta_t}{2R_{\max}\sigma e^{\sigma x_{\max}}x_{\max}},1\right\}\right\}\right\}.\tag{EC.90}$$

Then, conditioning on $\mathcal{E}_{5,(i,j),t}(w)$, the right hand side of (EC.89) can be transformed into

$$\begin{aligned}&\sum_{j\neq i}\mathbb{P}\left(\|\Delta_j\|_1+\|\Delta_i\|_1\geq\frac{w\delta_t}{R_{\max}\sigma e^{\sigma x_{\max}\max\{\|\Delta_j\|_1,\|\Delta_i\|_1\}}x_{\max}}\right)\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j})\\=&\sum_{j\neq i}\mathbb{P}\left(\left\{\|\Delta_j\|_1+\|\Delta_i\|_1\geq\frac{w\delta_t}{R_{\max}\sigma e^{\sigma x_{\max}\max\{\|\Delta_j\|_1,\|\Delta_i\|_1\}}x_{\max}}\right\}\cap\mathcal{E}_{5,(i,j),t}(w)\right)\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j})\\+&\sum_{j\neq i}\mathbb{P}\left(\left\{\|\Delta_j\|_1+\|\Delta_i\|_1\geq\frac{w\delta_t}{R_{\max}\sigma e^{\sigma x_{\max}\max\{\|\Delta_j\|_1,\|\Delta_i\|_1\}}x_{\max}}\right\}\cap(\mathcal{E}_{5,(i,j),t}(w))^{\text{c}}\right)\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j})\\\leq&\sum_{j\neq i}\mathbb{P}\left(\left\{\|\Delta_j\|_1+\|\Delta_i\|_1\geq\frac{w\delta_t}{R_{\max}\sigma e^{\sigma x_{\max}\max\{\|\Delta_j\|_1,\|\Delta_i\|_1\}}x_{\max}}\right\}\cap\mathcal{E}_{5,(i,j),t}(w)\right)\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j})\\+&\sum_{j\neq i}\mathbb{P}\left(\left\{\|\Delta_j\|_1+\|\Delta_i\|_1\geq\frac{w\delta_t}{R_{\max}\sigma e^{\sigma x_{\max}}x_{\max}}\right\}\cap(\mathcal{E}_{5,(i,j),t}(w))^{\text{c}}\right)\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j})\\=&\sum_{j\neq i}\mathbb{P}\left(\left\{\|\Delta_j\|_1+\|\Delta_i\|_1\geq\frac{w\delta_t}{R_{\max}\sigma e^{\sigma x_{\max}\max\{\|\Delta_j\|_1,\|\Delta_i\|_1\}}x_{\max}}\right\}\cap\mathcal{E}_{5,(i,j),t}(w)\right)\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j})\\\leq&\sum_{j\neq i}\mathbb{P}(\mathcal{E}_{5,(i,j),t}(w))\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j}).\end{aligned}\tag{EC.91}$$

We first bound $\mathbb{P}(\mathcal{E}_{5,(i,j),t}(w))$. As $\mathcal{E}_{5,(i,j),t}(w)$ holds automatically for $w=0$ (i.e., $\mathbb{P}(\mathcal{E}_{5,(i,j),t}(0))=1$), we will discuss the remaining cases where $w\geq1$. As the optimal decision set $\Pi_t$ only contains arms in the optimal decision subset $\mathcal{K}_o$, from Proposition 6, for $t\geq T_0$, we have the following inequality for $k\in\mathcal{K}_o$:

$$\mathbb{P}\left(\|\Delta_k\|_1\geq\frac{128s\zeta}{p^*\kappa}+\frac{128s\rho^{\text{whole}}_{\mathcal{S}^k/\mathcal{S}^k_{1,t}}}{p^*\kappa}\lambda_{2,t}\right)\leq5\delta_0(t,t_0)+\frac{10}{(t+1)^2}+2s\exp\left(-\frac{t\zeta^2}{2\sigma^2x_{\max}^2}\right).\tag{EC.92}$$

Combining (EC.92) and the choice of $T_0$ (i.e., $T_0\geq\max\left\{\left(\frac{512s\lambda_{2,0}}{p^*\kappa}\right)^2\log d,3\left(\frac{512s\lambda_{2,0}}{p^*\kappa}\right)^2\log\left(\frac{512s\lambda_{2,0}}{p^*\kappa}\right)^2\right\}$), we can ensure $\max_i\|\Delta_i\|_1\leq1$ for all $i\in\mathcal{K}_o$ with high probability for $t>T_0$. To see, note that by setting $\zeta=\lambda_{2,t}$ and using the fact that $\rho^{\text{whole}}_{\mathcal{S}^k/\mathcal{S}^k_{1,t}}\leq1$, we can show that (EC.92) implies

$$\mathbb{P}\left(\|\Delta_k\|_1\geq\frac{256s\lambda_{2,t}}{p^*\kappa}\right)\leq5\delta_0(t,t_0)+\frac{10}{(t+1)^2}+2s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2x_{\max}^2}\right)$$

$$\Rightarrow \mathbb{P}\left(\|\Delta_k\|_1 \geq \frac{256s\lambda_{2,0}}{p^*\kappa}\sqrt{\frac{\log d+\log t}{t}}\right) \leq 5\delta_0(t,t_0)+\frac{10}{(t+1)^2}+2s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)$$
$$\Rightarrow \mathbb{P}\left(\|\Delta_k\|_1 \geq \frac{256s\lambda_{2,0}}{p^*\kappa}\left(\sqrt{\frac{\log d}{t}}+\sqrt{\frac{\log t}{t}}\right)\right) \leq 5\delta_0(t,t_0)+\frac{10}{(t+1)^2}+2s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)$$
$$\Rightarrow \mathbb{P}\left(\|\Delta_k\|_1 \geq 1\right) \leq 5\delta_0(t,t_0)+\frac{10}{(t+1)^2}+2s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right), \tag{EC.93}$$

where (EC.93) uses the facts that $\frac{256s\lambda_{2,0}}{p^*\kappa}\sqrt{\frac{\log d}{t}} \leq \frac{1}{2}$ (because $t > T_0 \geq \left(\frac{512s\lambda_{2,0}}{p^*\kappa}\right)^2 \log d$) and $\frac{256s\lambda_{2,0}}{p^*\kappa}\sqrt{\frac{\log t}{t}} \leq \frac{1}{2}$ (by setting $\alpha = \left(\frac{p^*\kappa}{512s\lambda_{2,0}}\right)^2$ and then using Lemma EC.4 on $t$ for $t > T_0 \geq 3\left(\frac{512s\lambda_{2,0}}{p^*\kappa}\right)^2 \log\left(\frac{512s\lambda_{2,0}}{p^*\kappa}\right)^2$).

We then consider the case with upper bound $\frac{w\delta_t}{2R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}}$ instead of 1 on $\|\Delta_k\|_1$. If we set $\zeta = C_\rho s^{-1} w\delta_t$, where $C_\rho = \frac{p^*\kappa}{256R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}(1+\rho_{\max})}$, $\delta_t = C_{\rho_1} s\sqrt{\frac{\log t+\log d}{t}}$, and $C_{\rho_1} = \frac{512R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}\lambda_{2,0}}{p^*\kappa}$, then we can show that the right-hand-side within the $\mathbb{P}(\cdot)$ term in (EC.92) can be upper bounded as follows:

$$\begin{aligned}
\frac{128s\zeta}{p^*\kappa}+\frac{128s\rho^{\text{whole}}_{\mathcal{S}^k/\mathcal{S}^k_{1,t}}}{p^*\kappa}\lambda_{2,t} &\leq \frac{128s\zeta}{p^*\kappa}+\frac{128s\rho_{\max}}{p^*\kappa}\lambda_{2,t} \\
&= \frac{128sC_\rho s^{-1} w C_{\rho_1} s\sqrt{(\log t+\log d)/t}}{p^*\kappa}+\frac{128s\rho_{\max}}{p^*\kappa}\lambda_{2,t} \\
&= \frac{128sw}{p^*\kappa}\cdot\frac{2}{1+\rho_{\max}}\cdot\lambda_{2,0}\sqrt{\frac{\log t+\log d}{t}}+\frac{128s\rho_{\max}}{p^*\kappa}\cdot\lambda_{2,0}\sqrt{\frac{\log t+\log d}{t}} \\
&= \frac{128}{p^*\kappa}\left(\frac{2w}{1+\rho_{\max}}+\rho_{\max}\right)\cdot\lambda_{2,0}\cdot s\sqrt{\frac{\log t+\log d}{t}} \\
&= \frac{128}{p^*\kappa}\left(\frac{2w}{1+\rho_{\max}}+\rho_{\max}\right)\cdot\lambda_{2,0}\cdot C_{\rho_1}^{-1}\cdot\delta_t \\
&= \left(\frac{1}{1+\rho_{\max}}+\frac{\rho_{\max}}{2w}\right)\cdot\frac{w\delta_t}{2R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}}
\end{aligned}$$

Note that $\frac{1}{1+\rho_{\max}}+\frac{\rho_{\max}}{2w}$, where $\rho_{\max}\in[0,1]$ and $w \geq 1$, can be upper bounded by 1. To see, we first take the derivative of $\frac{1}{1+\rho_{\max}}+\frac{\rho_{\max}}{2w}$ w.r.t $\rho_{\max}$ to have $-\frac{1}{(1+\rho_{\max})^2}+\frac{1}{2w}$. If $w \geq 2$, then $-\frac{1}{(1+\rho_{\max})^2}+\frac{1}{2w}$ is non-positive for $\rho_{\max}\in[0,1]$, which means that $\rho_{\max}=0$ is the maximizer for $\frac{1}{1+\rho_{\max}}+\frac{\rho_{\max}}{2w}$, which gives the maximum value of 1; if $1 \leq w < 2$, then $-\frac{1}{(1+\rho_{\max})^2}+\frac{1}{2w}$ will be first negative and then positive for $\rho_{\max}\in[0,1]$, which means that $\frac{1}{1+\rho_{\max}}+\frac{\rho_{\max}}{2w}$ will be maximized at either $\rho_{\max}=0$ or $\rho_{\max}=1$, both of which give the maximum value of 1. Therefore, $\frac{1}{1+\rho_{\max}}+\frac{\rho_{\max}}{2w}$ is upper bounded by 1, which implies that

$$\frac{128s\zeta}{p^*\kappa}+\frac{128s\rho^{\text{whole}}_{S^k/S^k_{1,t}}}{p^*\kappa}\lambda_{2,t} \leq \frac{w\delta_t}{2R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}}. \tag{EC.94}$$

Then, using (EC.93) and (EC.94) for both $\|\Delta_j\|_1$ and $\|\Delta_i\|_1$ for $w \geq 1$, we will have that for $w \geq 1$,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{5,(i,j),t}(w)) &\leq \min\left\{1, 10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+\max\left\{4s\exp\left(-\frac{C_\rho^2 t w^2\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\right), 4s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)\right\}\right\} \\
&\leq \min\left\{1, 10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-\frac{C_\rho^2 t w^2\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\right)+4s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)\right\}. \quad \text{(EC.95)}
\end{aligned}$$

Note that (EC.95) also holds for the case when $w=0$, as we have $\mathbb{P}(\mathcal{E}_{5,(i,j),t}(0))=1$ and $10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-\frac{C_\rho^2 t w^2\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\right) \geq 4s\exp\left(-\frac{C_\rho^2 t w^2\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\right) = 4s > 1$.

Furthermore, by assumption A.2, we have

$$\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j}) \le CR_{\max}(1+w)\delta_t. \tag{EC.96}$$

Hence, via (EC.91), (EC.95), and (EC.96), we can show that

$$\begin{aligned}(*) &\le \sum_{j\neq i}\mathbb{P}(\mathcal{E}_{5,(i,j),t}(w))\mathbb{P}(\mathcal{E}(t,w,\delta_t)_{4,j})\\ &\le \sum_{j\neq i}\min\left\{1, 10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-\frac{C_\rho^2 t w^2\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\right)+4s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)\right\}\cdot CR_{\max}(1+w)\delta_t.\end{aligned}$$

Accordingly, the regret bound at time $t$ can be rewritten as follows:

$$\begin{aligned}\text{regret}_t &\le \sum_{w=0}^{w_{1,t}}(w+1)\delta_t\left(\sum_{j\neq i}\min\left\{1, 10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-\frac{C_\rho^2 t w^2\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\right)\right.\right.\\ &\qquad\left.\left.+4s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)\right\}\cdot CR_{\max}(1+w)\delta_t\right)\\ &\le CR_{\max}|\mathcal{K}|\sum_{w=0}^{w_{1,t}}(w+1)^2\delta_t^2\left(10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)+\min\left\{1,4s\exp\left(-\frac{C_\rho^2 t w^2\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\right)\right\}\right)\\ &\le CR_{\max}|\mathcal{K}|\left(\overbrace{\left(10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)\right)\sum_{w=0}^{w_{1,t}}(1+w)^2\delta_t^2}^{(a)}+\overbrace{\sum_{w=0}^{w_{0,t}}(1+w)^2\delta_t^2}^{(b)}\right.\\ &\qquad\left.+\overbrace{\sum_{w=w_{0,t}+1}^{w_{1,t}}4(1+w)^2\delta_t^2 s\exp\left(-\frac{C_\rho^2 t w^2\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\right)}^{(c)}\right),\end{aligned} \tag{EC.97}$$

where $w_{0,t}=\left\lfloor\sqrt{\frac{2\log(4s)s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}}\right\rfloor$. Next, we will bound part (a), (b) and (c) separately:

$$\begin{aligned}(a) &< \left(10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-\frac{t\lambda_{2,t}^2}{2\sigma^2 x_{\max}^2}\right)\right)(1+w_{1,t})(1+w_{1,t})^2\delta_t^2\\ &= \left(10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-\lambda_{2,0}^2\frac{\log d+\log t}{2\sigma^2 x_{\max}^2}\right)\right)(1+w_{1,t})(1+w_{1,t})^2\delta_t^2\\ &= \left(10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-8\frac{(p^*\kappa a)^2}{(p^*\kappa a-288s)^2}(\log d+\log t)\right)\right)(1+w_{1,t})(1+w_{1,t})^2\delta_t^2\\ &= \left(10\delta_0(t,t_0)+\frac{20}{(t+1)^2}+4s\exp\left(-32(\log d+\log t)\right)\right)(1+w_{1,t})(1+w_{1,t})^2\delta_t^2\\ &\le \left(10\delta_0(t,t_0)+\frac{24}{(t+1)^2}\right)\delta_t^2(1+w_{1,t})^3 \qquad\text{(EC.98)}\\ &\le \left(10\delta_0(t,t_0)+\frac{24}{(t+1)^2}\right)\delta_t^2\left(1+\frac{R_{\max}}{\delta_t}+1\right)^3\\ &\le \left(10\delta_0(t,t_0)+\frac{24}{(t+1)^2}\right)(3R_{\max})^3\delta_t^{-1} \qquad\text{(EC.99)}\\ &\le \frac{540R_{\max}^3(t_0+1)^4}{e^4(t+1)^4\delta_t}+\frac{648R_{\max}^3}{(t+1)^2\delta_t}, \qquad\text{(EC.100)}\end{aligned}$$

where (EC.98) uses $s \le d$ and $t^2 \ge t+1$ when $t \ge 1$, and (EC.99) uses $\frac{R_{\max}}{\delta_t} \ge 1$ for $t \ge T_0$, which can be shown in the following analysis

$$\begin{aligned}
\delta_t &= C_{\rho_1} s\sqrt{\frac{\log d + \log t}{t}} \\
&\le C_{\rho_1} s\sqrt{\frac{\log d + \log T_0}{T_0}} \\
&\le C_{\rho_1} s\sqrt{\frac{\log d}{T_0}} + C_{\rho_1} s\sqrt{\frac{\log T_0}{T_0}} \\
&\le \frac{512 R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}\lambda_{2,0}}{p^*\kappa} s\sqrt{\frac{\log d}{T_0}} + \frac{512 R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}\lambda_{2,0}}{p^*\kappa} s\sqrt{\frac{\log T_0}{T_0}} \\
&\le \frac{R_{\max}}{2} + \frac{512 R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}\lambda_{2,0}}{p^*\kappa} s\sqrt{\frac{\log T_0}{T_0}} \qquad \text{(EC.101)} \\
&\le \frac{R_{\max}}{2} + \frac{512 R_{\max}\sigma e^{\sigma x_{\max}} x_{\max}\lambda_{2,0}}{p^*\kappa} s\left(\frac{p^*\kappa}{1024\sigma e^{\sigma x_{\max}} x_{\max} s\lambda_{2,0}}\right) \qquad \text{(EC.102)} \\
&\le \frac{1}{2}R_{\max} + \frac{1}{2}R_{\max} = R_{\max},
\end{aligned}$$

where (EC.101) uses the fact that $T_0 \ge (\frac{1024\sigma e^{\sigma x_{\max}} x_{\max} s\lambda_{2,0}}{p^*\kappa})^2 \log d$ and (EC.102) uses Lemma EC.4 (by using the fact that $T_0 \ge 3\Big(\frac{1024\sigma e^{\sigma x_{\max}} x_{\max} s\lambda_{2,0}}{p^*\kappa}\Big)^2 \log\Big(\frac{1024\sigma e^{\sigma x_{\max}} x_{\max} s\lambda_{2,0}}{p^*\kappa}\Big)^2$ and setting $\alpha = \Big(\frac{p^*\kappa}{1024\sigma e^{\sigma x_{\max}} x_{\max} s\lambda_{2,0}}\Big)^2$ to show $\alpha T_0 \ge \log T_0$). Next, we can further upper bound part (b) and part (c) in (EC.97) as follows:

$$(b) < (1+w_{0,t})(1+w_{0,t})^2\delta_t^2 = (1+w_{0,t})^3\delta_t^2. \qquad \text{(EC.103)}$$

$$\begin{aligned}
(c) &\le 16 s\delta_t^2 \sum_{w=w_{0,t}+1}^{w_{1,t}} w^2 \exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot w^2\right) \\
&\le 16 s\delta_t^2 \left( \underbrace{(w_{0,t}+1)^2 \exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot (w_{0,t}+1)^2\right)}_{(c_1)} + \underbrace{\sum_{w=w_{0,t}+2}^{w_{1,t}} w^2 \exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot w^2\right)}_{(c_2)} \right). \qquad \text{(EC.104)}
\end{aligned}$$

Let's consider $(c_1)$ and $(c_2)$, separately. As $w_{0,t} = \left\lfloor \sqrt{\frac{2\log(4s)s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}} \right\rfloor$, we have $\sqrt{\frac{2\log(4s)s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}} \le w_{0,t}+1$, which implies

$$\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot(w_{0,t}+1)^2 \ge \frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot\left(\sqrt{\frac{2\log(4s)s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}}\right)^2 = \log(4s) > 1. \qquad \text{(EC.105)}$$

Combining (EC.105) with the fact that the function $x\exp(-x)$ is monotonically decreasing for $x \ge 1$, we can show that

$$\begin{aligned}
&\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}(w_{0,t}+1)^2 \exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}(w_{0,t}+1)^2\right) \le \log(4s)\exp(-\log 4s) = \frac{\log(4s)}{4s} \\
\Rightarrow &(c_1) = (w_{0,t}+1)^2 \exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}(w_{0,t}+1)^2\right) \le \frac{s\sigma^2 x_{\max}^2 \log(4s)}{2C_\rho^2 t\delta_t^2}. \qquad \text{(EC.106)}
\end{aligned}$$

Similarly, as the function $x^2\exp(-x^2)$ is monotonically decreasing for $x \geq 1$, we can upper bound the $(c_2)$ term as follows:

$$
\begin{aligned}
(c_2) &= \sum_{w=w_{0,t}+2}^{w_{1,t}} w^2 \exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot w^2\right) \\
&\leq \int_{w_{0,t}+1}^{\infty} w^2 \exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot w^2\right) dw \\
&= \int_{w=w_{0,t}+1}^{w=\infty} -\frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\cdot w \cdot d\left[\exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot w^2\right)\right] \\
&= \frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\cdot(w_{0,t}+1)\exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot(w_{0,t}+1)^2\right) + \frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\int_{w_{0,t}+1}^{\infty}\exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot w^2\right) dw \qquad \text{(EC.107)} \\
&\leq \frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\left(\frac{w_{0,t}+1}{4s} + \int_{w_{0,t}+1}^{+\infty}\frac{w}{w_{0,t}+1}\exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot w^2\right) dw\right) \qquad \text{(EC.108)} \\
&= \frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\left(\frac{w_{0,t}+1}{4s} + \frac{1}{w_{0,t}+1}\frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\exp\left(-\frac{C_\rho^2 t\delta_t^2}{2s^2\sigma^2 x_{\max}^2}\cdot(w_{0,t}+1)^2\right)\right) \\
&\leq \frac{s\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\left(\frac{w_{0,t}+1}{4s} + \frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\cdot\frac{1}{4s}\right) \\
&= \frac{\sigma^2 x_{\max}^2}{4C_\rho^2 t\delta_t^2}\left(w_{0,t}+1+\frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\right), \qquad \text{(EC.109)}
\end{aligned}
$$

where (EC.107) uses the integration by parts and (EC.108) uses (EC.105) and $w \geq w_{0,t}+1 \geq 1$. Combining (EC.104), (EC.106), and (EC.109), we have

$$
\begin{aligned}
(c) &\leq 16s\delta_t^2\left(\frac{s\log(4s)\sigma^2 x_{\max}^2}{2C_\rho^2 t\delta_t^2} + \frac{\sigma^2 x_{\max}^2}{4C_\rho^2 t\delta_t^2}\left(w_{0,t}+1+\frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\right)\right) \\
&= \frac{4s\sigma^2 x_{\max}^2}{C_\rho^2 t}\left(2s\log(4s)+w_{0,t}+1+\frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\right). \qquad \text{(EC.110)}
\end{aligned}
$$

Then, combining (EC.97), (EC.100), and (EC.103), (EC.110) with $\delta_t = C_{\rho_1} s\sqrt{\frac{\log t+\log d}{t}}$, we can show that

$$
\begin{aligned}
\text{regret}_t &\leq CR_{\max}|\mathcal{K}|\left(\frac{540R_{\max}^3(t_0+1)^4}{e^4(t+1)^4\delta_t} + \frac{648R_{\max}^3}{(t+1)^2\delta_t} + (1+w_{0,t})^3\delta_t^2 + \frac{4s\sigma^2 x_{\max}^2}{C_\rho^2 t}\left(2s\log(4s)+w_{0,t}+1+\frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 t\delta_t^2}\right)\right) \\
&= CR_{\max}|\mathcal{K}|\left(\frac{540R_{\max}^3(t_0+1)^4}{e^4(t+1)^4\cdot C_{\rho_1} s\sqrt{\frac{\log t+\log d}{t}}} + \frac{648R_{\max}^3}{(t+1)^2\cdot C_{\rho_1} s\sqrt{\frac{\log t+\log d}{t}}} + \frac{(1+w_{0,t})^3\cdot C_{\rho_1}^2 s^2(\log t+\log d)}{t}\right. \\
&\quad \left. + \frac{4s\sigma^2 x_{\max}^2}{C_\rho^2 t}\left(2s\log(4s)+w_{0,t}+1+\frac{s^2\sigma^2 x_{\max}^2}{C_\rho^2 tC_{\rho_1}^2 s^2(\frac{\log d+\log t}{t})}\right)\right) \\
&\leq CR_{\max}|\mathcal{K}|\left(\frac{540R_{\max}^3(t_0+1)^4}{e^4 C_{\rho_1} s\sqrt{\log d}}\cdot\frac{t^{1/2}}{(t+1)^4} + \frac{648R_{\max}^3}{C_{\rho_1} s\sqrt{\log d}}\cdot\frac{t^{1/2}}{(t+1)^2} + \frac{(1+w_{0,t})^3\cdot C_{\rho_1}^2 s^2(\log t+\log d)}{t}\right. \\
&\quad \left. + \frac{4s\sigma^2 x_{\max}^2}{C_\rho^2 t}\left(2s\log(4s)+w_{0,t}+1+\frac{\sigma^2 x_{\max}^2}{C_\rho^2 C_{\rho_1}^2\log d}\right)\right) \\
&\leq CR_{\max}|\mathcal{K}|\left(\frac{540R_{\max}^3(t_0+1)^4}{e^4 C_{\rho_1} s\sqrt{\log d}}\cdot\frac{t^{1/2}}{t^4} + \frac{648R_{\max}^3}{C_{\rho_1} s\sqrt{\log d}}\cdot\frac{t^{1/2}}{t^{3/2}\cdot\sqrt{T_0+1}} + \frac{(1+w_{0,t})^3 C_{\rho_1}^2\cdot s^2(\log t+\log d)}{t}\right. \\
&\quad \left. + \frac{4s\sigma^2 x_{\max}^2}{C_\rho^2 t}\left(2s\log(4s)+w_{0,t}+1+\frac{\sigma^2 x_{\max}^2}{C_\rho^2 C_{\rho_1}^2\log d}\right)\right)
\end{aligned}
$$

$$\leq R_{\max}|\mathcal{K}|\left(\frac{540CR_{\max}^3(t_0+1)^4}{e^4C_{\rho_1}s(\log^{1/2}d)t^{7/2}}+\frac{C_3+CC_{\rho_1}^2(1+w_{0,t})^3s^2\log t}{t}\right), \tag{EC.111}$$

where

$$C_3=\frac{648CR_{\max}^3}{C_{\rho_1}s(\log^{1/2}d)\sqrt{T_0+1}}+CC_{\rho_1}^2(1+w_{0,t})^3s^2\log d+\frac{4Cs\sigma^2x_{\max}^2}{C_\rho^2}\left(2s\log(4s)+w_{0,t}+1+\frac{\sigma^2x_{\max}^2}{C_\rho^2C_{\rho_1}^2\log d}\right). \tag{EC.112}$$

Hence, the third part of the regret can be bounded as follows:

$$\begin{aligned}
R_3(T)&\leq\sum_{t=T_0+1}^{T}\text{regret}_t\cdot\mathbb{P}(\mathcal{E}_2(t))\\
&\leq\sum_{t=T_0+1}^{T}R_{\max}|\mathcal{K}|\left(\frac{540CR_{\max}^3(t_0+1)^4}{e^4C_{\rho_1}s\log^{1/2}dt^{7/2}}+\frac{C_3+CC_{\rho_1}^2(1+\max_{t\leq T}w_{0,t})^3s^2\log t}{t}\right)\\
&\leq\int_{T_0}^{T}R_{\max}|\mathcal{K}|\left(\frac{540CR_{\max}^3(t_0+1)^4}{e^4C_{\rho_1}s\log^{1/2}dt^{7/2}}+\frac{C_3+CC_{\rho_1}^2(1+\max_{t\leq T}w_{0,t})^3s^2\log t}{t}\right)dt\\
&\leq R_{\max}|\mathcal{K}|\left(-\frac{1080CR_{\max}^3(t_0+1)^4}{5e^4C_{\rho_1}s(\log^{1/2}d)t^{5/2}}+C_3\log t+CC_{\rho_1}^2(1+\max_{t\leq T}w_{0,t})^3s^2\log^2t\right)\Bigg|_{T_0}^{T}\\
&\leq R_{\max}|\mathcal{K}|\left(\frac{1080CR_{\max}^3(t_0+1)^4}{5e^4C_{\rho_1}s(\log^{1/2}d)(T_0)^{5/2}}+C_3\log T+CC_{\rho_1}^2(1+\max_{t\leq T}w_{0,t})^3s^2\log^2T\right)\\
&\leq R_{\max}|\mathcal{K}|\left(C_4(t_0+1)+C_3\log T+C_5\log^2T\right)\\
&\leq R_{\max}|\mathcal{K}|\left(2C_4T_0+C_3\log T+C_5\log^2T\right),
\end{aligned} \tag{EC.113}$$

where we set

$$C_4=\frac{216CR_{\max}^3(t_0+1)^3}{e^4C_{\rho_1}s(\log^{1/2}d)(T_0)^{5/2}} \tag{EC.114}$$

$$C_5=CC_{\rho_1}^2(1+\max_{t\leq T}w_{0,t})^3s^2 \tag{EC.115}$$

and $t_0\leq T_0$. As we set $t_0=2C_0|\mathcal{K}|$ and $C_0=\mathcal{O}(s^2\log d)$, which implies $C_4=\mathcal{O}(1)$. Moreover, as $w_{0,t}=\left\lfloor\sqrt{\frac{2s^2\log(4s)\sigma^2x_{\max}^2}{C_\rho^2t\delta_t^2}}\right\rfloor\leq\left\lfloor\sqrt{\frac{\log(4s)\sigma^2x_{\max}^2}{2\lambda_{2,0}^2\log d}}\right\rfloor=\mathcal{O}(1)$ and $C_\rho=\mathcal{O}((1+\rho_{\max})^{-1})$, we can directly show that $C_3\leq\tilde{\mathcal{O}}((1+\rho_{\max})^2s^2\log d)$ and $C_5=\mathcal{O}(s^2)$.

Next, we consider the other case where $T>T_1$. Via proposition 6, we know that when $T>T_1$, for all $k\in\mathcal{K}$, $\mathcal{S}_{1,t}^k=\mathcal{S}^k\Rightarrow\rho_{\mathcal{S}^k/\mathcal{S}_{1,t}^k}=0$. In this case, we can restate (EC.92) as follows

$$\mathbb{P}\left(\|\Delta_k\|_1\geq\frac{128s\zeta}{p^*\kappa}\right)\leq5\delta_0(t,t_0)+\frac{10}{(t+1)^2}+2s\exp\left(-\frac{t\zeta^2}{2\sigma^2x_{\max}^2}\right).$$

Then, by setting $\zeta=C_\rho s^{-1}w\delta_t$, via the similar analysis to (EC.94), we can show

$$\begin{aligned}
\frac{128s\zeta}{p^*\kappa}&=\frac{128s}{p^*\kappa}\cdot\frac{p^*\kappa}{256R_{\max}\sigma e^{\sigma x_{\max}}x_{\max}(1+\rho_{\max})}\cdot s^{-1}w\delta_t\\
&=\frac{1}{1+\rho_{\max}}\cdot\frac{w\delta_t}{2R_{\max}\sigma e^{\sigma x_{\max}}x_{\max}}\\
&\leq\frac{w\delta_t}{2R_{\max}\sigma e^{\sigma x_{\max}}x_{\max}},
\end{aligned}$$

which implies that (EC.95) still holds and we have the same separation as in (EC.97). The analyses for parts (a), (b), and (c) remain unchanged. As we choose $\delta_t=C_{\rho_1}s\sqrt{\frac{\log d}{t}}$, which is different from the choice in the

$T \le T_1$ case where $\delta_t = C_{\rho_1} s\sqrt{\frac{\log d + \log t}{t}}$, the regret$_t$ calculation will be slightly different from the analysis for (EC.111). In particular, we can show that

$$\begin{aligned}
\text{regret}_t &\le CR_{\max}|\mathcal{K}|\left(\frac{540R_{\max}^3(t_0+1)^4}{e^4(t+1)^4\delta_t} + \frac{540R_{\max}^3}{(t+1)^2\delta_t} + (1+w_{0,t})^3\delta_t^2 + \frac{4s\sigma^2x_{\max}^2}{C_\rho^2 t}\left(2s\log(4s) + w_{0,t} + 1 + \frac{s^2\sigma^2x_{\max}^2}{C_\rho^2 t\delta_t^2}\right)\right) \\
&= CR_{\max}|\mathcal{K}|\left(\frac{540R_{\max}^3(t_0+1)^4}{e^4(t+1)^4\cdot C_{\rho_1}s\sqrt{\frac{\log d}{t}}} + \frac{540R_{\max}^3}{(t+1)^2\cdot C_{\rho_1}s\sqrt{\frac{\log d}{t}}} + \frac{(1+w_{0,t})^3\cdot C_{\rho_1}^2 s^2\log d}{t}\right. \\
&\left. + \frac{4s\sigma^2x_{\max}^2}{C_\rho^2 t}\left(2s\log(4s) + w_{0,t} + 1 + \frac{s^2\sigma^2x_{\max}^2}{C_\rho^2 tC_{\rho_1}^2 s^2(\frac{\log d}{t})}\right)\right) \\
&\le R_{\max}|\mathcal{K}|\left(\frac{540CR_{\max}^3(t_0+1)^4}{e^4C_{\rho_1}s(\log^{1/2}d)t^{7/2}} + \underbrace{\frac{C_3 + CC_{\rho_1}^2(1+w_{0,t})^3s^2}{t}}_{(d)}\right). \qquad \text{(EC.116)}
\end{aligned}$$

When comparing (EC.111) to (EC.116), we can show that the $\log t$ term disappeared in the (d) term. Therefore, following similar analysis as in (EC.113), the third part of the regret, when $T > T_1$, can be upper bounded as follows:

$$R_3(T) \le R_{\max}|\mathcal{K}|(2C_4T_0 + (C_3+C_5)\log T + C_5\log^2 T_1).$$

Finally, the total regret bound can be obtained by combining the bounds from all three parts: when $T \le T_1$, we have

$$\begin{aligned}
R_1(T) + R_2(T) + R_3(T) &\le 3C_0R_{\max}|\mathcal{K}|\log T + 5R_{\max}|\mathcal{K}|T_0 + 2R_{\max}|\mathcal{K}|T_0 + R_{\max}|\mathcal{K}|(2C_4T_0 + C_3\log T + C_5\log^2 T) \\
&= R_{\max}|\mathcal{K}|\left[(3C_0+C_3)\log T + C_5\log^2 T + (7+2C_4)T_0\right] \\
&= \tilde{\mathcal{O}}(s^2(\log d + \log T)\log T);
\end{aligned}$$

when $T > T_1$, we have

$$\begin{aligned}
&R_1(T) + R_2(T) + R_3(T) \\
&\le 3C_0R_{\max}|\mathcal{K}|\log T + 5R_{\max}|\mathcal{K}|T_0 + 2R_{\max}|\mathcal{K}|T_0 + R_{\max}|\mathcal{K}|(2C_4T_0 + (C_3+C_5)\log T + C_5\log^2 T_1) \\
&= R_{\max}|\mathcal{K}|\left[(3C_0+C_3+C_5)\log T + (7+2C_4)T_0 + C_5\log^2 T_1\right] \\
&= \tilde{\mathcal{O}}(s^2\log d\log T).
\end{aligned}$$

**Proof of Theorem 2** We adopt the FISTA method in Beck and Teboulle (2009) as the Lasso solver in the 2sWL procedure. For completeness, we first present the FISTA method in our settings.

**FISTA Method**:

**Require**: Loss function $\mathcal{L}(\boldsymbol{\beta})$, penalty parameter $\lambda$, total iteration number $k_0 \ge 1$, initial solution $\boldsymbol{\beta}_0$, and step-size $l_0$.

**Step 0**: Set $\boldsymbol{y}_1 = \boldsymbol{\beta}_0$, $t_1 = 1$, and $k = 1$.

**While** $k \le k_0$:

$$\boldsymbol{\beta}_k = \arg\min_{\boldsymbol{\beta}}\left\{\lambda\|\boldsymbol{\beta}\|_1 + \tfrac{l_0}{2}\left\|\boldsymbol{\beta} - \left(\boldsymbol{y}_k - \tfrac{1}{l_0}\nabla\mathcal{L}(\boldsymbol{y}_k)\right)\right\|_2^2\right\} \quad (*)$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$$

$$\boldsymbol{y}_{k+1} = \boldsymbol{\beta}_k + \tfrac{t_k - 1}{t_{k+1}}\cdot(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1})$$

$$k = k+1$$

The major computation cost in FISTA is from part $(*)$, in which we need the full gradient $\nabla\mathcal{L}(\boldsymbol{y_k})$. By definition, we have $\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{j=1}^{n} f(R_j|\boldsymbol{X}_j^T\boldsymbol{\beta})$. Hence, to evaluate each $\nabla\mathcal{L}(\boldsymbol{y_k})$, we need to compute $\mathcal{O}(dn)$ scalers multiplication. Let $\boldsymbol{\beta}^*$ be the optimal solution, $\boldsymbol{\beta}_0$ be the initial solution, and the Theorem 4.4 in Beck and Teboulle (2009) implies that for any $k \geq 1$

$$\mathcal{L}(\boldsymbol{\beta}_k) + \lambda\|\boldsymbol{\beta}_k\|_1 - \mathcal{L}(\boldsymbol{\beta}^*) + \lambda\|\boldsymbol{\beta}^*\|_1 \leq \mathcal{O}\left(\frac{L\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2}{(k+1)^2}\right), \tag{EC.117}$$

where $L$ is the Lipschitz constant of $\mathcal{L}(\boldsymbol{\beta})$ function and will be on the order of $\mathcal{O}(x_{\max}bd)$. Therefore, to achieve $\epsilon$-optimal solution, the required total iterations $k_0$ will be on the order of $\mathcal{O}(x_{\max}b\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2\epsilon^{-1/2}) = \mathcal{O}(x_{\max}b^3\epsilon^{-1/2})$, where we use $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 \leq 4b^2d^2$ by using $\|\boldsymbol{\beta}\|_1 \leq b$ for all feasible $\boldsymbol{\beta}$ in assumption A.1. Therefore, the total computational cost for running FISTA becomes $\mathcal{O}(x_{\max}b^3dn\epsilon^{-1/2})$. Note that at step $T$, each arm can not be pulled more than $T$ times, so the maximum computation cost of FISTA will be $\mathcal{O}(x_{\max}b^3d^4T\epsilon^{-1/2})$.

Next, we will upper bound the total number of the FISTA method called by time $T$. At each step, the G-MCP-Bandit algorithm will require to update $\boldsymbol{\beta}_k^{\text{random}}$ and $\boldsymbol{\beta}_k^{\text{whole}}$ by 2sWL for $k \in \mathcal{K}$ and each 2sWL procedure will need to run FISTA two times. So the average computation cost will be

$$\text{Average computation cost} \leq \mathcal{O}\left(\frac{1}{T}\cdot\sum_{k\in\mathcal{K}}\sum_{t=1}^{T} 2x_{\max}b^3dt\epsilon^{-1/2}\right) = \mathcal{O}(|\mathcal{K}|x_{\max}b^3d^4T\epsilon^{-1/2}). \tag{EC.118}$$

Next, we consider the long-run computation cost. We can reduce the computation cost with a warm start from the previous step. Via Proposition 3, we can show that with high probability, for $T \geq \max\{T_1, t_0^2\}$ and $\zeta = \frac{\epsilon^2 p^*\kappa}{16s}$, the following inequality holds:

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \epsilon^{\frac{1}{4}}\right) \geq 1 - \frac{10}{(T+1)^2} - 2\exp\left(-\frac{(T+1)(\epsilon^{1/4}p^*\kappa)^2}{512s^2\sigma^2x_{\max}^2} + \log s\right). \tag{EC.119}$$

Moreover, when $T \geq \max\left\{\frac{1024s^2\log(s)\sigma^2x_{\max}^2}{\epsilon^{1/2}(p^*\kappa)^2}, \frac{6144s^2\sigma^2x_{\max}^2}{\epsilon^{1/2}(p^*\kappa)^2}\log\left(\frac{2048s^2\sigma^2x_{\max}^2}{\epsilon^{1/2}(p^*\kappa)^2}\right)\right\} = \mathcal{O}(s^2\log(s)\epsilon^{-1/2})$, via Lemma EC.4, we have

$$\begin{aligned}&\frac{(T+1)\epsilon^{1/2}(p^*\kappa)^2}{1024s^2\sigma^2x_{\max}^2} \geq \log(s) \text{ and } \frac{(T+1)\epsilon^{1/2}(p^*\kappa)^2}{1024s^2\sigma^2x_{\max}^2} \geq 2\log(T+1)\\ \Rightarrow &2\exp\left(-\frac{(T+1)(\epsilon^{1/4}p^*\kappa)^2}{512s^2\sigma^2x_{\max}^2} + \log s\right) \leq \frac{1}{(T+1)^2}.\end{aligned} \tag{EC.120}$$

Combining (EC.119) and (EC.120), we have

$$\mathbb{P}\left(\|\boldsymbol{\beta}^{\text{MCP}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \epsilon^{1/4}\right) \geq 1 - \mathcal{O}(T^{-2}) \tag{EC.121}$$

and via Proposition 5 and Lemma EC.2 in E-Companion, we can show that for $T \geq \mathcal{O}(s^2\log d\epsilon^{-1/2})$, similar result holds

$$\mathbb{P}(\|\boldsymbol{\beta}^{\text{lasso}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \epsilon^{1/4}) \geq 1 - \mathcal{O}(T^{-2}). \tag{EC.122}$$

Note that (EC.121) and (EC.122) imply that for large enough $T$, both previous step solution and current step solution are close to $\boldsymbol{\beta}^{true}$ with high probability. If we use the previous step solution to initialize the FISTA algorithm, then we have

$$\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_2^2 \leq \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|_1^2 = \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^{\text{true}} + \boldsymbol{\beta}^{\text{true}} - \boldsymbol{\beta}^*\|_1^2$$

$$\begin{aligned} &\leq 2\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^{\text{true}}\|_1^2 + 2\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{\text{true}}\|_1^2 \\ &\leq 4\epsilon^{1/2}, \end{aligned} \tag{EC.123}$$

where the first inequality uses $(a+b)^2 \leq 2a^2 + 2b^2$, and the last inequality is because $\boldsymbol{\beta}_0$ is initialized with previous step's solution and $\boldsymbol{\beta}^*$ is the current step's MCP solution. Therefore, the results in (EC.118) can be improved to $\mathcal{O}(|\mathcal{K}|x_{\max}bd^2T)$.

## EC.2. Appendix: Supplemental Lemmas and Proofs

LEMMA EC.1. *Let $n$ be the size of the whole sample set and $\mathcal{A}$ be the random i.i.d. sample set consisting of $\boldsymbol{X} \in \mathbf{R}^d$ for $k \in \mathcal{K}_s$ and $\boldsymbol{X} \in U_k$ for $k \in \mathcal{K}_o$. Under assumptions A.1, A.4, and A.5, when $|\mathcal{A}| \geq C_1^{-1} \log d$, the follow inequality holds for all feasible $\boldsymbol{\xi}$ and $\boldsymbol{u}$ such that $\|\boldsymbol{u}_{\mathcal{S}^c}\|_1 \leq 3\|\boldsymbol{u}_{\mathcal{S}}\|_1$:*

$$\mathbb{P}\left(\frac{|\mathcal{A}|\kappa}{2ns}\|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \boldsymbol{u}^\top \nabla^2 \mathcal{L}(\boldsymbol{\xi})\boldsymbol{u}\right) \geq 1 - \exp(-C_1|\mathcal{A}|), \tag{EC.124}$$

*where*

$$C_1 = \min\left\{1, \kappa^2/\left(192s\sigma_2 x_{\max}^2(2+\sqrt{\sigma_2}x_{\max})\right)^2\right\}. \tag{EC.125}$$

*Proof of EC.1* Let $\mathcal{L}_{\mathcal{A}}(\boldsymbol{\beta})$ be the loss function with the sample set $\mathcal{A}$. Denote $\boldsymbol{Z}_j = \boldsymbol{X}_j\sqrt{f''_{yy}(R_j|\boldsymbol{X}_j^\top\boldsymbol{\xi})}$, where we replace $r$ and $y$ in $f''_{yy}(r|y)$ by $R_j$ and $\boldsymbol{X}_j^\top\boldsymbol{\xi}$ respectively. We then can present $\nabla^2\mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi})$ as follows:

$$\nabla^2\mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi}) = \frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\boldsymbol{X}_j\boldsymbol{X}_j^\top f''_{yy}(R_j|\boldsymbol{X}_j^\top\boldsymbol{\xi}) = \frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\boldsymbol{Z}_j\boldsymbol{Z}_j^\top.$$

As all realization of $\boldsymbol{X}$ is element-wise bounded by $x_{\max}$ (see assumption A.1) and $f''_{yy}(r_j|\boldsymbol{x}_j^\top\boldsymbol{\xi}) \leq \sigma_2$ (see assumptions A.4), $\boldsymbol{Z}_j$ is element-wise bounded by $\|\boldsymbol{Z}_j\|_\infty = \left\|\boldsymbol{X}_j\sqrt{f''_{yy}(R_j|\boldsymbol{X}_j^\top\boldsymbol{\xi})}\right\|_\infty \leq \sqrt{\sigma_2}x_{\max} \doteq z_{\max}$. Then, we use Bühlmann and Van De Geer (2011) to build the connection between the sample matrix $\frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\boldsymbol{Z}_j\boldsymbol{Z}_j^\top$ and its population counterpart $\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{Z}_j\boldsymbol{Z}_j^\top]$. By setting $K = z_{\max}$ and $\sigma_0 = \sqrt{2}z_{\max}$ in the exercise 14.3 in Bühlmann and Van De Geer (2011), for $t > 0$, we have

$$P\left\{\left\|\frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\boldsymbol{Z}_j\boldsymbol{Z}_j^\top - \mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{Z}_j\boldsymbol{Z}_j^\top]\right\|_\infty \geq 2z_{\max}^2 t + 4z_{\max}^2\sqrt{t} + \sqrt{8}z_{\max}^2\lambda\left(\frac{\sqrt{2}}{2}, |\mathcal{A}|, \binom{d}{2}\right)\right\} \leq \exp\left(-|\mathcal{A}|t\right), \tag{EC.126}$$

where $\lambda\left(\frac{\sqrt{2}}{2}, |\mathcal{A}|, \binom{d}{2}\right) = \sqrt{\frac{2\log(d(d-1))}{|\mathcal{A}|}} + \frac{z_{\max}\log(d(d-1))}{|\mathcal{A}|}$.

When $|\mathcal{A}| \geq C_1^{-1}\log d$ and $t = C_1$ in (EC.126), the following inequalities hold:

$$\begin{aligned}
2z_{\max}^2 t + 4z_{\max}^2\sqrt{t} &= 2z_{\max}^2 C_1 + 4z_{\max}^2\sqrt{C_1} \\
&\leq 6z_{\max}^2\sqrt{C_1} \qquad \text{(EC.127)} \\
\sqrt{8}z_{\max}^2\lambda\left(\frac{\sqrt{2}}{2}, |\mathcal{A}|, \binom{d}{2}\right) &\leq \sqrt{8}z_{\max}^2\left(\sqrt{\frac{2\log(d^2)}{|\mathcal{A}|}} + \frac{z_{\max}\log(d^2)}{|\mathcal{A}|}\right) \\
&= \sqrt{8}z_{\max}^2\left(\sqrt{\frac{4\log d}{|\mathcal{A}|}} + \frac{2z_{\max}\log d}{|\mathcal{A}|}\right) \\
&\leq \sqrt{8}z_{\max}^2\left(2\sqrt{C_1} + 2z_{\max}C_1\right) \\
&\leq 4\sqrt{2}z_{\max}^2(1+z_{\max})\sqrt{C_1}, \qquad \text{(EC.128)}
\end{aligned}$$

where in (EC.127) and (EC.128) we use the fact that when $C_1 \leq 1$, we have $\sqrt{C_1} \geq C_1$. Combining (EC.127) and (EC.128), we have

$$\begin{aligned}
2z_{\max}^2 t + 4z_{\max}^2\sqrt{t} + \sqrt{8}z_{\max}^2\lambda\left(\frac{\sqrt{2}}{2}, |\mathcal{A}|, \binom{d}{2}\right) &\leq 2z_{\max}^2\left(3 + 2\sqrt{2}(1+z_{\max})\right)\sqrt{C_1} \\
&< 6z_{\max}^2\left(2+z_{\max}\right))\sqrt{C_1} \leq \frac{\kappa}{32s}, \qquad \text{(EC.129)}
\end{aligned}$$

where (EC.129) uses $\sqrt{2} \leq \frac{3}{2}$ and $C_1 \leq \kappa^2 / \left(192 s \sigma_2 x_{\max}^2 \left(2 + \sqrt{\sigma_2} x_{\max}\right)\right)^2$. Combining (EC.126) and (EC.129), we have

$$\mathbb{P}\left\{ \left\| \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} \boldsymbol{Z}_j \boldsymbol{Z}_j^\top - \mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{Z}_j \boldsymbol{Z}_j^\top] \right\|_\infty \leq \frac{\kappa}{32 s} \right\} \geq 1 - \exp\left(-C_1 |\mathcal{A}|\right). \tag{EC.130}$$

By the definition of $\boldsymbol{Z}_j$, we can verify that $\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{Z}_j \boldsymbol{Z}_j^\top] = \mathbb{E}_{\boldsymbol{X},\epsilon}[f''_{yy}(R_j | \boldsymbol{\xi}^\top \boldsymbol{X}_j) \boldsymbol{X}_j \boldsymbol{X}_j^\top]$. Via assumption A.5, we have restricted eigenvalue condition holds for $\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{Z}_j \boldsymbol{Z}_j^\top]$ with parameter $\kappa$. Combining (EC.130) with the Corollary 6.8 in Bühlmann and Van De Geer (2011), we set $\tilde{\lambda} = \frac{\kappa}{32 s}$ in Corollary 6.8 to show $\frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} \boldsymbol{Z}_j \boldsymbol{Z}_j^\top$ also has restricted eigenvalue condition with parameter $\kappa/2$, which implies

$$\begin{aligned} &\mathbb{P}\left( \frac{\kappa}{2s} \|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} \boldsymbol{Z}_j \boldsymbol{Z}_j^\top \right) \geq 1 - \exp(-C_1 |\mathcal{A}|) \\ \Rightarrow &\mathbb{P}\left( \frac{\kappa}{2s} \|\boldsymbol{u}_{\mathcal{S}}\|_1^2 \leq \boldsymbol{u}^\top \nabla^2 \mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi}) \boldsymbol{u} \right) \geq 1 - \exp(-C_1 |\mathcal{A}|). \end{aligned} \tag{EC.131}$$

Note that for any realizations $\{\boldsymbol{x}_j, r_j\}$ of $\{\boldsymbol{X}_j, R_j\}$, we have

$$\begin{aligned} \boldsymbol{u}^T \nabla^2 \mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi}) \boldsymbol{u} &= \boldsymbol{u}^T \left[ \frac{1}{n} \sum_{j \in \mathcal{A}} x_j x_j^\top f''_{yy}(r_j | \boldsymbol{x}_j^T \boldsymbol{\xi}) \right] \boldsymbol{u} + \boldsymbol{u}^T \left[ \frac{1}{n} \sum_{j \in (\mathcal{A})^c} x_j x_j^\top f''_{yy}(r_j | \boldsymbol{x}_j^T \boldsymbol{\xi}) \right] \boldsymbol{u} \\ &\geq \boldsymbol{u}^T \left[ \frac{1}{n} \sum_{j \in \mathcal{A}} x_j x_j^\top f''_{yy}(r_j | \boldsymbol{x}_j^T \boldsymbol{\xi}) \right] \boldsymbol{u} = \frac{|\mathcal{A}|}{n} \boldsymbol{u}^\top \nabla^2 \mathcal{L}_{\mathcal{A}}(\boldsymbol{\xi}) \boldsymbol{u}. \end{aligned} \tag{EC.132}$$

The desirable result follows directly by combining (EC.131) and (EC.132).

LEMMA EC.2. *Let $n$ be the size of the whole sample set and $\mathcal{A}$ be the random i.i.d. sample set consisting of $\boldsymbol{X} \in \mathbf{R}^d$ for $k \in \mathcal{K}_s$ and $\boldsymbol{X} \in U_k$ for $k \in \mathcal{K}_o$. Per assumptions A.1, A.4 and A.5, when $|\mathcal{A}| \geq C_1^{-1} \log d$, the follow result holds with probability at least $1 - \exp\left(-C_1 |\mathcal{A}|\right) - 2d \exp\left(-\frac{n\lambda^2}{8\sigma^2 x_{\max}^2}\right)$:*

$$\|\boldsymbol{\beta}^{\text{lasso}} - \boldsymbol{\beta}^{\text{true}}\|_1 \leq \frac{24 n s \lambda}{|\mathcal{A}| \kappa}, \tag{EC.133}$$

*where $\boldsymbol{\beta}^{\text{lasso}}$ is the lasso estimator and $C_1$ is defined in* (EC.125)

*Proof of lemma EC.2* We first show that $\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^c} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^c}\|_1 \leq 3 \|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}} - \boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1$ holds. As $\boldsymbol{\beta}^{\text{lasso}}$ is the optimal solution, we have

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}^{\text{lasso}}) + \lambda \|\boldsymbol{\beta}^{\text{lasso}}\|_1 &\leq \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) + \lambda \|\boldsymbol{\beta}^{\text{true}}\|_1 \\ \mathcal{L}(\boldsymbol{\beta}^{\text{lasso}}) - \mathcal{L}(\boldsymbol{\beta}^{\text{true}}) + \lambda \|\boldsymbol{\beta}^{\text{lasso}}\|_1 &\leq \lambda \|\boldsymbol{\beta}^{\text{true}}\|_1 \qquad \text{(EC.134)} \\ \nabla \mathcal{L}(\boldsymbol{\beta}^{\text{true}})^T (\boldsymbol{\beta}^{\text{lasso}} - \boldsymbol{\beta}^{\text{true}}) + \lambda \|\boldsymbol{\beta}^{\text{lasso}}\|_1 &\leq \lambda \|\boldsymbol{\beta}^{\text{true}}\|_1 \qquad \text{(EC.135)} \\ -\|\nabla \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty \|\boldsymbol{\beta}^{\text{lasso}} - \boldsymbol{\beta}^{\text{true}}\|_1 + \lambda \|\boldsymbol{\beta}^{\text{lasso}}\|_1 &\leq \lambda \|\boldsymbol{\beta}^{\text{true}}\|_1, \qquad \text{(EC.136)} \end{aligned}$$

where (EC.135) uses the convexity of $\mathcal{L}(\boldsymbol{\beta}^{\text{lasso}})$. Denote event $\mathcal{E}_0$ as follows:

$$\mathcal{E}_0 = \left\{ \|\nabla \mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty < \frac{1}{2} \lambda \right\}. \tag{EC.137}$$

Under $\mathcal{E}_0$, (EC.136) can be further simplified into

$$\begin{aligned}
-\frac{1}{2}\lambda\|\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}}\|_1+\lambda\|\boldsymbol{\beta}^{\text{lasso}}\|_1 &\le \lambda\|\boldsymbol{\beta}^{\text{true}}\|_1\\
-\frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}}\|_1+\|\boldsymbol{\beta}^{\text{lasso}}\|_1 &\le \|\boldsymbol{\beta}^{\text{true}}\|_1\\
-\frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1-\frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^{\text{c}}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^{\text{c}}}\|_1+\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}\|_1+\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^{\text{c}}}\|_1 &\le \|\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1+\|\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^{\text{c}}}\|_1. \qquad \text{(EC.138)}
\end{aligned}$$

As $\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^{\text{c}}}=\mathbf{0}$ by definition, we then have

$$\begin{aligned}
-\frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1-\frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^{\text{c}}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^{\text{c}}}\|_1+\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}\|_1+\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^{\text{c}}}-\mathbf{0}\|_1 &\le \|\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1+0\\
-\frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1-\frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^{\text{c}}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^{\text{c}}}\|_1+\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}\|_1+\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^{\text{c}}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^{\text{c}}}\|_1 &\le \|\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1+0\\
\frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^{\text{c}}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^{\text{c}}}\|_1 &\le \frac{1}{2}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1+\|\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1-\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}\|_1\\
\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}^{\text{c}}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}^{\text{c}}}\|_1 &\le 3\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1 \qquad \text{(EC.139)}
\end{aligned}$$

Then, by Lemma EC.1, we obtain

$$\mathbb{P}\left((\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}})^\top\nabla^2\mathcal{L}(\boldsymbol{\xi})(\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}})\ge\frac{|\mathcal{A}|\kappa}{2ns}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1^2\right)\ge 1-\exp\left(-C_1|\mathcal{A}|\right). \qquad \text{(EC.140)}$$

Now, we turn back to (EC.134) and use the Taylor expansion on $\mathcal{L}(\boldsymbol{\beta}^{\text{lasso}})$ at $\boldsymbol{\beta}^{\text{true}}$. Then, the following inequality holds for some $\boldsymbol{\xi}$ between $\boldsymbol{\beta}^{\text{true}}$ and $\boldsymbol{\beta}^{\text{lasso}}$:

$$\nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top(\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}})+\frac{1}{2}(\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}})^\top\nabla^2\mathcal{L}(\boldsymbol{\xi})(\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}})+\lambda\|\boldsymbol{\beta}^{\text{lasso}}\|_1\le\lambda\|\boldsymbol{\beta}^{\text{true}}\|_1. \qquad \text{(EC.141)}$$

Combining (EC.140) and (EC.141), we know that with probability at least $1-\exp(-C_1 n)$, the following results hold:

$$\begin{aligned}
\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1^2 &\le -\nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})^\top(\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}})+\lambda\left(\|\boldsymbol{\beta}^{\text{true}}\|_1-\|\boldsymbol{\beta}^{\text{lasso}}\|_1\right)\\
\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\|_1^2 &\le \sum_{i\in\mathcal{S}\cup\mathcal{S}^{\text{c}}}\left[-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})(\beta_i^{\text{lasso}}-\beta_i^{\text{true}})+\lambda\left(|\beta_i^{\text{true}}|-|\beta_i^{\text{lasso}}|\right)\right]. \qquad \text{(EC.142)}
\end{aligned}$$

We then separately consider $i\in\mathcal{S}$ and $i\in\mathcal{S}^{\text{c}}$ as follow:

$$\begin{aligned}
&\sum_{i\in\mathcal{S}}\left[-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})(\beta_i^{\text{lasso}}-\beta_i^{\text{true}})-\lambda\left(|\beta_i^{\text{lasso}}|-|\beta_i^{\text{true}}|\right)\right]\\
\le&\left\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right\|_\infty\left\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\right\|_1+\lambda\left\|\boldsymbol{\beta}^{\text{lasso}}_{\mathcal{S}}-\boldsymbol{\beta}^{\text{true}}_{\mathcal{S}}\right\|_1 \qquad \text{(EC.143)}
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i\in\mathcal{S}^{\text{c}}}\left[-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})(\beta_i^{\text{lasso}}-\beta_i^{\text{true}})-\lambda\left(|\beta_i^{\text{lasso}}|-|\beta_i^{\text{true}}|\right)\right]\\
\le&\sum_{i\in\mathcal{S}^{\text{c}}}\left[-\nabla_i\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\beta_i^{\text{lasso}}-\lambda|\beta_i^{\text{lasso}}|\right]\\
\le&\sum_{i\in\mathcal{S}^{\text{c}}}\left(|\nabla_i\mathcal{L}(\beta^{\text{true}})|-\lambda\right)|\beta_i^{\text{lasso}}|\le 0, \qquad \text{(EC.144)}
\end{aligned}$$

where the last inequality uses $\|\nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\|_\infty\le\frac{1}{2}\lambda$ in $\mathcal{E}_0$.

Combining (EC.142), (EC.143) and (EC.144), we can show that

$$\begin{aligned}
&\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}_{\mathcal{S}}^{\text{lasso}}-\boldsymbol{\beta}_{S}^{\text{true}}\|_1^2 \le \left\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right\|_\infty \left\|\boldsymbol{\beta}_{\mathcal{S}}^{\text{lasso}}-\boldsymbol{\beta}_{\mathcal{S}}^{\text{true}}\right\|_1+\lambda\left\|\boldsymbol{\beta}_{\mathcal{S}}^{\text{lasso}}-\boldsymbol{\beta}_{\mathcal{S}}^{\text{true}}\right\|_1\\
&\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}_{\mathcal{S}}^{\text{lasso}}-\boldsymbol{\beta}_{S}^{\text{true}}\|_1 \le \left\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right\|_\infty+\lambda\\
&\frac{|\mathcal{A}|\kappa}{4ns}\|\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}}\|_1 \le 4\left(\left\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right\|_\infty+\lambda\right) \qquad \text{(EC.145)}\\
&\|\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}}\|_1 \le \frac{16ns}{|\mathcal{A}|\kappa}\left(\left\|\nabla_{\mathcal{S}}\mathcal{L}(\boldsymbol{\beta}^{\text{true}})\right\|_\infty+\lambda\right), \qquad \text{(EC.146)}
\end{aligned}$$

where (EC.145) uses $\|\boldsymbol{\beta}_{\mathcal{S}^c}^{\text{lasso}}-\boldsymbol{\beta}_{\mathcal{S}^c}^{\text{true}}\|_1 \le 3\|\boldsymbol{\beta}_{\mathcal{S}}^{\text{lasso}}-\boldsymbol{\beta}_{\mathcal{S}}^{\text{true}}\|_1$ in (EC.139). Under event $\mathcal{E}_0$, (EC.146) can be further reduced to:

$$\|\boldsymbol{\beta}^{\text{lasso}}-\boldsymbol{\beta}^{\text{true}}\|_1 \le \frac{24ns}{|\mathcal{A}|\kappa}\lambda. \qquad \text{(EC.147)}$$

Now, we assess the probability of event $\mathcal{E}_0$. The $i$-th element of $\nabla\mathcal{L}(\boldsymbol{\beta}^{\text{true}})$ is $\frac{1}{n}\sum_{j=1}^n X_{j,i}f_y'(R_j|\boldsymbol{X}_j^\top\boldsymbol{\beta}^{\text{true}})$. Under assumptions A.1 and A.4, $X_{j,i}f_y'(R_j|\boldsymbol{X}_j^\top\boldsymbol{\beta}^{\text{true}})$ are $x_{\max}^2\sigma^2$-subguassian random variables for given sample $\boldsymbol{X}_j$. We can use the Hoeffding's inequality and union bound to build the following probability bound.

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^n X_{j,i}f_y'(R_j|\boldsymbol{X}_j^\top\boldsymbol{\beta}^{\text{true}})\right|\ge t\right) \le 2\exp\left(-\frac{nt^2}{2\sigma^2x_{\max}^2}\right)\\
\Rightarrow&\mathbb{P}\left(\max_i\left|\frac{1}{n}\sum_{j=1}^n X_{j,i}f_y'(R_j|\boldsymbol{X}_j^\top\boldsymbol{\beta}^{\text{true}})\right|\le t\right) \ge 1-2d\exp\left(-\frac{nt^2}{2\sigma^2x_{\max}^2}\right), \qquad \text{(EC.148)}
\end{aligned}$$

Setting $t=\frac{1}{2}\lambda$, we will have event $\mathcal{E}_0$ defined in (EC.137) holds with at least probability $1-2d\exp\left(-\frac{n\lambda^2}{8\sigma^2x_{\max}^2}\right)$. The desirable result directly follows by (EC.146), (EC.147), and (EC.148).

LEMMA EC.3. *Under assumptions A.1, A.3, A.4, and A.5, for any feasible $\boldsymbol{x}$, $\boldsymbol{\beta}_i$ and $i\in\mathcal{K}$, the following two statements hold.*

1. $|\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}}]-\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i]| \le R_{\max}e^{\sigma x_{\max}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^{\text{true}}\|_1}\sigma x_{\max}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^{\text{true}}\|_1$
2. *Moreover, if* $\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^{\text{true}}\|_1 \le \min\left\{\frac{1}{\sigma x_{\max}},\frac{h}{4e\sigma R_{\max}x_{\max}}\right\}$, *then we have* $\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] > \max_{j\neq i}\mathbb{E}_\epsilon[R_j|\boldsymbol{x}^\top\boldsymbol{\beta}_j]+\frac{h}{2}$ *for* $i\in\mathcal{K}_o$ *and* $\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] < \max_{j\neq i}\mathbb{E}_\epsilon[R_j|\boldsymbol{x}^\top\boldsymbol{\beta}_j]-\frac{1}{2}h$ *for* $i\in\mathcal{K}_s$.

*Proof of Lemma EC.3* ***To show the part 1.*** We first expand the left-hand-side as follows.

$$\begin{aligned}
&\left|\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}}]-\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i]\right|\\
=&\left|\int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}})}dr_i-\int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i)}dr_i\right| \qquad \text{(EC.149)}\\
=&\left|\int_{-\infty}^{+\infty} r_i\left(e^{-f(r_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}})}-e^{-f(r_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i)}\right)dr_i\right|.\\
=&\left|\int_{-\infty}^{+\infty} -r_i\left(e^{-f(r_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i)}\right)'\Big|_{\boldsymbol{\beta}_i=\boldsymbol{\beta}_i^{\text{true}}+\boldsymbol{\delta}}\boldsymbol{x}^\top(\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^{\text{true}})dr_i\right|, \qquad \text{(EC.150)}
\end{aligned}$$

where (EC.149) uses $f$ being the sample-wise negative log-likelihood loss function and $\boldsymbol{\delta}$ in (EC.150) is between $\mathbf{0}$ and $\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}$. We then pull $\boldsymbol{x}^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}})$ out of the integral:

$$\begin{aligned}
&\left| \int_{-\infty}^{+\infty} -r_i \left( e^{-f(r_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i)} \right)' \Big|_{\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta}} \boldsymbol{x}^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}) dr_i \right| \\
=&\left| \boldsymbol{x}^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}) \int_{-\infty}^{+\infty} -r_i \left( e^{-f(r_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i)} \right)' \Big|_{\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta}} dr_i \right| \\
\leq&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top (\boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta}))} f_y'(r_i|\boldsymbol{x}^\top (\boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta})) dr_i \right| x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1. \qquad \text{(EC.151)}
\end{aligned}$$

As we assume $|f_y'(\cdot|\cdot)|$ is bounded by $\sigma$ in assumption A.4, (EC.151) is upper bounded by

$$\begin{aligned}
&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top (\boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta}))} f_y'(r_i|\boldsymbol{x}^\top (\boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta})) dr_i \right| x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \\
\leq&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top (\boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta}))} dr_i \right| \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1. \qquad \text{(EC.152)}
\end{aligned}$$

We then expand term $f(r_i|\boldsymbol{x}^\top(\boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta}))$ in (EC.152), and there exists a $\boldsymbol{\xi}$ between $\boldsymbol{\beta}^{\text{true}}$ and $\boldsymbol{\beta}^{\text{true}} + \boldsymbol{\delta}$ such that

$$\begin{aligned}
&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top (\boldsymbol{\beta}_i^{\text{true}} + \boldsymbol{\delta}))} dr_i \right| \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \\
=&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}}) - f_y'(r_i|\boldsymbol{x}^\top \boldsymbol{\xi}) \boldsymbol{x}^\top \boldsymbol{\delta}} dr_i \right| \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j^{\text{true}}\|_1 \\
\leq&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}}) + |f_y'(r_i|\boldsymbol{x}^\top \boldsymbol{\xi})| \|\boldsymbol{x}\|_\infty \|\boldsymbol{\delta}\|_1} dr_i \right| \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \\
\leq&\left| \int_{-\infty}^{+\infty} r_i e^{-f(r_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}})} dr_i \right| e^{\sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1} \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \qquad \text{(EC.153)} \\
=&|\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}}]| e^{\sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1} \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \qquad \text{(EC.154)}
\end{aligned}$$

where (EC.153) uses that $\boldsymbol{\delta}$ is between $\mathbf{0}$ and $\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}$, which implies $\|\boldsymbol{\delta}\|_1 \leq \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1$, and (EC.154) comes from the definition of $\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}}]$. Combining $\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}}] \in (0, R_{\max}]$ in assumption A.1 and (EC.154), we have:

$$\left| \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i] \right| \leq R_{\max} e^{\sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1} \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1. \qquad \text{(EC.155)}$$

***To show the part 2.*** If $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \leq \frac{1}{\sigma x_{\max}}$, then we can show that

$$e^{\sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1} \leq e. \qquad \text{(EC.156)}$$

Combining (EC.156) and (EC.155), we obtain

$$\begin{aligned}
\left| \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i] \right| \leq& R_{\max} e^{\sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1} \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \\
\leq& R_{\max} e \sigma x_{\max} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \qquad \text{(EC.157)}
\end{aligned}$$

Let $j_1 = \arg\max_{j \neq i} \mathbb{E}_\epsilon[R_j|\boldsymbol{x}^\top \boldsymbol{\beta}_j^{\text{true}}]$. We first consider the case with $i \in \mathcal{K}_o$. Under assumption A.3, for any $x \in U_i$, $i \in \mathcal{K}_o$, the following inequalities hold:

$$\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top \boldsymbol{\beta}_i^{\text{true}}] > \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top \boldsymbol{\beta}_{j_1}^{\text{true}}] + h$$

$$
\begin{aligned}
\Rightarrow \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] > & \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}^{\text{true}}] - \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] \\
& + \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] + h \\
\Rightarrow \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] - \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] > & -\left|\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}})]\right| \\
& - \left|\mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}^{\text{true}}] - \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}]\right| + h. \qquad \text{(EC.158)}
\end{aligned}
$$

If $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}}\|_1 \le \frac{h}{4e\sigma R_{\max} x_{\max}}$, then we have

$$
\|R_{\max} e\sigma x_{\max}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^{\text{true}})\|_1 \le \frac{h}{4}. \qquad \text{(EC.159)}
$$

Combining (EC.157), (EC.158), and (EC.159), we will have

$$
\begin{aligned}
\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] - \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] > & -\frac{h}{4} - \frac{h}{4} + h \\
\Rightarrow \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] > & \max_{j\neq i} \mathbb{E}_\epsilon[R_j|\boldsymbol{x}^\top\boldsymbol{\beta}_j] + \frac{h}{2}, \qquad \text{(EC.160)}
\end{aligned}
$$

where the last inequality uses $j_1 = \arg\max_{j\neq i} \mathbb{E}_\epsilon[R_j|\boldsymbol{x}^\top\boldsymbol{\beta}_j^{\text{true}}]$.

We then consider the case where $i \in \mathcal{K}_s$. Under assumption A.3, for any suboptimal arm $i \in \mathcal{K}_s$, we have

$$
\begin{aligned}
\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}}] + h < & \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}^{\text{true}}] \\
\Rightarrow \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] + h < & \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}^{\text{true}}] - \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] \\
& + \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] \\
\Rightarrow \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] > & -|\mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i^{\text{true}}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i]| \\
& - |\mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}^{\text{true}}] - \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}]| + h \\
\Rightarrow \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] - \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] > & -\frac{1}{4}h - \frac{1}{4}h + h \\
\Rightarrow \mathbb{E}_\epsilon[R_i|\boldsymbol{x}^\top\boldsymbol{\beta}_i] < & \mathbb{E}_\epsilon[R_{j_1}|\boldsymbol{x}^\top\boldsymbol{\beta}_{j_1}] - \frac{1}{2}h = \max_{j\neq i} \mathbb{E}_\epsilon[R_j|\boldsymbol{x}^\top\boldsymbol{\beta}_j] - \frac{1}{2}h,
\end{aligned}
$$

where the second-to-last inequality uses (EC.157) and (EC.159).

LEMMA EC.4. *Let $\alpha$ be a positive number. When $x > \max\{3\alpha^{-1}\log\alpha^{-1}, 0\}$, we have $\alpha x \ge \log x$.*

*Proof of Lemma EC.4* Let $f(x) = \alpha x - \log x$. We first prove that for the case where $\alpha < e^{-1}$, $f(x)$ is non-negative for $x > \max\{3\alpha^{-1}\log\alpha^{-1}, 0\}$ via solving the following equation:

$$
\alpha x - \log x = 0 \Leftrightarrow \frac{\exp(\alpha x)}{x} = 1 \Leftrightarrow x\exp(-\alpha x) = 1 \Leftrightarrow -\alpha x\exp(-\alpha x) = -\alpha, \qquad \text{(EC.161)}
$$

whose non-negative solution is $x = -\alpha^{-1}W_{-1}(-\alpha)$ where $W_{-1}(\cdot)$ is the Lambert $W$ function. Combining this result with the monotonicity of $f(x)$, we have $f(x) \ge 0$ for all $x \ge -\alpha^{-1}W_{-1}(-\alpha)$. Next, we will show that $-\alpha^{-1}W_{-1}(-\alpha) \le \max\{3\alpha^{-1}\log\alpha^{-1}, 0\}$. By setting $u = -\log\alpha - 1$ in Theorem 1 of Chatzigeorgiou (2013), we have

$$
W_{-1}(-e^{-(-\log\alpha-1)-1}) \ge -1 - \sqrt{2(-\log\alpha-1)} - (-\log\alpha - 1)
$$

$$\Rightarrow W_{-1}(-\alpha) \geq -\sqrt{2(-\log\alpha - 1)} + \log\alpha$$
$$\Rightarrow W_{-1}(-\alpha) \geq -2\sqrt{-\log\alpha} + \log\alpha$$
$$\Rightarrow -\alpha^{-1} W_{-1}(-\alpha) \leq \alpha^{-1}(2\sqrt{-\log\alpha} - \log\alpha)$$
$$\Rightarrow -\alpha^{-1} W_{-1}(-\alpha) \leq -3\alpha^{-1}\log a = 3\alpha^{-1}\log\alpha^{-1} \leq \max\{3\alpha^{-1}\log\alpha^{-1}, 0\}, \tag{EC.162}$$

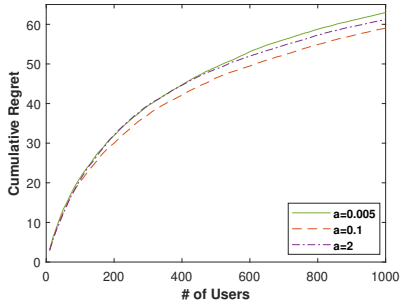
where the last inequality uses $-\log\alpha \geq \log e = 1$ when $\alpha \leq e^{-1}$. Therefore, we have $f(x) \geq 0$ for all $x \geq \max\{3\alpha^{-1}\log\alpha^{-1}, 0\}$.

Next, we consider the case where $\alpha \geq e^{-1}$. We can verify that $f(x)$ is convex with the minimum value $(1 + \log\alpha)$. If $\alpha \geq e^{-1}$, then $f(x)$ is non-negative, which implies that $\alpha x \geq \log x$ for all $x > 0$. Finally, the lemma follows directly by combining both the $\alpha < e^{-1}$ case and the $\alpha \geq e^{-1}$ case.

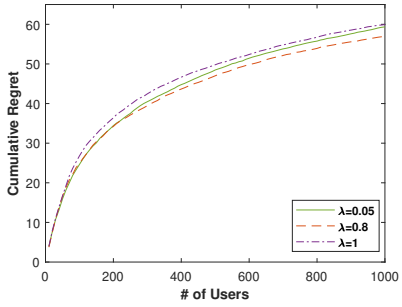
## EC.3. Appendix: Sensitivity Analyses

In this section, we conduct sensitivity analyses for input parameters (i.e., $a$, $\lambda_{1,0}$, $\lambda_{2,0}$, $h$, $t_0$) in the G-MCP-Bandit algorithm and the upper-bound for significant covariates dimension (i.e., $s$). In particular, we will hold the baseline case, where $d=50$, $s=10$, $\lambda_1=\lambda_{2,0}=1$, $h=1$, $t_0=4$, $a=0.1$, and $k=2$, largely unchanged while varying one of $a\in\{0.005,0.1,2\}$, $\lambda=\lambda_1=\lambda_{2,0}\in\{0.05,0.8,1\}$, $h\in\{0.25,0.5,2\}$, $t_0\in\{0.3,0.5,1\}$, and $s\in\{3,4,5\}$. For each sensitivity analysis, we perform 100 trials and report the average cumulative regret for the G-MCP-Bandit algorithm up to 1000 users.

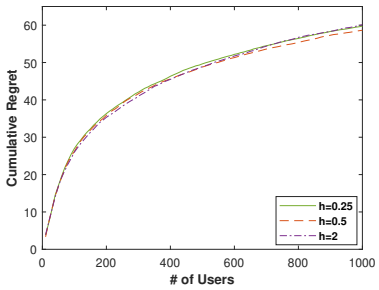
**Figure EC.1** **Sensitivity analysis, where $d=50$, $s=10$, $\lambda_1=\lambda_{2,0}=1$, $h=1$, $t_0=4$, $a=0.1$, and $k=2$ as the baseline.**



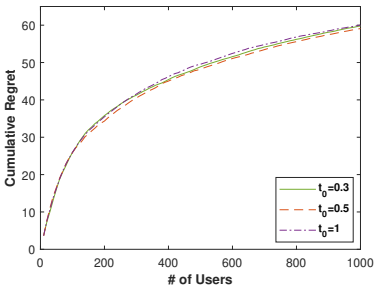

(a) Influence of $a$



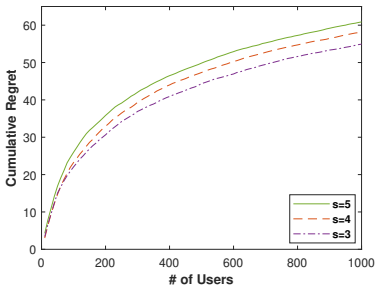

(b) Influence of $\lambda_1=\lambda_{2,0}=\lambda$




(c) Influence of $h$




(d) Influence of $t_0$




(e) Influence of $s$

We observe that the cumulative regret for the G-MCP-Bandit algorithm is robust with respect to the choices of its input parameters. Specifically, in Figure EC.1(a), (b), (c), and (d), the cumulative regret remains largely unchanged, when we vary the G-MCP-Bandit algorithm's input parameters (i.e., $a$, $\lambda_1$, $\lambda_{2,0}$, $h$, $t_0$). Furthermore, we find that the cumulative regret exhibits a non-monotonic pattern with respect to these input parameters changes and that the cumulative regret seems to be minimized for the median values of these parameters in EC.1(a), (b), (c), and (d). Hence, despite the mild improvement in the cumulative regret, actively tuning parameters may continue to be beneficial for decision-makers in practice.

At last, Figure EC.1(e) reports the influence of the upper bound for the significant covariates dimension ($s$) on the cumulative regret performance. In particular, we observe that the cumulative regret is monotonically increasing in $s$. Note that decreasing $s$ suggests a higher sparsity level and a smaller number of significant

covariates. Hence, as expected, with less non-zero parameters needed to be estimated, the G-MCP-Bandit algorithm will have better estimation accuracy, which in turn improves its regret performance.

### EC.3.1. The Knowledge of the Sparsity Level $s$

To establish the G-MCP-Bandit algorithm's regret upper bound, some input parameters may need to be selected based on the sparsity level $s$. For example, the parameter $a$ is chosen to ensure the condition $a > \frac{144s}{\kappa}$ holds in Proposition 1 and Theorem 1. Such a selection condition is standard in the high-dimensional statistics literature (e.g., Corollary 4 and Corollary 6 of Fan et al. 2014b and Lemma 5.3 of Wang et al. 2014) and the high-dimensional bandit literature (e.g., Proposition 1 of Bastani and Bayati 2020). In practice, however, decision-makers may not know the sparsity level $s$, especially without sufficient data at the beginning. Therefore, in this subsection, we will investigate the question of if decision-makers don't know the sparsity level $s$, then how the suboptimal parameter selection will influence G-MCP-Bandit's regret performance.

In particular, we use $\hat{s}$ to represent decision-makers' guess or estimation of the true sparsity level $s$. Without knowing the true $s$ value, decision-makers will tune the G-MCP-Bandit algorithm's parameters by using their estimated sparsity level $\hat{s}$. In Figure EC.2, we report five linear two-armed bandit experiments[5], in which the covariate dimension $d = 100$ and the true sparsity level $s \in \{5, 20, 30, 40, 50\}$. For each experiment, we perform 30 trials and report the average cumulative regret of five G-MCP-Bandit algorithms[6] that are tuned by using/assuming $\hat{s} = 5$, 20, 30, 40, and 50, respectively[7]. Therefore, in each experiment, only one G-MCP-Bandit algorithm's parameters are tuned by the true sparsity level $s$, and the other remaining four G-MCP-Bandit algorithms used the suboptimal parameters tuned by the wrong estimated sparsity level $\hat{s}$.

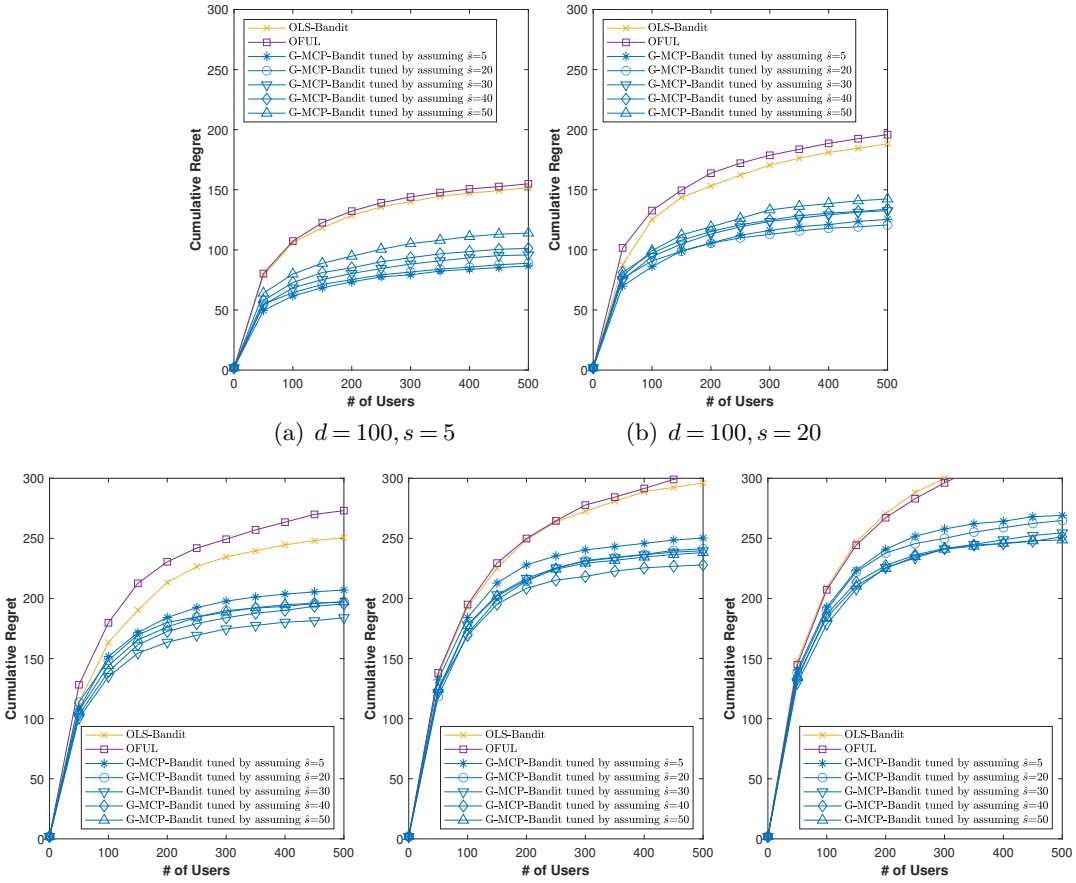
We observe that the G-MCP-Bandit algorithm's cumulative regret will be minimized when it is tuned by using the correct sparsity value (i.e., if $\hat{s} = s$). For example, in Figure EC.2(a), where the true sparsity level $s = 5$, the G-MCP-Bandit algorithm will have the lowest cumulative regret, if it was tuned by using $\hat{s} = 5$ (see the blue line with asterisk marks). In addition, we find that the larger the distance between $\hat{s}$ and $s$, the worse the G-MCP-Bandit algorithm will perform. And, the regret differences among the G-MCP-Bandit algorithms tuned by different $\hat{s}$ values tend to be narrowed as the true sparsity level $s$ increases. The regret differences among G-MCP-Bandit algorithms tuned by different $\hat{s}$ values highlight the importance of accurately estimating $s$. Therefore, to improve the G-MCP-Bandit algorithm's performance, decision-makers should, whenever possible, (1) use their earlier experience or data previously obtained from similar scenarios to improve the accuracy of estimating the sparsity level and (2) dynamically adjust the tuning parameters for the G-MCP-Bandit algorithm, when more data become available to support a better estimation of the $s$ value.

[5] Other experiments exhibit similar pattern and therefore are omitted.

[6] The Lasso-Bandit algorithm also requires the knowledge of the sparsity level to tune parameters, and therefore its cumulative regret performance depends on the estimated sparsity level $\hat{s}$. In our experiments, we observe that the G-MCP-Bandit algorithm continues to outperform the Lasso-Bandit algorithm, if both were tuned under the same $\hat{s}$. In addition, as the impact of the estimated sparsity level $\hat{s}$ on the Lasso-Bandit algorithm is nearly identical to that on the G-MCP-Bandit algorithm, we omit the Lasso-Bandit algorithm for better clarity in the figure.

[7] We also tried $\hat{s}$ value that is higher than 50, but the cumulative regret performance for cases with $\hat{s} > 50$ is very close to the $\hat{s} = 50$ case and therefore will be omitted to avoid duplication.

**Figure EC.2** **The influence of unknown $s$, where parameters for the G-MCP-Bandit algorithm are optimally tuned by assuming decision-makers' estimated sparsity levels to be $\hat{s}=5,20,30,40,$ and $50$.**




(a) $d=100, s=5$ (b) $d=100, s=20$

(c) $d=100, s=30$ (d) $d=100, s=40$ (e) $d=100, s=50$

REMARK EC.1. To theoretically remove the dependence of parameter $a$ on the sparsity level $s$, we will need to revise the existing assumption for a stronger version or introduce new assumptions. The dependence of $a$ on $s$ comes from the proof of Proposition 3, in which we want to ensure the penalty weight $w_i$ for $i \in \mathcal{S}^c$ to be positive. According to (EC.16), we must set $a=\mathcal{O}(s)$. One way to break such a dependence is to further separate $\mathcal{S}^c$ into two subsets $\mathcal{S}_3$ and $\mathcal{S}_4$. We then count the index in $\mathcal{S}^c$ with small enough penalty weight in $\mathcal{S}_3$ and $\mathcal{S}_4=\mathcal{S}^c/\mathcal{S}_3$. As the element in $\mathcal{S}_3$ indicates that the magnitude of the Lasso estimator is large while the ground truth is 0, it will happen with low probability. Thus, the cardinality of $\mathcal{S}_3$ will not be large. In fact, we can prove that $|\mathcal{S}_3| \le s$ under some additional mild conditions. Then, by setting $\hat{\mathcal{S}}=\mathcal{S}\cup\mathcal{S}^3$ and $\hat{\mathcal{S}}^c=\mathcal{S}^4$, for all $i$ in $\hat{\mathcal{S}}^c$, the penalty weights will be large enough. Therefore, if we can further introduce a stronger restricted eigenvalue condition so that $\frac{\kappa}{s}\|\boldsymbol{u}_{\hat{\mathcal{S}}}\|_1^2 \le \boldsymbol{u}^\top \mathbb{E}[\nabla^2 \mathcal{L}(\boldsymbol{\xi})]\boldsymbol{u}$ holds for the index set $\hat{\mathcal{S}}$ with $|\hat{\mathcal{S}}| \le 2s$, the proof of Proposition 3 will be able to be established without $a$ dependence on $s$.

The knowledge of the $s$ value in $\lambda_1$ and $C_0$ can be resolved by replacing $s$ with $\hat{s}\sqrt{\log t}$, where $\hat{s}$ is a guess or estimation on $s$. In our setting, $s$ is defined as an upper bound for the cardinality of the significant index sets for all arms, so our analysis works for the setting with an over-estimation on $s$. If we set $s=\hat{s}\sqrt{\log t}$

in the algorithm, for a large enough $t$, we will enter the over-estimating scenario and be able to recover the desired statistical properties of our algorithm, even when the initial parameter $\hat{s}$ is incorrectly specified to be small. However, the regret during the initial time periods may suffer as a result. We exclude the proof for brevity.

We can also remove the dependence on $s$ by introducing additional assumptions on covariates diversity/balance from the nearly exploration-free bandit literature (e.g., Assumption 3 in Bastani et al. 2021 and Assumption 6 in Oh et al. 2021). With these assumptions, we can directly ensure that Proposition 6 holds even without enough random samples or with a wrong construction for the optimal decision set $\Pi_t$ in the G-MCP-Bandit algorithm. The intuition is that by introducing these assumptions, we can ensure that all dimensions are explored with a nearly equal chance so that the MCP estimator under 2sWL will have a high probability to reach $\mathcal{S}_1 = \mathcal{S}$. Therefore, even if we use wrong $C_0$ and $\lambda_1$ from under-estimating $s$, the desirable regret bounds will continue to hold.

## EC.4. Appendix: Additional Experiments on Tencent dataset

In this section, we extend the Tencent search advertising experiments in §6.2 by considering the impacts of a large number of ads and the robustness of the G-MCP-bandit algorithm under the model misspecification, where the underlying reward function is not within the family of GLMs.

### EC.4.1. The impact of a large number of ads

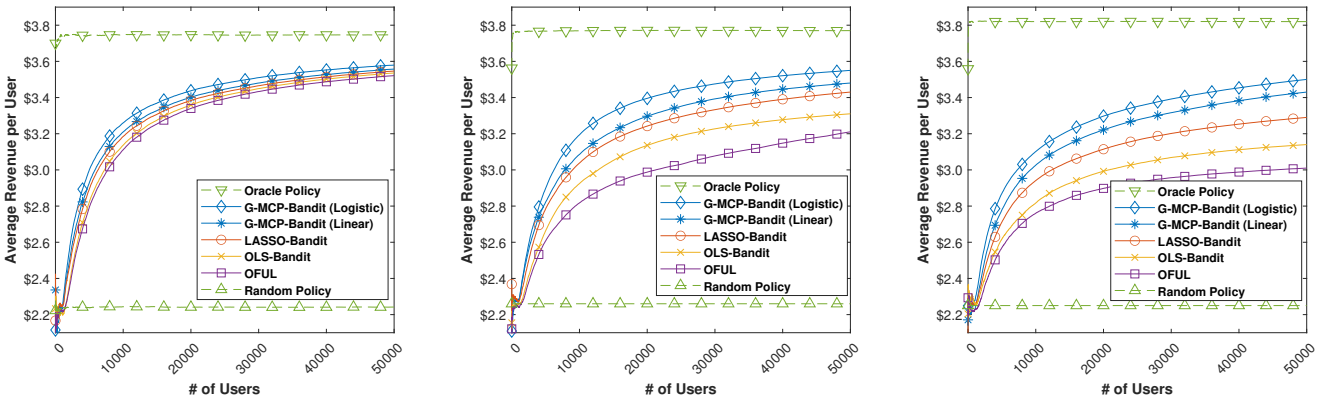
In this subsection, we expand the Tencent search advertising experiment to understand the impacts of a large number of ads. To be able to accurately estimate the true parameter vectors for the oracle policy, it is necessary to include ads with large session entries in all experiments. Hence, we first rank all ads that have CTR higher than 1% by their frequencies and then pick the top 10, 100, and 1000 ads for three experiments. The ads with the lowest frequency in these three experiments (i.e., $K = 10$, $K = 100$, and $K = 1000$) have 188997, 28954, and 2235 session entries, respectively, to provide estimations for parameter vectors under the oracle policy with reasonable accuracy. The reward for each clicked ad is initialized at the beginning of each experiment and randomly assigned to be \$1, \$5, or \$10 with equal probability.

First, as expected, we observe that the computational time increases in the number of ads and the number of users. In particular, with the Intel Xeon Platinum 8163 CPU (2.50GHz, 7 cores), the average computational time (in seconds) for the G-MCP-Bandit algorithm to complete $20,000$ users is 203 for $K = 10$, 226 for $K = 100$, and 349 for $K = 1000$. When the number of users is increased to $40,000$, the average computational time will increase to 583, 661, and 1375 seconds, respectively. Similar to §6.2, we benchmark the G-MCP-Bandit algorithm to OFUL, OLS-Bandit, Lasso-Bandit, the random policy, and the oracle policy. For each experiment, we perform 10 trials for each algorithm and report the average revenue with up to $50,000$ users.

Similar to the three-ad experiment in §6.2, we observe that the G-MCP-Bandit algorithm outperforms other algorithms in terms of the average revenue performance; see Figure EC.3. When the number of ads is comparatively small (e.g., $K = 10$), it does not need many users for all algorithms to identify the significant covariates and/or to estimate parameter vectors to eventually select the optimal ads for incoming users. Hence, the revenue improvement of the G-MCP-Bandit algorithm over other algorithms is most significant when the number of users is not too large. For example, when $T < 10000$, the revenue improvement of the G-MCP-Bandit algorithm over Lasso-Bandit is around $3\% - 4\%$. As the number of users increases, all algorithms eventually learn to accurately estimate parameter vectors to identify the revenue-maximizing ads. Therefore, the average revenue performance of all algorithms begins to converge.

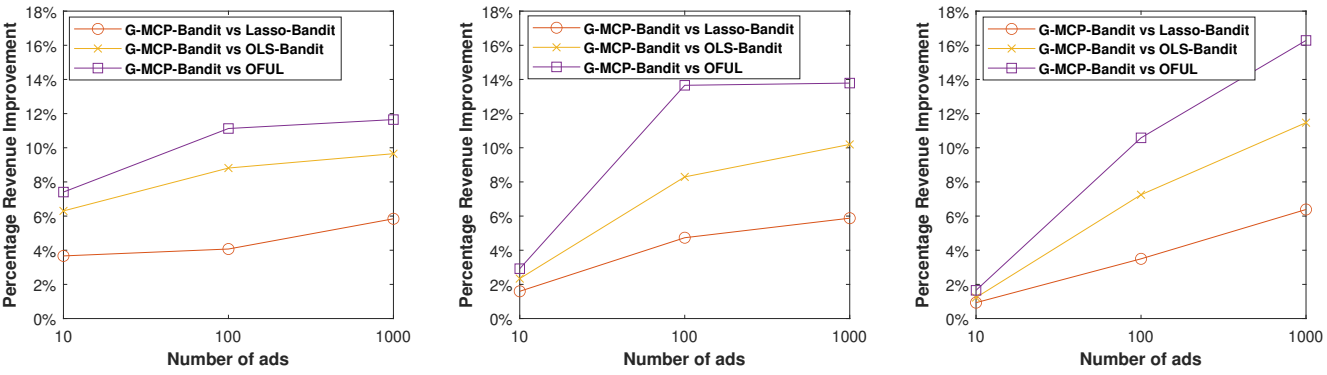
As the number of ads increases (e.g., $K = 100$ and $K = 1000$), accurately learning the parameter vectors and identifying the optimal ad require more users, which is where the G-MCP-Bandit algorithm shines the most. In particular, we observe that the revenue improvement of the G-MCP-Bandit algorithm over other algorithms tends to grow with the number of ads. Figure EC.4 plots the percentage revenue improvement of the G-MCP-Bandit algorithm over other benchmarking algorithms by fixing $T = 5000$, $T = 20000$, and $T = 50000$. In all three scenarios, the benefits of the G-MCP-Bandit algorithm increase with the number of ads. This observation suggests that the G-MCP-Bandit algorithm becomes more favorable in practice, especially when there are large pools of available ads for decision-makers to choose from.

**Figure EC.3 The impact of the number of ads $K$ on average revenue.**




(a) K=10 (b) K=100 (c) K=1000

**Figure EC.4 The percentage revenue improvement of G-MCP-Bandit (Logistic) over other algorithms.**




(a) T=5000 (b) T=20,000 (c) T=50,000

## EC.4.2. Robustness under model misspecification

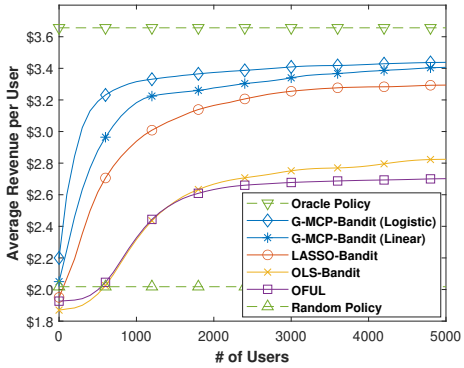
In §6.2 and Appendix EC.4.1, we have examined the robustness of the G-MCP-Bandit algorithm under the model misspecification, where all algorithms assume the linear model, but the true underlying reward function actually follows the logistic model. Under such a misspecified setting, the G-MCP-Bandit algorithm under the linear model outperforms all other algorithms in terms of average revenue performance.

In this subsection, we conduct another experiment to further test the robustness of the G-MCP-Bandit algorithm under the model misspecification, where the true underlying model does not belong to the GLMs family. In particular, we consider the scenario where the true underlying model follows the form of a two-component Gaussian Mixture Model (GMM), which does not belong to the GLMs family. Theoretically, GMM has better representation power than GLMs, and for the Tencent dataset, it actually fits the Tencent data better[8] than both the linear model and the logistic model. Analogous to Figure 3 in the main paper,

[8] We train the GMM model with around three hundred covariates with the highest frequency.

we consider the same three-ad experiment. For each algorithm, we perform 10 trials and report the average revenue with up to 5000 users in Figure EC.5.

**Figure EC.5** **The robustness of the G-MCP-Bandit algorithm under the model misspecification, where the true underlying model follows a two-component Gaussian Mixture Model.**




Consistent with all previous experiments, Figure EC.5 shows that the G-MCP-Bandit algorithm, under both the linear model and the logistic model, continues to outperform other algorithms in terms of average revenue performance. In addition, we observe that all algorithms in Figure EC.5 seem to generate less average revenue that what is shown in Figure 3. This observation may be due to the fact that it is much more difficult to use a GLM model (e.g., a linear or logistic model) to approximate the GMM model than to use the linear model to approximate the logistic model, so the impacts of the model misspecification on the average revenue performance are much more severe in Figure EC.5 than in Figure 3. Despite the negative impacts of the model misspecification, the G-MCP-Bandit algorithm continues to outperform Lasso-Bandit by 7.30% (under the logistic model) and 4.21% (under the linear model), and such an improvement is even larger when compared to OLS-Bandit and OFUL-Bandit.