Sequential Decision Making with Combinatorial Actions and High-Dimensional Contexts

Min-hwan Oh

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

In interactive sequential decision-making systems, the learning agent needs to react to new information both in the short term and in the long term, and learn to generalize through repeated interactions with the environment. Unlike in offline learning environments, the new data that arrives is typically a function of previous actions taken by the agent. One of the key challenges is to efficiently use and generalize from data that may never reappear. Furthermore, in many real-world applications, the agent only receives partial feedback on the decisions it makes. This necessitates a balanced exploration-exploitation approach, where the agent needs to both efficiently collect relevant information in order to prepare for future arrivals of feedback, and produce the desired outcome in the current periods by exploiting the already collected information. In this thesis, we focus on two classes of fundamental sequential learning problems:

Contextual bandits with combinatorial actions and user choice (Chapter 2 and Chapter 3): We investigate the dynamic assortment selection problem by combining statistical estimation of choice models and generalization using contextual information. For this problem, we design and analyze both UCB and Thomson sampling algorithms with rigorous performance guarantees and tractability.

High-dimensional contextual bandits (Chapter 4): We investigate policies that can efficiently exploit the structure in high-dimensional data, e.g., sparsity. We design and analyze an efficient sparse contextual bandit algorithm that does not require to know the sparsity of the underlying parameter – information that essentially all existing sparse bandit algorithms to date require.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First, I would like to express my sincere appreciation to my advisor Garud Iyengar for his guidance and mentorship throughout my PhD. His knowledge in various topics and inquisitive mind have greatly inspired and shaped me as a researcher. From his incredible intuition and deep understanding of problems, I learned how to formulate and approach research problems. His unreserved support and trust have enabled me to explore various ideas. I was very fortunate to have Garud as my thesis advisor and to work with and learn from him for the last five years. Also, I would like to extend my sincere gratitude to my co-advisor Assaf Zeevi, whose work has been a great inspiration to my research even before we started our collaboration. Working with him on high-dimensional contextual bandits, which became the third part of this thesis, was thrilling and rewarding. I have learned immensely from Assaf, and from his insights, constructive feedback, and mentorship. It was truly my honor and privilege to work with both Garud and Assaf.

I am deeply grateful to the committee members, Shipra Agrawal, Omar Besbes, and Adam Elmachtoub, for their willingness to serve on the committee and their valuable feedback. They are exceptional researchers for whom I have great respect.

Some of the most memorable experiences at Columbia were the courses I took and interactions with excellent instructors: Jose Blanchet, David Blei, Krzysztof Choromanski, Donald Goldfarb, Arian Maleki, John Paisley, Liam Paninski, Cliff Stein, and David Yao. Special thanks to Alekh Agarwal, Shipra Agrawal, Daniel Russo, and Alex Slivkins for exposing me to the fundamentals of bandit algorithms and reinforcement learning, from which my thesis work started. It was also a rewarding experience to serve as teaching assistant for Adam Elmachtoub, Ali Hirsa, and Karl Sigman.

I spent two summers during my PhD at IBM T.J. Watson Research Center under the supervision of Naoki Abe, where I had a chance to work with Peder Olsen, Karthikeyan N. Ramamurthy, Karthikeyan Shanmugam, and Toyo Suzumura. Their expertise in research and genuine friendship were the great encouragement during my internships and are the reason why I would always love to visit Yorktown. I am also grateful for the chance to collaborate with Naver. Special thanks to Jaeho Choi and Changbong Kim.

I am very thankful to the IEOR staff for their support throughout my PhD. Special thanks to Lizbeth for making my PhD life and logistics easier. Also, thanks to Carmen, Gerald, Jaya, Jenny, Kristen, Mindi, Shi Yee, and Yosimir.

I would also like to extend my gratitude to fellow PhD students at Columbia, with whom I have interacted for many technical discussions and countless non-technical conversations. Thanks to Apurv, Chaoxu, Francois, Goutam, Hal, Harsh, Jalaj, Mike, Omar, Raghav, Randy, Ryan, Sid, Suraj, Wenbo, Xiao, Yunhao, and all other PhD students who made my time in the department much more enjoyable. Also, thanks to Cheng, Praveen, Rajan, and Sheng for their friendship.

I am forever indebted to my family for their incredible and boundless support. I am blessed to have my parents who gave me more than I could ever hope for. Finally, my wife Hakyung deserves my deepest gratitude. None of this work would have been possible without her unconditional support and belief in me. To her, I dedicate this thesis.

To Hakyung and my daughters, Serim and Yerim

Chapter 1

Introduction

In interactive sequential decision-making systems, the learning agent needs to react to new information both in the short term and in the long term, and learn to generalize through repeated interactions with the environment. In each interaction, the agent adaptively takes an action based on the information given from the environment and then receives feedback. Typically, the feedback arrives as a form of reward (or loss) that the agent aims to maximize (or minimize) and as *partial* feedback only for the action chosen by the agent, rather than full feedback for all actions. The *multi-armed bandit* (Thompson, 1933; Lai and Robbins, 1985; Lattimore and Szepesvári, 2019) is a classic model for sequential decision making with partial feedback. The *contextual bandit* is a general extension of the multi-armed bandit that incorporates the contextual information given by the environment (Langford and Zhang, 2008). The contextual bandit is a fundamental reinforcement learning problem that possesses the full complexity of statistical learning. There have been successful applications of contextual bandits in various domains, such as recommendation systems and healthcare (Li et al., 2010; Tewari and Murphy, 2017).

In the contextual bandit problem, unlike in *offline* learning environments, the new information that arrives is often a function of the previous observations including the ac-

tions previously taken by the agent. This necessitates a balanced exploration-exploitation approach, where the agent needs to both efficiently collect relevant information in order to prepare for future reward (exploration) and maximize the reward in the current periods based on the already collected information (exploitation). Another key challenge is to efficiently utilize and generalize various contextual information that may never reappear.

Therefore, it is crucial to design tractable algorithms for this fundamental sequential decision-making problem, with provable guarantees on their statistical and computational performances. In this thesis, we study the following two classes of sequential learning problems that arise in various real-world applications: (i) contextual bandits with combinatorial actions and user choice consideration and (ii) high-dimensional contextual bandits. We describe these problems in detail in the following sections.

1.1 Background

1.1.1 Multi-Armed Bandits

The multi-armed bandit framework addresses the fundamental problem of sequential decision making under uncertainty with partial feedback. The agent has a set of arms to choose from. Each arm represents an action or a decision with a random reward (in stochastic settings), and the agent receives a sample from the random reward when an arm is pulled; however, the agent receives no feedback on the arms that were not pulled. The goal of the agent is to maximize the cumulative reward across the time horizon. Often, the performance of the agent is measured in comparison with the best competitor in hindsight (or the oracle that knows the true expected rewards of the arms). Hence, an equivalent goal is to minimize the cumulative *regret*, which is defined as the difference between the cumulative reward of the best competitor and that of the agent.

This model exemplifies the exploration-exploitation dilemma: if one pulls a myopically optimal arm, i.e. the optimal arm based on previous observations, one will receive a good reward but will forego the opportunity of discovering potentially a better arm. Typically, the (non-contextual) multi-armed bandit models the reward of each arm independent of each other and random i.i.d. noise around an arm's mean reward. Therefore, pulling an arm does not provide information for the other arms. While there is a rich literature on the classical multi-armed bandit problem (Auer, Cesa-Bianchi, and Fischer, 2002; Bubeck, Cesa-Bianchi, et al., 2012), many applications in the real world have a large number of actions needed to be considered and require much richer classes of decisions. Therefore, the basic multi-armed bandit may not be suitable in these real-world scenarios.

1.1.2 Contextual Bandits

Often, the real-world problems have a very large number of actions but also come with additional information about the actions. These large action sets are usually dealt with by introducing a structure that allows the learning agent to generalize from one action to another (Lattimore and Szepesvári, 2019). For example, a recommender system, where each action represents an item to be recommended, often has feature information about the items and may also have access to contextual information about users. Hence, it would be efficient to utilize the feedback for a recommended item to infer the user's preference on a similar item that was not recommended. The similarity between actions can be captured by their proximity in the feature space.

The linear contextual bandit has been widely studied (Abe and Long, 1999; Auer, 2002; Dani, Hayes, and Kakade, 2008; Li et al., 2010; Rusmevichientong, Shen, and Shmoys, 2010; Abbasi-Yadkori, Pál, and Szepesvári, 2011; Agrawal and Goyal, 2013) and provides the key characteristics of the contextual bandit problem in the most succinct way. In each round of interactions, the environment reveals feature vectors for each

action that the agent can choose from. The agent selects an action and observes a reward of the arm, whose expected reward is given by the inner product of the action's feature vector and the underlying parameter vector. The underlying parameter is unknown to the agent. Therefore, the agent is initially unsure of which action is best. However, as observations accumulate through repeated interactions, the agent is able to learn over time the underlying parameter as well as effective actions to take.

Beyond the linear reward model, there are other approaches to incorporate a nonlinear relationship between the feature vector and the reward, using generalized linear models (Filippi et al., 2010; Li, Lu, and Zhou, 2017), decision trees (Elmachtoub et al., 2017) and neural networks (Riquelme, Tucker, and Snoek, 2018). It is also worth mentioning model-agnostic approaches (Langford and Zhang, 2008; Agarwal et al., 2014) which do not specify any parametric form of the reward model. Despite the differences in the model assumption, all of the contextual bandit frameworks still possess the exploration-exploitation dilemma due to partial bandit feedback. Additionally, since the same context or features may not appear again, the agent needs to efficiently utilize the contextual information for learning the underlying reward model (or the policy itself). Therefore, efficient *generalization* across different interactions with the environment is desired.

1.2 Problems

1.2.1 Contextual Bandits with Combinatorial Actions

In many of today's human-AI interactions, a learning *agent* (AI) makes sequential decisions and receives *user* (human) feedback *only* about the specific decisions it takes. Therefore, one can model such problem instances as a multi-armed bandit or contextual bandit where the agent selects a sequence of actions while interacting with users. How-

ever, in most of these interactive systems, such as search engines, e-commerce, streaming video services, news websites, etc., the agent does not select just a single item, as in the basic multi-armed bandit setting, but rather selects a *set* of items – e.g., a list of search results, an assortment of products, a slate of recommended movies, or relevant news articles. Then the agent offers this set of items to the user, and the user may choose one item from the offered set (or may choose none). The agent receives a reward associated with the chosen item, if any. The specific choice of an item by a user is often a function of the *contextual* information about both the user and the items in the offer set. However, a naive implementation of contextual bandit algorithms in this setting, treating each feasible subset of items as an independent action, would be prohibitive due to the combinatorial nature of the action selection.

In Chapter 2 and Chapter 3, we address this combinatorial contextual bandit problem. We assume that the user choice is described by the multinomial logit (MNL) choice model (McFadden, 1978), where the expected utility of an item is given by an inner product of contextual information of the item and the unknown underlying parameter. This problem inherits the challenges of exploration-exploitation tradeoff from the contextual bandit problem, where we not only need to learn the users' choice behavior but also maximize reward by exploiting the information we already have, as well as generalization of contextual information that may not reappear. Furthermore, the substitution effect of items within an offer set makes the problem much more difficult. Thus, despite the fact that this problem setting is prevalent in practice, existing decision-making algorithms for this problem either lacked a theoretical guarantee or lacked tractability and practicality.

1.2.2 High-Dimensional Contextual Bandits

In many application domains, such as recommender systems and healthcare analytics, a large amount of contextual information is often available for both personalization as well

as generalization. This results in high-dimensional feature space; however, typically only a very small subset of the features influences the expected reward. That is, the unknown parameter vector is *sparse* with only the elements corresponding to the relevant features being non-zero. There is an emerging body of work on multi-armed bandit problems with sparse linear reward functions which propose methods to exploit the sparse structure under various regularity assumptions (Abbasi-Yadkori, Pal, and Szepesvari, 2012; Gilton and Willett, 2017; Bastani and Bayati, 2020; Wang, Wei, and Yao, 2018; Kim and Paik, 2019). All of these existing approaches suffer from a crucial drawback: these algorithms require the prior knowledge of sparsity index s_0 (i.e., the number of the non-zero elements in the unknown parameter). This information is almost never available in practice. In the absence of such knowledge, the existing algorithms fail to fully leverage the sparse structure, and their performance does not guarantee the improvements in dimensionality-dependence which can be realized in the sparse problem setting. Furthermore, the misspecification of this sparsity parameter can lead to severe deterioration in the performances of algorithms. Hence, designing an algorithm that operates in a sparsity-agnostic manner, and providing its performance guarantees has been an important open problem.

1.3 Summary of Contributions

Chapter 2: UCB Algorithms for MNL Contextual Bandits

In Chapter 2, we propose upper confidence bound (UCB) algorithms that combine exploration of the combinatorially large action space and exploitation of context information to maximize reward. The algorithms maintain a confidence set for the unknown parameter of the MNL model and select an optimal set of items under an optimistic reward function using the principle of optimism in the face of uncertainty. We propose the first

polynomial-time algorithm for this problem. We establish the *near-optimal* regret bounds that do not depend on the number of available items, where the regret is defined as the discrepancy between the reward of the optimal set to offer if the true parameter was known, and that of the agent's offer-set selection. Furthermore, to overcome computational challenges, we exploit the structure of the MNL model and show that an online Newton step is sufficient to maintain a tight confidence region. Hence, we achieve both statistical and computational efficiency. Then we study *provably optimal* algorithms with a finite total number of items, in particular, establishing regret bounds sublinear in the feature dimension. We show that a practical algorithm can still achieve sublinear dependence on the feature dimension. The effectiveness of our proposed algorithms is further supported by numerical experiments.

Chapter 3: Thompson Sampling for MNL Contextual Bandits

In Chapter 3, we investigate Thompson sampling (TS) algorithms for the MNL contextual bandit problem. TS methods are known to be empirically superior to UCB based methods for many multi-armed bandit variants (Chapelle and Li, 2011). However, TS algorithms are generally difficult to analyze, and this challenge is further exacerbated by the combinatorially large item space in the MNL contextual bandit problem, and the fact that there is no conjugate prior for the MNL model. We provide both a Bayesian regret bound and a worst-case regret bound for TS-based algorithms for the MNL contextual bandit problem. A key element in our approach is an optimistic sampling scheme to address the challenges that arise in the worst-case regret analysis. We also show that our proposed TS algorithm has superior numerical performances. To the best of our knowledge, this is the *first worst-case* theoretical guarantee for a TS algorithm in contextual bandits with combinatorial actions in general. The techniques developed here can be applied to analyzing other combinatorial contextual bandits.

Chapter 4: Sparsity-Agnostic High-dimensional Bandit

We demonstrate that a relatively simple contextual bandit algorithm, which exploits Lasso (ℓ_1-regularized regression) estimation in a sparsity-agnostic manner, has provably near-optimal regret (under suitable regularity). We also show that, in empirical tests, our proposed algorithm significantly outperforms all state-of-the-art alternative methods that rely on a priori knowledge of sparsity. To the best of our knowledge, this is the *first general sparse bandit method* for a general set of arms that *does not* require prior knowledge of the sparsity index, overcoming the critical drawback of the existing methods. The scalability in both the ambient feature dimension and the sparse support dimension matches the equivalent terms in the *offline* Lasso convergence results, which implies that our established result is best possible in *online* learning settings.

Our new results provide insights on how the previously proposed approaches for this fundamental problem result in inevitable inefficiency and show that a surprisingly simple solution can provide provably efficiency as well as significantly superior empirical performances. In high-stake decision-making domains such as healthcare, where a good performance of a policy may represent an increased number of saved lives and vice versa, we strongly believe our result can make a significant impact.

Chapter 2

Upper Confidence Bound Algorithms for MNL Contextual Bandits

In this chapter and Chapter 3, we study a sequential assortment selection problem which is a combinatorial version of the contextual bandit problem. The goal of the decision-making agent is to offer a sequence of assortments of at most K items from a set of N possible items. The sequence is chosen as a function of the contextual information of items, and possibly users, in order to minimize the expected regret, which is defined as the gap between the expected revenue generated by the algorithm and the optimal expected revenue when the true parameter is known. The contextual information in the form of d-dimensional feature vectors for each of the N items is revealed in each round t, i.e., the feature information about the items are allowed to change over time. The feedback here is the particular item chosen by the user from the offered assortment. We assume that the item choice follows a multinomial logistic (MNL) distribution (McFadden, 1978). This is one of the most widely used model in dynamic assortment optimization literature (Caro and Gallien, 2007; Rusmevichientong, Shen, and Shmoys, 2010; Sauré and Zeevi, 2013; Agrawal et al., 2019; Agrawal et al., 2017; Aouad, Levi, and Segev, 2018).

Chapter 2: UCB Algorithms for MNL Contextual Bandits

For sequential decision-making with contextual information, (generalized) linear bandits (Abe and Long, 1999; Auer, 2002; Filippi et al., 2010; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori, Pál, and Szepesvári, 2011; Chu et al., 2011; Li, Lu, and Zhou, 2017) and their variants have been widely studied. However, these methods are only limited to a single item selection which is increasingly rarer in practice as compared to multiple item offering that we consider in this work. There is a line of work on combinatorial variants of contextual bandit problems (Qin, Chen, and Zhu, 2014; Wen, Kveton, and Ashkan, 2015; Kveton et al., 2015; Zong et al., 2016) mostly with semi-bandit feedback or cascading feedback. However, these methods do not take the user choice into account. Hence, substitution effect is not modeled. In contrast to these contextual bandit problems and their combinatorial variants, in the multinomial logit (MNL) contextual bandit, the item choice (feedback) is a function of all items in the offered assortment. The key challenges here are to design an algorithm that offers assortments to simultaneously learn the unknown parameter and maximize the total expected revenue on a sequence interactions with users; and to establish a bound on the performance of the algorithm. There has been an emerging body of literature on the MNL bandits in both non-contextual and contextual settings (Agrawal et al., 2017; Agrawal et al., 2019; Cheung and Simchi-Levi, 2017b; Ou et al., 2018; Chen, Wang, and Zhou, 2018). However, designing a practical algorithm that achieves provable guarantees poses a greater challenge. In this chapter, we study upper confidence bound (UCB) algorithms for the MNL contextual bandit problem. An overview of this chapter is as follows:

(a) In Section 2.2, we formulate the MNL contextual bandit problem, and briefly discuss the maximum likelihood estimation for the MNL model.

(b) In Section 2.3, we propose a UCB-based algorithm, `UCB-MNL` (Algorithm 1), for the MNL contextual bandits. To our knowledge, is the first polynomial-time algorithm

that achieves an N independent $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret.[1] This result matches the previous best upper bound (up to logarithmic factors). We also propose a variant of `UCB-MNL` (Algorithm 2) that updates the MNL parameter in an online fashion and show that this modified `UCB-MNL` algorithm has the same regret performance but is substantially more computationally efficient.

(c) In Section 2.4, we prove a non-asymptotic confidence bound (Theorem 2.3) for the maximum likelihood estimator of the MNL model, which may be of independent interest. This sharp confidence bound is used in Section 2.5 and Section 2.6 to analyze algorithms that have improved regret bound in terms of the dependence on the feature dimension.

(d) In Section 2.5, we propose `supCB-MNL` (Algorithm 4) and utilize the sharper convergence result in Theorem 2.3 to establish $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret. This improves on the best previous result by $\sqrt{d}$ factor, and matches the lower bound for the MNL bandit problem within logarithmic factors. However, as with the existing provably optimal bandit algorithms that rely on a framework proposed in Auer (2002), `supCB-MNL` is not a practical algorithm.

(e) In Section 2.6, we propose a practical algorithm, `DBL-MNL` (Algorithms 5), which achieves $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret when item revenue is uniform, i.e., the goal is to maximize the click-through rate for offered assortments. `DBL-MNL` does *not* rely on the framework of Auer (2002), and has state-of-the-art computational efficiency. Thus, this work is the first one to provide a practical algorithm with provable $\sqrt{d}$ dependence on the feature dimension.

[1] $\widetilde{\mathcal{O}}$ suppresses logarithmic dependence on problem parameters.

2.1 Related Work

The MNL model (Plackett, 1975; McFadden, 1978; Luce, 2012) is one of the most widely used choice models for assortment selection problems. The problem of computing the optimal assortment (*static* assortment optimization problem), when the MNL parameters, i.e., user preferences, are known a priori, is well-studied (Talluri and Van Ryzin, 2004; Davis, Gallego, and Topaloglu, 2014; Désir, Goyal, and Zhang, 2014). Our work belongs to the literature on *dynamic* assortment optimization. Caro and Gallien (2007) consider the setting where the demand for each of the items in an assortment is independent. Rusmevichientong, Shen, and Shmoys (2010) and Sauré and Zeevi (2013) consider the problem of minimizing regret under the MNL choice model and present an "explore first then exploit later" approach. Rusmevichientong, Shen, and Shmoys (2010) showed $\mathcal{O}(N^2 \log^2 T)$ regret bound, where N is the number of total candidate items. Sauré and Zeevi (2013) later improved the bound to $\mathcal{O}(N \log T)$. However, these methods require a priori knowledge of "separability" between the true optimal assortment and the other sub-optimal alternatives.

More recent work by Agrawal et al. (2019), Agrawal et al. (2017), and Cheung and Simchi-Levi (2017a) and Chen and Wang (2017) also incorporated the MNL models into dynamic assortment optimization and formulated the problem into an online regret minimization problem without requiring a priori knowledge on separability. Agrawal et al. (2019) proposed a UCB-type algorithm which shows $\widetilde{\mathcal{O}}(\sqrt{NT})$ regret bound. Agrawal et al. (2017) achieve the same order of $\widetilde{\mathcal{O}}(\sqrt{NT})$ regret bound using a Thompson sampling (Thompson, 1933) approach with improved empirical performances. Chen and Wang (2017) show a matching lower bound of $\Omega(\sqrt{NT})$. All of these previous works mentioned so far assume each item is associated with a unique parameter, i.e., one cannot learn across items nor can incorporate multi-dimensional feature information which may be available

to the decision-making agent. In our work, we consider the setting where there is information about items with d features and these features can be time-varying. When the total number of items N is much larger than the feature dimension $d \ll \sqrt{N}$, utilizing the feature information and learning across items allows one to reduce the regret bound from $\widetilde{\mathcal{O}}(\sqrt{NT})$ to $\widetilde{\mathcal{O}}(d\sqrt{T})$. However, one cannot directly incorporate (time-varying) feature information into the previous work (Agrawal et al., 2019; Agrawal et al., 2017)) since these methods require that the same assortment be offered repeatedly for a random number of rounds until an outside choice (no purchase) is observed. Chen, Wang, and Zhou (2018) proposed a UCB method which establishes $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret bound for the contextual version of the MNL bandit problem. There is a fundamental difference between the algorithm proposed in Chen, Wang, and Zhou (2018) and our proposed algorithms. Chen, Wang, and Zhou (2018) enumerates the exponentially many $\binom{N}{K}$ assortments and builds confidence bounds for each of them. Hence, the resulting algorithm is an exponential-time algorithm in the number of total items. We circumvent this computational bottleneck by constructing confidence bounds for each item rather than each assortment (see Section 2.3). It is also worth mentioning the previous work in the personalized MNL-bandit problem (Cheung and Simchi-Levi, 2017b; Bernstein, Modaresi, and Sauré, 2018; Kallus and Udell, 2020). These works consider each item utility separately and learn N different parameters; hence there is no generalization across different items, which is different from our problem setting.

Linear bandits have been widely studied (Abe and Long, 1999; Auer, 2002; Dani, Hayes, and Kakade, 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori, Pál, and Szepesvári, 2011; Chu et al., 2011; Agrawal and Goyal, 2013). Filippi et al. (2010), Li, Lu, and Zhou (2017), and Kveton et al. (2020) extend linear bandits to scalar, monotone, generalized linear bandits. Filippi et al. (2010) established $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret bound. Li, Lu, and Zhou (2017) improved the regret bound to $\widetilde{\mathcal{O}}(\sqrt{dT})$ by establishing a new finite-

sample confidence bound for MLE in generalized linear models. However, these results in (generalized) linear bandits do not apply directly to our problem, since the choice probability of an item in an assortment is non-linear and non-monotone in the parameter of the MNL model. It is also worthwhile to mention a line of work in other combinatorial bandit problems (Qin, Chen, and Zhu, 2014; Wen, Kveton, and Ashkan, 2015; Kveton et al., 2015; Zong et al., 2016; Li et al., 2016) mostly with semi-bandit feedback or cascading feedback. Our work is distinct from these combinatorial bandit problems since in cascading or semi-bandit settings, the mapping from the item feature to the user feedback is still independent of other items in an offered set, ignoring the substitution effect of items. On the other hand, MNL choice feedback that we consider in this work is a function of an entire assortment which makes our analysis much more challenging.

2.2 Problem Formulation

2.2.1 Notations

For a vector $x \in \mathbb{R}^d$, we use $\|x\|$ to denote its ℓ_2-norm. The weighted ℓ_2-norm associated with a positive-definite matrix V is defined by $\|x\|_V := \sqrt{x^\top V x}$. The minimum and maximum eigenvalues of a symmetric matrix V are written as $\lambda_{\min}(V)$ and $\lambda_{\max}(V)$, respectively. The trace of a matrix V is trace(V). For two symmetric matrices V and W of the same dimensions, $V \succeq W$ means that $V - W$ is positive semi-definite. For a positive integer n, we define $[n] = \{1, ..., n\}$.

2.2.2 MNL Contextual Bandits

The MNL contextual bandits problem is defined as follows. The decision-making agent has a set of N distinct items. We define $\mathcal{S}$ to be the set of candidate assortments with size

constraint at most K, i.e. $\mathcal{S} = \{S \subset [N] : |S| \leq K\}$. Although we treat $\mathcal{S}$ as stationary for ease of exposition, we can allow $\mathcal{S}$ (as well as the item set $[N]$) to change over time.

In each round t, the agent observes feature vectors $x_{ti} \in \mathbb{R}^d$ for every item $i \in [N]$. Each feature vector x_{ti} combines the information of the user in round t and the corresponding item i. For example, suppose the user in round t is characterized by a feature vector v_t and the item i has a feature vector w_{ti} (note that we allow feature vectors for an item and a user to change over time), then we can use $x_{ti} = \text{vec}(v_t w_{ti}^\top)$, the vectorized outer-product of v_t and w_{ti}, as the combined feature vector of item i in round t. If v_t is not available, we can use item-dependent features only $x_{ti} = w_{ti}$. Given this contextual information, in every round t, the agent offers an assortment $S_t = \{i_1, \ldots, i_\ell\} \in \mathcal{S}$, $\ell \leq K$, and observes the user purchase decision $c_t \in S_t \cup \{0\}$, where $\{0\}$ denotes "outside option" which means the user did not choose any item offered in S_t. This selection is given by a multinomial logit (MNL) choice model (McFadden, 1978) under which the choice probability for item $i_k \in S_t \cup \{0\}$ is defined as

$$p_t(i_k|S_t, \theta^*) := \begin{cases} \dfrac{\exp\{x_{ti_k}^\top \theta^*\}}{1 + \sum_{i_j \in S_t} \exp\{x_{ti_j}^\top \theta^*\}}, & \text{if } i_k \in S_t \\ \dfrac{1}{1 + \sum_{i_j \in S_t} \exp\{x_{ti_j}^\top \theta^*\}}, & \text{if } i_k = 0 \end{cases} \tag{2.1}$$

where $\theta^* \in \mathbb{R}^d$ is a time-invariant parameter unknown to the agent. The choice response for each item $i_k \in S_t$ is defined as $y_{ti_k} := \mathbb{1}(c_t = i_k) \in \{0, 1\}$ and $y_{t0} := \mathbb{1}(c_t = 0)$ for the outside option. Hence the choice response variable $y_t = (y_{t0}, y_{ti_1}, ..., y_{ti_\ell})$ is a sample from this multinomial distribution:

$$y_t \sim \text{multinomial}\Big\{1, (p_t(0|S_t, \theta^*), ..., p_t(i_\ell|S_t, \theta^*))\Big\}$$

where the parameter 1 indicates that y_t is a single-trial sample, i.e., $y_{t0} + \sum_{k=1}^{\ell} y_{ti_k} = 1$. For

$i \in S_t \cup \{0\}$, we define the noise $\epsilon_{ti} := y_{ti} - p_t(i|S_t, \theta^*)$. Since each ϵ_{ti} is a bounded random variable in $[0, 1]$, ϵ_{ti} is σ^2-sub-Gaussian with $\sigma^2 = 1/4$; however, ϵ_{ti} is *not* independent across $i \in S_t$ due to the substitution effect in the MNL model. The revenue parameter r_{ti} for each item is also given at round t. r_{ti} is the revenue if item i is chosen in round t. Without loss of generality, assume $|r_{ti}| \leq 1$ for all i and t. Then, the expected revenue of the assortment S_t is given by

$$R_t(S_t, \theta^*) := \sum_{i \in S_t} r_{ti} p_t(i|S_t, \theta^*) \tag{2.2}$$

Note that for a very broad class of applications for the MNL bandit problem, including search engines and media recommendations, the goal of the agent is to maximize the click-through rate. Therefore, in this case, the item revenue is uniform, i.e., $r_{ti} = r$ for all i and t. Note that the substitution effect of items in an assortment still exists even with the uniform revenue parameter. Hence, while the optimization procedure becomes easier under the uniform revenue setting, the challenges in terms of statistical learning still remains the same; still more difficult than other combinatorial contextual bandit problems.

We define S_t^* to be the optimal assortment in round t when θ^* is known a priori, i.e. the true MNL probabilities $p_t(i|S, \theta^*)$ are known a priori:

$$S_t^* := \operatorname*{argmax}_{S \subset \mathcal{S}} R_t(S, \theta^*). \tag{2.3}$$

Note that S_t^* is also potentially time-varying since feature vectors $\{x_{ti}\}$ can change over time. Consider a time horizon of T rounds, during which the agent sequentially chooses an assortment to offer. The agent does not know the value of θ^*, and therefore, can only choose the assortment S_t in period t based on the choices S_τ for periods $\tau < t$, and the observed responses. We measure the performance of the agent by the regret $\mathcal{R}(T)$ for the

time horizon T, which is the cumulative gap between the expected revenue generated by the assortment chosen by the agent and that of the optimal assortment, i.e.,

$$\mathcal{R}(T) := \mathbb{E}\left[\sum_{t=1}^{T}\Big(R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*)\Big)\right] \tag{2.4}$$

where $R_t(S_t^*, \theta^*)$ is the expected revenue corresponding to the optimal assortment in round t, i.e., the highest revenue which can be obtained with the knowledge of θ^*. Hence, maximizing the cumulative expected revenue is equivalent to minimizing the cumulative expected regret. Note that the expectation in (2.4) is taken over two sources of stochasticity in our problem: the feature vector x_{ti} and the noise ϵ_{ti} with corresponding probability measures $\mathbb{P}_x$ and $\mathbb{P}_\epsilon$. Throughout this chapter, all the expectations and probabilities are with respect to the product measure $\mathbb{P}_x \times \mathbb{P}_\epsilon$. For notational brevity, we denote $\mathbb{E}$ and $\mathbb{P}$ as "expectation" and "probability" with respect to this product measure.

2.2.3 MLE for Multinomial Logistic Regression

We briefly discuss the maximum likelihood estimate (MLE) of the MNL model which we utilize for parameter estimation throughout this chapter. The MLE for the unknown parameter θ^* of the MNL model defined in (2.1) can be computed as follows. First, recall that $y_t \in \{0,1\}^{|S_t|+1}$ is the user choice where y_{ti} is the i-th component of y_t. Then, the likelihood function under parameter θ is then given by

$$\mathcal{L}(\mathcal{D}_n|\theta) = \prod_{t=1}^{n}\prod_{i\in S_t\cup\{0\}}(p_t(i|S_t,\theta))^{y_{ti}}$$

where $\mathcal{D}_n = \{X_t, S_t, y_t\}_{t=1}^n$ and $X_t = \{x_{ti}\}_{i\in[N]}$. Taking the negative logarithm gives

$$\ell_n(\theta) = -\log\mathcal{L}(\mathcal{D}_n|\theta) = -\sum_{t=1}^{n}\sum_{i\in S_t\cup\{0\}} y_{ti}\log p_t(i|S_t,\theta)$$

which is known as the cross-entropy error function for the multi-class classification problem. Taking the gradient of this negative log-likelihood with respect to θ, we obtain

$$\nabla_\theta \ell(\theta) = \sum_{t=1}^{n} \sum_{i \in S_t} (p_t(i|S_t, \theta) - y_{ti}) x_{ti}$$

As the sample size n goes to infinity, it is known from the classical likelihood theory (Lehmann and Casella, 2006) that the MLE $\hat{\theta}_n$ is asymptotically normal. In particular, $(\hat{\theta}_n - \theta^*) \to \mathcal{N}(0, \mathcal{I}_{\theta^*}^{-1})$ where $\mathcal{I}_{\theta^*}$ is the Fisher information matrix. We show in the proof of Theorem 2.3 that $\mathcal{I}_{\theta^*}$ is lower bounded by $\sum_t \sum_{i \in S_t} p_t(i|\theta^*) p_t(0|\theta^*) x_{ti} x_{ti}^\top$. Hence, if $p_t(i|\theta^*) p_t(0|\theta^*)$ is bounded below away from zero, then we can ensure that $\mathcal{I}_{\theta^*}$ is invertible and prevent asymptotic variance of $x^\top \hat{\theta}$ from going to infinity for any x.

2.3 UCB Algorithms for MNL Contextual Bandits

The basic idea of UCB algorithms is to maintain a confidence set for the parameter θ^*. The techniques of upper confidence bounds (UCB) have been widely known to be effective in balancing the exploration and exploitation trade-off in many bandit problems, including K-arm bandits (Auer, Cesa-Bianchi, and Fischer, 2002; Lattimore and Szepesvári, 2019), linear bandits (Auer, 2002; Dani, Hayes, and Kakade, 2008; Abbasi-Yadkori, Pál, and Szepesvári, 2011; Chu et al., 2011) and generalized linear bandits (Filippi et al., 2010; Li, Lu, and Zhou, 2017).

For each round t, the confidence set $\mathcal{C}_t$ for θ^* is constructed from the feature vectors $\{x_{t'i} : i \in S_{t'}, t' < t\}$, and the observed feedback of selected items $\{y_{t'}, t' < t\}$ from all previous rounds. Let $\hat{\theta}_t$ denote the estimate of the unknown parameter θ^* after t periods, and suppose we are guaranteed that θ^* lies within the confidence set $\mathcal{C}_t$ centered at MLE $\hat{\theta}_t$ with radius $\alpha_t > 0$ with high probability (see Lemma 2.2). The confidence radius α_t has to be chosen carefully: larger α_t induces more exploration; however, a too large

value of α_t can cause regret to increase. In the MNL bandit setting, exploitation is to offer $\text{argmax}_{S\in\mathcal{S}} R_t(S, \hat{\theta}_t)$ which is a greedy action with respect to the current estimate, whereas exploration is to choose a set S that has the potential for a high expected revenue $R_t(S, \theta)$ as θ varies over $\mathcal{C}_t$. Thus, a direct way to introduce optimism, and induce exploration, is to define an optimistic revenue for each $\binom{N}{K}$ assortments. This is the approach taken in Chen, Wang, and Zhou (2018); however, this enumeration has exponential complexity when N is large and K is relatively small. We show that one can induce sufficient exploration by defining an optimistic expected utility z_{ti} for each item, and defining the optimistic revenue for any assortment S using the optimistic utility. We define

$$z_{ti} := x_{ti}^\top \hat{\theta}_{t-1} + \alpha_{t-1}\|x_{ti}\|_{V_{t-1}^{-1}} \tag{2.5}$$

where $V_t = \sum_{t'=1}^{t}\sum_{i\in S_t} x_{t'i}x_{t'i}^\top \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix. The optimistic utility z_{ti} consists of two components: mean utility estimate $x_{ti}^\top\hat{\theta}_{t-1}$ and confidence interval $\alpha_{t-1}\|x_{ti}\|_{V_{t-1}^{-1}}$ with suitable confidence radius α_{t-1}. In the proof of the regret bound of our proposed algorithm, we show that z_{ti} is, indeed, an upper bound of $x_{ti}^\top\theta^*$ if θ^* lies within in the confidence ellipsoid centered at $\hat{\theta}_{t-1}$ (see Lemma 2.3).

2.3.1 Algorithm: `UCB-MNL`

We now have all the ingredients for our first UCB algorithm for the MNL contextual bandit problem, `UCB-MNL` (Algorithm 1). During the initialization phase of Algorithm 1, the agent first randomly chooses an assortment S_t with exactly K items (note that after the initialization, the size of S_t can be smaller than K) to ensure a unique MLE estimate. For this, the initialization duration T_0, specified in Theorem 2.1, is chosen to guarantee that $\lambda_{\min}(V_{T_0})$ is large enough.[2] After the initialization phase, the algorithm chooses an

[2]This also implies $\lambda_{\min}(V_{T_0}) > 0$ after the initialization. Therefore, V_t is invertible for $t \geq T_0$.

assortment based on upper confidence bounds of expected utility $\{z_{ti}\}$. That is, based on z_{ti}, we construct the following optimistic estimate of the expected revenue

$$\widetilde{R}_t(S) := \frac{\sum_{i \in S} r_{ti} \exp(z_{ti})}{1 + \sum_{j \in S} \exp(z_{tj})}. \tag{2.6}$$

The algorithm offers assortment S_t that maximizes this optimistic expected revenue, and the receives a user choice feedback y_t. We assume an access to an optimization method which returns the assortment choice in round t, $S_t = \arg\max_{S \subset \mathcal{S}} \widetilde{R}_t(S)$. There are efficient polynomial-time algorithms available to solve this optimization problem that we can use (Rusmevichientong, Shen, and Shmoys, 2010; Davis, Gallego, and Topaloglu, 2013).

Algorithm 1 UCB-MNL

Input: initialization T_0, confidence radius α_t

Initialization: **for** $t \in [T_0]$

  Randomly choose S_t with $|S_t| = K$

  $V_t \leftarrow V_{t-1} + \sum_{i \in S_t} x_{ti} x_{ti}^\top$

for all $t = T_0 + 1$ to T **do**

  Compute $z_{ti} = x_{ti}^\top \hat{\theta}_{t-1} + \alpha_{t-1} \|x_{ti}\|_{V_{t-1}^{-1}}$ for all i

  Compute $S_t = \text{argmax}_{S \subset \mathcal{S}} \widetilde{R}_t(S)$

  Offer S_t and observe y_t (user choice in round t)

  Update $V_t \leftarrow V_{t-1} + \sum_{i \in S_t} x_{ti} x_{ti}^\top$

  Compute MLE $\hat{\theta}_t$ by solving

$$\sum_{t'=1}^{t} \sum_{i \in S_{t'}} \left(p_{t'}(i | S_{t'}, \hat{\theta}_t) - y_{t'i} \right) x_{t'i} = \mathbf{0} \tag{2.7}$$

end for

2.3.2 Regret Bound for `UCB-MNL` Algorithm

We present the an upper bound on the regret of `UCB-MNL` under the following assumptions on the context process and the MNL model, both standard in the literature.

Assumption 2.1. *Each feature vector set* $\{x_{ti} \in \mathbb{R}^d, i \in [N]\}$ *is drawn i.i.d. from an unknown distribution* p_X *with* $\|x_{ti}\| \leq 1$ *all* t, i *and there exists a constant* $\sigma_0 > 0$ *such that* $\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{N}\sum_{i\in[N]} x_{ti}x_{ti}^\top\right]\right) \geq \sigma_0$.

The boundedness is used to make the regret bounds scale-free. The i.i.d. assumption is also used in generalized linear bandit (Li, Lu, and Zhou, 2017) and MNL contextual bandit (Cheung and Simchi-Levi, 2017b; Chen, Wang, and Zhou, 2018) literature. Note that the i.i.d. assumption is on each set of feature vectors across different rounds and we allow feature vectors x_{ti} and x_{tj} for $i \neq j$ to be correlated. Therefore, this is a weaker assumption than the i.i.d. assumption in Chen, Wang, and Zhou (2018) which further imposes independence between items. Also, the i.i.d. assumption on feature vectors is in fact only required during the initialization phase to ensure that after the initialization the MLE $\hat{\theta}$ is sufficiently close to θ^*, i.e., $\|\hat{\theta} - \theta^*\| \leq 1$. We discuss this aspect further in the proof of Theorem 2.1.

Assumption 2.2. *There exists* $\kappa > 0$ *such that for every item* $i \in S$ *and any* $S \in \mathcal{S}$ *and all round* t, $\min_{\|\theta-\theta^*\|\leq 1} p_t(i|S,\theta)p_t(0|S,\theta) \geq \kappa$.

As discussed in Section 2.2.3, this assumption is necessary for the asymptotic normality of the MLE. This is a standard assumption in MNL contextual bandits (Cheung and Simchi-Levi, 2017b; Chen, Wang, and Zhou, 2018), which is also equivalent to the standard assumption on the link function in generalized linear contextual bandits (Filippi et al., 2010; Li, Lu, and Zhou, 2017) to ensure the Fisher information matrix is invertible.

Theorem 2.1 (Regret bound of `UCB-MNL`)**.** *Suppose Assumptions 2.1 and 2.2 hold, and we run* `UCB-MNL` *for T rounds with confidence width $\alpha_t = \frac{1}{2\kappa}\sqrt{d\log\left(\frac{t}{d}\right) + 2\log t}$ and initial sampling for $T_0 = \Theta\left(\frac{d+\log T}{K\sigma_0^2} + \frac{\lambda_0}{K\sigma_0}\right)$ where $\lambda_0 = \max\left\{\frac{1}{4\kappa^2}\left[d\log\left(\frac{T}{d}\right) + 2\log T\right], K\right\}$. Then the cumulative expected regret of* `UCB-MNL` *is upper-bounded by*

$$\mathcal{R}(T) = \mathcal{O}\left(d\sqrt{T}\log(T/d)\right).$$

Discussion of Theorem 2.1. In terms of key problem primitives, Theorem 2.1 demonstrates $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret bound for `UCB-MNL` which is independent of N; hence, it is applicable to the case of very large set of candidate items. Chen, Wang, and Zhou (2018) established the lower bound result $\Omega(d\sqrt{T}/K)$ for the MNL contextual bandits. When K is small, which is typically true in many applications, the regret bound in Theorem 2.1 demonstrates that `UCB-MNL` is near-optimal. The established regret of `UCB-MNL` improves the previous regret bound of Chen, Wang, and Zhou (2018) in both logarithmic and additive factors. Moreover, although having the same rate of $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret up to logarithmic factors, the UCB method in Chen, Wang, and Zhou (2018) has exponential computational complexity, since it needs to enumerate all of the possible $\binom{N}{K}$ assortments.[3] Therefore, to our knowledge, `UCB-MNL` is the first polynomial-time algorithm that achieves $\widetilde{\mathcal{O}}(d\sqrt{T})$ worst-case regret.

Remark 2.1. *Besides this computational advantage, the item-wise upper confidence bounds used in* `UCB-MNL` *as well as the other algorithms proposed in this chapter can also provide additional insights for practitioners. That is, practitioners can directly check which item has more (or less) uncertainty in terms of estimated utility.*

[3]Chen, Wang, and Zhou (2018) recognize this computational issue and also an approximate optimization algorithm to somewhat remedy it; however, not completely. Consider the simple (but widely used in practice) problem setting where each item has unit revenue. In this case, assortment selection under `UCB-MNL` reduces to sorting items based upper confidence bounds and therefore the run time is independent of K, whereas the UCB algorithm proposed in Chen, Wang, and Zhou (2018) still has to construct upper confidence bounds for all the $\binom{N}{K}$ assortments.

2.3.3 Proof of Theorem 2.1

In this section, we present the proof of Theorem 2.1 and the key lemmas for our analysis, whose proofs are presented in the appendix. Recall that the initialization duration T_0, specified in Theorem 2.1, is chosen to ensure that $\|\hat{\theta}_t - \theta^*\| \leq 1$ for $t \geq T_0$. This is done by ensuring that $\lambda_{\min}(V_{T_0})$ is large enough at the end of the initialization. The first lemma specifies how large $\lambda_{\min}(V_{T_0})$ should be in order for $\hat{\theta}_t$ to be sufficiently close to θ^*.

Lemma 2.1. *Let T_0 be any round such that $\lambda_{\min}(V_{T_0}) \geq \frac{1}{4\kappa^2}\left[d\log(T/d) + 2\log T\right]$. Then for any $t \geq T_0$, we have $\mathbb{P}\left(\|\hat{\theta}_t - \theta^*\| > 1\right) \leq \frac{1}{T}$.*

Lemma 2.1 ensures $\|\hat{\theta}_t - \theta^*\|$ with high probability for large enough $\lambda_{\min}(V_{T_0})$. Besides ensuring the concentration of $\hat{\theta}_t$, we also require another lower bound on $\lambda_{\min}(V_{T_0})$, i.e., $\lambda_{\min}(V_{T_0}) \geq K$ to bound the self-normalized process $\sum_{t'=1}^{t}\sum_{i\in S_{t'}}\|x_{t'i}\|_{V_{t'-1}^{-1}}$ in Lemma 2.6. Therefore, at the end of the random initialization period T_0, we need to have

$$\lambda_{\min}(V_{T_0}) \geq \lambda_0 := \max\left\{\frac{1}{4\kappa^2}\left[d\log(T/d) + 2\log T\right], K\right\}.$$

The following proposition, which is a direct adaptation of Proposition 1 in Li, Lu, and Zhou (2017), allows us to find such T_0.

Proposition 1 (Li, Lu, and Zhou 2017, Proposition 1)**.** *Suppose Assumption 2.1 holds. Define $V_{T_0} = \sum_{t'=1}^{T_0}\sum_{i\in S_{t'}} x_{t'i}x_{t'i}^\top$, where T_0 is the length of random initialization. Suppose we run a random initialization with assortment size K for duration T_0 which satisfies*

$$T_0 \geq \frac{1}{K}\left(\frac{C_1\sqrt{d} + C_2\sqrt{\log T}}{\sigma_0}\right)^2 + \frac{2B}{K\sigma_0}$$

for some positive, universal constants C_1 and C_2. Then, $\lambda_{\min}(V_{T_0}) \geq B$ with probability at least $1 - \frac{1}{T}$.

Proposition 1 ensures that $\lambda_{\min}(V_{T_0}) \geq \lambda_0$ holds with high probability if we run the initialization for $\Omega\Big(\frac{d+\log T}{K\sigma_0^2} + \frac{\lambda_0}{K\sigma_0}\Big)$ rounds. Therefore, we only need a logarithmic number (in T) of initial rounds to satisfy the minimum eigenvalue requirement of $\lambda_{\min}(V_{T_0})$. Similar to Filippi et al. (2010) and Li, Lu, and Zhou (2017), the i.i.d. assumption (in Assumption 2.1) on feature vectors $\{x_{ti}\}$ is only needed to ensure the minimum eigenvalue condition for V_{T_0} at the end of the initialization phase. In the rest of the regret analysis of `UCB-MNL`, we do not require this stochastic assumption. Hence, after the initialization period, $\{x_{ti}\}$ can even be chosen adversarially as long as each $\|x_{ti}\|$ is bounded.

The rest of the proof involves bounding the parameter estimation error $\|\hat{\theta}_t - \theta^*\|_{V_t}$ and $\|x_{ti}\|_{V_{t-1}^{-1}}$ as well as the optimism guarantees. Lemma 2.2 shows that the unknown parameter θ^* lies within an ellipsoid centered at $\hat{\theta}_t$ with a suitable confidence radius under the ℓ_2 norm weighted by V_t with high probability. Note that the condition of Lemma 2.2 is ensured with high probability by combining Lemma 2.1 and Proposition 1.

Lemma 2.2. *Suppose* $\|\hat{\theta}_t - \theta^*\| \leq 1$ *for* $t \geq T_0$ *and* $\lambda_{\min}(V_{T_0}) \geq K$. *Then*

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \frac{1}{2\kappa}\sqrt{d\log(t/d) + 2\log t} \tag{2.8}$$

holds for all $t \geq T_0$ *with a probability* $1 - \frac{1}{t^2}$.

Lemma 2.2 is a finite-sample parameter estimation error bound for the MLE of the MNL model. Based on this result, we can construct the optimistic utility estimate using the confidence radius $\alpha_t = \frac{1}{2\kappa}\sqrt{d\log(t/d) + 2\log t}$. The following lemma shows our optimistic utility estimate z_{ti} is an upper confidence bound for the expected utility $x_{ti}^\top\theta^*$ if the underlying parameter θ^* is contained in the confidence ellipsoid centered at $\hat{\theta}_t$.

Lemma 2.3. *Let* $z_{ti} = x_{ti}^\top\hat{\theta}_{t-1} + \alpha_{t-1}\|x_{ti}\|_{V_{t-1}^{-1}}$. *If* (2.8) *holds, then we have*

$$0 \leq z_{ti} - x_{ti}^\top\theta^* \leq 2\alpha_{t-1}\|x_{ti}\|_{V_{t-1}^{-1}}.$$

The following lemma shows that the optimistic expected revenue $\widetilde{R}_t(S_t)$ is an upper bound of the expected revenue of the optimal assortment $R_t(S_t^*, \theta^*)$ under the true unknown parameter θ^*. The lemma is an adaptation of Lemma 4.2 in Agrawal et al. (2019) which was established in the non-contextual setting.

Lemma 2.4. *Suppose S_t^* is the optimal assortment as defined in* (2.3)*, and suppose $S_t = \arg\max_{S \subset \mathcal{S}} \widetilde{R}_t(S)$. If for every item $i \in S_t^*$, $z_{ti} \geq x_i^\top \theta^*$, then the revenues satisfy the following inequalities for all round t:*

$$R_t(S_t^*, \theta^*) \leq \widetilde{R}_t(S_t^*) \leq \widetilde{R}_t(S_t).$$

It is important to note that Lemma 2.4 does not claim that the expected revenue is generally a monotone function, but only the value of the expected revenue corresponding to the optimal assortment increases with an increase in the MNL parameters (Agrawal et al., 2019). Then we show that the expected revenue has the Lipschitz property and bound the immediate regret with the maximum variance over the assortment.

Lemma 2.5. *Suppose that $0 \leq z_{ti} - x_{ti}^\top \theta^* \leq 2\alpha_{t-1} \|x_{ti}\|_{V_{t-1}^{-1}}$ holds for $i \in S_t$ where S_t is the chosen assortment in round t. Then, we have*

$$\widetilde{R}_t(S_t) - R_t(S_t, \theta^*) \leq 2\alpha_{t-1} \max_{i \in S_t} \|x_{ti}\|_{V_{t-1}^{-1}}.$$

The next technical lemma bounds the sum of weighted squared norms. Note that we later apply the Cauchy-Schwarz inequality to eventually bound $\sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_{t-1}^{-1}}$ by $\widetilde{\mathcal{O}}(\sqrt{dT})$.

Lemma 2.6. *Define $V_{T_0} = \sum_{t=1}^{T_0} \sum_{i \in S_{t'}} x_{ti} x_{ti}^\top$ and $V_t = V_{T_0} + \sum_{t'=V_0+1}^{t} \sum_{i \in S_{t'}} x_{t'i} x_{t'i}^\top$. If*

$\lambda_{\min}(V_{T_0}) \geq K$, *then we have*

$$\sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|^2_{V_{t-1}^{-1}} \leq 2d \log(T/d) \ .$$

Hence, each of Lemma 2.2 and Lemma 2.6 contributes $\sqrt{d}$ factor separately to the overall regret, resulting in d factor in Theorem 2.1. Now we can combine the results to show the cumulative regret bound. First we define the joint high probability event for the concentration of the MLE and the concentration after the random initialization.

Definition 2.1. *Define the following event:*

$$\hat{\mathcal{E}} := \left\{ \|\hat{\theta}_t - \theta^*\| \leq 1, \ \ \|\hat{\theta}_t - \theta^*\|_{V_t} \leq \alpha_t, \forall t \geq T_0 \right\} \ .$$

Note that by Proposition 1 with $T_0 = \Omega\Big(\frac{d+\log T}{K\sigma_0^2} + \frac{\lambda_0}{K\sigma_0}\Big)$, we can show

$$\lambda_{\min}(V_{T_0}) \geq \lambda_0 = \max\left\{ \frac{1}{4\kappa^2} \left[d \log\left(\frac{T}{d}\right) + 2 \log T \right], K \right\}$$

with high probability, which in turn allows us to ensure $\|\hat{\theta}_t - \theta^*\| \leq 1$ by Lemma 2.1. Using the union bound, we can show $\|\hat{\theta}_t - \theta^*\| \leq 1$ holds with probability at least $1 - \frac{2}{T}$. We then break the regret into the initialization phase and the learning phase:

$$\begin{aligned}
\mathcal{R}(T) &= \mathbb{E}\left[\sum_{t=1}^{T_0} \left(R(S_t^*, \theta^*) - R(S_t, \theta^*) \right) \right] + \mathbb{E}\left[\sum_{t=T_0+1}^{T} \left(R(S_t^*, \theta^*) - R(S_t, \theta^*) \right) \right] \\
&\leq T_0 + \mathbb{E}\left[\sum_{t=T_0+1}^{T} \left(\widetilde{R}_t(S_t) - R(S_t, \theta^*) \right) \right]
\end{aligned}$$

where the last inequality comes from optimistic revenue estimation by Lemma 2.4. Now, we further decompose the regret of the learning phase further into two components – when the high probability event holds in Lemma 2.2 and in Lemma 2.1 (i.e., $\hat{\mathcal{E}}$ holds) and

when either of the events does not hold, (i.e., $\hat{\mathcal{E}}^c$).

$$\begin{aligned}
\mathcal{R}(T) &\leq T_0 + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(\widetilde{R}_t(S_t) - R_t(S_t,\theta^*)\right)\mathbb{1}(\hat{\mathcal{E}})\right] + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(\widetilde{R}_t(S_t) - R_t(S_t,\theta^*)\right)\mathbb{1}(\hat{\mathcal{E}}^c)\right] \\
&\leq T_0 + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(\widetilde{R}_t(S_t) - R_t(S_t,\theta^*)\right)\mathbb{1}(\hat{\mathcal{E}})\right] + \sum_{t=1}^{T}\left(\frac{1}{t^2}+\frac{2}{T}\right) \\
&\leq T_0 + \sum_{t=1}^{T} 2\alpha_{t-1}\max_{i\in S_t}\|x_{ti}\|_{V_{t-1}^{-1}} + \mathcal{O}(1) \\
&\leq T_0 + 2\alpha_T\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|_{V_{t-1}^{-1}} + \mathcal{O}(1)
\end{aligned}$$

where the third inequality is from Lemma 2.5. Applying the Cauchy-Schwarz inequality to the second term, it follows that

$$\mathcal{R}(T) \leq T_0 + 2\alpha_T\sqrt{T\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{V_{t-1}^{-1}}} + \mathcal{O}(1)\,.$$

Applying Lemma 2.6 for $\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{V_{t-1}^{-1}}$,

$$\mathcal{R}(T) \leq T_0 + 2\alpha_T\sqrt{2dT\log(T/d)} + \mathcal{O}(1)\,.$$

Finally, our choice of α_t gives $\alpha_T = \frac{1}{2\kappa}\sqrt{d\log(T/d) + 2\log T}$, we have

$$\mathcal{R}(T) \leq T_0 + \frac{d}{\kappa}\sqrt{2T}\log(T/d) + \frac{2}{\kappa}\sqrt{dT\log T\log(T/d)} + \mathcal{O}(1)\,.$$

2.3.4 Online Parameter Update

`UCB-MNL` is simple to implement and more practical compared to previously known methods in the MNL bandit problem. The algorithm is statistically efficient, as established in Theorem 2.1 and is also computationally much more efficient than the previously known method (e.g., Chen, Wang, and Zhou 2018). However, as t increases, the computational

cost of `UCB-MNL` becomes more expensive. In each round t, the MLE $\hat{\theta}_t$ is computed using $\Theta(tK)$ samples, i.e., the per-round computational complexity grows linearly with t for a straightforward implementation of the algorithm. Note that this issue is not unique to `UCB-MNL`. Chen, Wang, and Zhou (2018) also suffers from the same issue in addition to its computationally expensive procedure of the upper confidence construction for all assortments which we discussed in Section 2.3.2. In fact, this bottleneck makes many bandit algorithms including those in generalized linear bandits (Filippi et al., 2010; Li, Lu, and Zhou, 2017) inappropriate for online implementations in real-world applications since the entire learning history is stored in memory and used for parameter estimation in each round.

In this section, we discuss a modification of `UCB-MNL` which incorporates an efficient online update. This modification effectively exploits particular structures of the MNL model. In stead of computing the exact solution for MLE which does not scale well in time and space complexity, we propose an online update scheme to find an approximate solution. First, we define the per-round loss for the MNL model and its gradient.

Definition 2.2. *Define the per-round loss $f_t(\theta)$ and its gradient $G_t(\theta)$ as the following:*

$$f_t(\theta) := -\sum_{i \in S_t} y_{ti} \log p_t(i|S_t, \theta) = -\sum_{i \in S_t} y_{ti} x_{ti}^\top \theta + \log\left(1 + \sum_{j \in S_t} \exp(x_{tj}^\top \theta)\right)$$
$$G_t(\theta) := \nabla_\theta f_t(\theta) = \sum_{i \in S_t} \left(p_t(i|S_t, \theta) - y_{ti}\right) x_{ti}$$

The important observation here is that the loss for the MNL model in each round t is strongly convex over bounded domain, which enables us to apply a variant of the online Newton step (Hazan, Agarwal, and Kale, 2007), in particular inspired by (Hazan, Koren, and Levy, 2014; Zhang et al., 2016) that proposed online algorithms for the logistic model. Specifically, we propose to find an approximate solution by solving the following

optimization problem

$$\hat{\theta}_t = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\theta - \hat{\theta}_{t-1}\|_{V_t}^2 + (\theta - \hat{\theta}_{t-1})^\top G_{t-1}(\hat{\theta}_{t-1}) \right\} \tag{2.9}$$

where $V_t = V_{t-1} + \frac{\kappa}{2} \sum_{i \in S_t} x_{ti} x_{ti}^\top$.

Algorithm 2 `UCB-MNL` with online parameter update

Input: total rounds T, initialization rounds T_0 and confidence radius $\widetilde{\alpha}_t$

Initialization: **for** $t \in [T_0]$

  Randomly choose S_t with $|S_t| = K$

  $V_t \leftarrow V_{t-1} + \frac{\kappa}{2} \sum_{i \in S_t} x_{ti} x_{ti}^\top$

for all $t = T_0 + 1$ to T **do**

  Compute $\widetilde{z}_{ti} = x_{ti}^\top \hat{\theta}_{t-1} + \widetilde{\alpha}_{t-1} \|x_{ti}\|_{V_{t-1}^{-1}}$ for all $i \in [N]$

  Compute $S_t = \operatorname{argmax}_{S \subset \mathcal{S}} \widetilde{R}_t(S)$ based on $\{\widetilde{z}_{ti}\}$

  Offer S_t and observe y_t (user choice in round t)

  Update $V_t \leftarrow V_{t-1} + \frac{\kappa}{2} \sum_{i \in S_t} x_{ti} x_{ti}^\top$

  Compute $\hat{\theta}_t$ by solving the problem

$$\hat{\theta}_t = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\theta - \hat{\theta}_{t-1}\|_{V_t}^2 + (\theta - \hat{\theta}_{t-1})^\top G_{t-1}(\hat{\theta}_{t-1}) \right\}$$

end for

The modified algorithm is summarized in Algorithm 2. The key differences between Algorithm 1 and Algorithm 2 are the parameter update rule in (2.9) and the choice of the confidence radius. During the learning phase, the algorithm builds a upper confidence utility estimate $\widetilde{z}_{ti} = x_{ti}^\top \hat{\theta}_{t-1} + \widetilde{\alpha}_{t-1} \|x_{ti}\|_{V_{t-1}^{-1}}$ using a new confidence radius $\widetilde{\alpha}_t$ which is specified in Theorem 2.2 (based on the confidence bound in Lemma 2.7). For parameter estimation, only $\Theta(K)$ samples are needed for both computation and space per each round, compared to $\Theta(tK)$ in Algorithm 1 which grows linearly with each round t.

The following lemma provides the confidence bound for the online parameter update.

Lemma 2.7. *Suppose* $\|\hat{\theta}_t - \theta^*\| \leq 1$ *for* $t \geq T_0$ *and* $\lambda_{\min}(V_{T_0}) \geq K$*. Then*

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \sqrt{T_0 + \frac{8}{\kappa} d \log(t/d) + \left(\frac{8}{\kappa} + \frac{16}{3}\right) \log(\lceil 2\log_2(tK/2)\rceil t^4) + 4}$$

holds for all $t \geq T_0$ *with a probability* $1 - \frac{1}{t^2}$*.*

The proof of Lemma 2.7 relies on exploiting the structure of the loss of the MNL model and concentration inequalities for martingales. Since we use fewer samples (less information) per each parameter update in the modified online update compared to the *full* MLE computation, one might expect the confidence bound to increase with the online update modification. Nevertheless, Lemma 2.7 shows that the confidence bound of the parameter estimation scales with $\mathcal{O}\Big(\sqrt{d\log(t/d)}\Big)$, which is of the same order as the bound shown in Lemma 2.2 – although there are extra additive terms and potentially a larger constant. This suggests that the total regret bound for the modified `UCB-MNL` algorithm should be also of the same order (up to logarithmic factors) as the regret bound of the original `UCB-MNL` algorithm (Algorithm 1). In Theorem 2.2, we present the regret bound for the `UCB-MNL` with online parameter update (Algorithm 2).

Theorem 2.2. *Suppose Assumptions 2.1 and 2.2 hold, and we run* `UCB-MNL` *for* T *rounds with "online parameter update" with initial sampling for* $T_0 = \Theta\Big(\frac{d+\log T}{K\sigma_0^2} + \frac{\lambda_0}{K\sigma_0}\Big)$ *where* $\lambda_0 = \max\Big\{\frac{1}{4\kappa^2}\left[d\log(T/d) + 2\log T\right], K\Big\}$ *and with confidence width*

$$\widetilde{\alpha}_t = \sqrt{T_0 + \frac{8}{\kappa} d \log(t/d) + \left(\frac{8}{\kappa} + \frac{16}{3}\right) \log(\lceil 2\log_2(tK/2)\rceil t^4) + 4}\,.$$

Then the cumulative expected regret of the algorithm is upper-bounded by

$$\mathcal{R}(T) \leq T_0 + \mathcal{O}(1) + \widetilde{\alpha}_T \sqrt{dT\log(T/d)} = \mathcal{O}\Big(d\sqrt{T\log(1+T/d)\log(T/d)}\Big)\,.$$

Theorem 2.2 still achieves a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T})$ which matches the bound in Theorem 2.1 for `UCB-MNL` while improving both the time and space complexities of the algorithm. This result suggests that the modified `UCB-MNL` is appropriate for online implementation, achieving both statistical and computational efficiency. The proof of Theorem 2.2 follows the similar steps as that of Theorem 2.1 and is presented in Section A.2.

2.4 Non-asymptotic Normality of the MLE for MNL Models

We have shown that `UCB-MNL` is both statistically and computationally efficient. The algorithm also shows state-of-the-art practical performances as we report in Section 2.8. However, the regret bound in Theorem 2.1 has a linear dependence on feature dimension d and, therefore, is not very attractive when the feature vectors are high dimensional. We next investigate whether a sublinear dependence on d is possible. In the regret analysis for `UCB-MNL`, we upper-bound the prediction error $x^\top(\theta^* - \hat{\theta}_t)$ using Hölder's inequality, $|x^\top\hat{\theta}_t - x^\top\theta^*| \leq \|x\|_{V_t^{-1}}\|\hat{\theta}_t - \theta^*\|_{V_t}$, where we show each of the terms on the right hand side is bounded by $\widetilde{\mathcal{O}}(\sqrt{d})$, hence resulting in a linear dependence on d when combined. A potential solution to circumvent this challenge is to control the prediction error directly without bounding two terms separately and establish a sublinear dependence on d.

In Theorem 2.3 we propose a non-asymptotic normality bound for the MLE for the MNL model in order to establish a sharper concentration result for $|x^\top(\hat{\theta}_t - \theta^*)|$. This is a generalization of Theorem 1 in Li, Lu, and Zhou (2017) to the MNL model. To the best of our knowledge, there was no existing finite-sample normality results for the prediction error of the utility for the MNL model. This result can be of independent interest beyond the bandit problems.

Theorem 2.3 (Non-asymptotic normality of MLE)**.** *Suppose we have independent responses $y_1, ..., y_n$ conditioned on feature vectors $\{x_{ti}\}_{t=1,i=1}^{n,K}$. Define $V_n = \sum_{t=1}^{n} \sum_{i \in S_t} x_{ti} x_{ti}^\top$, and let $\delta > 0$ be given. Furthermore, assume that $\lambda_{\min}(V_n) \geq \max\Big\{\frac{9\mathcal{D}^4}{\kappa^4 \log(1/\delta)}, \frac{144\mathcal{D}^2}{\kappa^4}\Big\}$ where $\mathcal{D} := \min\Big\{4\sqrt{2d + \log\frac{1}{\delta}}, \sqrt{d\log(n/d) + 2\log\frac{1}{\delta}}\Big\}$. Then, for any $x \in \mathbb{R}^d$, the maximum likelihood estimator $\hat{\theta}_n$ of the MNL model satisfies with probability at least $1 - 3\delta$ that*

$$|x^\top \hat{\theta}_n - x^\top \theta^*| \leq \frac{5}{\kappa}\sqrt{\log\frac{1}{\delta}} \|x\|_{V_n^{-1}} .$$

Hence, the prediction error can be bounded by $\widetilde{\mathcal{O}}(\sqrt{d})$ with high probability as long as the conditions on independence of samples and the minimum eigenvalue are satisfied. Note that although the statement of Theorem 2.3 is similar to that of the generalized linear model version in Li, Lu, and Zhou (2017), the extension to the MNL model is non-trivial because choice probability for any given item $i \in S_t$ is function of the all the items in the assortment S_t, and hence the analysis is much more involved. Theorem 2.3 implies that we can control the behavior of the MLE in every direction allowing us to handle the prediction error in a tighter fashion.

2.5 Generating Independent Samples and Provable Optimality

Unfortunately, we cannot directly apply the tight confidence bound of the MLE shown in Theorem 2.3 to `UCB-MNL` since Theorem 2.3 requires independent samples (as well as the minimum eigenvalue being large enough, but this condition can be satisfied by initial random exploration). `UCB-MNL` is not guaranteed to produce independent samples since the algorithm chooses assortments based on previous observations, causing dependence between collected samples — this is a common characteristic of most bandit algorithms.

This issue can be handled by generating independent samples using a framework proposed in Auer (2002), which we denote as "Auer-framework." This Auer-framework has been previously used in several variants of (generalized) linear bandits (Chu et al., 2011; Li, Lu, and Zhou, 2017; Zhou, Xu, and Blanchet, 2019). In what follows, we propose an adaptation of the Auer-framework for the MNL contextual bandit problem, and establish a provably optimal regret bound for the algorithm.

Algorithm 3 `baseCB-MNL`

1: **Input**: confidence radius β, index set Ψ, set A, features $\{x_{ti}\}$

2: Compute MLE $\hat{\theta}$ by solving the equation

$$\sum_{t' \in \Psi} \sum_{i \in S_{t'}} \left(p_{t'}(i|S_{t'}, \theta) - y_{t'i}\right) x_{t'i} = \mathbf{0}$$

3: Update $V_\Psi = \sum_{t' \in \Psi} \sum_{i \in S_{t'}} x_{t'i} x_{t'i}^\top$

4: Compute the following:

$$m_{ti} = x_{ti}^\top \hat{\theta} \quad \text{for all } i \in \mathcal{I}$$
$$w_{ti} = \beta \|x_{ti}\|_{V_\Psi^{-1}} \quad \text{for all } i \in \mathcal{I}$$
$$\mathcal{W}_t = 2 \max_{i \in \mathcal{I}} w_{ti}$$

where $\mathcal{I} = \{i \in S : S \in A\}$

We adapt the decomposition of the algorithm introduced in Auer (2002). That is, we design a method which consists of two parts: (i) a subroutine algorithm `baseCB-MNL` (Algorithm 3) to compute the MLE and the maximum confidence interval of expected utility among the items in the candidate set (assuming statistical independence among the samples), and (ii) a master algorithm `supCB-MNL` (Algorithm 4) to ensure that the independence assumption holds. `supCB-MNL` operates on the radius of the confidence

bound, independent of estimated expected utility, to perform exploration. `supCB-MNL` maintains $\{\Psi_\ell\}_{\ell=0}^{L}$, the sets of time indices which are the partitions of the entire horizon $[T] = \{1, ..., T\}$. The purpose of this partitioning is to ensure that the choice responses y_t in each index set Ψ_ℓ are independent, so that we can apply the normality result of the MLE in Theorem 2.3 to samples in each index set Ψ_ℓ separately.

Algorithm 4 `supCB-MNL`

Input: T, initialization T_0, confidence radius β
Initialization: **for** $t \in [T_0]$
  randomly choose S_t with $|S_t| = K$
set $L = \lfloor \frac{1}{2} \log_2 T \rfloor$, and $\Psi_0 = \cdots = \Psi_L = \varnothing$.
for all $T_0 = \tau + 1$ to T **do**
  Initialize $A_1 = \mathcal{S}$ and $\ell = 1$
  while S_t is empty **do**
    (a). Run `baseCB-MNL` with A_ℓ, β and $\Psi_\ell \cup [T_0]$ to compute $\hat{\theta}_t^{(\ell)}$, $m_{ti}^{(\ell)}$, $w_{ti}^{(\ell)}$, $\mathcal{W}_t^{(\ell)}$
    (b). **If** $\mathcal{W}_t^{(\ell)} \le \frac{1}{\sqrt{T}}$,
      set $S_t = \text{argmax}_{S \in A_\ell} R_t(S, \hat{\theta}_t^{(\ell)})$ based on $\{m_{ti}^{(\ell)}\}$
      update $\Psi_0 = \Psi_0 \cup \{t\}$
    (c). **Else if** $\mathcal{W}_t^{(\ell)} > 2^{-\ell}$,
      set $S_t = \text{argmax}_{S \subseteq A_\ell} \sum_{i \in S} w_{ti}^{(\ell)}$
      update $\Psi_\ell = \Psi_\ell \cup \{t\}$
    (d). **Else if** $\mathcal{W}_t^{(\ell)} \le 2^{-\ell}$,
      compute $\mathcal{M}_t^{(\ell)} = \max_{S \in A_\ell} R_t(S, \hat{\theta}_t^{(\ell)})$ based on $\{m_{ti}^{(\ell)}\}$
      $A_{\ell+1} = \left\{ S \in A_\ell : R_t(S, \hat{\theta}_t^{(\ell)}) \ge \mathcal{M}_t^{(\ell)} - 2^{-\ell+1} \right\}$
      $\ell \leftarrow \ell + 1$
  end while
end for

In each round of `supCB-MNL` (Algorithm 4), the decision-making agent screens the

candidate assortments based on the value of $w_{ti} = \alpha \|x_{ti}\|_{V_t^{-1}}$ for items in assortments in A_ℓ through epochs $\ell = 1, ..., L$ until an assortment S_t is chosen. We describe each step of the inner-loop over these epochs for a given round.

- *Sub-routine:* In step (a), we run `baseCB-MNL` (Algorithm 3) which uses the normality result of the MLE (Theorem 2.3) to compute $m_{ti}^{(\ell)}$ and $w_{ti}^{(\ell)}$ for all i, $\mathcal{W}_t^{(\ell)}$, and $\hat{\theta}_t^{(\ell)}$. We can apply Theorem 2.3 here since samples in each index set $\{y_t, t \in \Psi_\ell\}$ are independent of each other given the feature vectors for each Ψ_ℓ (see Lemma 2.8).

- *Exploitation:* In step (b), if the maximal confidence interval $\mathcal{W}_t^{(\ell)}$ is sufficiently small, i.e., smaller than $\frac{1}{\sqrt{T}}$, for all possible candidate sets, then we perform pure exploitation. This step's contribution to the total regret will be small since we have well-concentrated estimated utilities for all items.

- *Exploration:* In step (c), if there is a set that has a large enough confidence interval (larger than $2^{-\ell}$), then we choose a set has the maximal uncertainty. Then we update the index set Ψ_ℓ to include the time-stamp t.

- *Pruning:* Finally, step (d) is a pruning step, where we remove clearly sub-optimal sets and keep the sets which are possibly optimal.

If the algorithm does not choose S_t in epoch ℓ, then it moves on to the next epoch $\ell + 1$ and repeats the process until S_t is chosen either via the exploitation step in (b) or via the exploration step in (c). Note that when maximizing the expected revenue $R_t(S, \hat{\theta})$ in step (b) or in step (d), it uses the expected revenue defined in (2.2) replacing θ^* with the current estimator $\hat{\theta}_t^{(\ell)}$ — notice that we use the expected revenue $R_t(S)$ in `supCB-MNL`, rather than the optimistic expected revenue $\widetilde{R}_t(S)$ used in `UCB-MNL` (Algorithm 1).

The following result (which is adapted from Lemma 14 of Auer (2002)) shows that the samples collected from Algorithm 4 in each index set Ψ_ℓ are independent.

Lemma 2.8 (Lemma 4 in Li, Lu, and Zhou (2017)). *For all $\ell \in [L]$ and $t \in [T]$, given the set of feature vectors in index set Ψ_ℓ, $\{[x_{ti}]_{i\in S_t}, t \in \Psi_\ell\}$, the corresponding choice responses $\{y_t, t \in \Psi_\ell\}$ are independent random variables.*

2.5.1 Regret Bound for `supCB-MNL` Algorithm

Independent samples ensured by the master algorithm `supCB-MNL` and Lemma 2.8 enable us to apply the non-asymptotic normality result in Theorem 2.3 separately to samples in each index set Ψ_ℓ. We present the following regret bound of `supCB-MNL` (Algorithm 4).

Theorem 2.4 (Regret bound of `supCB-MNL`). *Suppose Assumptions 2.1 and 2.2 hold, we run* `supCB-MNL` *for $T \geq \widetilde{T}$ rounds, where*

$$\widetilde{T} = \Omega\left(\frac{\log^2\left(TN\log_2 T\right)}{K^2\kappa^8 d} + \frac{d^3}{K^2\kappa^8}\right) \tag{2.10}$$

with initialization $T_0 = \sqrt{dT}$ and confidence width $\beta = \frac{5}{\kappa}\sqrt{\log(TN\log_2 T)}$. Then, the cumulative expected regret of the algorithm is upper-bounded by

$$\mathcal{R}(T) = \mathcal{O}\left(\sqrt{dT\log(T/d)\log(TN\log_2 T)\log_2 T}\right).$$

Discussion of Theorem 2.4. Theorem 2.4 establishes the regret bound of $\widetilde{\mathcal{O}}(\sqrt{dT})$ for `supCB-MNL`. $\Omega(\sqrt{NT})$ lower bound was shown in Chen and Wang (2017) for the non-contextual MNL bandits. This lower bound can be translated to $\Omega(\sqrt{dT})$ if each item is represented as one-hot encoding. Hence, the regret bound in Theorem 2.4 matches the lower bound for the MNL bandit problem with finite items. To our knowledge, this is the first result that achieves the rate of $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret and establishes the provable optimality (up to logarithmic fators) in the MNL contextual bandit problem. Comparing with Theorem 2.1 for `UCB-MNL` (Algorithm 1) as well as its online update variant (Algorithm 2),

which are near-optimal in the case of infinitely large item set (or exponentially large N), the improvement of $\sqrt{d}$ factor comes from directly controlling the utility estimation error using Theorem 2.3. Note that the regret bound in Theorem 2.4 has a logarithmic dependence on N, therefore `supCB-MNL` is not applicable to a case where there are an infinite number of total items. While provably optimal, however, `supCB-MNL` is not a practical algorithm as one may expect from the design of the algorithm (common issue for almost all $\sqrt{dT}$ regret algorithms that follows the framework of Auer (2002)). We discuss this aspect further in Section 2.6.

2.5.2 Proof Outline of Theorem 2.4

Note that we want to ensure that the concentration result of the prediction error in Theorem 2.3 holds for all items $i \in [N]$ and for all rounds $t \in [T]$ including the inner loop (epochs) in Algorithm 4; hence for all ℓ up to $L = \mathcal{O}(\log_2 T)$. Thus, we choose the confidence radius to be $\beta = \frac{5}{\kappa}\sqrt{\log(TN\log_2 T)}$. Then with probability at least $1 - \frac{3}{TN\log_2 T}$, we would have for each i, ℓ, and t,

$$|m_{ti}^{(\ell)} - x_{ti}\theta^*| \leq w_{ti}^{(\ell)} \tag{2.11}$$

if the independence condition and the minimum eigenvalue condition of Theorem 2.3 are satisfied. Then we can use the union bound to show this concentration holds jointly for all items and all rounds (including the inner loops) with high probability. Now, we know that the independence requirement is satisfied by `supCB-MNL` (as shown in Lemma 2.8) that produces independent sample for each index set Ψ_ℓ. For the minimum eigenvalue condition, we need to ensure that

$$\lambda_{\min}(V_{T_0}) \geq \Omega\left(\max\left\{\frac{(d+\log(TN\log_2 T))^2}{\kappa^4 \log(TN\log_2 T)}, \frac{d+\log(TN\log_2 T)}{\kappa^4}\right\}\right) \tag{2.12}$$

for sufficiently large T. For $T \geq \widetilde{T}$ satisfying the condition in (2.10), we can run the random initialization for $T_0 = \Theta(\sqrt{dT})$ to ensure (2.12) with high probability. Given this minimum eigenvalue guarantee and the concentration result in (2.11), we decompose the regret into to two disjoint parts — the regret incurred when an assortment is selected for exploitation (step (b) in Algorithm 4) and the regret for exploration (step (c) in Algorithm 4). We show the cumulative regret coming from step (b) is small since the utility estimates are already "accurate" in this case. We also show that even when we take an exploratory action in step (c), the regret incurred by such an action is not too large due to the concentration result accompanied by the pruning procedure in step (d). See Appendix A.4 for the full proof.

2.6 Practical Algorithm with Sublinear Dependence on Feature Dimension

We have shown that `supCB-MNL` establishes the provable optimality with $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret in the MNL contextual bandit problem. However, this comes at a cost. `supCB-MNL` (Algorithm 4), although provably optimal, is not practical (see Section 2.8). In fact, this is true for *all* methods (Chu et al., 2011; Li, Lu, and Zhou, 2017; Zhou, Xu, and Blanchet, 2019) that rely on the Auer-framework (Auer, 2002) because the framework wastes too many samples with random exploration — we verify this issue for `supCB-MNL` in the numerical experiments in Section 2.8.[4] Furthermore, the adaptation of the Auer-framework to the MNL contextual bandit problem creates an additional computational bottleneck where pruning sub-optimal candidate assortments (step (d) in Algorithm 4) can be computationally expensive.

[4]The previous methods (Chu et al., 2011; Li, Lu, and Zhou, 2017; Zhou, Xu, and Blanchet, 2019) that use techniques in Auer (2002) do not provide numerical evaluations.

Chapter 2: UCB Algorithms for MNL Contextual Bandits

In this section, we investigate whether $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret can be achieved in a practical manner for a class of the MNL contextual bandit problem where the revenue for each item is uniform. As briefly mentioned earlier, this is one of the most common classes used in various applications including web search engines and most recommender systems.

Algorithm 5 `DBL-MNL`

Input: sampling parameter q_k, confidence radius β_k
Set $\tau_1 \leftarrow d$, $t \leftarrow 1$, $V_0 \leftarrow \mathbf{0}_{d\times d}$
Initialization: **for** $t \in [d]$
  Randomly choose $S_t \in \mathcal{S}$ with $|S_t| = K$
  $V_t \leftarrow V_{t-1} + \sum_{i\in S_t} x_{ti}x_{ti}^\top$
for each episode $k = 2, 3, ...$ **do**
  Set the last round of k-th episode: $\tau_k \leftarrow 2^{k-1}$
  Compute MLE $\hat{\theta}_k$ by solving $\sum_{t=\tau_{k-2}+1}^{\tau_{k-1}} \sum_{i\in S_t} \Big(p_t(i|S_t, \hat{\theta}_k) - y_{ti}\Big)x_{ti} = \mathbf{0}$
  Update $W_{k-1} \leftarrow V_{\tau_{k-1}+1}$; Reset $V_{\tau_{k-1}+1} \leftarrow 0_{d\times d}$
  for each round $t = \tau_{k-1} + 1, ..., \tau_k$ **do**
    if $\tau_k - t \le q_k$ and $\lambda_{\min}(V_t) \le \frac{Kq_k\sigma_0}{2}$ **then**
      Randomly choose $S_t \in \mathcal{S}$ with $|S_t| = K$
    else
      Offer $S_t = \text{argmax}_{S\in\mathcal{S}} \widetilde{R}_t(S)$
    end if
    Update $V_{t+1} \leftarrow V_t + \sum_{i\in S_t} x_{ti}x_{ti}^\top$
  end for
end for

2.6.1 Algorithm: DBL-MNL

We propose a new algorithm, DBL-MNL (Algorithm 5), that achieves a sublinear dependence on feature dimension d and is *practical*. The algorithm starts with a short initialization perioid to ensure the invertibility of V_t. Then, DBL-MNL operates in an episodic manner. At the beginning of each episode, the MLE is computed using the samples from a previous episode. Within an episode, the parameter is not updated, but the algorithm takes an UCB action based on the parameter computed at the beginning of the episode. In particular, for round t in the k-th episode, the optimistic utility estimate is computed as

$$\widetilde{z}_{ti} = x_{ti}^\top \hat{\theta}_k + \beta_k \|x_{ti}\|_{W_{k-1}^{-1}} \quad \text{where } W_{k-1} = \sum_{t'=\tau_{k-1}+1}^{\tau_{k-1}} \sum_{i \in S_{t'}} x_{t'i} x_{t'i}^\top \tag{2.13}$$

and τ_{k-1} is the last period of the $(k-1)$-th episode . Note that the Gram matrix V_t resets at the beginning of each episode. Under this action selection, samples within each episode are independent of each other. Episode lengths are doubled over time such that the length of the k-th episode is twice as large as that of the $(k-1)$-th episode for each k. This doubling technique is inspired by Jaksch, Ortner, and Auer (2010) and Javanmard and Nazerzadeh (2019), but surprisingly has not been used much in the contextual bandit literature. Towards the end of each episode, the algorithm checks whether $\lambda_{\min}(V_t)$ is suitably large. If not, it performs random exploration. Since episode lengths are growing exponentially and the threshold for $\lambda_{\min}(V_t)$ is only logarithmic in t, even in the worst case, the algorithm draws $\mathcal{O}(\log^2 T)$ random samples. Note that the algorithm may not even take these exploratory actions since $\lambda_{\min}(V_t)$ may already surpass the threshold for long enough episodes (this is clearly observed in numerical evaluations). This makes DBL-MNL much more practical since it would perform minimal random exploration. Furthermore, the algorithm is computationally efficient with only logarithmic number of parameter updates instead of updating in every period.

2.6.2 Regret Bound for DBL-MNL Algorithm

We analyze the regret bound of DBL-MNL for which we aim to establish $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret. For our analysis, we first present the following assumption which is similar to Assumption 2.1.

Assumption 2.3. *Each feature vector x_{ti} is drawn i.i.d. from an unknown distribution p_X with $\|x_{ti}\| \leq 1$ all t, i and there exists a constant $\sigma_0 > 0$ such that $\lambda_{\min}\left(\mathbb{E}\left[x_{ti}x_{ti}^\top\right]\right) \geq \sigma_0$.*

Assumption 2.3 assumes that feature vectors are i.i.d. across items in each round, which is slightly stronger than Assumption 2.1. As mentioned earlier, this assumption is the standard assumption in the MNL contextual bandit literature (e.g., Chen, Wang, and Zhou 2018). We also add the following assumption on feature vectors which encompasses many canonical distributions.

Assumption 2.4 (Relaxed symmetry)**.** *For a joint distribution p_X, there exists $\rho_0 < \infty$ such that $\frac{p_X(-x)}{p_X(x)} \leq \rho_0$ for all x.*

This assumption is also used in the analysis of sparse contextual bandits (in Chapter 4) which states that the joint distribution $p_{\mathcal{X}}$ can be skewed but this skewness is bounded. For symmetrical distributions, Assumption 2.4 holds with $\rho_0 = 1$. One can see that a large class of continuous and discrete distributions satisfy Assumption 2.4, e.g., Gaussian distribution, truncated Gaussian distribution, uniform distribution, and Rademacher distribution, and many more. Under this suitable regularity, we establish the following regret bound for DBL-MNL.

Theorem 2.5 (Regret bound of DBL-MNL)**.** *Suppose Assumptions 2.3-2.4 hold, $r_i \equiv r$ is uniform for all $i \in [N]$, and we run* DBL-MNL *with $\beta_k = \frac{5}{\kappa}\sqrt{\log(\tau_k N/2)}$ and*

$$q_k = \frac{2}{\sigma_0 K}\max\left\{\frac{9\mathcal{D}_k^4}{\kappa^4 \log(\tau_k N/2)}, \frac{144\mathcal{D}_k^2}{\kappa^4}\right\}$$

where $\mathcal{D}_k = \min\Big\{4\sqrt{2d + \log(\tau_k N/2)}, \sqrt{d\log(\tau_k/d) + 2\log(\tau_k N/2)}\Big\}$. *Then the cumulative expected regret of* `DBL-MNL` *over horizon* T *is upper-bounded by*

$$\mathcal{R}(T) = \mathcal{O}\Big(\sqrt{dT\log(T/d)\log(TN)\log_2 T}\Big).$$

Discussion of Theorem 2.5. `DBL-MNL` achieves $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret when the revenue for each item is uniform. Theorem 2.5 provides insights beyond the MNL contextual bandits: it shows that under the suitable regularity condition, it is possible for a practical algorithm to attain $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret. We expect this technique to yield practical provably optimal algorithms for other variants of contextual bandit problems. Similar to the regret bound of `supCB-MNL` (Theorem 2.4), `DBL-MNL` has a logarithmic dependence on N (as is common for many $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret algorithms (Chu et al., 2011; Li, Lu, and Zhou, 2017)). In fact, the numerical experiments in Section 2.8 suggest that the performance of `DBL-MNL` does have at least a logarithmic dependence on N.

Remark 2.2. *Note that* `DBL-MNL` *(Algorithm 5) can still be used in non-uniform revenue settings although we establish the regret bound for the uniform revenues in Theorem 2.5. The uniform revenue setting that we consider in the regret analysis of* `DBL-MNL` *is an important problem class that still embeds the full statistical complexity of the MNL contextual bandit problem. This setting can be also considered as the top-K selection problem (Cao et al., 2015) as far as action selection is concerned since the agent selects K items with the highest estimated utilities for the user. (Recall that, in the non-uniform revenue setting, the size of the optimal assortment may be smaller than the assortment size constraint K.) However, this uniform revenue setting is still more difficult than other variants of combinatorial (contextual) bandits such as semi-bandits and cascading bandits in that top-K offering in this version of the MNL contextual bandit problem still takes the substitution effect into account. This problem class is particularly important because of its*

wide range of applications. For example, in many recommender systems, the agent's goal is to maximize the click-through rate (CTR) on its service where each click is weighted uniformly. An interesting aspect of the uniform revenue setting is that the combinatorial optimization step for assortment selection reduces to a sorting task based on estimated utilities in our proposed algorithms, making the assortment selection procedure much more computationally efficient.

2.6.3 Proof Outline of Theorem 2.5

Since the length of an episode grows exponentially, the number of episodes up to round T is logarithmic in T. In particular, the T-th round belongs to the L-th episode with $L = \lfloor \log_2 T \rfloor + 1$. Let $\mathcal{T}_k := \{\tau_{k-1} + 1, ..., \tau_k\}$ denote an index set of rounds that belong to the k-th episode. Note that the length of the k-th episode is $|\mathcal{T}_k| = \tau_k/2$. Then, we let $\texttt{Reg}(k\text{-th episode})$ denote the cumulative regret of the k-th episode, i.e.,

$$\texttt{Reg}(k\text{-th episode}) := \mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \Big(R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*) \Big)\right]$$

so that the cumulative expected regret over T rounds is $\mathcal{R}(T) = \sum_{k=1}^{L} \texttt{Reg}(k\text{-th episode})$. Therefore, it suffices to bound each $\texttt{Reg}(k\text{-th episode})$. Now, for each episode $k \in [L]$, we consider the following two cases.

(i) $|\mathcal{T}_k| \leq q_k$: In this case, the length of an episode is not large enough to have the concentration of the prediction error due to the failure of ensuring the lower bound on $\lambda_{\min}(V_t)$. Therefore, we cannot control the regret in this case. However, the total number of such rounds is only logarithmic in T, hence the regret corresponding to this case contributes minimally to the total regret.

(ii) $|\mathcal{T}_k| > q_k$: We can apply the fast convergence result in Theorem 2.3 as long as the lower bound on $\lambda_{\min}(V_t)$ is guaranteed — note that the independence condition is

already satisfied since samples in each episode are independent of each other. We show that $\lambda_{\min}(V_t)$ grows linearly as t increases in each episode with high probability. In case of $\lambda_{\min}(V_t)$ not growing as fast as the rate we require, we perform random sampling to satisfy this criterion towards the end of each episode. Therefore, with high probability, the lower bound on $\lambda_{\min}(V_t)$ becomes satisfied.

For case (i), clearly $q_k \leq q_L$ for any $k \in \{1, ..., L\}$. $|\mathcal{T}_k|$ eventually grows to be larger than q_L for some k since q_L is logarithmic in T. Let k' be the first episode such that $|\mathcal{T}_{k'}| \geq q_L$. Hence, $|\mathcal{T}_{k'}| \leq 2q_L$. Thus, the cumulative regret prior to the k'-th episode is

$$\sum_{k=1}^{k'-1} \texttt{Reg}(k\text{-th episode}) \leq \sum_{k=1}^{k'-1} |\mathcal{T}_k| = |\mathcal{T}_{k'}| \leq 2q_L = \mathcal{O}\Big(\log d + d^2 + \log^2(TN)\Big) \,.$$

Then, letting k'' be the first episode such that $|\mathcal{T}_{k''}| \geq q_{k''}$ and noting that $k'' \leq k'$ gives

$$\sum_{k=1}^{k''-1} \texttt{Reg}(k\text{-th episode}) \leq \sum_{k=1}^{k'-1} \texttt{Reg}(k\text{-th episode}) \,.$$

Hence, the cumulative regret corresponding to case (i) is at most poly-logarithic in T.

For case (ii), it suffices to show random sampling ensures the growth of $\lambda_{\min}(V_t)$. We show that random sampling with duration q_k specified in Theorem 2.5 ensures the minimum eigenvalue condition for the Gram matrix, i.e., $\lambda_{\min}(V_{\tau_k}) \geq \max\left\{\frac{9\mathcal{D}_k^4}{\kappa^4 \log(\tau_k N/2)}, \frac{144\mathcal{D}_k^2}{\kappa^4}\right\}$ with high probability (see Lemma A.11) for each episode $k \in [L]$. We then apply the confidence bound in Theorem 2.3 to the k-th episode which requires samples in the $(k-1)$-th episode are independent and $\lambda_{\min}(V_{\tau_{k-1}})$ at the end of the $(k-1)$-th episode is large enough. That is, with a lower bound guarantee on $\lambda_{\min}(V_{\tau_{k-1}})$ and the fact that samples are independent of each other within each episode, we have with high probability

$$|x_{ti}^\top(\hat{\theta}_k - \theta^*)| \leq \beta_k \|x_{ti}\|_{W_{k-1}^{-1}}, \;\; \forall i \in [N], \forall t \in \mathcal{T}_k$$

with suitable confidence width β_k specified in Theorem 2.5. Therefore, the expected regret in the k-th episode can be bounded by $\widetilde{\mathcal{O}}(\sqrt{d\tau_k})$. Then we combine the results over all episodes to establish $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret.

2.7 Extensions to Position Dependent Offering

In many real-world applications, the choices of items are affected by not only their utilities but also the positions where they are displayed in the offered assortment (Ghose, Ipeirotis, and Li, 2014). For example, in an online recommendation platform or a web store, items displayed at the top of the user interface or the web page are more likely to be clicked or purchased than those displayed at the bottom. Similarly, in a brick-and-mortar store, items displayed in upper-shelf positions often receive more attention than those displayed in lower-shelf positions. The effect of the display positions is usually unknown a priori.

In our proposed framework, we can easily incorporate display position effect by including a categorical variable indicating the display position. Hence, we need to estimate parameters corresponding to each display position. Suppose there are K distinct display positions. Let z_{tik} denote the upper confidence utility for item i in round t in display position $k \in [K]$ and let $w_{tik} := \exp(z_{tik})$. Then the optimal assortment choice $S_t = \{(i,k) \in [N] \times [K] : \phi_{tik} = 1\}$ can be given by the solutions of the following optimization problem:

$$\begin{aligned} \max \quad & \sum_{i\in[N],k\in[K]} \frac{r_{ti} w_{tik}\phi_{tik}}{1+\sum_{ik} w_{tik}\phi_{tik}} \\ \text{s.t.} \quad & \sum_i \phi_{tik} \le 1 \qquad \forall k \in [N] \\ & \sum_k \phi_{tik} \le 1 \qquad \forall i \in [N] \\ & \phi_{tik} \in \{0,1\} \qquad \forall i \in [N], k \in [K] \end{aligned} \tag{2.14}$$

where ϕ_{tik} is the decision variable indicating item i is displayed at position k at round t. Note that the constraints satisfy that each position displays at most 1 item, and each item is displayed at most once.

Proposition 2 (Davis, Gallego, and Topaloglu 2013)**.** *The optimal position dependent assortment can be computed by solving a linear programming.*

Note that Lemma 2.4 and 2.5 continue to hold for this problem setting. Therefore, our proposed algorithms extend to the setting of position dependent offering with the same order of regret bounds.

Comparisons with previous methods on position-dependent offering. The non-contextual setting in (Agrawal et al., 2019; Agrawal et al., 2017) can be extended to incorporate position dependence; however, unlike in the setting here, the agent must offer every item in each position to learn the effect of display position. Therefore, the extension would create at least linearly increased amount of learning to their algorithms that are already not scalable for a large number of items, i.e., large N. On the other hand, our proposed methods are able to learn the position effect across items with a simple extension as discussed earlier. In Chen, Wang, and Zhou (2018), it is possible to include a categorical variable corresponding to display position as part of features; however, this will result in a further exponential increase in computational complexity. Moreover, their method cannot exploit that fact that the assortment optimization problem is a linear programming.

2.8 Numerical Experiments

In this section, we evaluate the performances of our proposed algorithms `UCB-MNL` (Algorithm 1), `supCB-MNL` (Alogirhtm 4), and `DBL-MNL` (Algorithm 5) in numerical experiments. In our evaluations, we report the cumulative regret for each round $t \in \{1, ..., T\}$. For each experimental configuration, we evaluate the algorithms on 20 independent instances and

report average performances. In each instance, the underlying parameter θ^* is sampled from the d-dimensional uniform distribution, with each element of θ^* uniformly distributed in $[0, 1]$. The underlying parameters are fixed during each problem instance but not known to the algorithms. For efficient evaluations, we consider uniform revenues, i.e., $r_{ti} = 1$ for all i and t. Therefore, the combinatorial optimization step to solve for the optimal assortment reduces to sorting items according to their utility estimate. Also, recall that the regret bound for `DBL-MNL` (Theorem 2.5) is derived under the uniform revenue assumption, therefore, the uniform revenue setting provides a suitable test bed for all methods considered in this chapter.

Comparison with the existing method. We compare our proposed algorithms with the existing UCB algorithm for the MNL contextual bandit algorithm `MLE-UCB` (Chen, Wang, and Zhou, 2018). Since `MLE-UCB` is an exponential-time algorithm that enumerates entire $\binom{N}{K}$ assortments, we conduct experiments with relatively small N and K for this comparison: $N \in \{20, 40\}$ and $K = 3$. For this experiment, we consider two multivariate distributions for feature vectors: a Gaussian distribution and a uniform distribution. For a multivariate Gaussian distribution, we draw each feature vector x_{ti} independently from $\mathcal{N}(\mathbf{0}_d, 0.5I_d)$. For a uniform distribution, we sample each feature vector x_{ti} uniformly at random from hypercube $[-1, 1]^d$.

In Figure 2.1, we report the cumulative regrets of the algorithms averaged over 20 runs. The error bars represent standard deviations. In the first row, the plots show the experiment results based on feature vectors drawn from the multivariate Gaussian distribution. The second row shows the results with feature vectors drawn from the uniform distribution. In terms of cumulative regret, the performances of `UCB-MNL` and `DBL-MNL` are comparable to the existing method, `MLE-UCB`. While `UCB-MNL` shows slightly superior performances compared to `MLE-UCB` and `DBL-MNL` with the uniformly distributed features, `DBL-MNL` shows superior performances in the setting of multivariate Gaussian

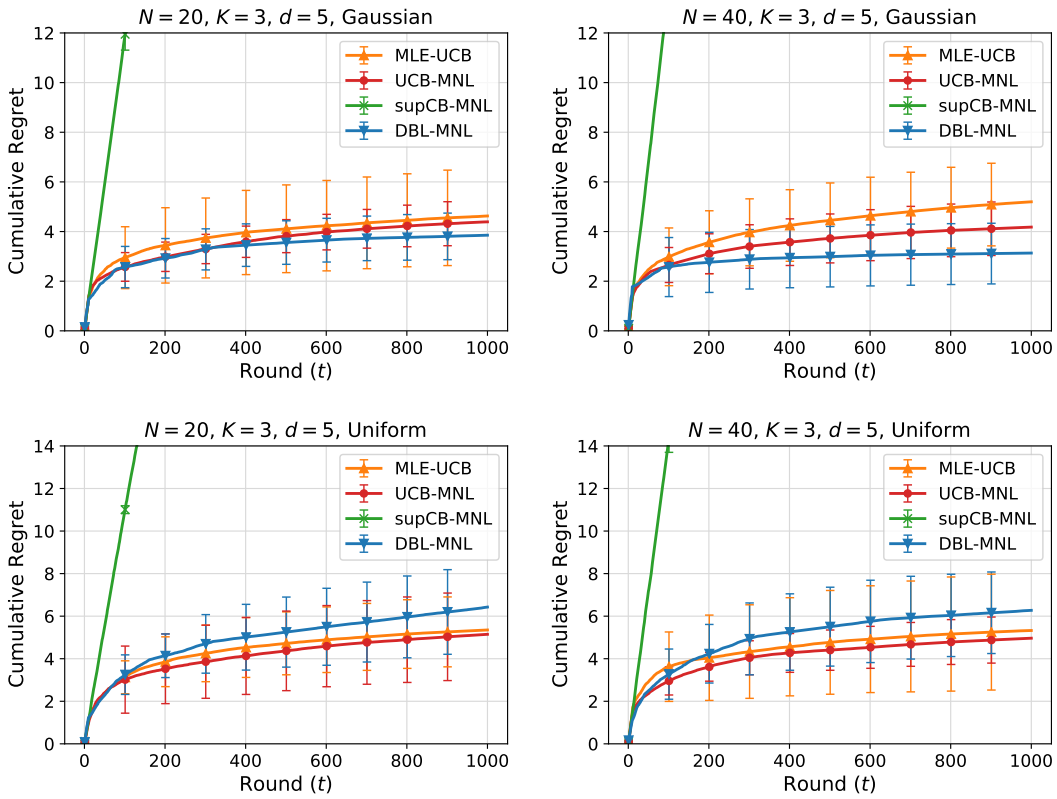

Figure 2.1: Evaluations of `UCB-MNL` (Algorithm 1), `supCB-MNL` (Algorithm 4), `DBL-MNL` (Algorithm 5), and `MLE-UCB` (Chen, Wang, and Zhou, 2018). Each plot shows the t-round cumulative regret as a function of t averaged over 20 runs. In the firs row, the features are drawn from a multivariate Gaussian distribution. In the second row, features are drawn from a uniform distribution in a hypercube.

features. However, the differences between the performances of the algorithms mostly appear to be within the standard deviations. As expected, `supCB-MNL` that relies on the framework of Auer (2002) is not competitive, wasting too many samples for random exploration. Therefore, `supCB-MNL` does not serve as a practical solution for this problem even though it is theoretically optimal.

We also conduct the run-time evaluations for the algorithms. We report the results in Table 2.1. We observe that `DBL-MNL` is significantly more efficient compared to the other methods computationally due to its logarithmic number of parameter updates. `UCB-MNL` also shows competitive performances in run-time. `MLE-UCB` suffers significantly in terms

	$N = 20$		$N = 40$	
Method	$t = 500$	$t = 1000$	$t = 500$	$t = 1000$
`MLE-UCB`	355.35	953.14	1932.04	4177.92
`UCB-MNL`	1.43	3.35	1.52	3.74
`supCB-MNL`	2.38	28.65	2.45	42.27
`DBL-MNL`	0.18	0.30	0.19	0.33

Table 2.1: Run-time evaluations (in seconds) with instances $N \in \{20, 40\}, K = 3, d = 5$. The reported run-times are averaged over 20 runs.

of computational cost even in these reasonably small-scale problem instances since it requires to compute upper confidence bounds for all combinatorially many assortments. Furthermore, as the number of items increases, we observe that the run-time of `MLE-UCB` grows exponentially, hence suggesting that the `MLE-UCB` algorithm is not suitable for large-scale problem instances. Overall, these experiments show that both `UCB-MNL` and `DBL-MNL` can learn to find the optimal policy efficiently in terms of both statistical and computational perspectives.

Impact of total number of items N. We have observed that our proposed algorithms, `UCB-MNL` and `DBL-MNL`, exhibit superior performances in the small-scale instances. We now examine whether their performance is consistent as we scale the problem setting. In particular, we examine the impact of the total number of items on the performance of the algorithms. We vary the value of $N \in \{800, 1600, 3200\}$ while keeping the other problem parameters fixed, $K = 5$ and $d = 5$. Figure 2.2 shows the performances of `UCB-MNL`, `DBL-MNL`, and `supCB-MNL`. The results suggest that `UCB-MNL` is most robust to the changes in the size of the item set N, showing almost no effect on the cumulative regret even for a very large N. `DBL-MNL`, on the other hand, is affected by the number of items, showing the deterioration in the cumulative regret performances, which appears to be consistent with our theoretical findings in the regret bound of `DBL-MNL` — recall that the regret bound of `DBL-MNL` in Theorem 2.5 has a logarithmic dependence on N.

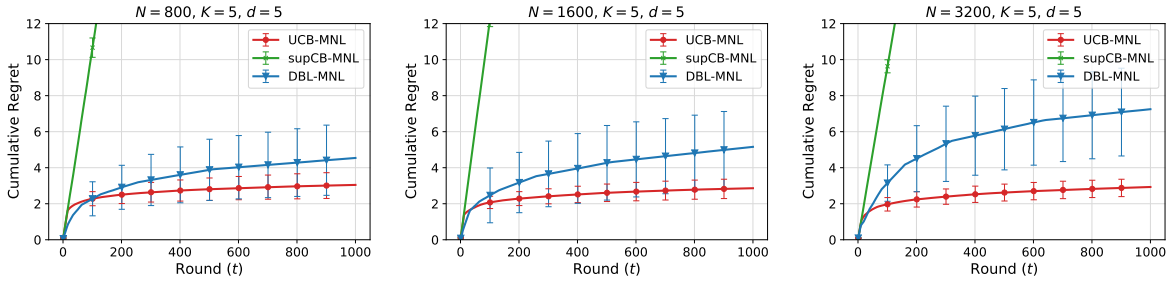

Figure 2.2: Evaluations of `UCB-MNL` (Algorithm 1), `supCB-MNL` (Algorithm 4), and `DBL-MNL` (Algorithm 5) in MNL contextual bandits. The plots show t-round regret as a function of t with varying $N \in \{800, 1600, 3200\}$.

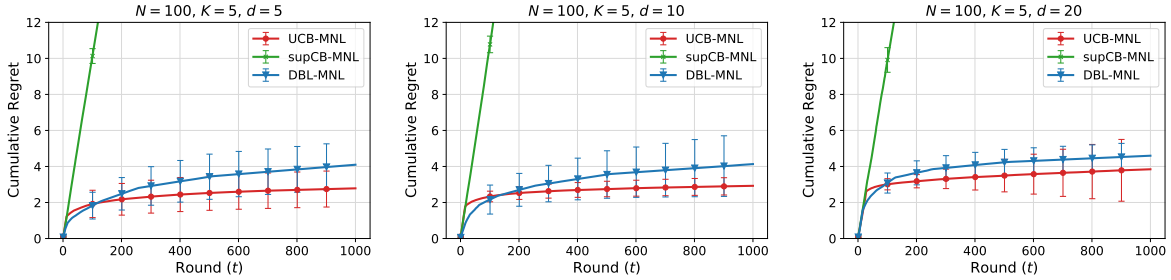

Figure 2.3: Evaluations of `UCB-MNL` (Algorithm 1), `supCB-MNL` (Algorithm 4), and `DBL-MNL` (Algorithm 5) in MNL contextual bandits. The plots show t-round regret as a function of t with varying $d \in \{5, 10, 20\}$.

Impact of feature dimension d. Now, we examine the impact of the feature dimension by varying $d \in \{5, 10, 20\}$ while keeping the other parameter fixed, $N = 100$ and $K = 5$. Figure 2.3 shows the cumulative regrets in these experiments. We again observe that `UCB-MNL` shows the best performance among the algorithms we compare. However, as the feature dimension increases, the performance of `DBL-MNL` stays almost the same showing robust performance with respect to changes in the feature dimension whereas the performance of `UCB-MNL` is slightly affected by an increase in the feature dimension. Nevertheless, we observe that `UCB-MNL` is still able to find the optimal policy in a few hundred rounds. A better scalability in d An interesting observation is that the regret of `UCB-MNL` appears to scale better than a linear dependence on d (as suggested by the regret bound of `UCB-MNL` in Theorem 1).

2.9 Concluding Remarks

In this chapter, we studied a sequential assortment selection problem, where the user choice is given by the MNL choice model whose parameter is unknown to the decision-making agent. This assortment selection problem is a fundamental sequential learning problem that embeds the core statistical and computational challenges of many real-world applications. Despite the wide applicability of the problem setting, the existing methods fall short in terms of tractability and scalability. We investigate practical UCB algorithms for this problem and establish near-optimal regret bounds. To establish a sharper regret bound, we present a non-asymptotic confidence bound for the maximum likelihood estimator of the MNL model that may be of independent interest as its own theoretical contribution. Our algorithms proposed in this chapter improve the state-of-the-art both in terms of statistical and computational efficiency, and are significantly morel practical than the existing methods.

Chapter 3

Thompson Sampling for MNL Contextual Bandits

In multi-armed bandit problems, UCB methods maintain a confidence set for the unknown parameter; and choose the most optimistic parameter from this set in each step, and pull the optimal arm corresponding to this optimistic parameter value. The confidence set is updated based on the reward feedback which is revealed after an arm is pulled. On the other hand, Thompson sampling (TS) assumes a prior distribution over the unknown parameter defining the reward distribution. At each step, a parameter value is sampled from the posterior distribution, and an optimal arm corresponding to a sampled parameter is pulled. Upon observing the reward for each round, the posterior distribution is updated via Bayes' rule. TS has been successfully applied in a wide range of settings (Strens, 2000; Chapelle and Li, 2011; Agrawal and Goyal, 2012; Russo et al., 2018).

While UCB algorithms are simple to implement and analyze, and come with good regret bounds (Li et al., 2010), TS has been shown to achieve better empirical performance in many simulated and real-world settings without sacrificing simplicity (Chapelle and Li, 2011; Kaufmann, Korda, and Munos, 2012). However, the analysis of TS is generally

considered more challenging than that of UCB algorithms. In order to bridge this gap, many recent studies have been focused on the analysis of worst-case regret and Bayesian regret in TS approaches for both contextual bandits and reinforcement learning settings (Agrawal and Goyal, 2013; Agrawal et al., 2017; Russo and Van Roy, 2014; Abeille, Lazaric, et al., 2017). The main technical difficulty in analyzing regret of a TS algorithm lies in controlling the deviation introduced by the randomness in the algorithm. This step is made more challenging by the combinatorial action selection of the MNL contextual bandit problem. In this chapter, we present TS algorithms for the MNL contextual bandit problem. To our knowledge, these are the first TS algorithms with regret guarantees for this problem. An overview of this chapter is as follows:

(a) In Section 3.3, we propose a TS algorithm, `TS-MNL` (Algorithm 6), that maintains a posterior distribution of the unknown parameter of the MNL model and establish $\widetilde{\mathcal{O}}(d\sqrt{T})$ Bayesian regret.

(b) In Section 3.4, we discuss challenges that arise in the worst-case regret analysis of TS based methods for the MNL contextual bandits. The challenges discussed here may also apply to other combinatorial bandits.

(c) In Section 3.5, we propose a modified algorithm, `TS-MNL` with "optimistic sampling" (Algorithm 7). This algorithm approximates the posterior by a Gaussian distribution and uses optimistic sampling procedure to address the issues that arise in worst-case regret analysis. We establish $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ worst-case (frequentist) regret bound for this algorithm.

The additional $\sqrt{d}$ factor in the worst-case regret of the second algorithm results from controlling the random sampling associated with TS, and is consistent with the results in TS methods for linear bandits (Agrawal and Goyal, 2013; Abeille, Lazaric, et al., 2017). Both regret bounds are independent of candidate item set size N, which implies that our

TS algorithms can be applied to a large item set. The TS algorithms that we propose in this chapter are efficient to implement as long as the assortment optimization step is solved efficiently, for which our TS algorithms can exploit efficient polynomial-time algorithms (Rusmevichientong and Tsitsiklis, 2010; Davis, Gallego, and Topaloglu, 2013). To our knowledge, the worst-case regret analysis in this chapter is the *first* result for a TS algorithm for any combinatorial variant of contextual bandits.

3.1 Related Work

In addition to the literature discussed in Section 2.1, we briefly discuss the literature on TS for contextual bandits. Agrawal and Goyal (2013) define TS for linear contextual bandit as a Bayesian algorithm where a Gaussian prior over the unknown parameter is updated according to the observed rewards, a random sample is drawn from the posterior, and the corresponding optimal arm is selected in each round. Agrawal and Goyal (2013) classify actions (or items in our case) as saturated and unsaturated depending on whether their standard deviation is smaller or bigger than their gap to the optimal action. While for unsaturated actions the regret is related to their standard deviation that decreases over time, they prove that TS has a small (but constant) probability to select saturated actions. They show $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ worst-case regret bound. Following the work of Agrawal and Goyal (2013), Abeille, Lazaric, et al. (2017) show that a TS algorithm does not need to sample from an actual Bayesian posterior distribution and that any distribution satisfying suitable concentration and anti-concentration properties guarantees a small regret and provide an alternative proof of TS achieving the same regret bound $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$. However, these results in (generalized) linear contextual bandits (either UCB or TS) do not apply directly to our MNL contextual bandit problem, since the choice probability of an item in an assortment is non-linear and non-monotone in the MNL parameter.

In the personalized MNL-bandit problem where independent parameters are assumed for each item, Cheung and Simchi-Levi (2017b) propose a TS approach. However, they only provide the Bayesian regret which is relatively easier to analyze compared to the worst-case regret (we discuss this aspect in Section 3.4), and their method (as well as other personalized MNL bandit methods) considers learning N separate parameters for each of the items; hence it is not scalable for a large item set (i.e., large N).

3.2 Worst-Case and Bayesian Regret

The problem setting in this chapter is identical to the setting in Chapter 2.[1] We refer the reader to Section 2.2 for the problem formulation of the MNL contextual bandits. As before, let S_t^* be the optimal assortment in round t under full information if θ^* is known:

$$S_t^* = \underset{S \in \mathcal{S}}{\operatorname{argmax}}\, R_t(S, \theta^*)\,.$$

The performance of an algorithm is measured by the regret, the gap between the expected revenue generated by the assortment S_t chosen by the algorithm and that of the optimal assortment S_t^* under the true parameter θ^*. We define the (worst-case) cumulative expected regret as

$$\mathcal{R}(T, \theta^*) = \sum_{t=1}^{T} \mathbb{E}\Big[R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*) \mid \theta^*\Big]$$

where $R_t(S_t^*, \theta^*)$ is the expected revenue corresponding to the optimal assortment in round t, and the expectation is taken over randomness in feature vectors and noise as well as possible randomization in a learning algorithm. When it is clear that we condition on a fixed θ^*, we denote $\mathcal{R}(T) := \mathcal{R}(T, \theta^*)$ in the rest of the chapter. In Bayesian settings, i.e., when θ^* is randomly generated or the learning agent has a prior belief in θ^*, the Bayesian

[1] For regret analysis, however, we use a slightly different set of assumptions. See Section 3.3.1.

cumulative regret (Russo and Van Roy, 2014) over T horizon is defined as

$$\mathcal{R}_{\text{Bayes}}(T) = \mathbb{E}_{\theta^*}\left[\mathcal{R}(T, \theta^*)\right] = \sum_{t=1}^{T} \mathbb{E}\Big[R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*)\Big]$$

where the expectation is taken also over the distribution of θ^*. In other words, $\mathcal{R}_{\text{Bayes}}(T)$ is a weighted average of $\mathcal{R}(T, \theta^*)$ under the prior on θ^*.

3.3 Algorithm: `TS-MNL`

In this section, we describe `TS-MNL`, our first TS algorithm for the MNL contextual bandit problem, and establish an upper bound on its Bayesian regret. `TS-MNL` follows the basic procedure of TS which maintains a posterior on the unknown parameter and updates the posterior when a new feedback is obtained.

We first provide the definition of the posterior distribution Q_t on the unknown parameter θ^*. At the beginning of the learning phase, the agent knows that θ^* is distributed according to Q_0, the prior distribution. Now, at each round t, the agent has access to the observations up to round t, $\mathcal{D}_t = \{X_\tau, y_\tau\}_{\tau=1}^{t-1}$ where $X_\tau = \{x_{\tau i}\}_{i \in S_\tau}$. Then the agent combines Q_0 and $\mathcal{D}_t$ to define the posterior distribution $Q_t(\theta)$:

$$Q_t(\theta) \propto Q_0(\theta) p(\mathcal{D}_t|\theta), \quad \text{where } p(\mathcal{D}_t|\theta) = \prod_{\tau=1}^{t-1} \prod_{i \in S_\tau \cup \{0\}} \left(p_{\tau i}(S_\tau, \theta)\right)^{y_{\tau i}} \tag{3.1}$$

and the "$\propto$" notation hides the partition function $\int_\phi Q_0(\phi) p(\mathcal{D}_t|\phi) d\phi$ in the denominator. In other words, the posterior distribution is proportional to the product of the prior distribution and the likelihood function. Note that there is no conjugate prior for the MNL model. Hence, sampling from Q_t is intractable. In order to overcome this intractability, one may draw an approximate sampling using Markov chain Monte Carlo (Andrieu et al., 2003). For ease of exposition, we assume that we can sample from posterior $Q_t(\theta)$ in

the Bayesian regret analysis. We will later provide a remedy for this intractability in the modification of our algorithm for the worst-case regret analysis.

Assumption 3.1. *We can sample from $Q_t(\theta)$.*

In each round t, `TS-MNL` algorithm takes three major steps. First, it randomly samples a parameter $\widetilde{\theta}_t$ from the posterior distribution Q_t. Second, it computes the assortment choice S_t under this sampled parameter $\widetilde{\theta}_t$. Finally, S_t is offered to the user and feedback y_t is observed. The pseudocode of `TS-MNL` is presented in Algorithm 6.

Algorithm 6 `TS-MNL`

1: **Input**: prior distribution Q_0
2: **for** all $t = 1$ to T **do**
3: Observe x_{ti} and r_{ti} for all $i \in [N]$
4: Sample $\widetilde{\theta}_t$ from the posterior distribution Q_t in (3.1)
5: Compute $S_t = \text{argmax}_{S \in \mathcal{S}} R_t(S, \widetilde{\theta}_t)$
6: Offer S_t and observe y_t (user choice at round t)
7: **end for**

Algorithm 6 has the combinatorial optimization step in Line 5. There are efficient polynomial-time algorithms available to solve this combinatorial optimization problem (Rusmevichientong, Shen, and Shmoys, 2010; Davis, Gallego, and Topaloglu, 2013) for given utility estimates under the sampled parameter. As in the case for the UCB algorithms in Chapter 2, we again assume an access to an optimization method which returns the assortment choice at time t, $S_t = \arg\max_{S \subset \mathcal{S}} R_t(S, \theta)$ for a given parameter θ.

3.3.1 Bayesian Regret of `TS-MNL`

In this section, we provide an upper bound on the Bayesian regret of `TS-MNL` under the following standard assumptions.

Assumption 3.2. $\|x_{ti}\| \leq 1$ *for all* t *and* i. *Also,* $\|\theta^*\| \leq 1$.

This assumption is used to make the regret bounds scale-free for convenience and is in fact standard in the bandit literature. If $\|x_{ti}\| \leq C$ and $\|\theta^*\| \leq C$ for some constant C instead, then our regret bounds would increase by a factor of C.

Assumption 3.3. *There exists* $\kappa > 0$ *such that, for every item* $i \in S$ *and any* $S \in \mathcal{S}$ *and all round* t, $\inf_{S\in\mathcal{S},\theta\in\mathbb{R}^d} p_{ti}(S,\theta)p_{t0}(S,\theta) \geq \kappa$.

This is a common assumption in other MNL contextual bandit literature Cheung and Simchi-Levi (2017b) and Chen, Wang, and Zhou (2018), and also is equivalent to a standard assumption in generalized linear contextual bandit literature (Filippi et al., 2010; Li, Lu, and Zhou, 2017) to ensure the Fisher information matrix is invertible and is adapted to suit our MNL setting.

Next, we state the Bayesian cumulative regret bound for Algorithm 6 in Theorem 3.1. We also provide a discussion and a proof outline for the regret bound.

Theorem 3.1 (Bayesian regret of `TS-MNL`)**.** *Suppose we run* ***`TS-MNL`*** *(Algorithm 6) for a total of* T *rounds with assortment size constraint* K. *Then the Bayesian regret of the algorithm is upper-bounded by*

$$\begin{aligned}\mathcal{R}_{Bayes}(T) &\leq \mathcal{O}(1) + \left[\frac{1}{\kappa}\sqrt{2d\log\left(1+\frac{TK}{d^2}\right)+2\log T}+\frac{\sqrt{d}}{\kappa}\right]\cdot\sqrt{2dT\log\left(1+\frac{TK}{d^2}\right)} \\ &= \mathcal{O}\left(d\sqrt{T}\log\left(1+\frac{TK}{d^2}\right)\right).\end{aligned}$$

Discussion of Theorem 3.1. Theorem 3.1 establishes $\widetilde{\mathcal{O}}(d\sqrt{T})$ Bayesian regret. Chen, Wang, and Zhou (2018) established the lower bound $\Omega(d\sqrt{T}/K)$ for the MNL contextual bandits under almost identical settings. When K is small and fixed (which is typically true in many applications), Theorem 3.1 demonstrates that `TS-MNL` is almost

optimal. Furthermore, the regret bound is completely free of N; hence `TS-MNL` is applicable to the case of a large number of items (large N). Also, if $K \leq d^2$, the regret bound becomes free of K. In Section 3.5, we introduce modifications to `TS-MNL` for the worst-case regret analysis which include the explicit use of regularized MLE for parameter estimation and sampling from the Gaussian distribution instead of maintaining the actual posterior to overcome the intractability. The concentration results derived for the Bayesian regret analysis in this section serve as a building block for the worst-case regret analysis for the modified algorithm.

The proof outline of Theorem 3.1 is motivated by Russo and Van Roy (2014) and Wen, Kveton, and Ashkan (2015). We let $\mathcal{F}_t$ denote the filtration that contains all observed information up to the beginning of the t-th round, prior to sampling a parameter in round t. Then, conditioning on the filtration $\mathcal{F}_t$, the sampled parameter $\widetilde{\theta}_t$ and the true parameter θ^* are i.i.d. with the posterior distribution Q_t in the Bayesian perspective. Also, the assortment selection step is a deterministic combinatorial optimization and the feature set $\{x_{ti}\}_{i\in[N]}$ are fixed given $\mathcal{F}_t$. Hence, conditioning on $\mathcal{F}_t$, S_t and S_t^* are also i.i.d. Therefore, there is no expected regret due to the random sampling in the Bayesian perspective. Thus, we only need to control the estimation error of θ^* for which we utilize the finite-sample concentration result for the MNL parameter.

3.3.2 Proof of Theorem 3.1: Bayesian Regret Analysis

Recall that $\widetilde{\theta}_t$ is independently drawn from the posterior distribution Q_t in Algorithm 6 and also note that in the Bayesian setting, conditioned on $\mathcal{F}_t$, the posterior distribution for θ^* is given by Q_t. Therefore, conditioned on $\mathcal{F}_t$, $\widetilde{\theta}_t$ and θ_t^* are i.i.d. samples from Q_t. Also note that our optimization oracle is a fixed combinatorial optimization algorithm and $\{x_{ti}\}_{i\in[N]}$ are fixed given $\mathcal{F}_t$. Hence, conditioning on $\mathcal{F}_t$, S_t and S^* are also i.i.d.

Confidence Bound for Expected Revenue

We define a upper confidence expected revenue as

$$U_t(S, \hat{\theta}_t) = \frac{\sum_{i \in S} r_{ti} \exp\left\{x_{ti}^\top \hat{\theta}_t + \alpha_t \|x_{ti}\|_{V_t^{-1}}\right\}}{1 + \sum_{j \in S} \exp\left\{x_{tj}^\top \hat{\theta}_t + \alpha_t \|x_{tj}\|_{V_t^{-1}}\right\}}$$

where $\alpha_t > 0$ is the confidence width and its value is specified later (Lemma 3.2). Also, we define $V_t = \sum_{\tau=1}^{t} \sum_{i \in S_\tau} x_{\tau i} x_{\tau i}^\top$. Note that this upper confidence expected revenue U_t is constructed for the sake of the analysis presented in this section and does not affect the proposed algorithm (or its assortment selection). We first decompose the immediate regret using U_t.

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(t) \mid \mathcal{F}_t] &= \mathbb{E}\Big[R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\Big] \\
&= \mathbb{E}\Big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t\Big] + \mathbb{E}\Big[U_t(S_t^*, \hat{\theta}_t) - U_t(S_t, \hat{\theta}_t) \mid \mathcal{F}_t\Big] \\
&\quad + \mathbb{E}\Big[U_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\Big].
\end{aligned}$$

Notice that $\mathbb{E}\Big[U_t(S_t^*, \hat{\theta}_t) - U_t(S_t, \hat{\theta}_t) \mid \mathcal{F}_t\Big] = 0$ since conditioning on $\mathcal{F}_t$, S_t and S^* are i.i.d. and U_t is a deterministic function. Hence, for the Bayesian cumulative regret, we are left bound the two quantities $\mathcal{R}^{(1)}_{\text{Bayes}}(T)$ and $\mathcal{R}^{(2)}_{\text{Bayes}}(T)$ as the following:

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{R}(t) \mid \mathcal{F}_t] = \underbrace{\sum_{t=1}^{T} \mathbb{E}\Big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t\Big]}_{\mathcal{R}^{(1)}_{\text{Bayes}}(T)} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\Big[U_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\Big]}_{\mathcal{R}^{(2)}_{\text{Bayes}}(T)}$$

In the following sections, we present the upper-bounds for $\mathcal{R}^{(1)}_{\text{Bayes}}(T)$ and $\mathcal{R}^{(2)}_{\text{Bayes}}(T)$. Then we combine the results to establish the Bayesian cumulative regret for `TS-MNL` (Algorithm 6).

Bounding $\mathcal{R}^{(1)}_{\textbf{Bayes}}(T)$

Before we present the upper bound for $\mathcal{R}^{(1)}_{\text{Bayes}}(T)$, we introduce the following lemma which utilizes the structure of the MNL model. Lemma 3.1 shows that the expected revenue R_t (and hence U_t) has a Lipschitz property, i.e., Lemma 3.1 ensures that we can control the difference between expected revenues by bounding with maximum difference in utilities.

Lemma 3.1. *For any two utility parameters $u_t = [u_{t1}, ..., u_{tN}]$ and $u'_t = [u'_{t1}, ..., u'_{tN}]$, we have*

$$\frac{\sum_{i\in S} r_{ti} \exp(u_{ti})}{1+\sum_{j\in S} \exp(u_{tj})} - \frac{\sum_{i\in S} r_{ti} \exp(u'_{ti})}{1+\sum_{j\in S} \exp(u'_{tj})} \leq \max_{i\in S} |u_{ti} - u'_{ti}| .$$

In particular, if $u_{ti} \geq u'_{ti}$ for all i, then

$$\frac{\sum_{i\in S} r_{ti} \exp(u_{ti})}{1+\sum_{j\in S} \exp(u_{tj})} - \frac{\sum_{i\in S} r_{ti} \exp(u'_{ti})}{1+\sum_{j\in S} \exp(u'_{tj})} \leq \max_{i\in S} (u_{ti} - u'_{ti}) .$$

Note that in the statement of Lemma 3.1 we use the explicit form of expected revenues (with generic utility parameters) in order to accommodate both R_t and U_t. Now, Lemma 3.2 below shows that the true parameter θ^* lies within an ellipsoid centered at $\hat{\theta}_t$ with confidence radius α_t. This is the result for the non-i.i.d. finite-sample confidence bound for the MNL parameter.

Lemma 3.2. *Define $\alpha_t := \frac{1}{2\kappa}\sqrt{d\log\left(1+\frac{tK}{d\lambda}\right)+4\log t} + \frac{\sqrt{\lambda}}{\kappa}$. If $\hat{\theta}_t$ is the solution to the regularized MLE in* (B.1) *at round t, then*

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \alpha_t$$

holds for all t with a probability $1-\mathcal{O}\left(\frac{1}{t^2}\right)$.

If θ^* is indeed within the confidence region for all t, i.e., if the high probability event of Lemma 3.2 holds, then one can show that $x_{ti}^\top \hat{\theta}_t + \alpha_t \|x_{ti}\|_{V_t^{-1}} \geq x_{ti}^\top \theta^*$ for all i. Hence,

$U_t(S_t^*, \hat{\theta}_t)$ is greater than $R_t(S_t^*, \theta^*)$. Then, $\mathcal{R}^{(1)}_{\text{Bayes}}(T)$ can be upper-bounded by 0. However, there is a small probability of failure for the confidence region which we need to take into consideration. The following lemma states the result formally.

Lemma 3.3. *Let the upper confidence expected revenue $U_t(S_t^*, \hat{\theta}_t)$ be defined with the confidence width $\alpha_t = \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{tK}{d\lambda}\right) + 4 \log t} + \frac{\sqrt{\lambda}}{\kappa}$. Then, we have*

$$\sum_{t=1}^{T} \mathbb{E}\Big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t\Big] = \mathcal{O}(1).$$

Bounding $\mathcal{R}^{(2)}_{\text{Bayes}}(T)$

This portion of the regret is controlled by the concentration of the upper confidence expected revenue $U_t(S_t, \hat{\theta}_t)$ to the true expected revenue $R_t(S_t, \theta^*)$. We can first use Lemma 3.1 to upper-bound $\mathcal{R}^{(2)}_{\text{Bayes}}(T)$ by the expected maximum difference in utilities. Now, suppose that θ^* resides within the confidence region with the radius α_t for all rounds t (Lemma 3.2). Then the same holds for the radius α_T since $\alpha_T \geq \alpha_t$. Using this fact and Cauchy-Schwartz inequality, we can further bound $\mathcal{R}^{(2)}_{\text{Bayes}}(T)$ by (3.2).

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\Big[U_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\Big] &\leq \sum_{t=1}^{T} \mathbb{E}\left[\max_{i \in S_t} \left(x_{ti}^\top \hat{\theta}_t + \alpha_t \|x_{ti}\|_{V_t^{-1}} - x_{ti}^\top \theta^*\right) \mid \mathcal{F}_t\right] \\ &\leq 2\alpha_T \sum_{t=1}^{T} \mathbb{E}\left[\max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] \qquad (3.2)\end{aligned}$$

Then, we are left to control the sum of the expectations in (3.2). Specifically, we provide a worst-case bound on $\sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}$ for any realization of random variables in Lemma 3.4, which presents a self-normalized bound.

Lemma 3.4. *Define $V_T = V + \sum_{t=1}^{T} \sum_{i \in S_t} x_{ti} x_{ti}^\top$ where $V = \lambda I_d$. Then we have*

$$\sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \leq \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)}.$$

Combining the results of Lemma 3.4 and (3.2), we have

$$\sum_{t=1}^{T} \mathbb{E}\Big[U_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \mid \mathcal{F}_t\Big] \leq 2\alpha_T \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \mathcal{O}(1)$$

where $\alpha_T = \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{TK}{d\lambda}\right) + 4 \log T} + \frac{\sqrt{\lambda}}{\kappa}$ and $\mathcal{O}(1)$ comes from the failure event of the concentration of $\hat{\theta}_t$ in Lemma 3.2.

Combining $\mathcal{R}^{(1)}_{\textbf{Bayes}}(T)$ and $\mathcal{R}^{(2)}_{\textbf{Bayes}}(T)$

Combining the bounds for $\mathcal{R}^{(1)}_{\text{Bayes}}(T)$ and $\mathcal{R}^{(2)}_{\text{Bayes}}(T)$, we have

$$\mathcal{R}_{\text{Bayes}}(T) \leq \mathcal{O}(1) + \left[\frac{1}{\kappa}\sqrt{2d \log\left(1 + \frac{TK}{d\lambda}\right) + 2 \log T} + \frac{\sqrt{\lambda}}{\kappa}\right] \cdot \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)}.$$

For completeness, we choose $\lambda = d$ to get the regret bound shown in Theorem 3.1 which gives the Bayesian regret $\mathcal{R}_{\text{Bayes}}(T) = \mathcal{O}\Big(d\sqrt{T} \log\left(1 + \frac{TK}{d^2}\right)\Big)$. Since Algorithm 6 itself does not use the regularized MLE for parameter estimation, one may optimize over the choice of λ in the regret bound.

3.4 Challenges in Worst-Case Regret Analysis

`TS-MNL` (Algorithm 6) is still valid under a frequentist setting, i.e., when the true parameter is not a random variable but a fixed parameter. However, when analyzing the worst-case regret (also known as frequentist regret) for the algorithm, the main technical difficulty lies in controlling the deviation in performance due to the random sampling of the algorithm. Note that in Bayesian regret analysis, the regret arising from the sampling is not addressed because $\widetilde{\theta}_t$ and θ^* are i.i.d. conditioned on $\mathcal{F}_t$. However, this does not hold anymore when θ^* is fixed; hence the worst-case regret analysis needs to ensure that the deviation due to sampling is small enough. To see this, we decompose the worst-case immediate regret

into a few components.

$$\begin{aligned}\mathcal{R}(t) &= \mathbb{E}[R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*)] \\ &= \mathbb{E}[R_t(S_t^*, \theta^*) - R_t(S_t^*, \widetilde{\theta}_t)] + \mathbb{E}[R_t(S_t^*, \widetilde{\theta}_t) - R_t(S_t, \widetilde{\theta}_t)] + \mathbb{E}[R_t(S_t, \widetilde{\theta}_t) - R_t(S_t, \theta^*)] \\ &\leq \mathbb{E}[R_t(S_t^*, \theta^*) - R_t(S_t^*, \widetilde{\theta}_t)] + \mathbb{E}[R_t(S_t, \widetilde{\theta}_t) - R_t(S_t, \theta^*)] \qquad (3.3)\end{aligned}$$

The inequality comes from the fact that our assortment choice at round t, S_t, is optimal under $\widetilde{\theta}_t$; hence $R_t(S_t^*, \widetilde{\theta}_t) \leq R_t(S_t, \widetilde{\theta}_t)$. The second term $\mathbb{E}[R_t(S_t, \widetilde{\theta}_t) - R_t(S_t, \theta^*)]$ in (3.3) is relatively easier to control. We can show that the term can be bounded by combining the upper-bound for the estimation error $|x^\top(\hat{\theta}_t - \theta^*)|$ and the concentration of the sampling probability of $\widetilde{\theta}_t$. However, controlling the first term $\mathbb{E}[R_t(S_t^*, \theta^*) - R_t(S_t^*, \widetilde{\theta}_t)]$ in (3.3) is more challenging in frequentist analysis. First, note that $\mathbb{E}[R_t(S_t^*, \theta^*) - R_t(S_t^*, \widetilde{\theta}_t)] = 0$ in the Bayesian setting since θ^* and $\hat{\theta}_t$ are i.i.d. conditioned on $\mathcal{F}_t$ as mentioned earlier. However, this is no longer true in the worst-case regret analysis. In the worst-case regret analysis of TS, this term is controlled by showing that a sampled parameter is optimistic frequently enough. In other words, we need to lower-bound the probability of the sampled parameter being optimistic, i.e., $\mathbb{P}\Big(R_t(S_t^*, \widetilde{\theta}_t) \geq R_t(S_t^*, \theta^*) \mid \mathcal{F}_t\Big) \geq p$ for some parameter free $p > 0$.

To describe the challenge in our MNL contextual bandit problem, we present the following lemma which shows that the expected revenue for the optimal assortment is monotonically increasing with an increase in the utility estimates.

Lemma 3.5 (Agrawal et al. (2019), Lemma 4.2)**.** *Suppose S_t^* is the optimal assortment under the true parameter θ^* at round t, i.e., $S_t^* = \arg\max_{S \in \mathcal{S}} R_t(S, \theta^*)$. Also suppose that $x_{ti}^\top \theta^* \leq x_{ti}^\top \theta'$ for all $i \in S_t^*$. Then $R_t(S_t^*, \theta^*) \leq R_t(S_t^*, \theta')$.*

Note that Lemma 3.5 shows the monotonicity of expected revenue only for the optimal assortment and it does not claim that the expected revenue is generally a monotone

function for all assortments. This lemma implies that we can lower-bound the probability of having an optimistic expected revenue under the sampled parameter.

$$\mathbb{P}\left(R_t(S_t^*, \widetilde{\theta}_t) \geq R_t(S_t^*, \theta^*) \mid \mathcal{F}_t\right) \geq \mathbb{P}\left(x_{ti}^\top \widetilde{\theta}_t \geq x_{ti}^\top \theta^*, \forall i \in S_t^* \mid \mathcal{F}_t\right)$$

However, this makes the probability of being optimistic exponentially small in the size of the assortment S_t^*, i.e., exponentially small in $\mathcal{O}(K)$, which in turn results in exponential dependence on $\mathcal{O}(K)$ in the worst-case regret bound. In order to overcome such an issue, we adopt a few modifications in the algorithm which we discuss in the following section.

3.5 `TS-MNL` with Optimistic Sampling

Motivated by the challenges in the worst-case analysis of `TS-MNL` discussed in Section 3.4, we present a variant of `TS-MNL`, which we call `TS-MNL` with "optimistic sampling." The main modifications in this variant of the algorithm are the posterior approximation by a Gaussian distribution and optimistic sampling by drawing multiple samples.

Sampling from Gaussian Distribution. We modify our TS algorithm to a generic randomized algorithm constructed on the regularized MLE rather than sampling from an actual Bayesian posterior. Abeille, Lazaric, et al. (2017) show that TS does not need to sample from an actual posterior distribution and that any distribution satisfying suitable concentration and anti-concentration properties guarantees a small regret. Specifically, instead of sampling from the posterior Q_t, we sample $\widetilde{\theta}_t$ from Gaussian distribution $\mathcal{N}\left(\hat{\theta}_t, \alpha_t^2 V_t^{-1}\right)$ where $\hat{\theta}_t$ is the regularized MLE, i.e., the solution of (3.4), and α_t is the confidence radius. This way, we ensure tractability of the sampling distribution. Furthermore, this Gaussian approximation allows us to adopt optimistic sampling (which we discuss below) in an efficient manner.

Optimistic Sampling. The optimistic sampling we present here is a key ingredient in avoiding the theoretical challenges present in the worst-case regret analysis. For optimistic sampling, instead of drawing a single sample $\widetilde{\theta}_t$, we draw M independent samples $\{\widetilde{\theta}_t^{(j)}\}_{j=1}^M$ from $\mathcal{N}\Big(\hat{\theta}_t, \alpha_t^2 V_t^{-1}\Big)$ (the exact value of M is specified in Theorem 3.2). Then we compute the optimistic utility estimate $\widetilde{u}_{ti}$ for each $i \in [N]$:

$$\widetilde{u}_{ti} = \max_j x_{ti}^\top \widetilde{\theta}_t^{(j)}.$$

We define $\widetilde{R}_t(S)$ to be the expected revenue of assortment S based on $\widetilde{u}_{ti}$:

$$\widetilde{R}_t(S) = \frac{\sum_{i\in S} r_{ti} \exp\{\widetilde{u}_{ti}\}}{1 + \sum_{j\in S} \exp\{\widetilde{u}_{tj}\}}$$

Note that this optimistic sampling scheme is different from that proposed in Agrawal et al. (2017). The setting in Agrawal et al. (2017) is non-contextual, and they use a 1-dimensional Gaussian random variable to correlate the samples of the utility of the K items in order to ensure the probability that all samples are simultaneously optimistic is a constant. This correlated sampling reduces the overall variance severely, hence they propose taking K samples instead of a single sample to increase the variance. In contrast, we take multiple samples of the multivariate Gaussian distribution to directly ensure that the probability of an optimistic sample is sufficiently large.

The pseudocode of the modified algorithm is presented in Algorithm 7. The modified algorithm now explicitly maintains the matrix V_t and computes the regularized MLE $\hat{\theta}_t$. Note that α_T can be replaced by $\alpha_t = \mathcal{O}\Big(\sqrt{d\log\left(1+\frac{tK}{d\lambda}\right) + 4\log t}\Big)$ at round t, if the horizon T is not known and the analysis holds for either case.

Algorithm 7 `TS-MNL` with Optimistic Sampling

Input: sample size M, confidence radius α_T, penalty parameter λ

for all $t = 1$ to T **do**

Observe x_{ti} and r_{ti} for all $i \in [N]$

Sample $\{\widetilde{\theta}_t^{(j)}\}_{j=1}^M$ independently from $\mathcal{N}(\hat{\theta}_t, \alpha_T^2 V_t^{-1})$

Compute $\widetilde{u}_{ti} = \max_j x_{ti}^\top \widetilde{\theta}_t^{(j)}$ for all $i \in [N]$

Compute $S_t = \text{argmax}_{S \in \mathcal{S}} \widetilde{R}_t(S)$

Offer S_t and observe y_t (user choice at round t)

Update $V_{t+1} \leftarrow V_t + \sum_{i \in S_t} x_{ti} x_{ti}^\top$

Compute the regularized MLE $\hat{\theta}_t$ by solving

$$\sum_{t=1}^{n} \sum_{i \in S_t} (p_{ti}(S_t, \theta) - y_{ti}) x_{ti} + \lambda \theta = \mathbf{0} \tag{3.4}$$

end for

3.5.1 Worst-Case Regret of `TS-MNL` with Optimistic Sampling

Theorem 3.2 (Regret of `TS-MNL` with optimistic sampling)**.** *Suppose we run `TS-MNL` with optimistic sampling (Algorithm 7) for a total of T rounds with the optimistic sample size $M = \lceil 1 - \frac{\log K}{\log\left(1 - 1/(4\sqrt{e\pi})\right)} \rceil$, the penalty parameter $\lambda \geq 1$ and assortment size constraint K. Then the worst-case regret of the algorithm is upper-bounded by*

$$\mathcal{R}(T) \leq \mathcal{O}(1) + 16\sqrt{e\pi}\beta_T \left(\sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda} \log 2T} \right) + (\alpha_T + \beta_T) \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)}$$

where $\alpha_T = \frac{1}{2\kappa} \sqrt{d \log\left(1 + \frac{TK}{d\lambda}\right) + 4 \log T} + \frac{\sqrt{\lambda}}{\kappa}$ and $\beta_T = \alpha_T \sqrt{2d \log(MT)}$.

Theorem 3.2 establishes $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ worst-case regret, which matches the regret bounds of TS methods for linear contextual bandits Agrawal and Goyal (2013) and Abeille,

Lazaric, et al. (2017) up to logarithmic factor. The regret bound shows no dependence on N, and has an additional $\mathcal{O}(\sqrt{\log \log K})$ dependence due to optimistic sampling which is very small for any reasonable assortment size K. Compared to Theorem 3.1, the additional factor $\sqrt{d}$ comes from the deviation of the random sampling which is addressed in the worst-case regret analysis.

The proof of Theorem 3.2 utilizes the anti-concentration property of the maximum of Gaussian random variables for ensuring frequent optimism. In particular, we show in the following lemma that the proposed optimistic sampling can ensure a constant probability of optimism.

Lemma 3.6 (Optimism)**.** *Suppose* $\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \frac{1}{2\kappa}\sqrt{d \log\left(1 + \frac{tK}{d\lambda}\right) + 4 \log t} + \frac{\sqrt{\lambda}}{\kappa}$ *and we take optimistic samples of size* $M = \lceil 1 - \frac{\log K}{\log\left(1 - 1/(4\sqrt{e\pi})\right)} \rceil$*. Then we have*

$$\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta^*) \mid \mathcal{F}_t\right) \geq \frac{1}{4\sqrt{e\pi}}\,.$$

The inverse of the lower-bounding probability $4\sqrt{e\pi}$ can be interpreted as the expected time between any two optimistic assortment selections. In other words, our modified algorithm is optimistic at least with a constant frequency even in the worst case. Then, using this frequent optimism, we can ensure that the cumulative regret due to the random sampling can be bounded. Along with this result, we show the concentrations of both regularized MLE and TS samples to establish the regret bound in Theorem 3.2. The proofs are left to Appendix 3.5.2.

3.5.2 Proof of Theorem 3.2: Worst-case Regret Analysis

We first decompose the cumulative regret, similar to the procedure in previous sections but this time using $\widetilde{R}_t(S_t)$. In the following sections, we derive the bounds for $\mathcal{R}^{(1)}(T)$

and $\mathcal{R}^{(2)}(T)$ separately.

$$\mathcal{R}(T) = \underbrace{\sum_{t=1}^{T} \mathbb{E}[R_t(S_t^*, \theta^*) - \widetilde{R}_t(S_t)]}_{\mathcal{R}^{(1)}(T)} + \underbrace{\sum_{t=1}^{T} \mathbb{E}[\widetilde{R}_t(S_t) - R_t(S_t, \theta^*)]}_{\mathcal{R}^{(2)}(T)}$$

Bounding $\mathcal{R}^{(2)}(T)$.

We can control $\mathcal{R}^{(2)}(T)$ by showing that both MLE $\hat{\theta}_t$ and TS parameters $\{\widetilde{\theta}_t\}$ concentrate appropriately. To show each of these concentration results, we first further decompose $\mathcal{R}^{(2)}(T)$:

$$\mathcal{R}^{(2)}(T) = \sum_{t=1}^{T} \mathbb{E}[\widetilde{R}_t(S_t) - R_t(S_t, \hat{\theta}_t)] + \sum_{t=1}^{T} \mathbb{E}[R_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*)]\,. \tag{3.5}$$

The second term deals with the estimation error and can be bounded by the concentration of $\hat{\theta}_t$ in Lemma 3.2 and the Lipschitz-like property in Lemma 3.1, i.e., with probability $1 - \mathcal{O}(t^{-2})$, we have

$$R_t(S_t, \hat{\theta}_t) - R_t(S_t, \theta^*) \le \max_{i \in S_t} \left| x_{ti}^\top (\hat{\theta}_t - \theta^*) \right| \le \alpha_t \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}\,. \tag{3.6}$$

The first term in (3.5) deals with the random sampling of $\{\widetilde{\theta}_t^{(j)}\}$. Again, we can bound the difference in expected revenue by the difference in utility estimates using Lemma 3.1: $\widetilde{R}_t(S_t) - R_t(S_t, \hat{\theta}_t) \le \max_{i \in S_t}(\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t)$. Then we are left to show that $\widetilde{u}_{ti}$ concentrates appropriately for all $i \in [N]$. The following lemma ensures the concentration of $\widetilde{u}_{ti}$.

Lemma 3.7. *Let* $\beta_t = \alpha_t \min\left(\sqrt{4d\log(Mt)}, \sqrt{2\log(2M)} + \sqrt{4\log(Nt)}\right)$. *Then for all* $i \in [N]$,

$$\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t \le \beta_t \|x_{ti}\|_{V_t^{-1}}$$

with probability $1 - \mathcal{O}\left(\frac{1}{t^2}\right)$.

Remark 3.1. *Lemma 3.7 shows that the confidence radius β_t is larger than α_t by the factor of at most $\sqrt{2d\log(Mt)}$. The additional $\sqrt{d}$ factor comes from the oversampling of TS, which also appears in other TS methods for linear contextual bandit problems (Agrawal and Goyal, 2013; Abeille, Lazaric, et al., 2017). $\sqrt{\log M}$ factor comes from drawing optimistic samples where $M = \mathcal{O}(\log K)$; hence the marginal increase of the regret bound due to optimistic sampling is very small.*

Hence for the first term in (3.5), we have $\widetilde{R}_t(S_t) - R_t(S_t, \hat{\theta}_t) \leq \beta_t \max_{i\in S_t} \|x_{ti}\|_{V_t^{-1}}$ with probability $1 - \mathcal{O}\left(\frac{1}{t^2}\right)$. We combine with (3.6) to derive the bound for $\mathcal{R}^{(2)}(T)$:

$$\mathcal{R}^{(2)}(T) \leq \sum_{t=1}^{T} (\alpha_t + \beta_t) \max_{i\in S_t} \|x_{ti}\|_{V_t^{-1}} + \sum_{t=1}^{T} \mathcal{O}\left(t^{-2}\right) \tag{3.7}$$

Bounding $\mathcal{R}^{(1)}(T)$.

As discussed in Section 3.4, a sufficient condition for ensuring the success of TS is to show the probability of TS samples being optimistic is high enough. Lemma 3.6 lower-bounds the probability that the expected revenue under sampled parameters is higher than the optimal expected revenue under the true parameter, which states that if we have sufficiently many samples with $M = \lceil 1 - \frac{\log K}{\log\left(1-1/(4\sqrt{e\pi})\right)} \rceil$, we have

$$\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t\right) \geq \frac{1}{4\sqrt{e\pi}}\,.$$

Using this frequent optimistic sampling, we can ensure that the regret due to the oversampling is not too large.

Lemma 3.8. *Let $\widetilde{p} = \frac{1}{4\sqrt{e\pi}}$. Then, we have*

$$\sum_{t=1}^{T} \mathbb{E}[R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)] \leq \frac{4\beta_T}{\widetilde{p}} \left(\sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda}\log 2T}\right) + \mathcal{O}(1)\,.$$

Combining the results

Applying Lemma 3.4 to the bound for $\mathcal{R}^{(2)}(T)$ in (3.7) and combining with Lemma 3.8, we have the final bound for the worst-case cumulative regret.

$$\mathcal{R}(T) \leq \mathcal{O}(1) + +16\sqrt{e\pi}\beta_T \left(\sqrt{2dT\log\left(1+\frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda}\log 2T}\right) + (\alpha_T + \beta_T)\sqrt{2dT\log\left(\frac{T}{d}\right)}.$$

3.6 Numerical Study

In this section, we perform numerical evaluations to analyze two variants of our proposed algorithm: `TS-MNL` with optimistic sampling (Algorithm 7) and `TS-MNL` with the Gaussian approximation for the posterior distribution. We perform synthetic experiments similar to the experiments in Chapter 2. We simulate instances of the MNL contextual bandit problem with varying parameter values. For each experiment, we report the (frequentist) cumulative regret for the algorithms, i.e., we compute a regret with respect to a fixed parameter θ^*. For each instance, we randomly draw θ^* from a multi-dimensional uniform distribution, where each component of θ^* is drawn uniformly at random in $[0, 1]$. Note that the parameter θ^* stays fixed for the entire horizon $t \in [T]$ in a given instance. For each experimental configuration, we evaluate the algorithms on 20 independent instances and report average performances. For each plot, the error bars represent standard deviations.

Impact of total number of items N. We first investigate the influence of the total number of items N on the performances of the algorithms. We vary $N \in \{100, 400, 1600\}$ while keeping the other problem parameters fixed, $K = 5$ and $d = 5$. In these experiments, we use feature vectors drawn from a multivariate Gaussian distribution. For each round $t \in \{1, ..., 1000\}$, we draw each x_{ti}, $i \in [N]$ independently from $\mathcal{N}(\mathbf{0}_d, 0.5I_d)$.

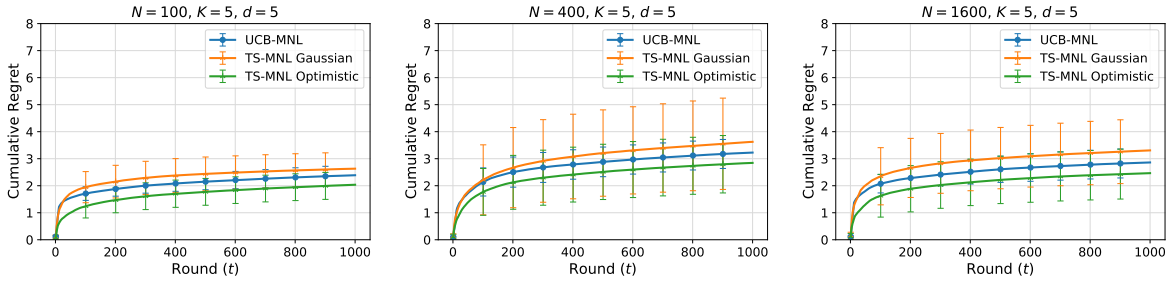

Figure 3.1: Evaluations of `TS-MNL` with optimistic sampling (Algorithm 7), `TS-MNL` with the Gaussian approximation only and `UCB-MNL` (Algorithm 1) in MNL contextual bandits. The plots show t-round regret as a function of t with varying $N \in \{100, 400, 1600\}$.

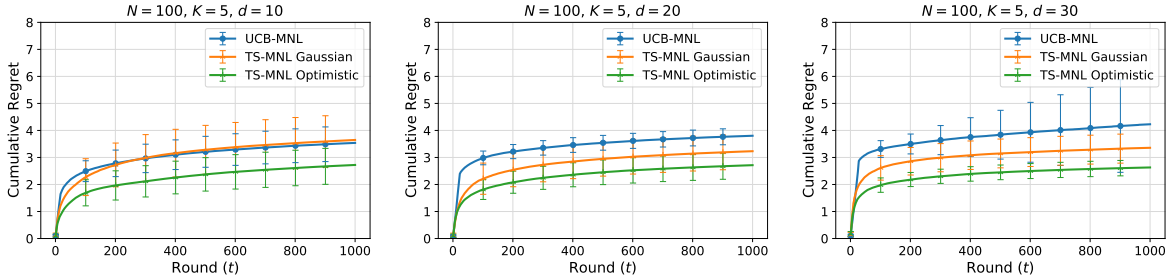

Figure 3.2: Evaluations of `TS-MNL` with optimistic sampling (Algorithm 7), `TS-MNL` with the Gaussian approximation only and `UCB-MNL` (Algorithm 1) in MNL contextual bandits. The plots show t-round regret as a function of t with varying $d \in \{10, 20, 30\}$.

For comparison, we evaluate the performances of our `TS-MNL` algorithms along with the performances of an efficient UCB method proposed in Chapter 2, `UCB-MNL` (Algorithm 1). Figure 3.1 shows that the three algorithms appear to scale well with an increase in the total number of items N. The performance of `TS-MNL` with optimistic sampling appears to be state-of-the-art, showing a slightly superior performance compared to the performance of `UCB-MNL`. Furthermore, `TS-MNL` with optimistic sampling consistently performs better than `TS-MNL` with the Gaussian approximation only. The results of these experiments support our theoretical analysis: `TS-MNL` with optimistic sampling takes advantage of the MNL structure and can guarantee a worst-case statistical efficiency. This is indeed consistent with our finding that single sample from a Gaussian distribution approximation does not provide optimism guarantees.

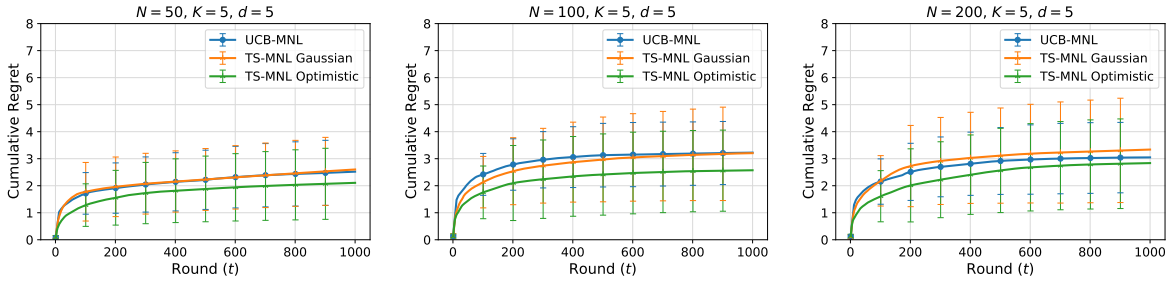

Figure 3.3: Evaluations of `TS-MNL` with optimistic sampling (Algorithm 7), `TS-MNL` with the Gaussian approximation only and `UCB-MNL` (Algorithm 1) in MNL contextual bandits. The plots show t-round regret with fixed feature vectors.

Impact of feature dimension d. We then evaluate the performances of the algorithms to test the impact of the feature dimension d on regret performances. Figure 3.2 reports the results averaged over 20 independent instances. Again, the performance of `TS-MNL` with optimistic sampling appears to be superior to the performances of the other methods while all of the three algorithms show favorable scalability in the feature dimension. An interesting observation is that the TS methods scale well with increases in the feature dimension even compared to the UCB method despite the fact that they have worse regret dependence (e.g., $\mathcal{O}(d^{3/2})$ worst-case dependence in the regret of `TS-MNL` with optimistic sampling). The numerical performances do not appear to suggest such dependence. These results suggest that the proposed `TS-MNL` algorithms are practical solutions even in potentially high-dimensional problem settings.

Experiments with fixed features. So far, we have evaluated the algorithms with features drawn randomly in each round. In this part of the experiments, we evaluate the performances of the algorithms with fixed features. That is, the features for each item stay fixed throughout the entire horizon. Figure 3.3 shows that both TS algorithms, `TS-MNL` with optimistic sampling and `TS-MNL` with the Gaussian approximation, and `UCB-MNL` perform well even with the fixed features. `TS-MNL` with optimistic sampling again outperforms the other methods in this set of experiments.

3.7 Concluding Remarks

In this chapter, we propose two TS algorithms for the MNL contextual bandits which learn the parameters of the underlying choice model while simultaneously maximizing the cumulative revenue. We provide their theoretical performance bounds and show attractive numerical performances in our experiments. We also discuss the challenges which arise in worst-case regret analysis for this combinatorial action selection problem under the MNL model. We believe that these challenges are potentially present in many other problems involving combinatorial action selections with feature information beyond the MNL model. To our knowledge, the worst-case regret analysis for our TS algorithm is the first frequentist regret guarantee of a TS method for contextual bandits with combinatorial action selection of any kind. We believe that our proposed optimistic sampling framework can be useful for other combinatorial contextual bandit problems.

Chapter 4

Sparsity-Agnostic High-Dimensional Bandit Algorithm

In classical multi-armed bandits, one of the arms is pulled in each round and a reward corresponding to the chosen arm is revealed to the decision-making agent. The rewards are, typically, independent and identically distributed samples from an arm-specific distribution. The goal of the agent is to devise a strategy for pulling arms that maximizes cumulative rewards, suitably balancing between exploration and exploitation. Linear contextual bandits (Abe and Long, 1999; Auer, 2002; Chu et al., 2011) and generalized linear contextual bandits (Filippi et al., 2010; Li, Lu, and Zhou, 2017) are more recent important extensions of the basic multi-armed bandit setting, where each arm a is associated with a known feature vector $x_a \in \mathbb{R}^d$, and the expected payoff of the arm is a (typically, monotone increasing) function of the inner product $x_a^\top \beta^*$ for a fixed and unknown parameter vector $\beta^* \in \mathbb{R}^d$. Unlike the traditional multi-armed bandit problem, pulling any arm provides some information about the unknown parameter vector, and hence, insight into the average reward of the other arms. These contextual bandit algorithms are applicable in a variety of problem settings, such as recommender systems, online retail, and healthcare

analytics (Li et al., 2010; Tewari and Murphy, 2017), where the contextual information can be used for personalization and generalization.

In most application domains highlighted above, the feature space is high-dimensional ($d \gg 1$), yet typically only a small subset of the features influence the expected reward. That is, the unknown parameter vector is *sparse* with only elements corresponding to the relevant features being non-zero, i.e., the *sparsity index* $s_0 = \|\beta^*\|_0 \ll d$, where the zero norm $\|x\|_0$ counts non-zero entries in the vector x. There is an emerging body of literature on contextual bandit problems with sparse linear reward functions (Abbasi-Yadkori, Pal, and Szepesvari, 2012; Gilton and Willett, 2017; Bastani and Bayati, 2020; Wang, Wei, and Yao, 2018; Kim and Paik, 2019) which propose methods to exploit the sparse structure under various conditions. However, there is a crucial shortcoming in almost all of these approaches: the algorithms require *prior* knowledge of the sparsity index s_0, information that is almost never available in practice. In the absence of such knowledge, the existing algorithms fail to fully leverage the sparse structure, and their performance does not guarantee the improvements in dimensionality-dependence which can be realized in the sparse problem setting (and can lead to extremely poor performance if s_0 is underspecified). The purpose of this work is to demonstrate that a relatively simple contextual bandit algorithm that exploits ℓ_1-regularized regression using Lasso (Tibshirani, 1996) in a sparsity-agnostic manner, is provably near-optimal insofar as its regret performance (under suitable regularity). Our contributions are as follows:

(a) We propose the first general[1] sparse bandit algorithm that does not require prior knowledge of the sparsity index s_0.

(b) We establish that the regret bound of our proposed algorithm is $\mathcal{O}(s_0\sqrt{T\log(dT)})$ for the two-armed case, which affords the most accessible exposition of the key

[1]Carpentier and Munos (2012) do not require to know sparsity, but both their algorithm and analysis are limited to the fixed ℓ_2 unit ball arm set. See more discussions in Section 4.1.

analytical ideas. (Extensions to the general K-armed case are discussed later.) The regret bound scale in s_0 and d matches the equivalent terms in the *offline* Lasso results (see the discussions in Section 4.4.2).

(c) We comprehensively evaluate our algorithm on numerical experiments and show that it consistently outperforms existing methods, even when these methods are granted prior knowledge of the correct sparsity index (and can greatly outperform them if this information is misspecified).

The salient feature of our algorithm is that it does not rely on *forced sampling* which was used by almost all previous work, e.g., Bastani and Bayati (2020), Wang, Wei, and Yao (2018), and Kim and Paik (2019), to satisfy certain regularity of the empirical Gram matrix. Forced sampling requires prior knowledge of s_0 because such schemes, the key ideas of which go back to Goldenshluger and Zeevi (2013), need to be fine-tuned using the *correct* sparsity index. (See further discussions in Section 4.1.2.)

The rest of the chapter is organized as follows. In Section 4.1, we review the related literature and discuss the reason why the previously proposed methods require the knowledge of the sparsity index s_0. In Section 4.2, we present the problem formulation. Section 4.3 describes our proposed algorithm. In Section 4.4, we describe the challenges when the sparsity information is unknown, and establish an upper bound on the cumulative regret for the two-armed sparse bandits. Section 4.5 contains the numerical experiments for the two-armed sparse bandits. In Section 4.6, we extend our analysis and numerical evaluations to the K-armed sparse bandits. Section 4.7 presents discussions and future directions. The complete proofs and additional numerical results are provided in the appendix.

4.1 Related Work

4.1.1 Review

Linear bandits and generalized linear bandits have been widely studied (Abe and Long, 1999; Auer, 2002; Dani, Hayes, and Kakade, 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori, Pál, and Szepesvári, 2011; Filippi et al., 2010; Chu et al., 2011; Agrawal and Goyal, 2013; Li, Lu, and Zhou, 2017; Kveton et al., 2020). However, when ported to the high-dimensional contextual bandit setting, these strategies have difficulty exploiting sparse structure in the unknown parameter vector, and hence may incur regret proportional to the full ambient dimension d rather than the sparse set of features of cardinality s_0. To exploit spare structure, Abbasi-Yadkori, Pal, and Szepesvari (2012) propose a framework to construct high probability confidence sets for online linear prediction and establish a regret bound of $\widetilde{\mathcal{O}}(\sqrt{s_0 dT})$, where $\widetilde{\mathcal{O}}$ hides logarithmic terms, when the sparsity index s_0 is *known*. Furthermore, their algorithm is not computationally efficient; an implementable version of their framework is not yet known (Section 23.5 in Lattimore and Szepesvári 2019). It is worth noting that the $\sqrt{d}$ dependence in the regret bound is unavoidable unless additional assumptions are imposed; see Theorem 24.3 in Lattimore and Szepesvári (2019). Gilton and Willett (2017) adapt Thompson sampling (Thompson, 1933) to sparse linear bandits; however, they also assume a priori knowledge of a small superset of the support for the parameter.

Bastani and Bayati (2020) address the contextual bandit problem with high-dimensional features using Lasso (Tibshirani, 1996) to estimate the parameter of each arm separately. To ensure compatibility of the empirical Gram matrices, they adapt the forced-sampling technique in Goldenshluger and Zeevi (2013) which is now tuned using the (a priori known) sparsity index, and is implemented for each arm at predefined time points. They

establish a regret bound of $\mathcal{O}(Ks_0^2[\log d + \log T]^2)$ where K is the number of arms. Note that they invoke several additional assumptions introduced in Goldenshluger and Zeevi (2013), including a margin condition that ensures that the density of the context distribution is bounded near the decision boundary, and arm-optimality which assumes a gap between the optimal and sub-optimal arms exists with some positive probability. In the same problem setting, Wang, Wei, and Yao (2018) propose an algorithm which uses forced-sampling along with the minimax concave penalty (MCP) estimator (Zhang, 2010) and improve the regret bound to $\mathcal{O}(Ks_0^2[s_0 + \log d] \log T)$. Note that Bastani and Bayati (2020) and Wang, Wei, and Yao (2018) achieve a poly-logarithmic dependence on T in the regret, exploiting the arm optimality condition which assumes a gap between the optimal and sub-optimal arms exists with some probability.[2] Since we do not assume such "separability" between arms, poly-logarithmic dependence on T is not attainable in our problem setting. Kim and Paik (2019) extend the method proposed in Bastani and Bayati (2020) to linear bandit settings and propose a different approach to address the non-compatibility of the empirical Gram matrices by using a doubly-robust technique (Bang and Robins, 2005) that originates with the missing data (imputation) literature. They achieve $\mathcal{O}(s_0\sqrt{T}\log(dT))$ regret.

All of the aforementioned algorithms require that the learning agent know the sparsity index s_0 of the unknown parameter (or a non-trivial upper-bound on sparsity which is strictly less than d).[3] That is, only when the algorithm knows s_0, it can guarantee the regret bounds mentioned above. Otherwise, the regret bounds would scale polynomially with d instead of s_0 or potentially scale linearly with T. To the best of our knowledge, the only work in sparse bandits which does not require this prior knowledge of the sparsity

[2]The regret bounds in both Bastani and Bayati (2020) and Wang, Wei, and Yao (2018) have additional dependence $\mathcal{O}(1/p_*^3)$ where p_* is the arm optimality lower bounding probability. Hence, in the worse case, the regret bounds have additional $\mathcal{O}(K^3)$ dependence.

[3]Besides sparsity, some algorithms require further knowledge, such as arm optimality lower bounding probability (Bastani and Bayati, 2020; Wang, Wei, and Yao, 2018), which is also not readily available in practice.

index is the work by Carpentier and Munos (2012) although their algorithm still requires to know the ℓ_2-norm of the unknown parameter. However, their analysis uses a non-standard definition of noise and is restricted to the case where the set of arms is the ℓ_2 unit ball and fixed over time, a structure they exploit in a significant manner, and which limits the scope of their algorithm.

4.1.2 Why do existing sparse bandit algorithms require prior knowledge of the sparsity index?

The primary reason that a priori knowledge of sparsity index s_0 is assumed throughout most of the literature is, roughly speaking, to ensure suitable "size" of the confidence bounds and concentration. For example, Abbasi-Yadkori, Pal, and Szepesvari (2012) require the parameter s_0 to explicitly construct a high probability confidence set with its radius proportional to s_0 rather than d. The recently proposed bandit algorithms of Bastani and Bayati (2020) and Kim and Paik (2019) and the variant with MCP estimator in Wang, Wei, and Yao (2018) employ a logic that is similar in spirit (though different in execution). Specifically, the compatibility condition is assumed to hold only for the theoretical Gram matrix, and the empirical Gram matrix may not satisfy such condition (the difficulty in controlling that is due to the non-i.i.d. adapted samples of the feature variables). As a remedy to this issue, Bastani and Bayati (2020) and Wang, Wei, and Yao (2018) utilize the forced-sampling technique of Goldenshluger and Zeevi (2013) to obtain a "sufficient" number of i.i.d. samples and use them to show that the empirical Gram matrices concentrate in the vicinity of the theoretical Gram matrix, and hence, satisfy the compatibility condition after a sufficient amount of forced-sampling. The forced-sampling duration needs to be predefined and scales at least polynomially in the sparsity index s_0 to ensure concentration of the Gram matrices. That is, if the algorithm does not know s_0,

the forced-sampling duration will have to scale polynomially in d. Kim and Paik (2019) propose an alternative to forced sampling that builds on doubly-robust techniques used in the missing data literature; however, their algorithm involves random arm selection with a probability that is calibrated using s_0, and initial uniform sampling whose duration requires knowledge of s_0 and scales polynomially with s_0 in order to establish their regret bounds. The sensitivity to the sparsity index specification is also evident in cases where its value is *misspecified*, which may result in severe deterioration in the performance of the algorithms (see further discussions in Section 5.1).

The key observation in our analysis is that i.i.d. samples, which are the key output of the forced samplings scheme, are, in fact, *not* required under some mild regularity conditions. We show that the empirical Gram matrix satisfies the compatibility condition after a sufficient number of rounds, provided the theoretical Gram matrix also satisfies the condition; the details of this analysis are in Section 4.4. Numerical experiments support these findings, and moreover, demonstrate that the performance of our proposed algorithm can be superior to forced-sampling-based schemes that are tuned with foreknowledge of the sparsity index s_0.

4.2 Preliminaries

4.2.1 Notation

For a vector $x \in \mathbb{R}^d$, we use $\|x\|_1$ and $\|x\|_2$ to denote its ℓ_1-norm and ℓ_2 norm respectively, the notation $\|x\|_0$ is reserved for the cardinality of the set of non-zero entries of that vector. The minimum and maximum singular values of a matrix V are written as $\lambda_{\min}(V)$ and $\lambda_{\max}(V)$ respectively. For two symmetric matrices V and W of the same dimensions, $V \succcurlyeq W$ means that $V - W$ is positive semi-definite. For a positive integer n, we define a

set of integers up to n as $[n] = \{1, ..., n\}$. For a real-valued differentiable function f, we use $\dot{f}$ to denote its first derivative.

4.2.2 Generalized Linear Contextual Bandits

We consider the stochastic generalized linear bandit problem with K arms. Let T be the problem horizon, namely the number of rounds to be played. In each round $t \in [T]$, the learning agent observes a context consisting of a set of K feature vectors $\mathcal{X}_t = \left\{X_{t,i} \in \mathbb{R}^d \mid i \in [K]\right\}$, where the tuple $\mathcal{X}_t$ is drawn i.i.d. over $t \in [T]$ from an unknown joint distribution with probability density $p_{\mathcal{X}}$ with respect to the Lebesgue measure. Note that the feature vectors for different arms are allowed to be correlated. Each feature vector $X_{t,i}$ is associated with an unknown stochastic reward $Y_{t,i} \in \mathbb{R}$. The agent then selects one arm, denoted by $a_t \in [K]$ and observes the reward $Y_t := Y_{t,a_t}$ corresponding to the chosen arm's feature $X_t := X_{t,a_t}$ as a bandit feedback. The policy consists of the sequence of actions $\pi = \{a_t : t = 1, 2, ...\}$ and is non-anticipating, namely each action only depends on past observations and actions.

In this work, we consider the generalized linear model (GLM) in which there is an unknown parameter $\beta^* \in \mathbb{R}^d$ and a fixed increasing function $\mu : \mathbb{R} \to \mathbb{R}$ (also known as *inverse link function*) such that the reward $Y_{t,i}$ of arm i is

$$Y_{t,i} = \mu(X_{t,i}^\top \beta^*) + \epsilon_{t,i}$$

where each $\epsilon_{t,i}$ is an independent zero-mean noise. Therefore, $\mathbb{E}[Y_{t,i}|X_{t,i} = x] = \mu(x^\top \beta^*)$ for all $i \in [K]$ and $t \in [T]$. Widely used examples for μ are $s\mu(z) = z$ which corresponds to the linear model, and $\mu(z) = 1/(1 + e^{-z})$ which corresponds to the logistic model. The parameter β^* and the feature vectors $\{x_{t,i}\}$ are potentially high-dimensional, i.e., $d \gg 1$, but β^* is *sparse*, that is, the number of non-zero elements in β^*, $s_0 = \|\beta^*\|_0 \ll d$. It is

important to note that the agent *does not* know s_0 or the support of β^*.

We assume that there is an increasing sequence of sigma fields $\{\mathcal{F}_t\}$ such that each $\epsilon_{t,i}$ is $\mathcal{F}_t$-measurable with $\mathbb{E}[\epsilon_{t,i}|\mathcal{F}_{t-1}] = 0$. In our problem, $\mathcal{F}_t$ is the sigma-field generated by random variables of chosen actions $\{a_1, ..., a_t\}$, their features $\{X_1, ..., X_t\}$, and the corresponding rewards $\{Y_1, ..., Y_t\}$. We assume the noise ϵ_t is sub-Gaussian with parameter σ, where σ is a positive absolute constant, i.e., $\mathbb{E}[e^{\alpha\epsilon_t}] \leq e^{\alpha^2\sigma^2/2}$ for all $\alpha \in \mathbb{R}$. In practice, for bounded reward $Y_{t,i}$, the noise $\epsilon_{t,i}$ is also bounded and hence satisfies the sub-Gaussian assumption with an appropriate σ value.

The agent's goal is to maximize the cumulative expected reward $\mathbb{E}[\sum_{t=1}^{T} \mu(X_{t,a_t}^\top \beta^*)]$ over T rounds. Let $a_t^* = \text{argmax}_{i\in[K]}\, \mu(X_{t,i}^\top \beta^*)$ denote the optimal arm for each round t. Then, the expected cumulative *regret* of policy $\pi = \{a_1, ..., a_T\}$ is defined as

$$\mathcal{R}^\pi(T) := \sum_{t=1}^{T} \mathbb{E}\left[\mu(X_{t,a_t^*}^\top \beta^*) - \mu(X_{t,a_t}^\top \beta^*)\right] .$$

Hence, maximizing the expected cumulative rewards of policy π over T rounds is equivalent to minimizing the cumulative regret $\mathcal{R}^\pi(T)$. Note that all the expectations and probabilities throughout the chapter are with respect to feature vectors and noise unless explicitly stated otherwise.

4.2.3 Lasso for Generalized Linear Models

Consider an offline setting where we have samples $Y_1, ..., Y_n$ and corresponding features $X_1, ..., X_n$. The log-likelihood function of β under the canonical GLM is

$$\log \mathcal{L}_n(\beta) := \sum_{j=1}^{n} \left[\frac{Y_j X_j^\top \beta - m(X_j^\top \beta)}{g(\eta)} - h(Y_j, \eta)\right] .$$

Here, $\eta \in \mathbb{R}^+$ is a known scale parameter, $m(\cdot)$, $g(\cdot)$ and $h(\cdot)$ are normalization functions, and $m(\cdot)$ is infinitely differentiable with the first derivative

$$\dot{m}(x^\top \beta^*) = \mathbb{E}[Y|X = x] = \mu(x^\top \beta^*)\,.$$

The Lasso (Tibshirani, 1996) estimate for the GLM can be defined as

$$\hat{\beta}_n \in \operatorname*{argmin}_{\beta} \left\{ \ell_n(\beta) + \lambda \|\beta\|_1 \right\} \tag{4.1}$$

where $\ell_n(\beta) := -\frac{1}{n}\sum_{j=1}^{n} \left[Y_j X_j^\top \beta - m(X_j^\top \beta) \right]$ and λ is a penalty parameter. Lasso is known to be an efficient (offline) tool for estimating the high-dimensional linear regression parameter. The "fast convergence" property of Lasso is guaranteed when the above data are i.i.d. and when the observed covariates are not "highly correlated." The restricted eigenvalue condition (Bickel, Ritov, Tsybakov, et al., 2009; Raskutti, Wainwright, and Yu, 2010), the compatibility condition (Van De Geer, Bühlmann, et al., 2009), and the restricted isometry property (Candes, Tao, et al., 2007) have all been used to ensure that such high correlations are avoided. In sequential learning settings, however, these conditions are often violated because the observations are adapted to the past, and the feature variables of the chosen arms converge to a small region of the feature space as the learning agent updates its arm selection policy.

4.3 Proposed Algorithm

Our proposed `Sparsity-Agnostic (SA) Lasso Bandit` algorithm for high-dimensional GLM bandits is summarized in Algorithm 8. As the name suggests, our algorithm does not require prior knowledge of the sparsity index s_0. It relies on Lasso for parameter estimation, and does not explicitly use exploration strategies or forced-sampling. Instead,

in each round, we choose an arm which maximizes the inner product of a feature vector and the Lasso estimate. After observing the reward, we update the regularization parameter λ_t and update the Lasso estimate $\hat{\beta}_t$ which minimizes the penalized negative log-likelihood function defined in (4.1).

`SA Lasso Bandit` requires only one input parameter λ_0. We show in Section 4.4 that $\lambda_0 = 2\sigma x_{\max}$ where $x_{\max}$ is a bound the ℓ_2-norm of the feature vectors $X_{t,i}$. Thus, λ_0 does *not* depend on the sparsity index s_0 or the underlying parameter β^*. (Note that, in comparison, Kim and Paik (2019) require three tuning parameters, and Bastani and Bayati (2020) and Wang, Wei, and Yao (2018) require four tuning parameters, most of which are functions of the unknown sparsity index s_0.) It is worth noting that tuning parameters, while helping achieve low regret, are challenging to specify in online learning settings. Therefore, our proposed algorithm is practical and easy to implement.

Algorithm 8 `SA Lasso Bandit`

Input parameter: λ_0
for all $t = 1$ to T **do**
  Observe $X_{t,i}$ for all $i \in [K]$
  Compute $a_t = \text{argmax}_{i\in[K]} X_{t,i}^\top \hat{\beta}_t$
  Pull arm a_t and observe Y_t
  Update $\lambda_t \leftarrow \lambda_0 \sqrt{\frac{4\log t + 2\log d}{t}}$
  Update $\hat{\beta}_{t+1} \leftarrow \text{argmin}_\beta \{\ell_t(\beta) + \lambda_t \|\beta\|_1\}$
end for

Discussion of the algorithm. Algorithm 8 may appear to be an *exploration-free* greedy algorithm (e.g., Bastani, Bayati, and Khosravi 2017), but this is not the case. To better see this, recall that upper-confidence bound (UCB) algorithms construct a high-probability confidence ellipsoid around a *greedy* estimate and choose the parameter value that maximizes the reward. Once the UCB estimate is chosen, the action selection is greedy with

respect to the parameter estimate.[4] The UCB algorithms carefully control the size of the confidence ellipsoid to ensure convergence, thus, exploration is loosely equivalent to regularizing the parameter estimate. The algorithm we propose also computes the parameter estimate by *regularizing* the MLE with a sparsifying norm, and then, as in UCB, takes a greedy action with respect to this regularized parameter estimate. We adjust the penalty associated with the sparsifying norm over time at a suitable rate *in order to* ensure that our estimate is consistent as we collect more samples. (This adjustment and specification do not require knowledge of sparsity s_0.) An inadequate choice of this penalty parameter would lead to large regret, which is analogous to poor choice of confidence widths in UCB.

4.4 Regret Analysis

4.4.1 Regularity Condition

In this section, we establish an upper bound on the expected regret of `SA Lasso Bandit` for the two-armed ($K = 2$) generalized linear bandits. We focus on the two-arm case primarily for clarity and accessibility of key analysis ideas, and later illustrate how this analysis extends to the K-armed case with $K \geq 3$ under suitable regularity (see Section 4.6). We first provide a few definitions and assumptions used throughout the analysis, starting with assumptions standard in the (generalized) linear bandit literature.

Assumption 4.1 (Feature set and parameter)**.** *There exists a positive constant $x_{\max}$ such that $\|x\|_2 \leq x_{\max}$ for all $x \in \mathcal{X}_t$ and all t, and a positive constant b such that $\|\beta^*\|_2 \leq b$.*

Assumption 4.2 (Link function)**.** *There exist $\kappa_0 > 0$ and $\kappa_1 < \infty$ such that the derivative $\dot{\mu}(\cdot)$ of the link function satisfies $\kappa_0 \leq \dot{\mu}(x^\top \beta) \leq \kappa_1$ for all x and β.*

[4] Likewise, in Thompson sampling (Thompson, 1933), the agent chooses the greedy action for the sampled parameter.

Clearly for the linear link function, $\kappa_0 = \kappa_1 = 1$. For the logistic link function, we have $\kappa_1 = 1/4$.

Definition 4.1 (Active set and sparsity index)**.** *The active set* $S_0 := \{j : \beta_j^* \neq 0\}$ *is the set of indices* j *for which* β_j^* *is non-zero, and the sparsity index* $s_0 = |S_0|$ *denotes the cardinality of the active set* S_0*.*

For the active set S_0, and an arbitrary vector $\beta \in \mathbb{R}^d$, we can define

$$\beta_{j,S_0} := \beta_j \mathbb{1}\{j \in S_0\}\,, \quad \beta_{j,S_0^c} := \beta_j \mathbb{1}\{j \notin S_0\}\,.$$

Thus, $\beta_{S_0} = [\beta_{1,S_0}, ..., \beta_{d,S_0}]^\top$ has zero elements outside the set S_0 and the components of $\beta_{S_0^c}$ can only be non-zero in the complement of S_0. Let $\mathbb{C}(S_0)$ denote the set of vectors

$$\mathbb{C}(S_0) := \{\beta \in \mathbb{R}^d \mid \|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1\}\,. \tag{4.2}$$

Let $\mathbf{X} \in \mathbb{R}^{K\times d}$ denote the design matrix where each row is a feature vector for an arm. (Although we focus on $K = 2$ case in this section, the definitions and the assumptions introduced here also apply to the case of $K \geq 3$.) Then, in keeping with the previous literature on sparse estimation and specifically on sparse bandits (Bastani and Bayati, 2020; Wang, Wei, and Yao, 2018; Kim and Paik, 2019), we assume that the following compatibility condition is satisfied for the theoretical Gram matrix $\Sigma := \frac{1}{K}\mathbb{E}[\mathbf{X}^\top\mathbf{X}]$.

Assumption 4.3 (Compatibility condition)**.** *For active set* S_0*, there exists compatibility constant* $\phi_0^2 > 0$ *such that*

$$\phi_0^2\|\beta_{S_0}\|_1^2 \leq s_0\beta^\top\Sigma\beta \quad \textit{for all } \beta \in \mathbb{C}(S_0)\,.$$

We add to this the following mild assumption that is more specific to our analysis.

Assumption 4.4 (Relaxed symmetry)**.** *For a joint distribution $p_{\mathcal{X}}$, there exists $\nu < \infty$ such that $\frac{p_{\mathcal{X}}(-\mathbf{x})}{p_{\mathcal{X}}(\mathbf{x})} \leq \nu$ for all $\mathbf{x}$.*

Discussion of the assumptions. Assumptions 4.1 and 4.2 are the standard regularity assumptions used in the GLM bandit literature (Filippi et al., 2010; Li, Lu, and Zhou, 2017; Kveton et al., 2020). It is important to note that unlike the existing GLM bandit algorithms which explicitly use the value of κ_0, our proposed algorithm does not use κ_0 or κ_1 — this information is only needed to establish the regret bound. The compatibility condition in Assumption 4.3 is analogous to the standard positive-definite assumption on the Gram matrix for the ordinary least squares estimator for linear models but is less restrictive. The compatibility condition ensures that truly active components of the parameter vector are not "too correlated." As mentioned above, the compatibility condition is a standard assumption in the sparse bandit literature (Bastani and Bayati, 2020; Wang, Wei, and Yao, 2018; Kim and Paik, 2019). Assumption 4.4 states that the joint distribution $p_{\mathcal{X}}$ can be skewed but this skewness is bounded. Obviously, if $p_{\mathcal{X}}$ is symmetrical, we have $\nu = 1$. Assumption 4.4 is satisfied for a large class of continuous and discrete distributions, e.g., elliptical distributions including Gaussian and truncated Gaussian distributions, multi-dimensional uniform distribution, and Rademacher distribution.

4.4.2 Regret Bound for `SA Lasso Bandit`

Theorem 4.1 (Regret bound for two arms)**.** *Suppose $K = 2$ and Assumptions 4.1-4.4 hold. Let $\lambda_0 = 2\sigma x_{\max}$. Then the expected cumulative regret of the* ***`SA Lasso Bandit`*** *policy π over horizon $T \geq 1$ is upper-bounded by*

$$\mathcal{R}^{\pi}(T) \leq 4\kappa_1 + \frac{4\kappa_1 x_{\max} b(\log(2d^2) + 1)}{C_0(s_0)^2} + \frac{32\kappa_1 \nu \sigma x_{\max} s_0 \sqrt{T \log(dT)}}{\kappa_0 \phi_0^2}$$

where $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 \nu x_{\max}^2}\right)$.

Discussion of Theorem 4.1. In terms of key problem primitives, Theorem 4.1 establishes $\mathcal{O}(s_0\sqrt{T\log(dT)})$ regret without any prior knowledge on s_0. The bound shows that the regret of our algorithm grows at most logarithmically in feature dimension d. The key takeaway from this theorem is that `SA Lasso Bandit` is sparsity-agnostic and is able to achieve "correct" dependence on parameters d and s_0. That is, based on the offline Lasso convergence results under the compatibility condition (e.g., Theorem 6.1 in Bühlmann and Van De Geer 2011), we believe that the dependence on d and s_0 in Theorem 4.1 is best possible.[5]

The regret bound in Theorem 4.1 is tighter than the previously known bound in the same problem setting (Kim and Paik, 2019) although direct comparison is not immediate, given the difference in assumptions involved — compared to Kim and Paik (2019), we require Assumption 4.4 whereas they assume the sparsity index s_0 is known. Having said that, the numerical experiments in Section 4.5 support our theoretical claims and provide additional evidence that our proposed algorithm compares very favorably to other existing methods (which are tuned with the knowledge of the correct s_0), and moreover, the performance is not sensitive to the assumptions that were imposed primarily for technical tractability purposes. Note that the input parameter $\lambda_0 = 2\sigma x_{\max}$ depends on σ and $x_{\max}$ which are parameters required by all parametric bandit methods, and hence our algorithm does not require any additional information.

As mentioned earlier, the previous work on sparse bandits (Bastani and Bayati, 2020; Wang, Wei, and Yao, 2018; Kim and Paik, 2019) require the knowledge of the sparsity index s_0. In the absence of such knowledge, if sparsity is underspecified, then these algorithms would suffer a regret linear in T. On the other hand, if the sparsity is overspecified,

[5]Since the horizon T does not exist in offline Lasso results, it is not straightforward to see whether $\sqrt{T}$ dependence can be improved comparing only with the offline Lasso results. Clearly, without an additional assumption on the separability of the arms, we know that poly-logarithmic scalability in T is not feasible. We briefly discuss our conjecture in comparison with the lower bound result in the non-sparse linear bandits in Section 4.4.4 where we discuss the regret bound under the RE condition.

the regret of these algorithms may scale with d instead of s_0. Our proposed algorithm does not require such prior knowledge, hence there is no risk of under-specification or over-specification, and yet our analysis provides a sharper regret guarantee. Furthermore, our result also suggests that even when the sparsity is known, random sampling to satisfy the compatibility condition, invoked by all existing sparse bandit algorithms to date, can be wasteful since said conditions may be already satisfied even in the absence of such sampling. This finding is also supported by the numerical experiments in Section 4.5 and Section 4.6.2. We provide the outline of the proof and the key lemmas in the following section.

4.4.3 Challenges and Proof Outlines

There are two essential challenges that prevent us from fully benefiting from the fast convergence property of Lasso:

(i) The samples induced by our bandit policy are not i.i.d., therefore the standard Lasso oracle inequality does not hold.

(ii) Empirical Gram matrices do not necessarily satisfy the compatibility condition even under Assumption 4.3. This is because the selected feature variables for which the rewards are observed do not provide an "even" representation for the entire distribution.

To resolve (i), we provide a Lasso oracle inequality for the GLM with non-i.i.d. adapted samples under the compatibility condition in Lemma 4.1. For (ii), we aim to provide a remedy without using the knowledge of sparsity or without using i.i.d. samples. Hence, this poses a greater challenge. In Section 4.4.3, we address this issue by showing that the empirical Gram matrix behaves "nicely" even when we choose arms adaptively without deliberate random sampling. In particular, we show that adapted Gram matrices can be

controlled by the theoretical Gram matrix and the empirical Gram matrix concentrates properly around the adapted Gram matrix as we collect more samples. Connecting this matrix concentration to the corresponding compatibility constants, we show that the empirical Gram matrix satisfies the compatibility condition with high probability.

Lasso Oracle Inequality for GLM with Non-i.i.d. Data.

We present an oracle inequality for the Lasso estimator for the GLM with non-i.i.d. data. This is a generalization of the standard Lasso oracle inequality (Bühlmann and Van De Geer, 2011; Geer et al., 2008) that allows adapted sequences of observations. This is also a generalization of Proposition 1 in Bastani and Bayati (2020) to the GLM. This convergence result may be of independent interest.

Lemma 4.1 (Oracle inequality)**.** *Let* $\{X_\tau : \tau \in [t]\}$ *be an adapted sequence such that each* X_τ *may depend on* $\{X_s : s < \tau\}$*. Suppose the compatibility condition holds for the empirical covariance matrix* $\hat{\Sigma}_t = \frac{1}{t}\sum_{\tau=1}^{t} X_\tau X_\tau^\top$ *with active set* S_0 *and compatibility constant* ϕ_t*. For* $\delta \in (0,1)$*, define the regularization parameter*

$$\lambda_t := 2\sigma x_{\max}\sqrt{\frac{2[\log(2/\delta) + \log d]}{t}}\,.$$

Then with probability at least $1-\delta$*, the Lasso estimate* $\hat{\beta}_t$ *defined in* (4.1) *satisfies*

$$\|\hat{\beta}_t - \beta^*\|_1 \le \frac{4s_0\lambda_t}{\kappa_0\phi_t^2}\,.$$

Note that here we assume that the compatibility condition holds for the empirical Gram matrix $\hat{\Sigma}_t$. In the next section, we show that this holds with high probability. The Lasso oracle inequality holds without further assumptions on the underlying parameter β^* or its support. Therefore, if we show that $\hat{\Sigma}_t$ satisfies the compatibility condition without

the knowledge of s_0, then the remainder of the result does not require this knowledge.

Compatibility Condition and Matrix Concentration.

We first define the generic compatibility constant for matrix M with respect to S_0.

Definition 4.2. *The compatibility constant of M over S_0 is*

$$\phi^2(M, S_0) := \min_{\beta} \left\{ \frac{s_0 \beta^\top M \beta}{\|\beta_{S_0}\|_1^2} : \|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1 \neq 0 \right\}.$$

Hence, it suffices to show $\phi^2(M, S_0) > 0$ in order to show that matrix M satisfies the compatibility condition. Although one can define a compatibility constant with respect to any index set, in this section, we will focus on the active index set S_0 of the parameter β^*. Also, note that the constant 3 in the inequality is for ease of exposition and may be replaced by a different value, but then one has to adjust the choice of the regularization parameter accordingly. Now, under Assumption 4.3, the theoretical Gram matrix $\Sigma = \frac{1}{K}\mathbb{E}[\mathbf{X}^\top \mathbf{X}]$ satisfies the compatibility condition i.e., $\phi_0^2 = \phi^2(\Sigma, S_0) > 0$.

Definition 4.3. *We define the adapted Gram matrix as $\Sigma_t := \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}[X_\tau X_\tau^\top | \mathcal{F}_{\tau-1}]$ and the empirical Gram matrix as $\hat{\Sigma}_t := \sum_{\tau=1}^{t} X_\tau X_\tau^\top$.*

For each term $\mathbb{E}[X_\tau X_\tau^\top | \mathcal{F}_{\tau-1}]$ in Σ_t, the past observations $\mathcal{F}_{\tau-1}$ affects how the feature vector X_τ is chosen. More specifically, our algorithm uses $\mathcal{F}_{\tau-1}$ to compute $\hat{\beta}_\tau$ and then chooses arm a_τ such that its feature x_{a_τ} maximizes $x_{a_\tau}^\top \hat{\beta}_\tau$. Hence, we can rewrite Σ_t as

$$\Sigma_t = \frac{1}{t}\sum_{\tau=1}^{t}\sum_{i=1}^{2} \mathbb{E}_{\mathcal{X}_\tau}\left[X_{\tau,i} X_{\tau,i}^\top \mathbb{1}\{X_{\tau,i} = \operatorname*{argmax}_{X \in \mathcal{X}_\tau} X^\top \hat{\beta}_\tau\} \mid \hat{\beta}_\tau \right].$$

Since the compatibility condition is satisfied only for the theoretical Gram matrix Σ and we need to show the empirical Gram matrix $\hat{\Sigma}_t$ satisfies the compatibility condition, the adapted Gram matrix Σ_t serves as a bridge between Σ and $\hat{\Sigma}_t$ in our analysis. We first

lower-bound the compatibility constant $\phi^2(\Sigma_t, S_0)$ in terms of $\phi^2(\Sigma, S_0)$ so that we can show that Σ_t satisfies the compatibility condition as long as Σ satisfies the compatibility condition. Then, we show that $\hat{\Sigma}_t$ concentrates around Σ_t with high probability and that such matrix concentration guarantees the compatibility condition of $\hat{\Sigma}_t$.

In Lemma 4.2, we show that the adapted Gram matrix Σ_t can be controlled in terms of the theoretical Gram matrix Σ, which allows us to link the compatibility constant of Σ to compatibility constant of Σ_t. Note that Lemma 4.2 shows the result for any fixed vector β; hence, it can be applied to $\mathbb{E}[X_\tau X_\tau^\top | \mathcal{F}_{\tau-1}]$.

Lemma 4.2. *For a fixed vector $\beta \in \mathbb{R}^d$, we have*

$$\sum_{i=1}^{2} \mathbb{E}_{\mathcal{X}_t}\Big[X_{t,i} X_{t,i}^\top \mathbb{1}\{X_{t,i} = \operatorname*{argmax}_{X \in \mathcal{X}_t} X^\top \beta\}\Big] \succcurlyeq \nu^{-1}\Sigma,$$

where ν the degree of asymmetry of the distribution $p_{\mathcal{X}}$ defined in Assumption 4.4.

Therefore, we have $\Sigma_t \succcurlyeq \nu^{-1}\Sigma$ which implies that $\phi^2(\Sigma_t, S_0) \geq \frac{\phi^2(\Sigma, S_0)}{\nu} > 0$, i.e., Σ_t satisfies the compatibility condition. Note that both Σ and Σ_t can be singular. In Lemma 4.3, we show that $\hat{\Sigma}_t$ concentrates to Σ_t with high probability. This result is crucial in our analysis since it allows the matrix concentration without using i.i.d. samples. The proof of Lemma 4.3 utilizes a new Bernstein-type inequality for adapted samples (Lemma C.5 in the appendix) which may be of independent interest.

Lemma 4.3 (Matrix concentration)**.** *For $t \geq \frac{2\log(2d^2)}{C_0(s_0)^2}$ where $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 \nu x_{\max}^2}\right)$, we have*

$$\mathbb{P}\left(\|\Sigma_t - \hat{\Sigma}_t\|_\infty \geq \frac{\phi_0^2}{32 s_0 \nu}\right) \leq \exp\left(-\frac{t C_0(s_0)^2}{2}\right).$$

Then, we invoke the following corollary to use the matrix concentration results to ensure the compatibility condition for $\hat{\Sigma}_t$.

Corollary 4.1 (Corollary 6.8, Bühlmann and Van De Geer (2011))**.** *Suppose that* Σ_0*-compatibility condition holds for the index set* S *with cardinality* $s = |S|$*, with compatibility constant* $\phi^2(\Sigma_0, S)$*, and that* $\|\Sigma_1 - \Sigma_0\|_\infty \leq \Delta$*, where* $32s\Delta \leq \phi^2(\Sigma_0, S)$*. Then, for the set* S*, the* Σ_1*-compatibility condition holds as well, with* $\phi^2(\Sigma_1, S) \geq \phi^2(\Sigma_0, S)/2$*.*

In order to satisfy the hypotheses in Lemma 4.3 and Corollary 4.1, we define the *initial* period $t < T_0 := \frac{2\log(2d^2)}{C_0(s_0)^2}$ during which the compatibility condition for the empirical Gram matrix is not guaranteed, and the event

$$\mathcal{E}_t := \left\{ \|\Sigma_t - \hat{\Sigma}_t\|_\infty \leq \frac{\phi_0^2}{32 s_0 \nu} \right\}.$$

Then for all $t \geq \lceil T_0 \rceil$ and Σ_t for which event $\mathcal{E}_t$ holds, we have

$$\phi_t^2 := \phi^2(\hat{\Sigma}_t, S_0) \geq \frac{\phi^2(\Sigma_t, S_0)}{2} \geq \frac{\phi_0^2}{2\nu} > 0\,.$$

Hence, the compatibility condition is satisfied for the empirical Gram matrix without using sparsity information.

Proof Sketch of Theorem 4.1

We combine the results above to analyze the regret bound of `SA Lasso Bandit` shown in Theorem 4.1. First, we divide the time horizon $[T]$ into three groups:

(a) $(t \leq T_0)$. Here the compatibility condition is not guaranteed to hold.

(b) $(t > T_0)$ such that $\mathcal{E}_t$ holds.

(c) $(t > T_0)$ such that $\mathcal{E}_t$ does not hold.

These sets are disjoint, hence we bound the regret contribution from each separately and obtain an upper bound on the overall regret. It is important to note that `SA Lasso Bandit`

Algorithm does not rely in any way on this partitioning – it is introduced purely for the purpose of analysis. Set (a) is the initial period over which we do not have guarantees for the compatibility condition. Therefore, we cannot apply the Lasso convergence result; hence we can incur $\mathcal{O}(s_0^2 \log d)$ regret. Set (b) is where the compatibility condition is satisfied; hence the Lasso oracle inequality in Lemma 4.1 can apply. In fact, this group can be further divided to two cases: (b-1) when the high-probability Lasso result holds and (b-2) when it does not, where the regret of (b-2) can be bounded by $\mathcal{O}(1)$. For (b-1), using the Lasso convergence result and summing the regret over the time horizon gives $\mathcal{O}(s_0\sqrt{T\log(dT)})$ regret, which is the leading factor in the regret bound of Theorem 4.1. Lastly, (c) contains the failure events of Lemma 4.3 whose regret is $\mathcal{O}(s_0^2)$. The proofs of the lemmas are in Appendix C.1, followed by the complete proof of Theorem 4.1 in Appendix C.2.

4.4.4 Regret under the Restricted Eigenvalue Condition

In our analysis so far, we have presented the main results under the compatibility condition in order to be consistent with previous results in the sparse bandit literature. In this section, we present the regret bound for `SA Lasso Bandit` under the restricted eigenvalue (RE) condition and briefly discuss its implication in terms of potentially matching lower bounds. Similar to the analysis under the compatibility condition, we assume that the RE condition is satisfied only for the theoretical Gram matrix $\Sigma = \frac{1}{K}\mathbb{E}[\mathbf{X}^\top \mathbf{X}]$.

Assumption 4.5 (RE condition)**.** *For active set S_0 and Σ, there exists restricted eigenvalue $\phi_1 > 0$ such that $\phi_1^2\|\beta\|_2^2 \leq \beta^\top \Sigma \beta$ for all $\beta \in \mathbb{C}(S_0)$ defined in* (4.2)*.*

The RE condition is very similar to the compatibility condition in Assumption 4.3 but uses the ℓ_2 norm instead of the ℓ_1 norm. Based on this condition, we can show the following regret bound.

Theorem 4.2 (Regret bound under RE condition)**.** *Suppose $K = 2$ and Assumptions 4.1, 4.2, 4.4, and 4.5 hold. Then the expected cumulative regret of the* `SA Lasso Bandit` *policy is $\mathcal{O}\Big(\sqrt{s_0 T \log(dT)}\Big)$.*

Theorem 4.2 establishes $\mathcal{O}\Big(\sqrt{s_0 T \log(dT)}\Big)$ regret without any prior knowledge on s_0. The regret upper-bound based on the RE condition still enjoys logarithmic dependence on d and furthermore sublinear dependence on s_0. Compared to Theorem 4.1, the regret bound in Theorem 4.2 is smaller by $\sqrt{s_0}$ factor, which is again consistent with the offline Lasso results under the RE condition (Theorem 7.19 in Wainwright 2019). The difference in the regret bounds in Theorem 4.1 and Theorem 4.2 is due to the RE condition being slightly stronger than the compatibility condition.

The RE condition is more directly analogous (as compared to the compatibility condition) to the standard positive-definiteness assumption for covariance matrices in GLM bandits (Li, Lu, and Zhou, 2017). That is, the RE condition is equivalent to positive-definite covariance when $s_0 = d$, i.e., non-sparse settings. Li, Lu, and Zhou (2017) showed $\mathcal{O}\Big((\log T)^{3/2}\sqrt{dT \log K}\Big)$ regret bound of for GLM bandits, which matches the $\Omega(\sqrt{dT})$ minimax lower bound established (Chu et al., 2011) for linear bandits with finite arms, up to logarithmic factors. Therefore, in sparse settings, we conjecture that $\mathcal{O}\Big(\sqrt{s_0 T \log(dT)}\Big)$ regret is *best possible* up to logarithmic factors under the RE condition (and so is $\mathcal{O}\Big(s_0\sqrt{T \log(dT)}\Big)$ regret under the compatibility condition). While we present these conjectures, we do not claim our results are minimax. In fact, we discuss in Section 4.7 that the entire notion of minimax regret is much more delicate in sparse contextual bandits.

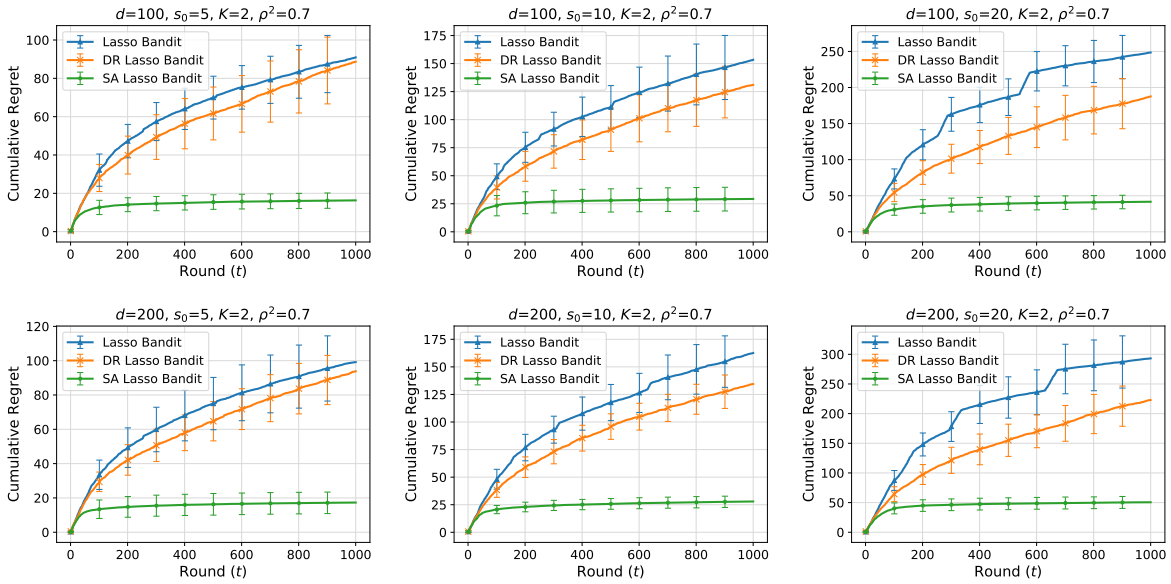

Figure 4.1: The plots show the t-round cumulative regret of `SA Lasso Bandit` (Algorithm 8), `DR Lasso Bandit` (Kim and Paik, 2019), and `Lasso Bandit` (Bastani and Bayati, 2020) for $K = 2$, $d = 100$ (first row) and $d = 200$ (second row) with varying sparsity $s_0 \in \{5, 10, 20\}$ under strong correlation, $\rho^2 = 0.7$.

4.5 Numerical Experiments

We conduct numerical experiments to evaluate `SA Lasso Bandit` and compare with existing sparse bandit algorithms: `DR Lasso Bandit` (Kim and Paik, 2019) and `Lasso Bandit` (Bastani and Bayati, 2020) in two-armed contextual bandits. We follow the experimental setup of Kim and Paik (2019) to evaluate algorithms under different levels of correlation between arms. Although we consider $K = 2$ case in this section, the experimental setup introduced here also applies to numerical evaluations for $K \geq 3$ armed case in Section 4.6. For each dimension $i \in [d]$, we sample each element of the feature vectors $[X_{t,1}^{(i)}, ..., X_{t,K}^{(i)}]$ from multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}_K, V)$ where covariance matrix V is defined as $V_{i,i} = 1$ for all diagonal elements $i \in [K]$ and $V_{i,j} = \rho^2$ for all off-diagonal elements $i \neq j \in [K]$. Hence, for $\rho^2 > 0$, feature vectors for each arm are allowed to be correlated. We consider different levels of correlation with $\rho^2 = 0.7$ (strong correlation) in Figure 4.1

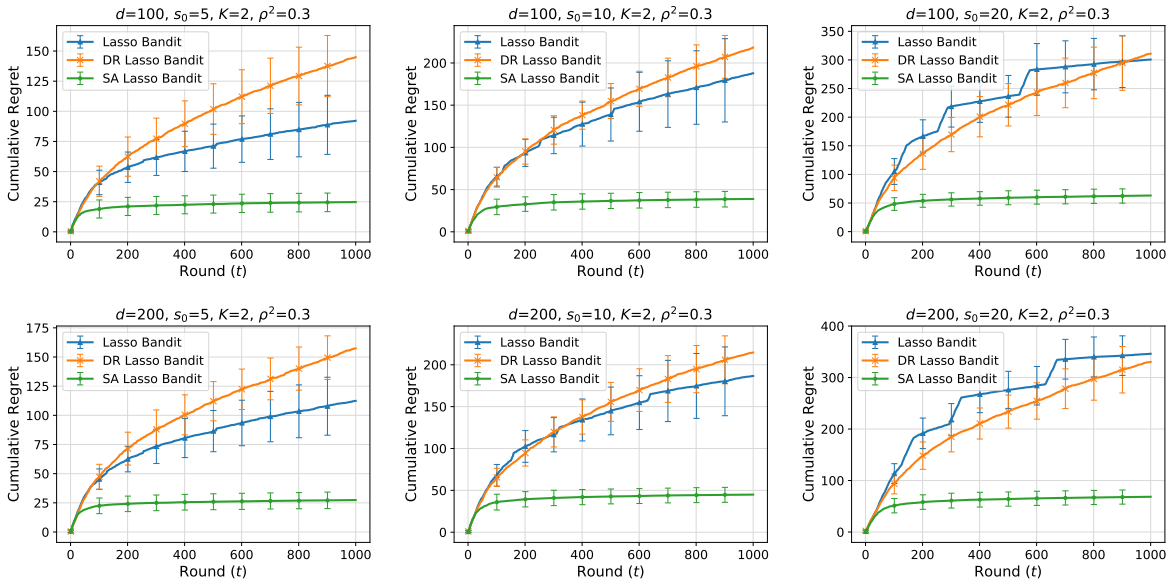

Figure 4.2: The plots show the t-round cumulative regret of `SA Lasso Bandit` (Algorithm 8), `DR Lasso Bandit` (Kim and Paik, 2019), and `Lasso Bandit` (Bastani and Bayati, 2020) for $K = 2$, $d = 100$ (first row) and $d = 200$ (second row) with varying sparsity $s_0 \in \{5, 10, 20\}$ under weak correlation, $\rho^2 = 0.3$.

and $\rho^2 = 0.3$ (weak correlation) in Figure 4.2 as well as $\rho^2 = 0$ (no correlation) in the appendix. In these sets of experiments, we consider feature dimensions $d = 100$ and $d = 200$. For comparison, we use a linear reward with the linear link function $\mu(z) = z$ since both `Lasso Bandit` and `DR Lasso Bandit` are proposed in linear reward settings. We generate β^* with varying sparsity $s_0 = \|\beta^*\|_0$. For a given s_0, we generate each non-zero element of β^* from a uniform distribution in $[0, 1]$. For noise, we sample $\epsilon_t \sim \mathcal{N}(0, 1)$ independently for all rounds. For each case with different experimental configurations, we conduct 20 independent runs, and report the average of the cumulative regret for each of the algorithms. The error bars represent the standard deviations.

`DR Lasso Bandit` is proposed for the same problem setting as ours. Therefore, it does not require any modifications for experiments. However, the problem setting of `Lasso Bandit` is different from ours: it assumes that the context variable is the same for all arms but each arm has a different parameter. We follow the setup in Kim and Paik (2019), and

adapt `Lasso Bandit to our setting by defining` a Kd-dimensional context vector $X_t = [X_{t,1}^\top, ..., X_{t,K}^\top]^\top \in \mathbb{R}^{Kd}$ and a Kd-dimensional parameter β_i^* for each arm i where $\beta_i^* = [\beta^{*\top}\mathbb{1}(i=1), ..., \beta^{*\top}\mathbb{1}(i=K)]^\top \in \mathbb{R}^{Kd}$; thus, $X_t^\top \beta_i^* = X_{ti}^\top \beta^*$s. Note that despite the concatenation, the effective dimension of the unknown parameter β_i^* remains the same as far as estimation is concerned. We defer the other details of the experimental setup and additional results to the appendix.

It is important to note that we report the performances of the benchmarks (`DR Lasso Bandit` and `Lasso Bandit`) assuming that they have access to correct sparsity index s_0; however, this information is hidden from our algorithm. Despite this advantage, the experiment results shown in Figure 4.1 and Figure 4.2 demonstrate that `SA Lasso Bandit` outperforms the other methods by significant margin consistently across various problem instances. We also verify that the performance of our proposed algorithm is the least sensitive to the details of the problem instances, and scales well with changes in the instance. The regret of our algorithm appears to scale linearly with the sparsity index s_0, while its dependence on the feature dimension d appears to be very minimal in most of the instances, which is consistent with our theoretical findings. We also observe that a higher correlation between arms (feature vectors) improves the overall performances of the algorithms. This finding is stronger in the experiments for the K-armed case. We discuss this phenomenon in detail in Section 4.6.

4.6 Extension to K Arms

Thus far, we have presented our main results in two-armed bandit settings which highlight the main challenges of sparse bandit problems without prior knowledge of sparsity. In this section, we extend our regret analysis to the case of $K \geq 3$ arms. Also, we present additional numerical experiments for K-armed bandits.

4.6.1 Regret Analysis for K Arms

Recall that `SA Lasso Bandit` is valid for any number of arms; hence, no modifications are required to extend the algorithm to $K \geq 3$ arms. The analysis of `SA Lasso Bandit` for the K-armed case tackles largely the same challenges described in Section 4.4.3: the need for a Lasso convergence result for adapted samples and ensuring the compatibility condition without knowing s_0 (and without relying on i.i.d. samples). The former challenge is again taken care of by the Lasso convergence result in Lemma 4.1. However, the latter issue is more subtle in the K-armed case than in the two-armed case. In particular, when controlling the adapted Gram matrix Σ_t with the theoretical Gram matrix Σ, the Gram matrix for the unobserved feature vectors could be incomparable with the Gram matrix for the observed feature vectors. For this issue, we introduce an additional regularity condition, which we denote as the "balanced covariance" condition.

Assumption 4.6 (Balanced covariance)**.** *Consider a permutation* $(i_1, ..., i_K)$ *of* $(1, ..., K)$*. For any integer* $k \in \{2, ..., K-1\}$ *and fixed vector* β*, there exists* $C_{\mathcal{X}} < \infty$ *such that*

$$\mathbb{E}\left[X_{i_k} X_{i_k}^\top \mathbb{1}\{X_{i_1}^\top \beta < ... < X_{i_K}^\top \beta\}\right] \preccurlyeq C_{\mathcal{X}} \mathbb{E}\left[(X_{i_1} X_{i_1}^\top + X_{i_K} X_{i_K}^\top) \mathbb{1}\{X_{i_1}^\top \beta < ... < X_{i_K}^\top \beta\}\right].$$

This balanced covariance condition implies that there is "sufficient randomness" in the observed features compared to non-observed features. The exact value of $C_{\mathcal{X}}$ depends on the joint distribution of $\mathcal{X}$ including the correlation between arms. In general, the more positive the correlation, the smaller $C_{\mathcal{X}}$ (obviously, with an extreme case of perfectly correlated arms having a constant $C_{\mathcal{X}}$ independent of any problem parameters). When the arms are independent and identically distributed, Assumption 4.6 holds with $C_{\mathcal{X}} = O(1)$ for both the multivariate Gaussian distribution and a uniform distribution on a sphere, and for an arbitrary independent distribution for each arm, Assumption 4.6 holds for

$C_{\mathcal{X}} = \binom{K-1}{K_0}$ where $K_0 = \lceil (K-1)/2 \rceil$. It is important to note that even in this pessimistic case, $C_{\mathcal{X}}$ does not exhibit dependence on dimensionality d or the sparsity index s_0. These are formalized in Proposition 4 in Appendix C.4.[6] This balanced covariance condition is somewhat similar to "positive-definiteness" condition for observed contexts in the bandit literature (e.g., Goldenshluger and Zeevi (2013) and Bastani, Bayati, and Khosravi (2017)). However, notice that we allow the covariance matrices on both sides of the inequality to be singular. Hence, the positive-definiteness condition for observed context in our setting may not hold even when the balanced covariance condition holds. While this condition admittedly originates from our proof technique, it also provides potential insights on learnability of problem instances. That is, $C_{\mathcal{X}}$ close to infinity implies that the distribution of feature vectors is heavily skewed toward a particular direction. Hence, learning algorithms may require many more samples to learn the unknown parameter, leading to larger regret. It is important to note that our algorithm does not require any prior information on $C_{\mathcal{X}}$. The regret bound for the K-armed sparse bandits under Assumption 4.6 is as follows.

Theorem 4.3 (Regret bound for K arms)**.** *Suppose $K \geq 3$ and Assumptions 4.1-4.4, and 4.6 hold. Let $\lambda_0 = 2\sigma x_{\max}$. Then the expected cumulative regret of the* `SA Lasso Bandit` *policy π over horizon $T \geq 1$ is upper-bounded by*

$$\mathcal{R}^{\pi}(T) \leq 4\kappa_1 + \frac{4\kappa_1 x_{\max} b(\log(2d^2)+1)}{C_1(s_0)^2} + \frac{64\kappa_1 \nu C_{\mathcal{X}} \sigma x_{\max} s_0 \sqrt{T\log(dT)}}{\kappa_0 \phi_0^2}$$

where $C_1(s_0) = \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 \nu C_{\mathcal{X}} x_{\max}^2}\right)$.

[6] While it is not our primary goal to derive general tight bounds on $C_{\mathcal{X}}$, we acknowledge that the bound on $C_{\mathcal{X}}$ for an arbitrary distribution for independent arms is very loose, and is the result of conservative analysis driven by lack of information on $p_{\mathcal{X}}$. Numerical evaluation on distributions other than Gaussian and uniform distributions, detailed in Section 4.6, buttress this point and indicate that the dependence on K is no greater than linear.

Theorem 4.3 establishes $\mathcal{O}\Big(s_0\sqrt{T\log(dT)}\Big)$ regret without prior knowledge on s_0, achieving the same rate as Theorem 4.1 in terms of the key problem primitives. The proof of Theorem 4.3 largely follows that of Theorem 4.1. The main difference is how we control the adapted Gram matrix Σ_t with the theoretical Gram matrix Σ. Under the balanced covariance condition, we can ensure the lower bound of the adapted Gram matrix as a function of the theoretical Gram matrix, which is analogous to the result in Lemma 4.2. In particular, we can show that for a fixed vector $\beta \in \mathbb{R}^d$,

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{X}_t}\Big[X_{t,i}X_{t,i}^\top \mathbb{1}\{X_{t,i} = \operatorname*{argmax}_{X\in\mathcal{X}_t} X^\top\beta\}\Big] \succcurlyeq (2\nu C_{\mathcal{X}})^{-1}\Sigma\,.$$

The formal result is presented in Lemma C.7 in Appendix C.4 along with its proof. Next, we again invoke the matrix concentration result in Lemma 4.3 to connect the compatibility constant of empirical Gram matrix $\hat{\Sigma}_t$ to that of Σ_t, and eventually to the theoretical Gram matrix Σ. Thus, we ensure the compatibility condition of $\hat{\Sigma}_t$. The additional regret in the K-armed case as compared to the two-armed case is essentially a scaling by $C_{\mathcal{X}}$ to ensure the balanced covariance condition.

4.6.2 Numerical Experiments for K Arms

We now validate the performance of `SA Lasso Bandit` in K-armed sparse bandit settings via additional numerical experiments and provide comparison with the existing sparse bandit algorithms. The setup of the experiments is identical to the setup described in Section 4.5. We perform evaluations under various instances. In particular, we focus on the performances of algorithms as the number of arms increases. Additionally, to investigate the effect of the balanced covariance condition, we evaluate algorithms on features drawn from a non-Gaussian elliptical distribution, for which we do not have a tight bound of $C_{\mathcal{X}}$ as well as the multi-dimensional uniform distribution.

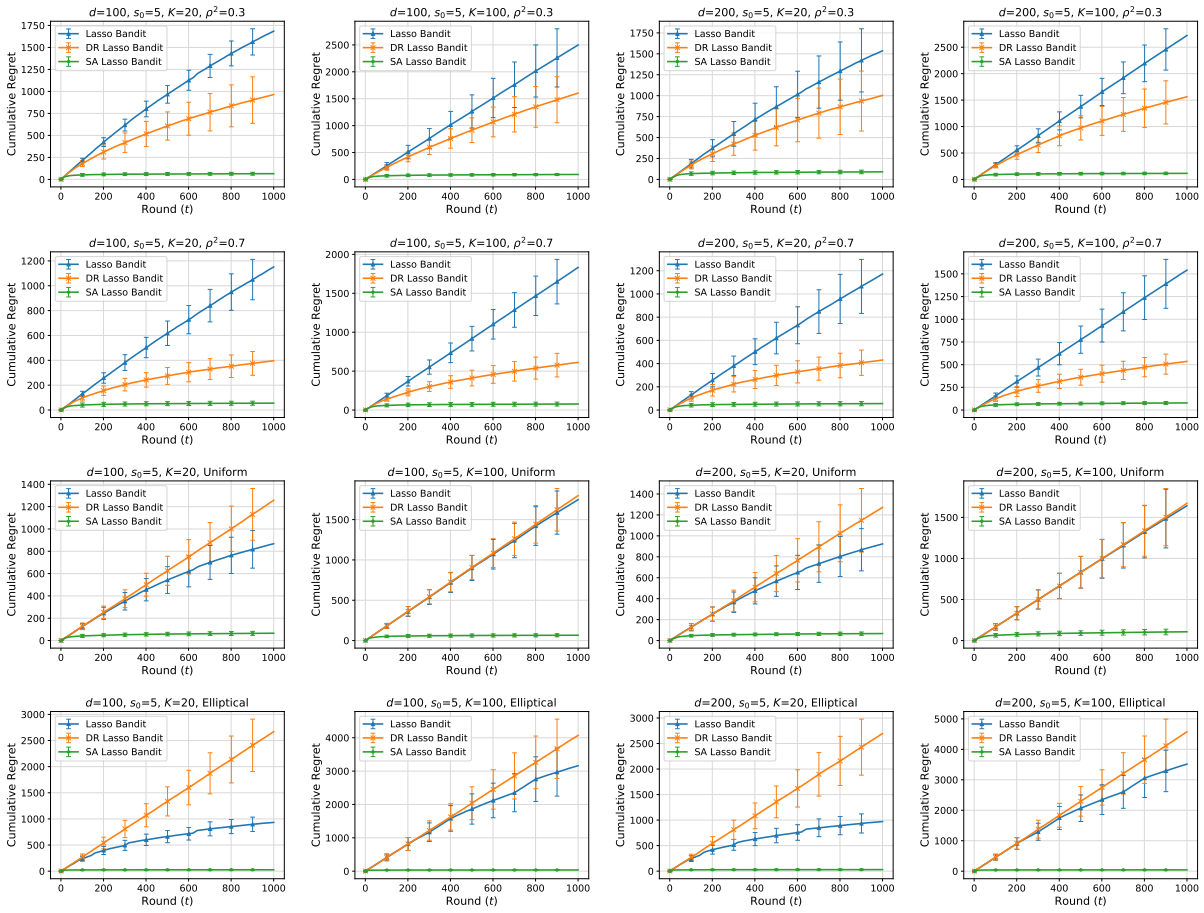

Figure 4.3: The plots show the t-round cumulative regret of `SA Lasso Bandit` (Algorithm 8), `DR Lasso Bandit` (Kim and Paik, 2019), and `Lasso Bandit` (Bastani and Bayati, 2020) with varying number of arms $K \in \{20, 100\}$, feature dimensions $d \in \{100, 200\}$, and different distributions. In the first two rows, features are drawn from a multivariate Gaussian distribution with weak and strong correlation levels. The third row shows evaluations with features drawn from the multi-dimensional uniform distribution. In the fourth row, features are drawn from a non-Gaussian elliptical distribution.

Figure 4.3 shows the sample results of the numerical evaluations (averaged over 20 independent runs per problem instance), and the additional results are also presented in the appendix. The experiment results provide the convincing evidence that the performance of our proposed algorithm is superior to the existing sparse bandit methods that we compare with. Again, `SA Lasso Bandit` outperforms the existing sparse bandit algorithms by significant margins, even though the correct sparsity index s_0 is revealed to these algorithms and kept hidden from `SA Lasso Bandit`. Furthermore, `SA Lasso Bandit` is much more practical and simple to implement with a minimal number of a hyperparameter.

In the experiments with Gaussian distributions shown in the first and second rows in Figure 4.3, we again observe that algorithms generally perform better under strong correlation compared to weak correlation instances. This is expected since strongly (positively) correlated arms imply a smaller discrepancy between expected payoffs of the arms. A strong correlation between the arms also implies a smaller $C_{\mathcal{X}}$, hence leading to a lower regret, as briefly discussed earlier when we introduce the balanced covariance condition. Thus, the balanced covariance condition appears to capture the essence of positive correlation between arms. It is important to note that there are two different notions of correlation: correlation between the arms and correlation between the features of an arm. A higher correlation between the features potentially decreases the value of compatibility constant. Thus, the regret may increase with an increase in correlation of the features as far as the compatibility condition is concerned. The plots in the third and fourth rows in Figure 4.3 show that when the feature vectors are drawn i.i.d. according to the uniform distribution and non-Gaussian elliptical distributions, the performance of existing algorithms (e.g., `DR Lasso Bandit` from Kim and Paik (2019)) deteriorates significantly; `SA Lasso Bandit` still exhibits superior performances. Thus, our proposed algorithm is very robust to the changes in the distribution of the feature vectors.

4.7 Concluding Remarks

In this chapter, we study high-dimensional contextual bandit problem with sparse structure. In particular, we address the fundamental issue that previously known learning algorithms for this problem require a priori knowledge of the sparsity index s_0 of the unknown parameter. We propose and analyze an algorithm that does not require this information. The proposed algorithm achieves a tight regret upper bound which depends on a logarithmic function of the feature dimension which matches the scaling of the offline Lasso convergence results. The algorithm attains this sharp result without knowing the sparsity of the unknown parameter, overcoming weaknesses of the existing algorithms. We demonstrate that our proposed algorithm significantly outperforms the benchmark, supporting the theoretical claims. We conclude by outlining some of future directions.

Minimax Regret in Sparse Bandits. Minimax regret in sparse bandits is more subtle to define than in (non-sparse) linear or GLM bandits. Consider the following setting. Suppose nature is allowed to freely choose $s_0 \in [d]$, it can force the regret for any sparse bandit algorithm to be polynomial in d by choosing $s_0 = d$. On the other hand, if we limit nature to choose $s_0 \in [1, s_{\max}]$, it will choose $s_0 = s_{\max}$, and therefore, sparse bandit algorithms can assume that the sparsity index s_0 is known, and set equal to $s_{\max}$. Thus, it is not clear how to define a minimax criterion in a manner that does not reveal the dominating choice for nature, and therefore, forces learning algorithm to play a strategy which hedges against a range of values of the sparsity index.

Reinforcement Learning with High-Dimensional Covariates. Another compelling direction is to extend our analysis and proposed approach to reinforcement learning with high-dimensional context or with high-dimensional function approximation. A main challenge in this direction appears to be the need for an algorithm to be optimistic. To our knowledge, almost all reinforcement learning algorithms with provable efficiency rely on

the principle of optimism. But, as we have discussed in this chapter, in order to be optimistic in the tightest sense under sparse structure, the knowledge on sparsity is generally needed.

Bibliography

Abbasi-Yadkori, Yasin, Dávid Pál, and Csaba Szepesvári (2011). "Improved algorithms for linear stochastic bandits". In: *Advances in Neural Information Processing Systems*, pp. 2312–2320.

Abbasi-Yadkori, Yasin, David Pal, and Csaba Szepesvari (2012). "Online-to-confidence-set conversions and application to sparse stochastic bandits". In: *Artificial Intelligence and Statistics*, pp. 1–9.

Abe, Naoki and Philip M Long (1999). "Associative reinforcement learning using linear probabilistic concepts". In: *International Conference on Machine Learning*, pp. 3–11.

Abeille, Marc, Alessandro Lazaric, et al. (2017). "Linear Thompson sampling revisited". In: *Electronic Journal of Statistics* 11.2, pp. 5165–5197.

Abramowitz, Milton and Irene A Stegun (1965). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Vol. 55. Courier Corporation.

Agarwal, Alekh, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire (2014). "Taming the monster: A fast and simple algorithm for contextual bandits". In: *International Conference on Machine Learning*, pp. 1638–1646.

Agrawal, Shipra, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi (2017). "Thompson Sampling for the MNL-Bandit". In: *Conference on Learning Theory*, pp. 76–78.

— (2019). "MNL-bandit: A dynamic learning approach to assortment selection". In: *Operations Research* 67.5, pp. 1453–1485.

Agrawal, Shipra and Navin Goyal (2012). "Analysis of thompson sampling for the multi-armed bandit problem". In: *Conference on learning theory*, pp. 39–1.

— (2013). "Thompson sampling for contextual bandits with linear payoffs". In: *International Conference on Machine Learning*, pp. 127–135.

Andrieu, Christophe, Nando De Freitas, Arnaud Doucet, and Michael I Jordan (2003). "An introduction to MCMC for machine learning". In: *Machine learning* 50.1-2, pp. 5–43.

Bibliography

Aouad, Ali, Retsef Levi, and Danny Segev (2018). "Greedy-like algorithms for dynamic assortment planning under multinomial logit preferences". In: *Operations Research* 66.5, pp. 1321–1345.

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

Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer (2002). "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2-3, pp. 235–256.

Bang, Heejung and James M Robins (2005). "Doubly robust estimation in missing data and causal inference models". In: *Biometrics* 61.4, pp. 962–973.

Bartlett, Peter L, Olivier Bousquet, Shahar Mendelson, et al. (2005). "Local rademacher complexities". In: *The Annals of Statistics* 33.4, pp. 1497–1537.

Bastani, Hamsa and Mohsen Bayati (2020). "Online decision making with high-dimensional covariates". In: *Operations Research* 68.1, pp. 276–294.

Bastani, Hamsa, Mohsen Bayati, and Khashayar Khosravi (2017). "Mostly exploration-free algorithms for contextual bandits". In: *arXiv preprint arXiv:1704.09011*.

Bernstein, Fernando, Sajad Modaresi, and Denis Sauré (2018). "A dynamic clustering approach to data-driven assortment personalization". In: *Management Science*.

Bickel, Peter J, Ya'acov Ritov, Alexandre B Tsybakov, et al. (2009). "Simultaneous analysis of Lasso and Dantzig selector". In: *The Annals of Statistics* 37.4, pp. 1705–1732.

Bubeck, Sébastien, Nicolò Cesa-Bianchi, et al. (2012). "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems". In: *Foundations and Trends in Machine Learning* 5.1, pp. 1–122.

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

Cambanis, Stamatis, Steel Huang, and Gordon Simons (1981). "On the theory of elliptically contoured distributions". In: *Journal of Multivariate Analysis* 11.3, pp. 368–385.

Candes, Emmanuel, Terence Tao, et al. (2007). "The Dantzig selector: Statistical estimation when p is much larger than n". In: *The annals of Statistics* 35.6, pp. 2313–2351.

Cao, Wei, Jian Li, Yufei Tao, and Zhize Li (2015). "On top-k selection in multi-armed bandits and hidden bipartite graphs". In: *Advances in Neural Information Processing Systems*, pp. 1036–1044.

Caro, Felipe and Jérémie Gallien (2007). "Dynamic assortment with demand learning for seasonal consumer goods". In: *Management Science* 53.2, pp. 276–292.

Carpentier, Alexandra and Rémi Munos (2012). "Bandit theory meets compressed sensing for high dimensional stochastic linear bandit". In: *Artificial Intelligence and Statistics*, pp. 190–198.

Chapelle, Olivier and Lihong Li (2011). "An empirical evaluation of thompson sampling". In: *Advances in neural information processing systems*, pp. 2249–2257.

Chen, Kani, Inchi Hu, Zhiliang Ying, et al. (1999). "Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs". In: *The Annals of Statistics* 27.4, pp. 1155–1163.

Chen, Xi and Yining Wang (2017). "A Note on Tight Lower Bound for MNL-Bandit Assortment Selection Models". In: *arXiv preprint arXiv:1709.06109*.

Chen, Xi, Yining Wang, and Yuan Zhou (2018). "Dynamic Assortment Optimization with Changing Contextual Information". In: *arXiv preprint arXiv:1810.13069*.

Cheung, Wang Chi and David Simchi-Levi (2017a). "Assortment Optimization under Unknown MultiNomial Logit Choice Models". In: *arXiv preprint arXiv:1704.00108*.

— (2017b). "Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models". In: *Available at SSRN 3075658*.

Chu, Wei, Lihong Li, Lev Reyzin, and Robert Schapire (2011). "Contextual bandits with linear payoff functions". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214.

Dani, Varsha, Thomas P Hayes, and Sham M Kakade (2008). "Stochastic linear optimization under bandit feedback". In: *Proceedings of the 21st Annual Conference on Learning Theory*, 355–366.

Davis, James, Guillermo Gallego, and Huseyin Topaloglu (2013). "Assortment planning under the multinomial logit model with totally unimodular constraint structures". In:

Davis, James M, Guillermo Gallego, and Huseyin Topaloglu (2014). "Assortment optimization under variants of the nested logit model". In: *Operations Research* 62.2, pp. 250–273.

Désir, Antoine, Vineet Goyal, and Jiawei Zhang (2014). "Near-optimal algorithms for capacity constrained assortment optimization". In: *Available at SSRN 2543309*.

Elmachtoub, Adam N., Ryan McNellis, Sechan Oh, and Marek Petrik (2017). "A Practical Method for Solving Contextual Bandit Problems Using Decision Trees". In: *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017*. AUAI Press.

Filippi, Sarah, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári (2010). "Parametric bandits: The generalized linear case". In: *Advances in Neural Information Processing Systems*, pp. 586–594.

Geer, Sara A Van de et al. (2008). "High-dimensional generalized linear models and the lasso". In: *The Annals of Statistics* 36.2, pp. 614–645.

Ghose, Anindya, Panagiotis G Ipeirotis, and Beibei Li (2014). "Examining the impact of ranking on consumer behavior and search engine revenue". In: *Management Science* 60.7, pp. 1632–1654.

Gilton, Davis and Rebecca Willett (2017). "Sparse linear contextual bandits via relevance vector machines". In: *2017 International Conference on Sampling Theory and Applications (SampTA)*. IEEE, pp. 518–522.

Goldenshluger, Alexander and Assaf Zeevi (2013). "A linear response bandit problem". In: *Stochastic Systems* 3.1, pp. 230–261.

Hazan, Elad, Amit Agarwal, and Satyen Kale (2007). "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69.2-3, pp. 169–192.

Hazan, Elad, Tomer Koren, and Kfir Y Levy (2014). "Logistic regression: Tight bounds for stochastic and online optimization". In: *Conference on Learning Theory*, pp. 197–209.

Jaksch, Thomas, Ronald Ortner, and Peter Auer (2010). "Near-optimal regret bounds for reinforcement learning". In: *Journal of Machine Learning Research* 11.Apr, pp. 1563–1600.

Javanmard, Adel and Hamid Nazerzadeh (2019). "Dynamic pricing in high-dimensions". In: *The Journal of Machine Learning Research* 20.1, pp. 315–363.

Kallus, Nathan and Madeleine Udell (2020). "Dynamic assortment personalization in high dimensions". In: *Operations Research*.

Kaufmann, Emilie, Nathaniel Korda, and Rémi Munos (2012). "Thompson sampling: An asymptotically optimal finite-time analysis". In: *International conference on algorithmic learning theory*. Springer, pp. 199–213.

Kim, Gi-Soo and Myunghee Cho Paik (2019). "Doubly-robust lasso bandit". In: *Advances in Neural Information Processing Systems*, pp. 5877–5887.

Kveton, Branislav, Csaba Szepesvari, Zheng Wen, and Azin Ashkan (2015). "Cascading bandits: Learning to rank in the cascade model". In: *International Conference on Machine Learning*, pp. 767–776.

Kveton, Branislav, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier (2020). "Randomized exploration in generalized linear bandits". In: *International Conference on Artificial Intelligence and Statistics*, pp. 2066–2076.

Lai, Tze Leung and Herbert Robbins (1985). "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1, pp. 4–22.

Langford, John and Tong Zhang (2008). "The epoch-greedy algorithm for multi-armed bandits with side information". In: *Advances in neural information processing systems*, pp. 817–824.

Lattimore, Tor and Csaba Szepesvári (2019). *Bandit Algorithms*. Cambridge University Press (preprint).

Lehmann, Erich L and George Casella (2006). *Theory of point estimation*. Springer Science & Business Media.

Li, Lihong, Wei Chu, John Langford, and Robert E Schapire (2010). "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th international conference on World wide web*. ACM, pp. 661–670.

Li, Lihong, Yu Lu, and Dengyong Zhou (2017). "Provably Optimal Algorithms for Generalized Linear Contextual Bandits". In: *International Conference on Machine Learning*, pp. 2071–2080.

Li, Shuai, Baoxiang Wang, Shengyu Zhang, and Wei Chen (2016). "Contextual Combinatorial Cascading Bandits." In: *ICML*. Vol. 16, pp. 1245–1253.

Luce, R Duncan (2012). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

McFadden, Daniel (1978). "Modeling the choice of residential location". In: *Transportation Research Record* 673.

Ou, Mingdong, Nan Li, Shenghuo Zhu, and Rong Jin (2018). "Multinomial Logit Bandit with Linear Utility Functions". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, 2602–2608.

Plackett, Robin L (1975). "The analysis of permutations". In: *Applied Statistics*, pp. 193–202.

Pollard, David (1990). "Empirical processes: theory and applications". In: *NSF-CBMS regional conference series in probability and statistics*. JSTOR, pp. i–86.

Qin, Lijing, Shouyuan Chen, and Xiaoyan Zhu (2014). "Contextual combinatorial bandit and its application on diversified online recommendation". In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, pp. 461–469.

Raskutti, Garvesh, Martin J Wainwright, and Bin Yu (2010). "Restricted eigenvalue properties for correlated Gaussian designs". In: *Journal of Machine Learning Research* 11.Aug, pp. 2241–2259.

Riquelme, Carlos, George Tucker, and Jasper Snoek (2018). "Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling". In: *International Conference on Learning Representations*.

Rusmevichientong, Paat, Zuo-Jun Max Shen, and David B Shmoys (2010). "Dynamic assortment optimization with a multinomial logit choice model and capacity constraint". In: *Operations research* 58.6, pp. 1666–1680.

Rusmevichientong, Paat and John N Tsitsiklis (2010). "Linearly parameterized bandits". In: *Mathematics of Operations Research* 35.2, pp. 395–411.

Russo, Daniel and Benjamin Van Roy (2014). "Learning to optimize via posterior sampling". In: *Mathematics of Operations Research* 39.4, pp. 1221–1243.

Russo, Daniel J, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. (2018). "A Tutorial on Thompson Sampling". In: *Foundations and Trends in Machine Learning* 11.1, pp. 1–96.

Sauré, Denis and Assaf Zeevi (2013). "Optimal dynamic assortment planning with demand learning". In: *Manufacturing & Service Operations Management* 15.3, pp. 387–404.

Strens, Malcolm (2000). "A Bayesian framework for reinforcement learning". In: *International Conference on Machine Learning*, pp. 943–950.

Talluri, Kalyan and Garrett Van Ryzin (2004). "Revenue management under a general discrete choice model of consumer behavior". In: *Management Science* 50.1, pp. 15–33.

Tewari, Ambuj and Susan A Murphy (2017). "From ads to interventions: Contextual bandits in mobile health". In: *Mobile Health*. Springer, pp. 495–517.

Thompson, William R (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3/4, pp. 285–294.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tropp, Joel A (2012). "User-friendly tail bounds for sums of random matrices". In: *Foundations of computational mathematics* 12.4, pp. 389–434.

Van De Geer, Sara A, Peter Bühlmann, et al. (2009). "On the conditions used to prove oracle results for the Lasso". In: *Electronic Journal of Statistics* 3, pp. 1360–1392.

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

Wang, Xue, Mingcheng Wei, and Tao Yao (2018). "Minimax Concave Penalized Multi-Armed Bandit Model with High-Dimensional Covariates". In: *International Conference on Machine Learning*, pp. 5200–5208.

Wen, Zheng, Branislav Kveton, and Azin Ashkan (2015). "Efficient learning in large-scale combinatorial semi-bandits". In: *International Conference on Machine Learning*, pp. 1113–1122.

Zhang, Cun-Hui (2010). "Nearly unbiased variable selection under minimax concave penalty". In: *The Annals of statistics* 38.2, pp. 894–942.

Zhang, Lijun, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou (2016). "Online stochastic linear optimization under one-bit feedback". In: *International Conference on Machine Learning*, pp. 392–401.

Zhou, Zhengyuan, Renyuan Xu, and Jose Blanchet (2019). "Learning in generalized linear contextual bandits with stochastic delays". In: *Advances in Neural Information Processing Systems*, pp. 5198–5209.

Zong, Shi, Hao Ni, Kenny Sung, Nan Rosemary Ke, Zheng Wen, and Branislav Kveton (2016). "Cascading Bandits for Large-scale Recommendation Problems". In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. UAI'16, pp. 835–844.

Appendix A: Upper Confidence Bound Algorithms for MNL Contextual Bandits

A.1 Proofs of Lemmas for Theorem 2.1

A.1.1 Proof of lemma 2.2

Proof. We first define the following:

$$J_t(\theta) = \sum_{t'=1}^{t} \sum_{i \in S_{t'}} \left(p_{t'}(i|S_{t'}, \theta) - p_{t'}(i|S_{t'}, \theta^*) \right) x_{t'i}$$

$$Z_t := J_t(\hat{\theta}_t) = \sum_{t'=1}^{t} \sum_{i \in S_{t'}} \epsilon_{t'i} x_{t'i} \,.$$

Then we follow the same arguments of the proof of Theorem 2.3 up to (A.8) and combine with $\|\hat{\theta}_t - \theta^*\| \leq 1$. Therefore, we have

$$\|Z_t\|_{V_t^{-1}} = \|J_t(\hat{\theta}_t)\|_{V_t^{-1}} \geq \kappa^2 \|\hat{\theta}_t - \theta^*\|_{V_t}^2 \,. \tag{A.1}$$

Then we are left to bound $\|Z_t\|_{V_t^{-1}}^2$. We can use Theorem 1 in Abbasi-Yadkori, Pál, and Szepesvári (2011), which states if the noise ϵ_{ti} is sub-gaussian with parameter σ, then

$$\|Z_t\|_{V_t^{-1}}^2 \leq 2\sigma^2 \log\left(\frac{\det(V_t)^{1/2} \det(V_{T_0})^{-1/2}}{\delta} \right)$$

with probability at least $1-\delta$. Then we combine with Lemma A.2. So it follows that

$$\|Z_t\|^2_{V_t^{-1}} \leq 2\sigma^2 \left[\frac{d}{2}\log\Big(\frac{tK}{d}\Big) - \frac{1}{2}\log\det(V_{T_0}) + \log\frac{1}{\delta}\right].$$

Since $\det(V_{T_0}) \geq (\lambda_{\min}(V_{T_0}))^d$, we have

$$\begin{aligned}\|Z_t\|^2_{V_t^{-1}} &\leq 2\sigma^2 \left[\frac{d}{2}\log\Big(\frac{tK}{d}\Big) - \frac{d}{2}\log\lambda_{\min}(V_{T_0}) + \log\frac{1}{\delta}\right] \\ &\leq 2\sigma^2 \left[\frac{d}{2}\log\Big(\frac{tK}{d}\Big) - \frac{d}{2}\log K + \log\frac{1}{\delta}\right] \\ &\leq \sigma^2 \left[d\log\Big(\frac{t}{d}\Big) + 2\log\frac{1}{\delta}\right] \qquad \text{(A.2)}\end{aligned}$$

where the second inequality is by $\lambda_{\min}(V_{T_0}) \geq K$. Then, using the fact that $\sigma^2 = \frac{1}{4}$ in our problem and combining with (A.1), we have that

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \frac{1}{2\kappa}\sqrt{d\log\Big(\frac{t}{d}\Big) + 2\log\frac{1}{\delta}}.$$

with probability at least $1-\delta$. □

A.1.2 Proof of Lemma 2.1

Proof. The proof of this lemma is the adaptation of Lemma 9 in Kveton et al. (2020), which follows the proof of Theorem 1 in Li, Lu, and Zhou (2017). Note from (A.7) that $J_t(\theta)$ is an injection and satisfies the conditions of Lemma A of Chen, Hu, Ying, et al. (1999). Therefore, we follow the same arguments of Theorem 1 of Li, Lu, and Zhou (2017) to use Lemma A of Chen, Hu, Ying, et al. (1999). For any T_0 such that $\lambda_{\min}(V_{T_0}) \geq 1$

and for $t \geq T_0$, we have

$$\|J_t(\hat{\theta}_t)\|_{V_t^{-1}} \leq \kappa\sqrt{\lambda_{\min}(V_{T_0})} \Longrightarrow \|J_t(\hat{\theta}_t)\|_{V_t^{-1}} \leq \kappa\sqrt{\lambda_{\min}(V_t)}$$
$$\Longrightarrow \|\hat{\theta}_t - \theta^*\| \leq 1\,.$$

Therefore, it suffices to show $\|J_t(\hat{\theta}_t)\|_{V_t^{-1}} \leq \kappa\sqrt{\lambda_{\min}(V_{T_0})}$ for large enough T_0. Then from (A.2), we have

$$\|J_t(\hat{\theta}_t)\|^2_{V_t^{-1}} \leq \frac{1}{4}\left[d\log(t/d) + 2\log\frac{1}{\delta}\right]$$

with probability at least $1-\delta$. Letting $\delta = \frac{1}{T}$, we have

$$\|J_t(\hat{\theta}_t)\|^2_{V_t^{-1}} \leq \frac{1}{4}\left[d\log(T/d) + 2\log T\right]\,.$$

Therefore, if $\lambda_{\min}(V_t)$ is large enough such that

$$\lambda_{\min}(V_t) \geq \frac{1}{4\kappa^2}\left[d\log(T/d) + 2\log T\right],$$

we have $\|\hat{\theta}_t - \theta^*\| \leq 1$ with probability at least $1 - \frac{1}{T}$. □

A.1.3 Proof of Lemma 2.6

The proof of Lemma 2.6 requires the following technical lemmas.

Lemma A.1. *Suppose $\|x_{ti}\| \leq 1$ for all i and t. Define $V_t = V_{T_0} + \sum_{t'=T_0+1}^{t}\sum_{i\in S_{t'}} x_{t'i}x_{t'i}^\top$. Suppose $\lambda_{\min}(V_{T_0}) \geq K$. Then*

$$\sum_{t'=T_0+1}^{t}\sum_{i\in S_{t'}} \|x_{t'i}\|^2_{V_{t'-1}^{-1}} \leq 2\log\left(\frac{\det(V_t)}{\lambda_{\min}(V_{T_0})^d}\right)$$

Proof. Let $\lambda_1, \lambda_2, ..., \lambda_d$ be the eigenvalues of $\sum_{i=1}^n x_{ti} x_{ti}^\top$. Since $\sum_{i=1}^n x_{ti} x_{ti}^\top$ is positive semi-definite, $\lambda_j \geq 0$ for all j. Hence, we have

$$\begin{aligned}
\det\left(I + \sum_{i \in S_t} x_{ti} x_{ti}^\top\right) &= \prod_{j=1}^d (1 + \lambda_j) \\
&\geq 1 + \sum_{j=1}^d \lambda_j = 1 - d + \sum_{j=1}^d (1 + \lambda_j) \\
&= 1 - d + \operatorname{trace}\left(I + \sum_{i \in S_t} x_{ti} x_{ti}^\top\right) = 1 + \sum_{i \in S_t} \|x_{ti}\|_2^2 \qquad \text{(A.3)}
\end{aligned}$$

Now, we lower-bound $\det(V_t)$.

$$\begin{aligned}
\det(V_t) &= \det\left(V_t + \sum_{i \in S_t} x_{ti} x_{ti}^\top\right) \\
&= \det(V_t) \det\left(I + \sum_{i \in S_t} V_t^{-1/2} x_{ti} (V_t^{-1/2} x_{ti})^\top\right) \\
&\geq \det(V_t) \left(1 + \sum_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}^2\right) \\
&\geq \det(V_{T_0}) \prod_{t'=T_0+1}^{t} \left(1 + \sum_{i \in S_t} \|x_{t'i}\|_{V_{t'-1}^{-1}}^2\right) \qquad \text{(A.4)}
\end{aligned}$$

The first inequality comes from (A.3). The second inequality comes from applying the first inequality repeatedly. Since $\lambda_{\min}(V_t)$ is increasing over time, i.e., $\lambda_{\min}(V_t) \geq \lambda_{\min}(V_{T_0})$ for $t \geq T_0$, it follows that

$$\|x_{ti}\|_{V_{t'-1}^{-1}}^2 \leq \frac{\|x_{ti}\|^2}{\lambda_{\min}(V_{\tau-1})} \leq \frac{1}{\lambda_{\min}(V_{T_0})} \leq \frac{1}{K}.$$

Hence $\sum_{i \in S_t} \|x_{ti}\|_{V_{t'-1}^{-1}}^2 \leq 1$ for all $t \geq T_0$. Then using the fact that $z \leq 2\log(1+z)$ for

any $z \in [0,1]$, we have

$$\begin{aligned}\sum_{t'=T_0+1}^{t} \sum_{i \in S_{t'}} \|x_{t'i}\|^2_{V_{t'-1}^{-1}} &\le 2 \sum_{t'=T_0+1}^{t} \log\left(1 + \sum_{i \in S_{t'}} \|x_{t'i}\|^2_{V_{t'-1}^{-1}}\right) \\ &= 2 \log \prod_{t'=T_0+1}^{t} \left(1 + \sum_{i \in S_{t'}} \|x_{t'i}\|^2_{V_{t'-1}^{-1}}\right) \\ &\le 2 \log\left(\frac{\det(V_t)}{\det(V_{T_0})}\right) \\ &\le 2 \log\left(\frac{\det(V_t)}{\lambda_{\min}(V_{T_0})^d}\right)\end{aligned}$$

The second inequality is from (A.4). □

Lemma A.2. *Suppose* $\|x_{ti}\| \le 1$ *for all* i *and* t. *Then* $\det(V_t)$ *is increasing with respect to* t *and*

$$\det(V_t) \le \left(\frac{tK}{d}\right)^d \tag{A.5}$$

Proof. For any symmetric positive definite matrix $\widetilde{V} \in \mathbb{R}^{d \times d}$ and column vector $x \in \mathbb{R}^d$, we have

$$\begin{aligned}\det(\widetilde{V} + xx^\top) &= \det(V) \det\left(I + \widetilde{V}^{-1/2} x x^\top \widetilde{V}^{-1/2}\right) \\ &= \det(\widetilde{V}) \det(1 + \|\widetilde{V}^{-1/2} x\|^2) \\ &\ge \det(\widetilde{V}).\end{aligned}$$

The second equality above is due to Sylvester's determinant theorem, which states that

$\det(I + BA) = \det(I + AB)$. Let $\lambda_1, ..., \lambda_d > 0$ be the eigenvalues of V_t. Then

$$\begin{aligned}
\det(V_t) &\leq \left(\frac{\lambda_1 + ... + \lambda_d}{d}\right)^d \\
&= \left(\frac{\mathrm{trace}(V_t)}{d}\right)^d \\
&= \left(\frac{\sum_{t'=1}^{t} \sum_{i \in S_{t'}} \mathrm{trace}(x_{t'i} x_{t'i}^\top)}{d}\right)^d \\
&= \left(\frac{\sum_{t'=1}^{t} \sum_{i \in S_{t'}} \|x_{t'i}\|^2}{d}\right)^d \\
&\leq \left(\frac{tK}{d}\right)^d .
\end{aligned}$$

□

Proof of Lemma 2.6

Proof. Combining Lemma A.1 and Lemma A.2,

$$\begin{aligned}
\sum_{t'=1}^{t} \max_{i \in S_{t'}} \|x_{t'i}\|^2_{V_{t'-1}^{-1}} &\leq \sum_{t'=1}^{t} \sum_{i \in S_{t'}} \|x_{t'i}\|^2_{V_{t'-1}^{-1}} \\
&\leq 2 \log\left(\frac{\det(V_t)}{\det(V_{T_0})}\right) \\
&\leq 2 \log\left(\frac{tK}{d\lambda_{\min}(V_{T_0})}\right)^d \\
&\leq 2d \log(t/d) \ .
\end{aligned}$$

where the last inequality is by $\lambda_{\min}(V_{T_0}) \geq K$. Then we complete the proof. □

A.1.4 Proof of Lemma 2.3

Proof.

$$\begin{aligned}
|x_{ti}^\top \hat{\theta}_{t-1} - x_{ti}^\top \theta^*| &= \left| \left[V_{t-1}^{-1/2} (\hat{\theta}_{t-1} - \theta^*) \right]^\top (V_{t-1}^{-1/2} x_{ti}) \right| \\
&\leq \|V_{t-1}^{-1/2} (\hat{\theta}_{t-1} - \theta^*)\|_2 \|(V_{t-1}^{-1/2} x_{ti})\|_2 \\
&= \|\hat{\theta}_{t-1} - \theta^*\|_{V_t} \|x_{ti}\|_{V_t^{-1}} \\
&\leq \alpha_{t-1} \|x_{ti}\|_{V_t^{-1}}
\end{aligned}$$

where the first inequality is by Hölder's inequality. Hence, it follows that

$$\left(x_{ti}^\top \hat{\theta}_{t-1} + \alpha_{t-1} \|x_{ti}\|_{V_t^{-1}} \right) - x_{ti}^\top \theta^* \leq 2\alpha \|x_{ti}\|_{V_t^{-1}}.$$

Also, From $|x_{ti}^\top \hat{\theta}_{t-1} - x_{ti}^\top \theta^*| \leq \alpha_{t-1} \|x_{ti}\|_{V_t^{-1}}$, we have

$$x_{ti}^\top \hat{\theta}_{t-1} - x_{ti}^\top \theta^* \geq -\alpha_{t-1} \|x_{ti}\|_{V_t^{-1}}$$

Hence, we have $\left(x_{ti}^\top \hat{\theta}_{t-1} + \alpha_{t-1} \|x_{ti}\|_{V_t^{-1}} \right) - x_{ti}^\top \theta^* \geq 0$ □

A.1.5 Proof of Lemma 2.5

Proof. Let $u_{ti} \geq u'_{ti}$ for all i. By the mean value theorem, there exists $\bar{u}_{ti} := (1-c)u_{ti} + cu'_{ti}$ for some $c \in (0,1)$ with

$$\begin{aligned}
&\frac{\sum_{i\in S} r_{ti} \exp(u_{ti})}{1+\sum_{j\in S}\exp(u_{tj})} - \frac{\sum_{i\in S} r_{ti}\exp(u'_{ti})}{1+\sum_{j\in S}\exp(u'_{tj})} \\
&= \frac{(\sum_{i\in S} r_{ti}\exp\{\bar{u}_{ti}\}(u_{ti}-u'_{ti}))(1+\sum_{i\in S}\exp\{\bar{u}_{ti}\})}{(1+\sum_{i\in S}\exp\{\bar{u}_{ti}\})^2} \\
&\quad - \frac{(\sum_{i\in S} r_{ti}\exp\{\bar{u}_{ti}\})(\sum_{i\in S}\exp\{\bar{u}_{ti}\}(u_{ti}-u'_{ti}))}{(1+\sum_{i\in S}\exp\{\bar{u}_{ti}\})^2} \\
&= \sum_{i\in S} r_{ti} p_{ti}(S,\bar{u}_t)(u_{ti}-u'_{ti}) - R_t(S,\bar{u}_t)\cdot\sum_{i\in S} p_{ti}(S,\bar{u}_t)(u_{ti}-u'_{ti}) \\
&= \sum_{i\in S}\Big(r_{ti} - R_t(S,\bar{u}_t)\Big)p_{ti}(S,\bar{u}_t)(u_{ti}-u'_{ti}) \\
&\leq \max_{i\in S}|u_{ti}-u'_{ti}| = \max_{i\in S}(u_{ti}-u'_{ti})
\end{aligned}$$

where the inequality is from $|r_{ti}| \leq 1$, and $p_{ti}(S,\bar{u}_t) \leq 1$ is a multinomial probability. □

A.2 Proofs for Lemma 2.7 and Theorem 2.2

The proof of Lemma 2.7 depends on the few technical lemma we present here in this section. Recall from Definition 2.2 for the per-round loss $f_t(\theta)$ and its gradient $G_t(\theta)$:

$$\begin{aligned}
f_t(\theta) &= -\sum_{i\in S_t\cup\{0\}} y_{ti}\log p_t(i|S_t,\theta) = -\sum_{i\in S_t} y_{ti}x_{ti}^\top\theta + \log\Big(1+\sum_{j\in S_t}\exp(x_{tj}^\top\theta)\Big) \\
G_t(\theta) &= \nabla_\theta f_t(\theta) = \sum_{i\in S_t}\left(p_t(i|S_t,\theta) - y_{ti}\right)x_{ti}
\end{aligned}$$

We will use these terms throughout this section. In addition to $f_t(\theta)$ and $G_t(\theta)$, we also define their conditional expectations which we will utilize in the proofs of this section.

Appendix A: UCB Algorithms for MNL Contextual Bandits

Definition A.1. *Define the conditional expectations over y of $f_t(\theta)$ and its gradient $G_t(\theta)$.*

$$\bar{f}_t(\theta) := \mathbb{E}_y\left[f_t(\theta)|\mathcal{F}_t\right] \qquad \bar{G}_t(\theta) := \mathbb{E}_y[G_t(\theta)|\mathcal{F}_t] = \mathbb{E}_y[\nabla f_t(\theta)|\mathcal{F}_t]$$

where $\mathcal{F}_t$ contains all the information up to the beginning of the t-th round.

Lemma A.3. *For any θ_1, θ_2, we have*

$$f_t(\theta_2) \geq f_t(\theta_1) + G_t(\theta_1)^\top(\theta_2 - \theta_1) + \frac{\kappa}{2}(\theta_2 - \theta_1)^\top\Big(\sum_{i\in S_t} x_{ti}x_{ti}^\top\Big)(\theta_2 - \theta_1)$$

Proof. Using the Taylor expansion, with $\bar{\theta} = c\theta_2 - (1-c)\theta_1$ for some $c \in (0,1)$

$$f_t(\theta_2) = f_t(\theta_1) + G_t(\theta_1)^\top(\theta_2 - \theta_1) + \frac{1}{2}(\theta_2 - \theta_1)^\top H_f(\bar{\theta})(\theta_2 - \theta_1)$$

where $H_f(\bar{\theta})$ is the Hessian matrix at $\bar{\theta}$. Following the proof of Theorem 2.3, the Hessian matrix can be lower-bounded as follows

$$\begin{aligned} H_f(\bar{\theta}) &= \sum_{i\in S_t} p_t(i|S_t,\bar{\theta})x_{ti}x_{ti}^\top - \sum_{i\in S_t}\sum_{j\in S_t} p_t(i|S_t,\bar{\theta})p_{tj}(S_t,\bar{\theta})x_{ti}x_{tj}^\top \\ &\succeq \sum_{i\in S_t} p_t(i|S_t,\bar{\theta})p_{t0}(\bar{\theta})x_{ti}x_{ti}^\top \end{aligned}$$

From Assumption 2.2, we have

$$H_f(\bar{\theta}) \succeq \kappa \sum_{i\in S_t} x_{ti}x_{ti}^\top$$

Therefore, we have

$$\begin{aligned} f_t(\theta_2) &= f_t(\theta_1) + G_t(\theta_1)^\top(\theta_2-\theta_1) + \frac{1}{2}(\theta_2-\theta_1)^\top H_f(\bar{\theta})(\theta_2-\theta_1) \\ &\geq f_t(\theta_1) + G_t(\theta_1)^\top(\theta_2-\theta_1) + \frac{\kappa}{2}(\theta_2-\theta_1)^\top\Big(\sum_{i\in S_t} x_{ti}x_{ti}^\top\Big)(\theta_2-\theta_1). \end{aligned}$$

□

Lemma A.4.

$$2G_t(\hat{\theta}_t)^\top(\hat{\theta}_t-\theta^*) \leq \|G_t(\theta_t)\|^2_{V_{t+1}^{-1}} + \|\hat{\theta}_t-\theta^*\|^2_{V_{t+1}} - \|\hat{\theta}_{t+1}-\theta^*\|^2_{V_{t+1}}$$

Proof. Note that $\hat{\theta}_{t+1}$ is the optimal solution to the problem

$$\hat{\theta}_{t+1} = \operatorname*{argmin}_{\theta}\left\{\frac{1}{2}\|\theta-\hat{\theta}_t\|^2_{V_{t+1}} + (\theta-\hat{\theta}_t)^\top G_t(\hat{\theta}_t)\right\}$$

Hence, from the first-order optimality condition, we have

$$\Big[G_t(\hat{\theta}_t) + V_{t+1}(\hat{\theta}_{t+1}-\hat{\theta}_t)\Big]^\top(\theta-\hat{\theta}_{t+1}) \geq 0, \forall\theta$$

which gives

$$\theta^\top V_{t+1}(\hat{\theta}_{t+1}-\hat{\theta}_t) \geq \hat{\theta}_{t+1}^\top V_{t+1}(\hat{\theta}_{t+1}-\hat{\theta}_t) - G_t(\hat{\theta}_t)(\theta-\hat{\theta}_{t+1}).$$

Then we can write

$$
\begin{aligned}
&\|\hat{\theta}_t - \theta^*\|^2_{V_{t+1}} - \|\hat{\theta}_{t+1} - \theta^*\|^2_{V_{t+1}} \\
&= \hat{\theta}_t^\top V_{t+1}\hat{\theta}_t - \hat{\theta}_{t+1}^\top V_{t+1}\hat{\theta}_{t+1} + 2{\theta^*}^\top V_{t+1}(\hat{\theta}_{t+1} - \hat{\theta}_t) \\
&\geq \hat{\theta}_t^\top V_{t+1}\hat{\theta}_t - \hat{\theta}_{t+1}^\top V_{t+1}\hat{\theta}_{t+1} + 2\hat{\theta}_{t+1}^\top V_{t+1}(\hat{\theta}_{t+1} - \hat{\theta}_t) - 2G_t(\hat{\theta}_t)(\theta^* - \hat{\theta}_{t+1}) \\
&= \hat{\theta}_t^\top V_{t+1}\hat{\theta}_t + \hat{\theta}_{t+1}^\top V_{t+1}\hat{\theta}_{t+1} - 2\hat{\theta}_{t+1}^\top V_{t+1}\hat{\theta}_t - 2G_t(\hat{\theta}_t)(\theta^* - \hat{\theta}_{t+1}) \\
&= \|\hat{\theta}_t - \hat{\theta}_{t+1}\|^2_{V_{t+1}} + 2G_t(\hat{\theta}_t)(\hat{\theta}_{t+1} - \hat{\theta}_t) + 2G_t(\hat{\theta}_t)(\hat{\theta}_t - \theta^*) \\
&\geq -\|G_t(\theta_t)\|^2_{V_{t+1}^{-1}} + 2G_t(\hat{\theta}_t)(\hat{\theta}_t - \theta^*)
\end{aligned}
$$

where the last inequality is from the fact that

$$
\begin{aligned}
\|\hat{\theta}_t - \hat{\theta}_{t+1}\|^2_{V_{t+1}} + 2G_t(\hat{\theta}_t)(\hat{\theta}_{t+1} - \hat{\theta}_t) &\geq \min_\theta \left\{\|\theta\|^2_{V_{t+1}} + 2G_t(\hat{\theta}_t)(\theta)\right\} \\
&= -\|G_t(\theta_t)\|^2_{V_{t+1}^{-1}}.
\end{aligned}
$$

□

Lemma A.5. *For all $\theta \in \mathbb{R}^d$, we have $\bar{f}_t(\theta) \geq \bar{f}_t(\theta^*)$.*

Proof.

$$
\begin{aligned}
\bar{f}_t(\theta) - \bar{f}_t(\theta^*) &= -\sum_{i \in S_t} p_t(i|S_t, \theta^*) \log p_t(i|S_t, \theta) + \sum_{i \in S_t} p_t(i|S_t, \theta^*) \log p_t(i|S_t, \theta^*) \\
&= \sum_{i \in S_t} p_t(i|S_t, \theta^*) \left[\log p_t(i|S_t, \theta^*) - \log p_t(i|S_t, \theta)\right] \\
&= \sum_{i \in S_t} p_t(i|S_t, \theta^*) \log \frac{p_t(i|S_t, \theta^*)}{p_t(i|S_t, \theta)} \\
&\geq 0
\end{aligned}
$$

where $\sum_{i \in S_t} p_t(i|S_t, \theta^*) \log \frac{p_t(i|S_t, \theta^*)}{p_t(i|S_t, \theta)}$ is the Kullback-Leibler divergence between two distri-

butions which is always non-negative. □

Lemma A.6. *For any positive-semidefinte matrix V,*

$$\|G_t(\theta)\|_V^2 \leq 4 \max_{i \in S_t} \|x_{ti}\|_V^2$$

Proof. For any positive-semidefinte matrix V

$$(z_i - z_j)^\top V (z_i - z_j)^\top = z_i^\top V z_i + z_j^\top V z_j - z_i^\top V z_j - z_j^\top V z_i \geq 0$$

which implies $z_i^\top V z_i + z_j^\top V z_j \geq z_i^\top V z_j + z_j^\top V z_i$. We let $z_i := (p_t(i|S_t, \theta) - y_{ti})\, x_{ti}$

$$\begin{aligned}
\|G_t(\theta)\|_V^2 &= \sum_{i \in S_t} \sum_{j \in S_t} (p_t(i|S_t, \theta) - y_{ti})\, (p_{tj}(S_t, \theta) - y_{tj})\, x_{ti}^\top V x_{tj} \\
&= \sum_{i \in S_t} (p_t(i|S_t, \theta) - y_{ti})^2\, x_{ti}^\top V x_{ti} \\
&\quad + \frac{1}{2} \sum_{i \in S_t} \sum_{j \in S_t} (p_t(i|S_t, \theta) - y_{ti})\, (p_{tj}(S_t, \theta) - y_{tj})\, (x_{ti}^\top V x_{tj} + x_{tj}^\top V x_{ti}) \\
&\leq \sum_{i \in S_t} (p_t(i|S_t, \theta) - y_{ti})^2\, x_{ti}^\top V x_{ti} \\
&\quad + \frac{1}{2} \sum_{i \in S_t} \sum_{j \in S_t} \left[(p_t(i|S_t, \theta) - y_{ti})^2\, x_{ti}^\top V x_{tj} + (p_{tj}(S_t, \theta) - y_{tj})^2\, x_{tj}^\top V x_{ti} \right] \\
&= \sum_{i \in S_t} (p_t(i|S_t, \theta) - y_{ti})^2\, x_{ti}^\top V x_{ti} + \sum_{i \in S_t} (p_t(i|S_t, \theta) - y_{ti})^2\, x_{ti}^\top V x_{ti} \\
&= 2 \sum_{i \in S_t} (p_t(i|S_t, \theta) - y_{ti})^2\, x_{ti}^\top V x_{ti} \\
&\leq 4 \max_{i \in S_t} x_{ti}^\top V x_{ti} \\
&= 4 \max_{i \in S_t} \|x_{ti}\|_V^2
\end{aligned}$$

□

Lemma A.7. *With a probability at least* $1-\delta$,

$$\begin{aligned}
&\sum_{t'=T_0+1}^{t} \left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*) \\
&\leq \frac{\kappa}{4} \sum_{t'=T_0+1}^{t} \|\theta^* - \hat{\theta}_{t'}\|^2_{W_{t'}} + \left(\frac{4}{\kappa} + \frac{8}{3}\right) \log\left(\frac{\lceil 2\log_2 \frac{tK}{2} \rceil t^2}{\delta}\right) + 2
\end{aligned}$$

Proof. First, notice that $\left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*)$ is a martingale difference sequence. Also, we have

$$\begin{aligned}
\left|\left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*)\right| &\leq \left|\left[\bar{G}_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*)\right| + \left|\left[G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*)\right| \\
&\leq \left\|\bar{G}_{t'}(\hat{\theta}_{t'})\right\| \left\|\hat{\theta}_{t'} - \theta^*\right\| + \left\|G_{t'}(\hat{\theta}_{t'})\right\| \left\|\hat{\theta}_{t'} - \theta^*\right\| \\
&\leq 2\sqrt{2}\|\hat{\theta}_{t'} - \theta^*\|
\end{aligned}$$

where the last inequality is from the fact that $\|G_t(\theta)\| = \|\sum_{i\in S_t} (p_t(i|S_t,\theta) - y_{ti})\, x_{ti}\| \leq \sqrt{2}$ for any θ. Also, note that for large enough t' (i.e. after the random initialization), we have $\|\hat{\theta}_{t'} - \theta^*\| \leq 1$. Hence, we have

$$\left|\left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*)\right| \leq 2\sqrt{2}.$$

We define the martingale $M_t := \sum_{t'=1}^{t} \left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*)$. And, we also define

Σ_t as

$$\begin{aligned}
\Sigma_t &:= \sum_{t'=1}^{t} \mathbb{E}_{y_{t'}} \left[\left(\left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'}) \right]^\top (\hat{\theta}_{t'} - \theta^*) \right)^2 \right] \\
&\leq \sum_{t'=1}^{t} \mathbb{E}_{y_{t'}} \left[\left(G_{t'}(\hat{\theta}_{t'})^\top (\hat{\theta}_{t'} - \theta^*) \right)^2 \right] \\
&\leq \sum_{t'=1}^{t} \sum_{i \in S_{t'}} \left(x_{t'i}^\top (\hat{\theta}_{t'} - \theta^*) \right)^2 \\
&= \sum_{t'=1}^{t} \|\hat{\theta}_{t'} - \theta^*\|_{W_{t'}}^2 := B_t
\end{aligned}$$

Note that B_t, the upper bound for Σ_t, is a random variable, so we cannot directly apply Bernstein's inequality to M_t. Instead, we consider two cases (i) $B_t \leq \frac{4}{tK}$ and (ii) $B_t > \frac{4}{tK}$.

Case (i)

Let's assume $B_t = \sum_{t'=1}^{t} \|\hat{\theta}_{t'} - \theta^*\|_{W_{t'}}^2 \leq \frac{4}{tK}$. Then we have

$$\begin{aligned}
M_t &= \sum_{t'=1}^{t} \left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'}) \right]^\top (\hat{\theta}_{t'} - \theta^*) \\
&= \sum_{t'=1}^{t} \sum_{i \in S_{t'}} (y_{t'i} - p(S_t, \theta^*)) x_{t'i}^\top (\hat{\theta}_{t'} - \theta^*) \\
&\leq \sum_{t'=1}^{t} \sum_{i \in S_{t'}} |x_{t'i}^\top (\hat{\theta}_t - \theta^*)| \\
&\leq \sqrt{tK \sum_{t'=1}^{t} \sum_{i \in S_{t'}} \left(x_{t'i}^\top (\hat{\theta}_t - \theta^*) \right)^2} \\
&\leq 2.
\end{aligned}$$

Case (ii)

Let's assume $B_t = \sum_{t'=1}^{t} \|\hat{\theta}_{t'} - \theta^*\|^2_{W_{t'}} > \frac{4}{tK}$. Note that we have both a lower and upper bounds for B_t, i.e., $\frac{4}{tK} < B_t \leq tK$. Then we can use the peeling process (Bartlett, Bousquet, Mendelson, et al., 2005).

$$\begin{aligned}
\mathbb{P}\left(M_t \geq 2\sqrt{\eta_t B_t} + \frac{8\eta_t}{3}\right) &= \mathbb{P}\left(M_t \geq 2\sqrt{\eta_t B_t} + \frac{8\eta_t}{3}, \frac{4}{tK} < B_t \leq tK\right) \\
&= \mathbb{P}\left(M_t \geq 2\sqrt{\eta_t B_t} + \frac{8\eta_t}{3}, \frac{4}{tK} < B_t \leq tK, \Sigma_t \leq B_t\right) \\
&\leq \sum_{j=1}^{m} \mathbb{P}\left(M_t \geq 2\sqrt{\eta_t B_t} + \frac{8\eta_t}{3}, \frac{4 \cdot 2^{j-1}}{tK} < B_t \leq \frac{4 \cdot 2^j}{tK}, \Sigma_t \leq B_t\right) \\
&\leq \sum_{j=1}^{m} \mathbb{P}\left(M_t \geq \sqrt{\eta_t \frac{8 \cdot 2^j}{tK}} + \frac{8\eta_t}{3}, \Sigma_t \leq \frac{4 \cdot 2^j}{tK}\right) \\
&\leq m \exp(-\eta_t)
\end{aligned}$$

where $m = \lceil 2 \log_2 \frac{tK}{2} \rceil$, and the last inequality is from Bernstein's inequality for martingales. Combining with the result in Cases (i) and (ii), letting $\eta_t = \log \frac{mt^2}{\delta} = \log \frac{\lceil 2 \log_2 \frac{tK}{2} \rceil t^2}{\delta}$ and taking the union bound over t, we have with probability at least $1 - \delta$

$$M_t = \sum_{t'=1}^{t} \left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*) \leq 2\sqrt{\eta_t \sum_{t'=T_0+1}^{t} \|\theta^* - \hat{\theta}_{t'}\|^2_{W_{t'}}} + \frac{8\eta_t}{3} + 2.$$

Then we apply $uv \leq cu^2 + v^2/(4c)$ to the second term on the right hand side with $c = \frac{2}{\kappa}$.

$$\sqrt{\eta_t \sum_{t'=T_0+1}^{t} \|\theta^* - \hat{\theta}_{t'}\|^2_{W_{t'}}} \leq \frac{2\eta_t}{\kappa} + \frac{\kappa}{8} \sum_{t'=T_0+1}^{t} \|\theta^* - \hat{\theta}_{t'}\|^2_{W_{t'}}$$

Then we have

$$\sum_{t'=T_0+1}^{t} \left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*) \leq \frac{\kappa}{4} \sum_{t'=T_0+1}^{t} \|\theta^* - \hat{\theta}_{t'}\|^2_{W_{t'}} + \left(\frac{4}{\kappa} + \frac{8}{3}\right)\eta_t + 2$$

□

A.2.1 Proof of Lemma 2.7

Proof. From Lemma A.3, we have

$$f_t(\hat{\theta}_t) \leq f_t(\theta^*) + G_t(\hat{\theta}_t)^\top(\hat{\theta}_t - \theta^*) - \frac{\kappa}{2}(\theta^* - \hat{\theta}_t)^\top\Big(\sum_{i\in S_t} x_{ti}x_{ti}^\top\Big)(\theta^* - \hat{\theta}_t)$$

Taking expectation over y gives

$$\bar{f}_t(\hat{\theta}_t) \leq \bar{f}_t(\theta^*) + \bar{G}_t(\hat{\theta}_t)^\top(\hat{\theta}_t - \theta^*) - \frac{\kappa}{2}(\theta^* - \hat{\theta}_t)^\top\Big(\sum_{i\in S_t} x_{ti}x_{ti}^\top\Big)(\theta^* - \hat{\theta}_t)$$

Note that $\nabla \bar{f}_t(\theta) = \mathbb{E}_y[\nabla f_t(\theta)|\mathcal{F}_t] = \bar{G}_t(\theta)$ by the Leibniz integral rule.
Also, let $W_t := \sum_{i\in S_t} x_{ti}x_{ti}^\top$. Since $\bar{f}_t(\theta) \geq \bar{f}_t(\theta^*)$ from Lemma A.5, we have

$$\begin{aligned}
0 &\leq \bar{f}_t(\hat{\theta}_t) - \bar{f}_t(\theta^*) \\
&\leq \bar{G}_t(\hat{\theta}_t)^\top(\hat{\theta}_t - \theta^*) - \frac{\kappa}{2}\|\theta^* - \hat{\theta}_t\|_{W_t}^2 \\
&= G_t(\hat{\theta}_t)^\top(\hat{\theta}_t - \theta^*) - \frac{\kappa}{2}\|\theta^* - \hat{\theta}_t\|_{W_t}^2 + \Big[\bar{G}_t(\hat{\theta}_t) - G_t(\hat{\theta}_t)\Big]^\top(\hat{\theta}_t - \theta^*)
\end{aligned}$$

From Lemma A.4, we have $2G_t(\hat{\theta}_t)^\top(\hat{\theta}_t - \theta^*) \leq \|G_t(\theta_t)\|_{V_{t+1}^{-1}}^2 + \|\hat{\theta}_t - \theta^*\|_{V_{t+1}}^2 - \|\hat{\theta}_{t+1} - \theta^*\|_{V_{t+1}}^2$. So we have

$$\begin{aligned}
0 &\leq \frac{1}{2}\|G_t(\theta_t)\|_{V_{t+1}^{-1}}^2 + \frac{1}{2}\|\hat{\theta}_t - \theta^*\|_{V_{t+1}}^2 - \frac{1}{2}\|\hat{\theta}_{t+1} - \theta^*\|_{V_{t+1}}^2 \\
&\quad - \frac{\kappa}{2}\|\theta^* - \hat{\theta}_t\|_{W_t}^2 + \Big[\bar{G}_t(\hat{\theta}_t) - G_t(\hat{\theta}_t)\Big]^\top(\hat{\theta}_t - \theta^*) \\
&\leq 2\max_{i\in S_t}\|x_{ti}\|_{V_{t+1}^{-1}}^2 + \frac{1}{2}\|\hat{\theta}_t - \theta^*\|_{V_{t+1}}^2 - \frac{1}{2}\|\hat{\theta}_{t+1} - \theta^*\|_{V_{t+1}}^2 \\
&\quad - \frac{\kappa}{2}\|\theta^* - \hat{\theta}_t\|_{W_t}^2 + \Big[\bar{G}_t(\hat{\theta}_t) - G_t(\hat{\theta}_t)\Big]^\top(\hat{\theta}_t - \theta^*)
\end{aligned}$$

where the last inequality is by Lemma A.6, $\|G_t(\theta)\|^2_{V_{t+1}^{-1}} \leq 4\max_{i\in S_t}\|x_{ti}\|^2_{V_{t+1}^{-1}}$. Note that since $V_{t+1} = V_t + \frac{\kappa}{2}\sum_{i\in S_t} x_{ti}x_{ti}^\top$, we have

$$\begin{aligned}\|\hat{\theta}_t - \theta^*\|^2_{V_{t+1}} &= \|\hat{\theta}_t - \theta^*\|^2_{V_t} + \frac{\kappa}{2}(\hat{\theta}_t - \theta^*)^\top\left(\sum_{i\in S_t} x_{ti}x_{ti}^\top\right)(\hat{\theta}_t - \theta^*)\\ &= \|\hat{\theta}_t - \theta^*\|^2_{V_t} + \frac{\kappa}{2}\|\hat{\theta}_t - \theta^*\|^2_{W_t}.\end{aligned}$$

Therefore, we can continue

$$\begin{aligned}0 &\leq 2\max_{i\in S_t}\|x_{ti}\|^2_{V_{t+1}^{-1}} + \frac{1}{2}\|\hat{\theta}_t - \theta^*\|^2_{V_t} + \frac{\kappa}{4}\|\hat{\theta}_t - \theta^*\|^2_{W_t} - \frac{1}{2}\|\hat{\theta}_{t+1} - \theta^*\|^2_{V_{t+1}}\\ &\quad - \frac{\kappa}{2}\|\theta^* - \hat{\theta}_t\|^2_{W_t} + \left[\bar{G}_t(\hat{\theta}_t) - G_t(\hat{\theta}_t)\right]^\top(\hat{\theta}_t - \theta^*)\\ &= 2\max_{i\in S_t}\|x_{ti}\|^2_{V_{t+1}^{-1}} + \frac{1}{2}\|\hat{\theta}_t - \theta^*\|^2_{V_t} - \frac{1}{2}\|\hat{\theta}_{t+1} - \theta^*\|^2_{V_{t+1}} - \frac{\kappa}{4}\|\theta^* - \hat{\theta}_t\|^2_{W_t}\\ &\quad + \left[\bar{G}_t(\hat{\theta}_t) - G_t(\hat{\theta}_t)\right]^\top(\hat{\theta}_t - \theta^*)\end{aligned}$$

Hence, we have

$$\begin{aligned}\|\hat{\theta}_{t+1} - \theta^*\|^2_{V_{t+1}} &\leq \|\hat{\theta}_t - \theta^*\|^2_{V_t} + 4\max_{i\in S_t}\|x_{ti}\|^2_{V_{t+1}^{-1}} - \frac{\kappa}{2}\|\theta^* - \hat{\theta}_t\|^2_{W_t}\\ &\quad + 2\left[\bar{G}_t(\hat{\theta}_t) - G_t(\hat{\theta}_t)\right]^\top(\hat{\theta}_t - \theta^*).\end{aligned}$$

Summing over t gives

$$\begin{aligned}\|\hat{\theta}_{t+1} - \theta^*\|^2_{V_{t+1}} &\leq \lambda_{\max}(V_{T_0}) + 4\sum_{t'=T_0+1}^{t}\max_{i\in S_{t'}}\|x_{t'i}\|^2_{V_{t'+1}^{-1}} - \frac{\kappa}{2}\sum_{t'=T_0+1}^{t}\|\theta^* - \hat{\theta}_{t'}\|^2_{W_{t'}}\\ &\quad + 2\sum_{t'=T_0+1}^{t}\left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top(\hat{\theta}_{t'} - \theta^*)\end{aligned}$$

Now, we can use Lemma A.7 which shows with a probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{t'=T_0+1}^{t} \left[\bar{G}_{t'}(\hat{\theta}_{t'}) - G_{t'}(\hat{\theta}_{t'})\right]^\top (\hat{\theta}_{t'} - \theta^*) \\
&\leq \frac{\kappa}{4} \sum_{t'=T_0+1}^{t} \|\theta^* - \hat{\theta}_{t'}\|^2_{W_{t'}} + \left(\frac{4}{\kappa} + \frac{8}{3}\right) \log\left(\frac{\lceil 2\log_2 \frac{tK}{2}\rceil t^2}{\delta}\right) + 2.
\end{aligned}$$

We have with a probability at least $1-\delta$

$$\begin{aligned}
\|\hat{\theta}_{t+1} - \theta^*\|^2_{V_{t+1}} &\leq T_0 + 4 \sum_{t'=T_0+1}^{t} \max_{i\in S_{t'}} \|x_{t'i}\|^2_{V^{-1}_{t'+1}} + \left(\frac{8}{\kappa} + \frac{16}{3}\right) \log\left(\frac{\lceil 2\log_2 \frac{tK}{2}\rceil t^2}{\delta}\right) + 4 \\
&\leq T_0 + \frac{8}{\kappa} d \log\left(\frac{t}{d}\right) + \left(\frac{8}{\kappa} + \frac{16}{3}\right) \log\left(\frac{\lceil 2\log_2 \frac{tK}{2}\rceil t^2}{\delta}\right) + 4
\end{aligned}$$

where we apply Lemma 2.6 to bound $\sum_{t'=1}^{t} \max_{i\in S_{t'}} \|x_{t'i}\|^2_{V^{-1}_{t'+1}}$ in the last inequality. Note that $V_t \preccurlyeq V_{t'}$ for any $t \leq t'$, which implies $\|x_{ti}\|^2_{V^{-1}_{t+1}} \leq \|x_{ti}\|^2_{V^{-1}_{t-1}}$. Therefore, we can apply Lemma 2.6 here. Also, note that V_t in Algorithm 1 and V_t in Algorithm 2 are different by the factor of $\frac{\kappa}{2}$, which results in additional $\frac{2}{\kappa}$ factor for the bound of $\sum_{t'=1}^{t} \max_{i\in S_{t'}} \|x_{t'i}\|^2_{V^{-1}_{t'+1}}$. □

A.2.2 Proof of Theorem 2.2

Proof. Similar to the proof of Theorem 2.1, we first define the high probability event

Definition A.2. *Define the following joint event for $t \geq T_0$:*

$$\widetilde{\mathcal{E}}_t = \left\{\lambda_{\min}(V_{T_0}) \geq K,\ \ \|\hat{\theta}_t - \theta^*\| \leq 1,\ \ \|\hat{\theta}_t - \theta^*\|_{V_t} \leq \widetilde{\alpha}_t, \forall t \geq T_0\right\}$$

where $\widetilde{\alpha}_t$ is defined as Theorem 2.2.

First, using Proposition 1 and Lemma 2.1 with the union bound, we can show that $\mathbb{P}\left(\lambda_{\min}(V_{T_0}) \geq K, \|\hat{\theta}_t - \theta^*\| \leq 1\right) \leq \frac{2}{T}$. Hence, the failure event of $\widetilde{\mathcal{E}}_t$ can be bounded with

the concentration result in Lemma 2.7. We begin with decomposition of the cumulative regret based on $\widetilde{\mathcal{E}}_t$.

$$\begin{aligned}
\mathcal{R}(T) &\leq T_0 + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(\widetilde{R}_t(S_t) - R_t(S_t,\theta^*)\right)\mathbb{1}(\widetilde{\mathcal{E}}_t)\right] + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(\widetilde{R}_t(S_t) - R_t(S_t,\theta^*)\right)\mathbb{1}(\widetilde{\mathcal{E}}_t^c)\right] \\
&\leq T_0 + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(\widetilde{R}_t(S_t) - R_t(S_t,\theta^*)\right)\mathbb{1}(\widetilde{\mathcal{E}}_t)\right] + \sum_{t=1}^{T}\mathcal{O}(t^{-2}) \\
&\leq T_0 + \sum_{t=1}^{T} 2\widetilde{\alpha}_T \max_{i\in S_t}\|x_{ti}\|_{V_t^{-1}} + \mathcal{O}(1)
\end{aligned}$$

Applying the Cauchy-Schwarz inequality and Lemma 2.6 for $\sum_{t=1}^{T}\max_{i\in S_t}\|x_{ti}\|^2_{V_t^{-1}}$, we have

$$\mathcal{R}(T) \leq T_0 + 2\widetilde{\alpha}_T\sqrt{2dT\log(T/d)} + \mathcal{O}(1)$$

where $\widetilde{\alpha}_T = \sqrt{T_0 + \frac{8}{\kappa}d\log(T/d) + \left(\frac{8}{\kappa}+\frac{16}{3}\right)\log(\lceil 2\log_2(TK/2)\rceil t^4) + 4}$. □

A.3 Proof of Theorem 2.3

In this section, we present a finite-sample version of the asymptotic normality of the MLE for the MNL model. It is a generalization of Theorem 1 in (Li, Lu, and Zhou, 2017) to a multinomial setting.

Proof. Recall that the gradient of the negative log-likelihood of the MNL model is given by

$$\nabla_\theta \ell_n(\theta) = \sum_{t=1}^{n}\sum_{i\in S_t}(p_t(i|S_t,\theta) - y_{ti})x_{ti}$$

We define its conditional expectation $J_n(\theta)$ and will use this term throughout this section

Definition A.3. *Define the conditional expectation* $\nabla_\theta \ell(\theta)$ *as*

$$J_n(\theta) := \mathbb{E}_y \left[\nabla_\theta \ell_n(\theta) | \mathcal{F}_t\right] = \sum_{t=1}^{n} \sum_{i \in S_t} \left(p_t(i|S_t, \theta) - p_t(i|S_t, \theta^*)\right) x_{ti}.$$

Notice that $J_n(\hat{\theta}_n) = \sum_{t=1}^{n} \sum_{i \in S_t} \epsilon_{ti} x_{ti}$ since the choice of $\hat{\theta}_n$ is given by the MLE. In other words, $\hat{\theta}_n$ is given by the solution to the following:

$$\sum_{t=1}^{n} \sum_{i \in S_t} \left(p_t(i|S_t, \hat{\theta}_n) - y_{ti}\right) x_{ti} = 0$$

Hence it follows that

$$\begin{aligned} J_n(\hat{\theta}_n) &= \sum_{t=1}^{n} \sum_{i \in S_t} \left(p_t(i|S_t, \hat{\theta}_n) - p_t(i|S_t, \theta^*)\right) x_{ti} \\ &= \sum_{t=1}^{n} \sum_{i \in S_t} \left(p_t(i|S_t, \hat{\theta}_n) - y_{ti}\right) x_{ti} + \sum_{t=1}^{n} \sum_{i \in S_t} \left(y_{ti} - p_t(i|S_t, \theta^*)\right) x_{ti} \\ &= 0 + \sum_{t=1}^{n} \sum_{i \in S_t} \epsilon_{ti} x_{ti} \end{aligned}$$

For convenience, define $Z_n := J_n(\hat{\theta}_n)$. For brevity, we will denote $p_{ti}(\theta) := p_t(i|S_t, \theta)$ when it is clear that S_t is the assortment chosen at round t.

A.3.1 Consistency of MLE

In this section, we show the consistency of MLE $\hat{\theta}_n$. For any $\theta_1, \theta_2 \in \mathbb{R}^d$, the mean value theorem implies that there exists $\bar{\theta} = c\theta_1 + (1-c)\theta_2$ with $c \in (0, 1)$.

$$\begin{aligned} J_n(\theta_1) - J_n(\theta_2) &= \left[\sum_{t=1}^{n} \sum_{i \in S_t} \sum_{j \in S_t} \nabla_j p_{ti}(\bar{\theta}) x_{ti} x_{tj}^\top\right] (\theta_1 - \theta_2) \\ &= \sum_{t=1}^{n} \left[\sum_{i \in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i \in S_t} \sum_{j \in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top\right] (\theta_1 - \theta_2) \end{aligned}$$

Let $H_t := \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i,j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top$. Notice H_t is a Hessian of a negative log-likelihood which is convex. Hence, H_t is positive semidefinite. Also note that

$$(x_i - x_j)(x_i - x_j)^\top = x_i x_i^\top + x_j x_j^\top - x_i x_j^\top - x_j x_i^\top \succeq 0$$

which implies $x_i x_i^\top + x_j x_j^\top \succeq x_i x_j^\top + x_j x_i^\top$. Therefore, it follows that

$$\begin{aligned}
H_t &= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \frac{1}{2}\sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) \left(x_{ti} x_{tj}^\top + x_{tj} x_{ti}^\top\right) \\
&\succeq \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \frac{1}{2}\sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) \left(x_{ti} x_{ti}^\top + x_{tj} x_{tj}^\top\right) \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{ti}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) \left(1 - \sum_{j\in S_t} p_{tj}(\bar{\theta})\right) x_{ti} x_{ti}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta}) x_{ti} x_{ti}^\top
\end{aligned}$$

where $p_{t0}(\bar{\theta})$ is the probability of choosing the no purchase option under parameter $\bar{\theta}$. Define $\mathcal{H}_n(\theta) := \sum_{t=1}^n \sum_{i\in S_t} p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta}) x_{ti} x_{ti}^\top$. Then, we can write

$$\begin{aligned}
J_n(\theta_1) - J_n(\theta_2) &= \left[\sum_{t=1}^n H_t\right](\theta_1 - \theta_2) \\
&\geq \left[\sum_{t=1}^n \sum_{i\in S_t} p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta}) x_{ti} x_{ti}^\top\right](\theta_1 - \theta_2) \\
&= \mathcal{H}_n(\bar{\theta})(\theta_1 - \theta_2) \qquad \text{(A.6)}
\end{aligned}$$

If $\bar{\theta} \in \mathcal{B}_\eta := \{\theta : \|\theta - \theta^*\| \leq \eta\}$ with some $\eta > 0$, then $p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta}) \geq \kappa_\eta$, where κ_η is

defined as $\kappa_\eta := \inf_{\theta\in\mathcal{B}_\eta, i\in S, S\in\mathcal{S}} p_{ti}(\theta)p_{t0}(\theta) > 0$. Then since $\mathcal{H}_n(\bar{\theta}) \succeq \kappa_\eta V_n$, we have

$$(\theta_1-\theta_2)^\top(J_n(\theta_1)-J_n(\theta_2)) \geq (\theta_1-\theta_2)^\top(\kappa_\eta V_n)(\theta_1-\theta_2) > 0 \tag{A.7}$$

for any $\theta_1 \neq \theta_2$. Therefore, $J_n(\theta)$ is an injection from $\mathbb{R}^d$ to $\mathbb{R}^d$. and so the inverse J^{-1} is a well-defined function. Note that $\mathcal{B}_\eta$ is a convex set. Hence, if $\theta_1, \theta_2 \in \mathcal{B}_\eta$, then also $\bar{\theta} \in \mathcal{B}_\eta$. Also, by the definition of $J_n(\theta)$, we have $J_n(\theta^*) = 0$. Then, for any $\theta \in \mathcal{B}_\eta$, it follows that

$$\begin{aligned}
\|J_n(\theta)\|^2_{V_n^{-1}} &= \|J_n(\theta) - J_n(\theta^*)\|^2_{V_n^{-1}} \\
&\geq (\theta-\theta^*)^\top \mathcal{H}_n(\bar{\theta}) V_n^{-1} \mathcal{H}_n(\bar{\theta})(\theta-\theta^*) \\
&\geq \kappa_\eta^2 \lambda_{\min}(V_n)\|\theta-\theta^*\|^2
\end{aligned} \tag{A.8}$$

where the first inequality is due to (A.6) and the second inequality is again from the fact that $\mathcal{H}_n(\bar{\theta}) \succeq \kappa_\eta V_n$. Now, we need an upper-bound for $\|J_n(\theta)\|_{V_n^{-1}}$. From Lemma A.8,

$$\|J_n(\hat{\theta}_n)\|_{V_n^{-1}} \leq 4\sqrt{2d + \log\frac{1}{\delta}} \tag{A.9}$$

with probability at least $1-\delta$. Also, from (A.2) in Lemma 2.2, we have

$$\|J_n(\hat{\theta}_n)\|_{V_n^{-1}} \leq \frac{1}{2}\sqrt{d\log\Big(\frac{n}{d}\Big) + 2\log\frac{1}{\delta}}$$

with probability at least $1-\delta$ for $n \geq T_0$ with $\lambda_{\min}(V_{T_0}) \geq K$. We let $\mathcal{D}$ denote a high probability upper bound on $\|J_n(\hat{\theta}_n)\|_{V_n^{-1}}$:

$$\mathcal{D} := \min\left\{4\sqrt{2d + \log\frac{1}{\delta}}, \sqrt{d\log\Big(\frac{n}{d}\Big) + 2\log\frac{1}{\delta}}\right\}$$

so that $\|J_n(\hat{\theta}_n)\|_{V_n^{-1}} \leq \mathcal{D}$ with probability at least $1-2\delta$. For $n \geq T_0$ such that $\lambda_{\min}(V_{T_0}) \geq \max\Big\{\frac{1}{4\kappa^2}\Big[d\log\Big(\frac{T}{d}\Big)+4\log T\Big], K\Big\}$, we can apply Lemma 2.1 to ensure $\|\hat{\theta}_n - \theta^*\| \leq 1$. Hence, using $\kappa \leq \min_{\|\theta-\theta^*\|\leq 1} p_{ti}(S,\theta)p_{t0}(S,\theta)$ in Assumption 2.2 and combining with (A.8), we have

$$\|\hat{\theta}_n - \theta^*\| \leq \frac{\mathcal{D}}{\kappa\sqrt{\lambda_{\min}(V_n)}}. \tag{A.10}$$

A.3.2 Normality of MLE

In this section, we show the normality result of MLE $\hat{\theta}_n$. For the rest of the section, we assumes (A.9) holds. First, we define F, L and E which are defined as:

$$\begin{aligned}
F(\theta) &:= \sum_{t=1}^{n}\sum_{i\in S_t} p_{ti}(\theta)x_{ti}x_{ti}^\top - \sum_{t=1}^{n}\sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\theta)p_{tj}(\theta)x_{ti}x_{tj}^\top \\
L &:= F(\theta^*) = \sum_{t=1}^{n}\sum_{i\in S_t} p_{ti}(\theta^*)x_{ti}x_{ti}^\top - \sum_{t=1}^{n}\sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\theta^*)p_{tj}(\theta^*)x_{ti}x_{tj}^\top \\
E &:= F(\widetilde{\theta}) - F(\theta^*)
\end{aligned}$$

where $\widetilde{\theta} := c\theta^* + (1-c)\hat{\theta}_n$ for some constant $c \in (0,1)$. Then, it follows that

$$\begin{aligned}
Z_n = J_n(\hat{\theta}_n) &= J_n(\hat{\theta}_n) - J_n(\theta^*) \\
&= (L+E)(\hat{\theta} - \theta^*).
\end{aligned}$$

Hence, for any $x \in \mathbb{R}^2$, we can write

$$\begin{aligned}
x^\top(\hat{\theta}_n - \theta^*) &= x^\top(L+E)^{-1}Z_n \\
&= x^\top L^{-1}Z_n - x^\top L^{-1}E(L+E)^{-1}Z_n.
\end{aligned} \tag{A.11}$$

Note that $(L+E)$ is a non-singular matrix, hence $(L+E)$ is invertible. Here, the key element is controlling the matrix E. Note that if $\hat{\theta}_n$ and θ^* are close (so $\widetilde{\theta}$ and θ^* are also close), elements in E are small.

A.3.3 Bounding Matrix E

First, we further decompose E into two summations, E_1 and E_2

$$E = \underbrace{\sum_{t=1}^{n}\sum_{i\in S_t}\left(p_{ti}(\widetilde{\theta})-p_{ti}(\theta^*)\right)x_{ti}x_{ti}^\top}_{E_1} - \underbrace{\sum_{t=1}^{n}\sum_{i\in S_t}\sum_{j\in S_t}\left(p_{ti}(\widetilde{\theta})p_{tj}(\widetilde{\theta})-p_{ti}(\theta^*)p_{tj}(\theta^*)\right)x_{ti}x_{tj}^\top}_{E_2} \tag{A.12}$$

We first bound the first summation E_1. Note that

$$\begin{aligned}
E_1 &= \sum_{t=1}^{n}\sum_{i\in S_t}\left(p_{ti}(\widetilde{\theta})-p_{ti}(\theta^*)\right)x_{ti}x_{ti}^\top \\
&= \sum_{t=1}^{n}\sum_{i\in S_t}\sum_{j\in S_t}\nabla_j p_{ti}(\theta_1)x_{tj}^\top(\hat{\theta}_n-\theta^*)x_{ti}x_{ti}^\top \\
&= \sum_{t=1}^{n}\sum_{i\in S_t}p_{ti}(\theta_1)x_{ti}^\top(\hat{\theta}_n-\theta^*)x_{ti}x_{ti}^\top - \sum_{t=1}^{n}\sum_{i\in S_t}\sum_{j\in S_t}p_{ti}(\theta_1)p_{tj}(\theta_1)x_{tj}^\top(\hat{\theta}_n-\theta^*)x_{ti}x_{ti}^\top
\end{aligned}$$

where the second equality is by the mean value theorem for some $\theta_1 := c_1\theta^* + (1-c_1)\hat{\theta}_n$ with $c_1 \in (0,1)$. Note that the mean value theorem is applied to $\widetilde{\theta}$ and θ^*, and since $\widetilde{\theta}$ is a convex combination of $\hat{\theta}_n$ and θ^*, we can find such c_1. Then it follows that

$$\begin{aligned}
E_1 &= \sum_{t=1}^{n}\sum_{i\in S_t}p_{ti}(\theta_1)\left(x_{ti}^\top(\hat{\theta}_n-\theta^*) - \sum_{j\in S_t}p_{tj}(\theta_1)x_{tj}^\top(\hat{\theta}_n-\theta^*)\right)x_{ti}x_{ti}^\top \\
&\le \sum_{t=1}^{n}\sum_{i\in S_t}p_{ti}(\theta_1)\left\|x_{ti}-\sum_{j\in S_t}p_{tj}(\theta_1)x_{tj}\right\|\|\hat{\theta}_n-\theta^*\|x_{ti}x_{ti}^\top \\
&\le \sum_{t=1}^{n}\sum_{i\in S_t}2p_{ti}(\theta_1)\|\hat{\theta}_n-\theta^*\|x_{ti}x_{ti}^\top
\end{aligned}$$

where we have used the assumption that $\|x_{ti}\| < 1$ for all i and t for the last inequality. Then, for any $x \in \mathbb{R}^d \setminus \{0\}$, we have

$$\begin{aligned} x^\top L^{-1/2} E_1 L^{-1/2} x &\le \sum_{t=1}^{n} \sum_{i \in S_t} 2 p_{ti}(\theta_1) \|\hat{\theta}_n - \theta^*\| \|x^\top L^{-1/2} x_{ti}\|^2 \\ &\le \sum_{t=1}^{n} \sum_{i \in S_t} 2 \|\hat{\theta}_n - \theta^*\| \|x^\top L^{-1/2} x_{ti}\|^2 \\ &\le 2\|\hat{\theta}_n - \theta^*\| \left(x^\top L^{-1/2} \left(\sum_{t=1}^{n} \sum_{i \in S_t} x_{ti} x_{ti}^\top \right) L^{-1/2} x \right) \\ &\le \frac{2}{\kappa} \|\hat{\theta}_n - \theta^*\| \|x\|^2 \end{aligned}$$

where the third inequality follows from the fact that $p_{ti}(\theta_1) \le 1$. Therefore, combining with (A.10) it follows that

$$\|L^{-1/2} E_1 L^{-1/2}\| \le \frac{2}{\kappa} \|\hat{\theta}_n - \theta^*\| \le \frac{2\mathcal{D}}{\kappa^2 \sqrt{\lambda_{\min}(V_n)}}. \tag{A.13}$$

Similarly, we can bound the second summation E_2 in (A.12). Again by the mean value theorem, for some $\theta_2 := c_2 \theta^* + (1 - c_2)\hat{\theta}_n$ with $c_2 \in (0, 1)$ we have

$$\begin{aligned} E_2 &= \sum_{t=1}^{n} \sum_{i \in S_t} \sum_{j \in S_t} \left(p_{ti}(\widetilde{\theta}) p_{tj}(\widetilde{\theta}) - p_{ti}(\theta^*) p_{tj}(\theta^*) \right) x_{ti} x_{tj}^\top \\ &= \sum_{t=1}^{n} \sum_{i \in S_t} \sum_{j \in S_t} \sum_{k \in S_t} \nabla_k [p_{ti}(\theta_2) p_{tj}(\theta_2)] x_{t,k}^\top (\hat{\theta}_n - \theta^*) x_{ti} x_{ti}^\top. \end{aligned}$$

Let $p_{ti} = p_{ti}(\theta_2)$ for brevity. Then, it follows that

$$\begin{aligned}
E_2 &= \sum_{t=1}^{n} \sum_{i\in S_t} \sum_{j\in S_t} \sum_{k\in S_t} \nabla_k [p_{ti}p_{tj}] x_{t,k}^\top (\hat{\theta}_n - \theta^*) x_{ti} x_{tj}^\top \\
&= \sum_{t=1}^{n} \sum_{i\in S_t} \sum_{j\in S_t} \left[p_{tj} \left(p_{ti} x_{ti} - \sum_{k\in S_t} p_{ti} p_{t,k} x_{t,k} \right) \right]^\top (\hat{\theta}_n - \theta^*) x_{ti} x_{tj}^\top \\
&\quad + \sum_{t=1}^{n} \sum_{i\in S_t} \sum_{j\in S_t} \left[p_{ti} \left(p_{tj} x_{tj} - \sum_{k\in S_t} p_{tj} p_{t,k} x_{t,k} \right) \right]^\top (\hat{\theta}_n - \theta^*) x_{ti} x_{tj}^\top \\
&= \sum_{t=1}^{n} \sum_{i\in S_t} \sum_{j\in S_t} p_{ti}p_{tj} \left[(x_{ti} + x_{tj}) - 2 \sum_{k\in S_t} p_{t,k} x_{t,k} \right]^\top (\hat{\theta}_n - \theta^*) x_{ti} x_{tj}^\top \\
&\le \sum_{t=1}^{n} \sum_{i\in S_t} \sum_{j\in S_t} p_{ti}p_{tj} \left\| (x_{ti} + x_{tj}) - 2 \sum_{k\in S_t} p_{t,k} x_{t,k} \right\| \|\hat{\theta}_n - \theta^*\| x_{ti} x_{tj}^\top \\
&\le \sum_{t=1}^{n} \sum_{i\in S_t} \sum_{j\in S_t} 4 p_{ti}p_{tj} \|\hat{\theta}_n - \theta^*\| x_{ti} x_{tj}^\top \\
&= \sum_{t=1}^{n} \sum_{i\in S_t} 4 p_{ti} \left(1 - p_{t0}\right) \|\hat{\theta}_n - \theta^*\| x_{ti} x_{ti}^\top
\end{aligned}$$

where $p_{t0} = p_{t0}(\theta_2)$ is a probability of choosing an outside option. Then, for any $x \in \mathbb{R}^d \setminus \{0\}$, we have

$$\begin{aligned}
x^\top L^{-1/2} E_2 L^{-1/2} x &\le \sum_{t=1}^{n} \sum_{i\in S_t} 4 p_{ti}(\theta_2) \left(1 - p_{t0}(\theta_2)\right) \|\hat{\theta}_n - \theta^*\| \|x^\top L^{-1/2} x_{ti}\|^2 \\
&\le \sum_{t=1}^{n} \sum_{i\in S_t} 4 \|\hat{\theta}_n - \theta^*\| \|x^\top L^{-1/2} x_{ti}\|^2 \\
&\le 4 \|\hat{\theta}_n - \theta^*\| \left(x^\top L^{-1/2} \left(\sum_{t=1}^{n} \sum_{i\in S_t} x_{ti} x_{ti}^\top \right) L^{-1/2} x \right) \\
&\le \frac{4}{\kappa} \|\hat{\theta}_n - \theta^*\| \|x\|^2 .
\end{aligned}$$

Similarly, combining with (A.10) it follows that

$$\|L^{-1/2} E_2 L^{-1/2}\| \le \frac{4}{\kappa} \|\hat{\theta}_n - \theta^*\| \le \frac{4\mathcal{D}}{\kappa^2 \sqrt{\lambda_{\min}(V_n)}} . \tag{A.14}$$

Hence, combining (A.13) and (A.14), we have for $\lambda_{\min}(V_n) \geq \frac{144}{\kappa^4}\mathcal{D}^2$

$$\begin{aligned}\|L^{-1/2}EL^{-1/2}\| &= \|L^{-1/2}(E_1 - E_2)L^{-1/2}\| \\ &\leq \|L^{-1/2}E_1L^{-1/2}\| + \|L^{-1/2}E_2L^{-1/2}\| \\ &\leq \frac{6\mathcal{D}}{\kappa^2\sqrt{\lambda_{\min}(V_n)}} \leq \frac{1}{2}. \end{aligned} \tag{A.15}$$

A.3.4 Bounding the Prediction Error $x^\top(\hat{\theta}_n - \theta^*)$

Recall from (A.11) that the prediction error for any $x \in \mathbb{R}^2$ can be written as

$$x^\top(\hat{\theta}_n - \theta^*) = x^\top L^{-1}Z_n - x^\top L^{-1}E(L+E)^{-1}Z_n.$$

First, we bound the first term $x^\top L^{-1}Z_n$ in (A.11). We start with providning the following definitions for the ease of our presentation:

$$\begin{aligned} X_t &:= [x_{t1}; x_{t2}; ...; x_{t|S_t|}]^\top \in \mathbb{R}^{|S_t|\times d} \\ \mathbf{X} &:= [X_1; X_2; ...; X_n]^\top \in \mathbb{R}^{(\sum_t |S_t|)\times d} \\ \mathcal{E}_t &:= [\epsilon_{t1}, \epsilon_{t2}, ..., \epsilon_{t|S_t|}]^\top \in \mathbb{R}^{|S_t|} \end{aligned}$$

Then we use the notations above to see $|x^\top L^{-1}Z_n| = \left|\sum_t x^\top L^{-1}X_t^\top \mathcal{E}_t\right|$. For independent samples, X_t and $\mathcal{E}_t$ are independent. Therefore, for each t

$$\mathbb{E}\left[x^\top L^{-1}X_t^\top \mathcal{E}_t\right] = \mathbb{E}\left[\sum_{i\in S_t} x^\top L^{-1}x_{ti}\epsilon_{ti}\right] = \sum_{i\in S_t}\mathbb{E}\left[x^\top L^{-1}x_{ti}\right]\mathbb{E}[\epsilon_{ti}] = 0$$

since $\mathbb{E}[\epsilon_{ti}] = 0$ for all t, i. Also, we have

$$\left|x^\top L^{-1}X_t^\top \mathcal{E}_t\right| \leq \|x^\top L^{-1}X_t^\top\|\|\mathcal{E}_t\| \leq \sqrt{2}\|x^\top L^{-1}X_t^\top\|$$

where we use $\|\mathcal{E}_t\| \leq \sqrt{2}$. We also know $\|x^\top L^{-1} X_t^\top\|$ is bounded since both X_t and x are bounded. Hence, each $x^\top L^{-1} X_t^\top \mathcal{E}_t$ is therefore a bounded random variable. This allows us to apply Hoeffding inequality for bounded random variables in Lemma A.15.

$$\begin{aligned}
\mathbb{P}\left(|x^\top L^{-1} Z_n| \geq \nu\right) &= \mathbb{P}\left(\left|\sum_{t=1}^{n} x^\top L^{-1} X_t^\top \mathcal{E}_t\right| \geq \nu\right) \\
&\leq 2\exp\left\{-\frac{2\nu^2}{\sum_{t=1}^{n}\left(2\sqrt{2}\|x^\top L^{-1} X_t^\top\|\right)^2}\right\} \\
&= 2\exp\left\{-\frac{\nu^2}{4\|x^\top L^{-1}\mathbf{X}^\top\|^2}\right\} \\
&\leq 2\exp\left\{-\frac{\kappa^2\nu^2}{4\|x\|^2_{V_n^{-1}}}\right\} \qquad \text{(A.16)}
\end{aligned}$$

where the second equality follows from the definition of $\mathbf{X}$, i.e.,

$$\sum_{t=1}^{n} \|x^\top L^{-1} X_t^\top\|^2 = \sum_{t=1}^{n} x^\top L^{-1} X_t^\top X_t L^{-1} x = x^\top L^{-1}\mathbf{X}^\top\mathbf{X} L^{-1} x = \|x^\top L^{-1}\mathbf{X}^\top\|^2 .$$

And, the last inequality follows from the fact that $L \succeq \kappa V = \kappa\mathbf{X}^\top\mathbf{X}$ and combining it with the following:

$$\|x^\top L^{-1} D^\top\|^2 = x^\top L^{-1}\mathbf{X}^\top\mathbf{X} L^{-1} x \leq \frac{1}{\kappa^2}\|x\|^2_{V_n^{-1}} .$$

Then, letting the right-hand side of (A.16) be 2δ and solving for ν, we obtain that with probability at least $1 - 2\delta$,

$$|x^\top L^{-1} Z| \leq \frac{2\sqrt{\log(1/\delta)}}{\kappa}\|x\|_{V_n^{-1}}. \qquad \text{(A.17)}$$

Then, the rest of the proof for the theorem largely follows the proof of Theorem 1 in Li,

Lu, and Zhou (2017). For the sake of completeness, we present the full proof.

$$\begin{aligned}|x^\top L^{-1}E(L+E)^{-1}Z_n| &\le \|x\|_{L^{-1}}\|L^{-1/2}E(L+E)^{-1}Z_n\| \\ &\le \|x\|_{L^{-1}}\|L^{-1/2}E(L+E)^{-1}L^{1/2}\|\|Z_n\|_{L^{-1}} \\ &\le \frac{1}{\kappa}\|x\|_{V_n^{-1}}\|L^{-1/2}E(L+E)^{-1}L^{1/2}\|\|Z_n\|_{V_n^{-1}} \end{aligned} \tag{A.18}$$

where the last inequality is from $L \succeq \kappa V_n$. Then it follows that

$$\begin{aligned}\|L^{-1/2}E(L+E)^{-1}L^{1/2}\| &= \|L^{-1/2}E(L^{-1}-L^{-1}E(L+E)^{-1})L^{1/2}\| \\ &= \|L^{-1/2}EL^{-1/2} - L^{-1/2}EL^{-1}E(L+E)^{-1}L^{1/2}\| \\ &\le \|L^{-1/2}EL^{-1/2}\| + \|L^{-1/2}EL^{-1/2}\|\|L^{-1/2}E(L+E)^{-1}L^{1/2}\|\end{aligned}$$

By solving this inequality, we get

$$\begin{aligned}\|L^{-1/2}E(L+E)^{-1}L^{1/2}\| &\le \frac{\|L^{-1/2}EL^{-1/2}\|}{1-\|L^{-1/2}EL^{-1/2}\|} \\ &\le 2\|L^{-1/2}EL^{-1/2}\| \\ &\le \frac{12\mathcal{D}}{\kappa^2\sqrt{\lambda_{\min}(V_n)}}\end{aligned}$$

where the second inequality is from (A.15) and the third inequality is from combining with (A.15). Combining with (A.18) and $\|Z_n\|_{V_n^{-1}} \le \mathcal{D}$ (which we assume to hold in this section), we have

$$\begin{aligned}|x^\top L^{-1}E(L+E)^{-1}Z_n| &\le \frac{1}{\kappa}\|x\|_{V_n^{-1}}\|L^{-1/2}E(L+E)^{-1}L^{1/2}\|\|Z_n\|_{V_n^{-1}} \\ &\le \frac{12\mathcal{D}^2}{\kappa^3\sqrt{\lambda_{\min}(V_n)}}\|x\|_{V_n^{-1}}\end{aligned} \tag{A.19}$$

Then combining the results from (A.17) and (A.19), we have

$$\begin{aligned}|x^\top(\hat{\theta}_n - \theta^*)| &\le |x^\top L^{-1} Z| + |x^\top L^{-1} E(L+E)^{-1} Z_n| \\ &\le \frac{\sqrt{\log\frac{1}{\delta}}}{\kappa}\|x\|_{V_n^{-1}} + \frac{12\mathcal{D}^2}{\kappa^3\sqrt{\lambda_{\min}(V_n)}}\|x\|_{V_n^{-1}}.\end{aligned}$$

Then it follows that $|x^\top(\hat{\theta}_n - \theta^*)| \le \frac{5}{\kappa}\sqrt{\log\frac{1}{\delta}}\|x\|_{V_n^{-1}}$ holds as long as $\lambda_{\min}(V_n) \ge \frac{9\mathcal{D}^4}{\kappa^4\log(1/\delta)}$ holds. Recall that in (A.15), we also use the condition $\lambda_{\min}(V_n) \ge \frac{144\mathcal{D}^2}{\kappa^4}$. Therefore, we require that $\lambda_{\min}(V_n) \ge \max\Big\{\frac{9\mathcal{D}^4}{\kappa^4\log(1/\delta)}, \frac{144\mathcal{D}^2}{\kappa^4}\Big\}$.

□

Lemma A.8. *For any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\|J_n(\hat{\theta}_n)\|_{V_n^{-1}} \le 4\sqrt{2d + \log\frac{1}{\delta}}. \tag{A.20}$$

Proof. This lemma is an extension of Lemma 7 in Li, Lu, and Zhou (2017). For convenience, let $Z = J_n(\hat{\theta}_n)$ and $V = V_n$. Let $\hat{\mathbb{B}}$ be a 1/2-net of the unit ball $\mathbb{B}^d$. Then $|\hat{\mathbb{B}}| \le 6^d$ (Pollard 1990, Lemma 4.1), and for any $x \in \mathbb{B}^d$, there is a $\hat{x} \in \hat{\mathbb{B}}$ such that $\|x - \hat{x}\| \le \frac{1}{2}$. Therefore, we have

$$\begin{aligned}x^\top V^{-1/2} Z &= \hat{x}^\top V^{-1/2} Z + (x-\hat{x})^\top V^{-1/2} Z \\ &= \hat{x}^\top V^{-1/2} Z + \|x-\hat{x}\| \cdot \frac{1}{\|x-\hat{x}\|}(x-\hat{x})^\top V^{-1/2} Z \\ &\le \hat{x}^\top V^{-1/2} Z + \frac{1}{2}\sup_{z\in\mathbb{B}^d} z^\top V^{-1/2} Z\,.\end{aligned}$$

Taking supremum on both sides, we get

$$\sup_{x\in\mathbb{B}^d} x^\top V^{-1/2} Z \le 2\max_{\hat{x}\in\hat{\mathbb{B}}} \hat{x}^\top V^{-1/2} Z\,.$$

Also, note that $\|Z\|_{V^{-1}} = \|V^{-1/2}Z\| = \sup_{\|x\|\le 1} x^\top V^{-1/2} Z$. Recall that $Z = \sum_{t=1}^n X_t^\top \mathcal{E}_t$. Then, it follows that

$$\begin{aligned}
\mathbb{P}\left(\|Z\|_{V^{-1}} \ge \nu\right) &\le \mathbb{P}\left(\max_{\hat{x}\in\hat{\mathbb{B}}} \hat{x}^\top V^{-1/2} Z > \frac{\nu}{2}\right) \\
&\le \sum_{\hat{x}\in\hat{\mathbb{B}}} \mathbb{P}\left(\hat{x}^\top V^{-1/2} Z > \frac{\nu}{2}\right) \\
&= \sum_{\hat{x}\in\hat{\mathbb{B}}} \mathbb{P}\left(\sum_{t=1}^n \hat{x}^\top V^{-1/2} X_t^\top \mathcal{E}_t \ge \frac{\nu}{2}\right).
\end{aligned}$$

Noting that $|\hat{x}^\top V^{-1/2} X_t^\top \mathcal{E}_t| \le \sqrt{2}\|\hat{x}^\top V^{-1/2} X_t^\top\|$, we again apply Hoeffding inequality (Lemma A.15) to a sum of bounded random variables $\hat{x}^\top V^{-1/2} X_t^\top \mathcal{E}_t$ as done in (A.16). Then, it follows that

$$\begin{aligned}
\mathbb{P}\left(\|Z\|_{V^{-1}} \ge \nu\right) &\le \sum_{\hat{x}\in\hat{\mathbb{B}}} \exp\left\{-\frac{2\nu^2}{32\sum_{t=1}^n \|\hat{x}^\top V^{-1/2} X_t^\top\|^2}\right\} \\
&= \sum_{\hat{x}\in\hat{\mathbb{B}}} \exp\left\{-\frac{\nu^2}{16\|\hat{x}^\top V^{-1/2}\mathbf{X}^\top\|^2}\right\} \\
&\le \exp\left\{-\frac{\nu^2}{16} + d\log 6\right\}
\end{aligned}$$

where the last inequality is by the fact that $|\hat{\mathbb{B}}| \le 6^d$ and the following bound on $\|\hat{x}^\top V^{-1/2}\mathbf{X}^\top\|^2$ with $V = \mathbf{X}^\top\mathbf{X}$

$$\|\hat{x}^\top V^{-1/2}\mathbf{X}^\top\|^2 = \hat{x}^\top V^{-1/2}\mathbf{X}^\top\mathbf{X}V^{-1/2}\hat{x} = \|\hat{x}\|^2 \le 1\,.$$

If we let $\nu = 4\sqrt{2d + \log(1/\delta)}$, then we have

$$\mathbb{P}\left(\|Z\|_{V^{-1}} \ge 4\sqrt{2d+\log(1/\delta)}\right) \le \exp\left\{-\frac{32d + 16\log(1/\delta)}{16} + d\log 6\right\} \le \delta.$$

□

A.4 Proof of Theorem 2.4

For suitably large $T \geq \widetilde{T} = \Omega\Big(\frac{\log^2(TN\log_2 T)}{K^2\kappa^8 d} + \frac{d^3}{K^2\kappa^8}\Big)$, setting the initialization during $T_0 = \sqrt{dT}$ would satisfy the minimum eigenvalue condition of Theorem 2.3, i.e., there exists some constant c such that $T_0 = \sqrt{dT} = \frac{c}{K\kappa^4}\sqrt{\log^2(TN\log_2 T) + d^4}$ satisfies $\lambda_{\min}(V_{T_0}) \geq \max\Big\{\frac{9\mathcal{D}^4}{\kappa^4\log(TN\log_2 T)}, \frac{144\mathcal{D}^2}{\kappa^4}\Big\}$. Note that here we choose $\delta = \frac{1}{TN\log_2 T}$. Also, it is important to note that the samples collected during the random initialization are used in the subroutine estimation for all index sets since they are also independent of samples from each index set. Therefore, once the samples from the initialization satisfies the minimum eigenvalue condition of Theorem 2.3, we can apply the confidence bound in Theorem 2.3 to each index set simultaneously satisfy the condition (since the independence condition is already ensured).

We now present two technical lemmas to help establish the cumulative expected regret in Theorem 2.4. The first lemma ensures that normality results (Theorem 2.3) holds with given confidence radius β for all items.

Lemma A.9. *Suppose that T satisfy the condition in* (2.10)*. Choose $T_0 = \sqrt{dT}$ and confidence width $\beta = \frac{5}{\kappa}\sqrt{\log(TN\log_2 T)}$. Define the following event:*

$$\mathcal{E}_t := \Big\{|m_{ti}^{(\ell)} - x_{ti}^\top \theta^*| \leq w_{ti}^{(\ell)},\ \ \forall i \in [N], \forall \ell \in [L]\Big\} \tag{A.21}$$

Then, event $\mathcal{E}_t$ holds with probability at least $1 - \mathcal{O}(T^{-1})$ for all $t \geq T_0$

The next lemma bounds the immediate regret of `supCB-MNL`, breaking down to two assortment selection scenarios — when an assortment is selected for exploitation (step (b)) or for exploration (step (c)) in Algorithm 3. Intuitively, the cumulative regret incurred by step (b) is small since the utility estimates are "accurate," that is, the uncertainty in estimated utilities are suffciently small for all items in this case. The challenge is to show

that even when we take an exploratory action in step (c), the regret incurred by such an action is not too large.

Lemma A.10. *Suppose that event $\mathcal{E}_t$ in (A.21) holds, and that in round t, the assortment S_t is chosen at stage ℓ_t. Then $S_t^* \in A_\ell$ for all $\ell \leq \ell_t$. Furthermore, we have*

$$R_t(S_t^*, \theta^*) - R_t(S_t, \theta^*) \leq \begin{cases} \dfrac{2}{\sqrt{T}}, & \text{if } S_t \text{ chosen in step } (b) \\ \dfrac{8}{2^{\ell_t}}, & \text{if } S_t \text{ chosen in step } (c) \end{cases}$$

Then, we follow the similar arguments of Li, Lu, and Zhou (2017) to show the cumulative expected regret bound. First, define $V_{\ell,t} = \sum_{t \in \Psi_\ell} \sum_{i \in S_t} x_{ti} x_{ti}^\top$, then by Lemma 2.6 and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \sum_{t \in \Psi_\ell} \max_{i \in S_t} w_{ti}^{(\ell)} &= \sum_{t \in \Psi_\ell} \max_{i \in S_t} \beta \|x_{ti}\|_{V_{\ell,t}^{-1}} \\ &\leq \beta \sqrt{2|\Psi_\ell| d \log(T/d)}. \end{aligned}$$

However, from the choices made at exploration steps (step (c)) of Algorithm 4, we know

$$2^{-\ell} |\Psi_\ell| \leq 2 \sum_{t \in \Psi_\ell} \max_{i \in S_t} w_{ti}^{(\ell)}$$

for $\ell \in \{1, ..., L\}$. Now, we combine the two inequalities above. Then it follows that

$$|\Psi_\ell| \leq 2^{\ell+1} \beta \sqrt{2|\Psi_\ell| d \log(T/d)}. \tag{A.22}$$

Note that each index set Ψ_ℓ is a disjoint set with $\cup_{\ell=0}^{L} \Psi_\ell = \{t+1, ..., T\}$. Then, we break the regret into three components – when event $\mathcal{E}_t$ in (A.21) holds, i.e., the concentration result holds, and when the event does not hold ($\mathcal{E}_t^c$), and the random initialization phase with length T_0. Note that we need the minimum eigenvalue of V_{T_0} to be larger than

the case in `UCB-MNL` but we can still use Proposition 1 to ensure such case with high probability.

$$\begin{aligned}
\mathcal{R}(T) &= \mathbb{E}\left[\sum_{t=1}^{T}\left(R(S^*,\theta^*) - R(S_t,\theta^*)\right)\right] \\
&\le T_0 + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(R(S^*,\theta^*) - R(S_t,\theta^*)\right)\mathbb{1}\left(\mathcal{E}_t\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=T_0+1}^{T}\left(R(S^*,\theta^*) - R(S_t,\theta^*)\right)\mathbb{1}\left(\mathcal{E}_t^c\right)\right]
\end{aligned}$$

We further decompose the regret into the disjoint stages recorded by Ψ_ℓ.

$$\begin{aligned}
\mathcal{R}(T) &\le T_0 + \mathbb{E}\left[\sum_{t\in\Psi_0}\left(R(S^*,\theta^*) - R(S_t,\theta^*)\right)\mathbb{1}\left(\mathcal{E}_t\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{t\in\Psi_\ell}\left(R(S^*,\theta^*) - R(S_t,\theta^*)\right)\mathbb{1}\left(\mathcal{E}_t\right)\right] + \mathcal{O}(1) \\
&\le T_0 + \frac{2}{\sqrt{T}}|\Psi_0| + \sum_{\ell=1}^{L}\frac{8}{2^\ell}|\Psi_\ell| + \mathcal{O}(1) \\
&\le T_0 + 2\sqrt{T} + \sum_{\ell=1}^{L} 16\beta\sqrt{2|\Psi_\ell| d\log(T/d)} + \mathcal{O}(1) \\
&\le T_0 + 2\sqrt{T} + 16\beta\sqrt{2dLT\log(T/d)} + \mathcal{O}(1)
\end{aligned}$$

where the third inequality uses (A.22) and the last inequality is by Cauchy-Schwartz inequality. Now, with our choices of confidence with $\beta = \frac{5}{\kappa}\sqrt{\log(TN\log_2 T)}$ and initialization $T_0 = \sqrt{dT}$ and epoch length $L = \lfloor \frac{1}{2}\log_2 T \rfloor \le \frac{1}{2}\log_2 T$, we complete the proof.

A.5 Proofs of Lemmas for Theorem 2.4

A.5.1 Proof of Lemma 2.8

Proof. Since a time-stamp t can only be added to $\Psi_\ell, \ell \geq 1$ in step (c) of Algorithm 4, the event $\{t \in \Psi_\ell\}$ only depends on the results of trials $t' \in \cup_{\ell' < \ell} \Psi_{\ell'}$ and on $\bar{w}_{ti}^{(\ell)}$. From the definition of $\bar{w}_{ti}^{(\ell)}$, we know it only depends on the sets of feature vectors $\{x_{u,i}\}_{i \in S_u}, u \in \Psi_\ell$ and on $\{x_{ti}\}_{i \in S_t}$. □

A.5.2 Proof of Lemma A.9

Proof. With $T_0 = \sqrt{dT}$ and $T \geq \widetilde{T}$ where $\widetilde{T}$ is defined as (2.10), at the end of random initialization, we can show that there exists a large enough constant c such that $T_0 = \frac{c}{K\kappa^4}\sqrt{\log^2(TN\log_2 T) + d^4}$ satisfies

$$\lambda_{\min}(V_{T_0}) \geq \max\left\{\frac{9\mathcal{D}^4}{\kappa^4 \log(TN\log_2 T)}, \frac{144\mathcal{D}^2}{\kappa^4}\right\}$$

with high probability using Proposition 1. Then, the condition on the minimum eigenvalue of V_t for any $t \geq T_0$ is also satisfied since $\lambda_{\min}(V_t) \geq \lambda_{\min}(V_{T_0})$ for all $t \geq T_0$. Then the minimum eigenvalue condition is satisfied for all sub-routine estimation since the samples in the initialization period is shared across all index sets, i.e., `baseCB-MNL` is run on samples in $\Psi_\ell \cup [T_0]$ for all ℓ in step (a). Therefore, applying Theorem 2.3 with confidence width $\beta = \frac{5}{\kappa}\sqrt{\log(TN\log_2 T)}$, we can show

$$|m_{ti}^{(\ell)} - x_{ti}\theta^*| \leq w_{ti}^{(\ell)}$$

holds for all $i \in [N]$, $\ell \in [L]$, and $t \in \{T_0 + 1, ..., T\}$ with probability at least $1 - \frac{3}{TN\log_2 T}$. Now, applying the union bound over all items and epochs, we complete the proof. □

A.5.3 Proof of Lemma A.10

Proof. Combining Lemma 2.4 and Lemma 2.5, we have

$$\left|R_t(S,\theta^*) - R_t(S,\hat{\theta}^{(\ell)})\right| \le \left|\widetilde{R}_t(S,\hat{\theta}^{(\ell)}) - R_t(S,\hat{\theta}^{(\ell)})\right| \le 2\max_{i\in S} w_{ti}^{(\ell)} \le \mathcal{W}_t^{(\ell)}.$$

We first show the optimal assortment $S_t^* \in A_\ell$ for all ℓ. We prove this by induction. For $\ell = 1$, the lemma automatically holds. As an inductive step, suppose $S_t^* \in A_\ell$ and we want to prove $S_t^* \in A_{\ell+1}$. Since the algorithm proceed to stage $\ell+1$, we know from step (c) in Algorithm 4 that

$$\left|R_t(S,\theta^*) - R_t(S,\hat{\theta}^{(\ell)})\right| \le \mathcal{W}_t^{(\ell)} \le 2^{-\ell}$$

for all $S \in A_\ell$. In particular, it holds for $S = S_t^*$ since $S_t^* \in A_\ell$ by the inductive step. Then the optimality of S_t^* implies

$$R_t(S_t^*,\hat{\theta}^{(\ell)}) \ge R_t(S_t^*,\theta^*) - 2^{-\ell} \ge R_t(S,\theta^*) - 2^{-\ell} \ge R_t(S,\hat{\theta}^{(\ell)}) - 2\cdot 2^{-\ell}$$

for $S \in A_\ell$. Hence, it follows that

$$R_t(S_t^*,\hat{\theta}^{(\ell)}) \ge \max_{S\in A_\ell} R_t(S,\hat{\theta}^{(\ell)}) - 2\cdot 2^{-\ell} = \mathcal{M}_t^{(\ell)} - 2\cdot 2^{-\ell}.$$

Therefore, we have $S_t^* \in A_{\ell+1}$ according to step (d). If S_t is selected in step (b), that it implies $R_t(S_t,\hat{\theta}^{(\ell_t)}) \ge R_t(S_t^*,\hat{\theta}^{(\ell_t)})$. Then if follows that

$$R_t(S_t,\theta^*) \ge R_t(S_t,\hat{\theta}^{(\ell_t)}) - \frac{1}{\sqrt{T}} \ge R_t(S_t^*,\hat{\theta}^{(\ell_t)}) - \frac{1}{\sqrt{T}} \ge R_t(S_t^*,\theta^*) - \frac{2}{\sqrt{T}}.$$

Suppose S_t is chose at stage ℓ_t in step (c) in Algorithm 4. The lemma holds automat-

ically for $\ell_t = 1$ since $R_t(S, \theta^*) \in [0, 1]$ for all S and t. If $\ell_t > 1$, S_t must have passed through steps (c) and (d) in the previous stage, $\ell_t - 1$. Also note that we have already shown that the optimal assortment $S_t^* \in A_{\ell_t}$. Hence, S_t^* also must have passed through steps (c) and (d) in stage $\ell_t - 1$. Therefore, passing through step (c) at stage $\ell_t - 1$ implies that we can bound

$$\left| R_t(S, \hat{\theta}^{(\ell_t - 1)}) - R_t(S, \theta^*) \right| \leq \mathcal{W}_t^{(\ell_t - 1)} \leq 2^{-(\ell_t - 1)}$$

for $S = S_t$ and $S = S_t^*$. Also, for step (d) at stage $\ell_t - 1$ implies that

$$R_t(S_t^*, \hat{\theta}^{(\ell_t - 1)}) - R_t(S_t, \hat{\theta}^{(\ell_t - 1)}) \leq 2 \cdot 2^{-(\ell_t - 1)}$$

Combining these inequalities above, we have

$$\begin{aligned} R_t(S_t, \theta^*) &\geq R_t(S_t, \hat{\theta}^{(\ell_t - 1)}) - 2^{-(\ell_t - 1)} \\ &\geq R_t(S_t^*, \hat{\theta}^{(\ell_t - 1)}) - 3 \cdot 2^{-(\ell_t - 1)} \\ &\geq R_t(S_t^*, \theta^*) - 4 \cdot 2^{-(\ell_t - 1)}. \end{aligned}$$

□

A.6 Proof of Theorem 2.5

Following the proof outline presented in Section 2.6.3, we first define the notations which are used throughout our analysis in this section.

Definition A.4. *Let L be the last episode over horizon, i.e., for a given time horizon T, $L := \lfloor \log_2 T \rfloor + 1$. We let $\mathcal{T}_k$ denote an index set of all rounds that belong to the k-th episode $\mathcal{T}_k := \{\tau_{k-1} + 1, ..., \tau_k\}$.*

Appendix A: UCB Algorithms for MNL Contextual Bandits

By the design of the DBL-MNL algorithm, the length of the k-th episode is $|\mathcal{T}_k| = \tau_k/2$ where τ_k is the last period of the k-th episode. We aim to bound the cumulative regret for each episode $\texttt{Reg}(k\text{-th episode})$ so that $\mathcal{R}(T) = \sum_{k=1}^{L} \texttt{Reg}(k\text{-th episode})$ is also bounded. As briefly discussed in Section 2.6.3, there are two scenarios for a given episode.

(i) $|\mathcal{T}_k| \leq q_k$: In this case, the length of an episode is not large enough to ensure the concentration of the prediction error due to the failure to ensure the lower bound on $\lambda_{\min}(V_t)$. Therefore, we cannot control the regret in this case. However, the number of such rounds is only logarithmic in T, hence the regret corresponding to this case contributes minimally to the total regret.

(ii) $|\mathcal{T}_k| > q_k$: We can apply the fast convergence result in Theorem 2.3 as long as the lower bound on $\lambda_{\min}(V_t)$ is guaranteed — note that the independence condition is already satisfied since samples in each episode are independent of each other. We show that $\lambda_{\min}(V_t)$ grows linearly as t increases in each episode with high probability. In case of $\lambda_{\min}(V_t)$ not growing as fast as the rate we require, we perform random sampling to satisfy this criterion towards the end of each episode. Therefore, with high probability, the lower bound on $\lambda_{\min}(V_t)$ becomes satisfied.

For case (i), clearly $q_k \leq q_L$ for any $k \in \{1, ..., L\}$. $|\mathcal{T}_k|$ eventually grows to be larger than q_L for some k since q_L is logarithmic in T. Let k' be the first episode such that $|\mathcal{T}_{k'}| \geq q_L$. Hence, $|\mathcal{T}_{k'}| \leq 2q_L$. Thus, the cumulative regret prior to the k'-th episode is

$$\sum_{k=1}^{k'-1} \texttt{Reg}(k\text{-th episode}) \leq \sum_{k=1}^{k'-1} |\mathcal{T}_k| = |\mathcal{T}_{k'}| \leq 2q_L = \mathcal{O}\Big(\log d + d^2 + \log^2(TN)\Big) \,.$$

Then, letting k'' be the first episode such that $|\mathcal{T}_{k''}| \geq q_{k''}$ and noting that $k'' \leq k'$ gives

$$\sum_{k=1}^{k''-1} \texttt{Reg}(k\text{-th episode}) \leq \sum_{k=1}^{k'-1} \texttt{Reg}(k\text{-th episode}) \,.$$

Hence, the cumulative regret corresponding to case (i) is at most poly-logarithic in T.

For case (ii), it suffices to show random sampling ensures the growth of $\lambda_{\min}(V_t)$. Lemma A.11 shows that random sampling with duration q_k specified in Theorem 2.5 ensures the lower bound of $\lambda_{\min}(V_t)$, i.e., $\lambda_{\min}(V_t) \geq \max\left\{\frac{9\mathcal{D}_k^4}{\kappa^4 \log(\tau_k N/2)}, \frac{144\mathcal{D}_k^2}{\kappa^4}\right\}$ with high probability.

Lemma A.11. *Suppose*

$$q_k = \frac{2}{\sigma_0 K} \max\left\{\frac{9\mathcal{D}_k^4}{\kappa^4 \log(\tau_k N/2)}, \frac{144\mathcal{D}_k^2}{\kappa^4}\right\}$$
$$\text{where } \mathcal{D}_k = \min\left\{4\sqrt{2d + \log(\tau_k N/2)}, \sqrt{d\log(\tau_k/d) + 2\log(\tau_k N/2)}\right\}.$$

Then, for the k-th episode, with probability at least $1 - d\exp\left\{-\frac{q_k\sigma_0}{10}\right\}$*, we have*

$$\lambda_{\min}(V_{\tau_k}) \geq \max\left\{\frac{9\mathcal{D}_k^4}{\kappa^4 \log(\tau_k N/2)}, \frac{144\mathcal{D}_k^2}{\kappa^4}\right\}. \tag{A.23}$$

Remark A.1. *We emphasize that the assumption* $K \leq \frac{18}{\kappa^4}$ *is not restrictive. In fact, we can instead use Proposition 1 to show that*

$$q_k = \frac{C}{K} \max\left\{\frac{d^2 + \log(\tau_k^2 N/4)}{\sigma_0 \kappa^4}, \frac{d + 2\log(\tau_k/2)}{\sigma_0^2}\right\}$$

for some constant C satisfies the threshold on $\lambda_{\min}(V_{\tau_k})$ *without assuming* $K \leq \frac{18}{\kappa^4}$*. However, we provide a specific value of* q_k *which does not depend on an additional unknown constant since* q_k *is an input to the algorithm. Furthermore, in many real-world applications, K is typically small; hence* $K \leq \frac{18}{\kappa^4}$ *(recall that* $\kappa \in (0, 1)$*) is a reasonable assumption.*

We then apply Theorem 2.3 to prediction error in the k-th episode which requires samples in the $(k-1)$-th episode are independent and $\lambda_{\min}(V_{\tau_{k-1}})$ at the end of the

$(k-1)$-th episode is large enough. With a lower bound guarantee on $\lambda_{\min}(V_{\tau_{k-1}})$ from Lemma A.11 and the fact that samples are independent of each other within each episode, we have with probability at least $1 - \frac{6}{\tau_k N}$

$$|x_{ti}^\top(\hat{\theta}_k - \theta^*)| \le \beta_k \|x_{ti}\|_{W_{k-1}^{-1}}$$

where $\beta_k = \frac{5}{\kappa}\sqrt{\log(\tau_k N/2)}$. Recall that $W_{k-1} = V_{\tau_{k-1}} = \sum_{t'=\tau_{k-1}+1}^{\tau_{k-1}} \sum_{i \in S_{t'}} x_{t'i} x_{t'i}^\top$ is the Gram matrix at the end of the $(k-1)$-th episode. Then, we can use the union bound to show this concentration result for all items and all rounds within the episode.

$$|x_{ti}^\top(\hat{\theta}_k - \theta^*)| \le \beta_k \|x_{ti}\|_{W_{k-1}^{-1}}, \ \ \forall i \in [N], \forall t \in \mathcal{T}_k \,. \tag{A.24}$$

Let $\widetilde{\mathcal{E}}_{k,1}$ and $\widetilde{\mathcal{E}}_{k,2}$ denote the event that the minimum eigenvalue condition in (A.23) holds (at the end of the $(k-1)$-th episode) and the event that the MLE concentration result in (A.24) holds respectively.

$$\begin{aligned}
\widetilde{\mathcal{E}}_{k,1} &:= \left\{ \lambda_{\min}(V_{\tau_{k-1}}) \ge \max\left\{ \frac{9\mathcal{D}_{k-1}^4}{\kappa^4 \log(\tau_{k-1} N/2)}, \frac{144 \mathcal{D}_{k-1}^2}{\kappa^4} \right\} \right\} \\
\widetilde{\mathcal{E}}_{k,2} &:= \left\{ |x_{ti}^\top(\hat{\theta}_k - \theta^*)| \le \beta_k \|x_{ti}\|_{W_{k-1}^{-1}}, \forall i \in [N], \forall t \in \mathcal{T}_k \right\} .
\end{aligned}$$

On the joint event $\widetilde{\mathcal{E}}_{k,1} \cap \widetilde{\mathcal{E}}_{k,2}$, by the definition of the upper confidence bound of an utility estimate $\widetilde{z}_{ti}$ and following the same arguments as Lemma 2.3, we have

$$0 \le \widetilde{z}_{ti} - x_{ti}^\top \theta^* \le 2\beta_k \|x_{ti}\|_{W_{k-1}^{-1}} \,.$$

Therefore, the optimistic expected revenue $\widetilde{R}_t(S)$ based on $\{\widetilde{z}_{ti}\}$ is computed the same way as (2.6). It is important to note that while the formation of the optimistic revenue $\widetilde{R}_t(S)$ is identical to (2.6), the actual values of $\widetilde{R}_t(S)$ are different for the two algorithms,

UCB-MNL and DBL-MNL. In particular, when feature dimension d is large, the confidence bound of $\widetilde{R}_t(S)$ in DBL-MNL can be much tighter than that of UCB-MNL since the confidence width β_k for DBL-MNL does not have dependence on d.

Let $S_t = \text{argmax}_{S\in\mathcal{S}} \widetilde{R}_t(S)$. Then, it follows that $\widetilde{R}_t(S_t) \geq R(S_t^*, \theta^*)$ following from Lemma 2.4. Thus, we can bound the regret in the k-th episode as follows:

$$\begin{aligned}\text{Regret}(k) &= \sum_{t\in\mathcal{T}_k} \left(R(S_t^*, \theta^*) - R(S_t, \theta^*)\right) \mathbb{1}(\widetilde{\mathcal{E}}_{k,1} \cap \widetilde{\mathcal{E}}_{k,2}) \\ &\leq \sum_{t\in\mathcal{T}_k} \left(\widetilde{R}(S_t) - R(S_t, \theta^*)\right) \mathbb{1}(\widetilde{\mathcal{E}}_{k,1} \cap \widetilde{\mathcal{E}}_{k,2})\end{aligned}$$

Then, by the Lipschitz property of the expected revenue of the MNL model shown in Lemma 2.5, it follows that

$$\begin{aligned}\sum_{t\in\mathcal{T}_k} \left(\widetilde{R}(S_t) - R(S_t, \theta^*)\right) \mathbb{1}(\widetilde{\mathcal{E}}_k) &\leq \sum_{t\in\mathcal{T}_k} \sum_{i\in S_t} \left| x_{ti}^\top (\hat{\theta}_k - \theta^*) + \beta_k \|x_{ti}\|_{W_{k-1}^{-1}} \right| \\ &\leq 2\beta_k \sum_{t\in\mathcal{T}_k} \sum_{i\in S_t} \|x_{ti}\|_{W_{k-1}^{-1}}\end{aligned}$$

where the last inequality is from (A.24). Then we use Lemma A.12 to bound using the norm using the current Gram matrix. This result utilizes the fact that the minimum eigenvalue of the Gram matrix grows linearly within each episode since the samples are independent from each other, allowing us to use the matrix Chernoff inequality to the sum of independent matrices. Furthermore, the fact that episode length difference is two-fold for adjacent episodes allows us to bound the difference between the Gram matrices.

Lemma A.12. *For $t \in \mathcal{T}_k$,*

$$\sum_{t\in\mathcal{T}_k} \sum_{i\in S_t} \|x_{ti}\|_{W_{k-1}^{-1}} \leq C_1 \sum_{t\in\mathcal{T}_k} \sum_{i\in S_t} \|x_{ti}\|_{V_{t-1}^{-1}}$$

with probability at least $1 - de^{-C_2(t-\tau_{k-1})}$ for some constants C_1 and C_2.

Let $\widetilde{\mathcal{E}}_{k,3}$ denote the event that Lemma A.12 holds for the k-th episode.

$$\widetilde{\mathcal{E}}_{k,3} := \left\{ \sum_{t\in\mathcal{T}_k} \sum_{i\in S_t} \|x_{ti}\|_{W_{k-1}^{-1}} \le C_1 \sum_{t\in\mathcal{T}_k} \sum_{i\in S_t} \|x_{ti}\|_{V_{t-1}^{-1}}, \forall t \in \mathcal{T}_k \right\}$$

On this event along with , it follows that

$$\begin{aligned}
\sum_{t\in\mathcal{T}_k} \left(\widetilde{R}(S_t) - R(S_t,\theta^*)\right) \mathbb{1}(\widetilde{\mathcal{E}}_{k,1}\cap\widetilde{\mathcal{E}}_{k,2}\cap\widetilde{\mathcal{E}}_{k,3}) &\le 2C_1\beta_k \sum_{t\in\mathcal{T}_k}\sum_{i\in S_t} \|x_{ti}\|_{V_{t-1}^{-1}} \\
&\le 2C_1\beta_k \sqrt{\frac{\tau_k}{2}\sum_{t\in\mathcal{T}_k}\sum_{i\in S_t} \|x_{ti}\|^2_{V_{t-1}^{-1}}} \\
&\le 2C_1\beta_k \sqrt{\tau_k d \log\left(\frac{\tau_k}{2d}\right)}
\end{aligned}$$

where we use the Cauchy-Schwarz inequality in the second inequality and apply the bound on the self-normalized process in Lemma 2.6 in the last inequality. Thus, when events $\widetilde{\mathcal{E}}_k$ and $\widetilde{\mathcal{E}}_{k,3}$ hold, the regret in the k-th episode is bounded by

$$\sum_{t\in\mathcal{T}_k} \left(R(S_t^*,\theta^*) - R(S_t,\theta^*)\right) \mathbb{1}(\widetilde{\mathcal{E}}_{k,1}\cap\widetilde{\mathcal{E}}_{k,2}\cap\widetilde{\mathcal{E}}_{k,3}) = \mathcal{O}\left(\sqrt{d\tau_k \log(\tau_k/d)\log(\tau_k N)}\right)$$

On the other hand, the regret in the episode under the failure events of $\widetilde{\mathcal{E}}_{k,1}$, $\widetilde{\mathcal{E}}_{k,2}$, and $\widetilde{\mathcal{E}}_{k,3}$ are bounded by

$$\begin{aligned}
\sum_{t\in\mathcal{T}_k} \left(R(S_t^*,\theta^*) - R(S_t,\theta^*)\right) \mathbb{1}(\widetilde{\mathcal{E}}_{k,1}^c) &= \widetilde{\mathcal{O}}(d) \\
\sum_{t\in\mathcal{T}_k} \left(R(S_t^*,\theta^*) - R(S_t,\theta^*)\right) \mathbb{1}(\widetilde{\mathcal{E}}_{k,2}^c) &= \widetilde{\mathcal{O}}(1) \\
\sum_{t\in\mathcal{T}_k} \left(R(S_t^*,\theta^*) - R(S_t,\theta^*)\right) \mathbb{1}(\widetilde{\mathcal{E}}_{k,3}^c) &= \widetilde{\mathcal{O}}(d)\,.
\end{aligned}$$

Therefore, summing over all episodes, the cumulative expected regret is given by

$$\mathcal{R}(T) = \mathcal{O}\left(\sqrt{dT \log(T/d) \log(TN) \log_2 T}\right)$$

A.6.1 Proof of Lemma A.11

Proof. By the design of Algorithm 5, it suffices to show that the random sampling for duration q_k provides sufficient growth of $\lambda_{\min}(V_{\tau_k})$. Let $\widetilde{\mathcal{T}}_k$ be the set of rounds in the k-th episode that random sampling is performed. Without loss of generality, assume that the random initialization is invoked for the full duration q_k (note that Algorithm 5 may not invoke random sampling at all if the minimum eigenvalue condition is already satisfied). Hence, $\widetilde{\mathcal{T}}_k = \{\tau_k - q_k + 1, \tau_k\}$ in this case. First, under random sampling of S_t, we have

$$\begin{aligned}
\lambda_{\min}\left(\sum_{t\in\widetilde{\mathcal{T}}_k}\sum_{i\in S_t}\mathbb{E}\left[x_{ti}x_{ti}^\top\right]\right) &= \lambda_{\min}\left(\sum_{t\in\widetilde{\mathcal{T}}_k} K\mathbb{E}\left[\frac{1}{N}\sum_{j\in[N]} x_{tj}x_{tj}^\top\right]\right) \\
&\geq \sum_{t\in\widetilde{\mathcal{T}}_k} K\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{N}\sum_{j\in[N]} x_{tj}x_{tj}^\top\right]\right) \\
&\geq q_k K\sigma_0 \qquad \text{(A.25)}
\end{aligned}$$

where the first inequality is from the fact that the minimum eigenvalue function $\lambda_{\min}(\cdot)$ is concave over positive semi-definite matrices. We also use the fact that in the uniform revenue setting, the size of the assortment is $|S_t| = K$ for all t. Then, since $\|x_{ti}\| \leq 1$ for all t and i, we can upper-bound the maximum eigenvalue

$$\lambda_{\max}\left(\sum_{i\in S_t} x_{ti}x_{ti}^\top\right) \leq K$$

for all t. Let $\widetilde{V}_k := \sum_{t\in\widetilde{\mathcal{T}}_k}\sum_{i\in S_t} x_{ti}x_{ti}^\top$. Then, we can use the matrix Chernoff inequality shown in Lemma A.16 (Corollary 5.2 of Tropp (2012)).

$$\begin{aligned}\mathbb{P}\left(\lambda_{\min}\left(\widetilde{V}_k\right) \le \frac{q_k K\sigma_0}{2}\right) &\le \mathbb{P}\left(\lambda_{\min}\left(\widetilde{V}_k\right) \le \frac{1}{2}\cdot\lambda_{\min}\left(\mathbb{E}\left[\widetilde{V}_k\right]\right)\right)\\ &\le d\left(\frac{e^{-1/2}}{(1/2)^{1/2}}\right)^{\lambda_{\min}\left(\mathbb{E}\left[\widetilde{V}_k\right]\right)/K}\\ &\le d\exp\left\{-\frac{\lambda_{\min}\left(\mathbb{E}\left[\widetilde{V}_k\right]\right)}{10K}\right\}\end{aligned}$$

where we use the fact that $-\frac{1}{2}-\frac{1}{2}\log\left(\frac{1}{2}\right) \le -\frac{1}{10}$ in the last inequality. Then using the lower bound of $\mathbb{E}\left[\widetilde{V}_k\right]$ in (A.25), it follows that

$$\mathbb{P}\left(\lambda_{\min}\left(\widetilde{V}_k\right) \le \frac{q_k K\sigma_0}{2}\right) \le d\exp\left\{-\frac{q_k K\sigma_0}{10K}\right\} = d\exp\left\{-\frac{q_k\sigma_0}{10}\right\}.$$

Then, by our choice of q_k with $q_k = \frac{2}{\sigma_0 K}\max\left\{\frac{9\mathcal{D}_k^4}{\kappa^4\log(\tau_k N/2)}, \frac{144\mathcal{D}_k^2}{\kappa^4}\right\}$, we have

$$\mathbb{P}\left(\lambda_{\min}\left(\widetilde{V}_k\right) \le \max\left\{\frac{9\mathcal{D}_k^4}{\kappa^4\log(\tau_k N/2)}, \frac{144\mathcal{D}_k^2}{\kappa^4}\right\}\right) \le d\exp\left\{-\frac{q_k\sigma_0}{10}\right\}.$$

□

A.6.2 Proof of Lemma A.12

Proof. Recall that W_{k-1} is the Gram matrix at the end of the $(k-1)$-th episode, i.e., $V_{\tau_{t-1}}$ before it resets at the beginning of the k-th episode. Since V_t resets at the beginning of each episode, we focus on how V_t grows in the k-th episode relative to W_{k-1}, the Gram matrix at the end of the previous episode. Clearly, if $CW_{k-1} \succcurlyeq V_t$, for all $t \in \{\tau_{k-1}+1, \tau_k\}$ for some constant C, then the claim holds. Then it suffices to show $\lambda_{\min}(V_t)$ grows linearly as t increases during the $(k-1)$-th episode. In fact, since $\mathcal{X}$ is time-invariant, we show the $\lambda_{\min}(V_t)$ grows linearly with t in all episodes.

Appendix A: UCB Algorithms for MNL Contextual Bandits

Let $\widetilde{\theta}_{k,t}$ be the parameter corresponding to the upper confidence reward at round t, $\max_{S\in\mathcal{S}} \widetilde{R}_t(S)$. Note that $\widetilde{\theta}_{k,t}$ is not the same as the MLE $\hat{\theta}_k$. Since we take an UCB action in Algorithm 5, this is equivalent to taking some optimistic parameter within the confidence ellipsoid centered at $\hat{\theta}_k$. It is important to note that since we do not update the MLE and confidence bound within each episode, the samples y_t's are still independent from each other in the same episode.

Consider $\{(i_1, ..., i_N)\}$, a set of all permutations of integers $\{1, , , N\}$. Without loss of generality, assume N is divisible by K. Then we can write

$$\begin{aligned}
\mathbb{E}\left[X_{ti}X_{ti}^\top\right] &= \frac{1}{N}\mathbb{E}\left[X_{t1}X_{t1}^\top + ... + X_{tN}X_{tN}^\top\right] \\
&= \frac{1}{N}\sum_{(i_1,...,i_N)} \mathbb{E}\left[(X_{t,i_1}X_{t,i_1}^\top + ... + X_{t,i_N}X_{t,i_N}^\top)\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t} < \cdots < X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right] \\
&\preccurlyeq \frac{1}{N}\sum_{(i_1,...,i_N)} \frac{N}{K}C_X\mathbb{E}\left[(\mathbf{V}_{t,\min}(\mathcal{I}) + \mathbf{V}_{t,\max}(\mathcal{I}))\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t} < \cdots < X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right]
\end{aligned}$$

where $\mathbf{V}_{t,\min}(\mathcal{I})$ and $\mathbf{V}_{t,\max}(\mathcal{I})$ are the first and last K sums respectively under ordering $\mathcal{I} = (i_1, ..., i_N)$. That is,

$$\begin{aligned}
\mathbf{V}_{t,\min}(\mathcal{I}) = \mathbf{V}_{t,\min}(i_1, ..., i_N) &:= X_{t,i_1}X_{t,i_1}^\top + ... + X_{t,i_K}X_{t,i_K}^\top \\
\mathbf{V}_{t,\max}(\mathcal{I}) = \mathbf{V}_{t,\max}(i_1, ..., i_N) &:= X_{t,i_{N-K+1}}X_{t,i_{N-K+1}}^\top + ... + X_{t,i_N}X_{t,i_N}^\top
\end{aligned}$$

Note that the last inequality holds since $C_X(\mathbf{V}_{\min}(\mathcal{I}) + \mathbf{V}_{\max}(\mathcal{I}))$ dominates any K sum in $\{X_{t,i_1}X_{t,i_1}^\top, ..., X_{t,i_N}X_{t,i_N}^\top\}$ which is shown in Lemma A.13. Note that Lemma A.13 shows the result under any vector θ, hence can be applied here.

$$\begin{aligned}
\mathbb{E}\left[X_{ti}X_{ti}^\top\right] &\preccurlyeq \frac{C_X}{K}\sum_{(i_1,\ldots,i_N)}\mathbb{E}\left[(\mathbf{V}_{t,\min}(\mathcal{I})+\mathbf{V}_{t,\max}(\mathcal{I}))\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t}<\cdots<X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right]\\
&\preccurlyeq \frac{C_X(\rho_0+1)}{K}\sum_{(i_1,\ldots,i_N)}\mathbb{E}\left[\mathbf{V}_{t,\max}(\mathcal{I})\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t}<\cdots<X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right]\\
&\preccurlyeq \frac{2C_X\rho_0}{K}\mathbb{E}\left[\sum_{X_{ti}\in\mathcal{X}_t}X_{ti}X_{ti}^\top\mathbb{1}\Big(X_{ti}\in S_t\Big)\right]
\end{aligned}$$

where the second inequality comes from utilizing the relaxed symmetry (Assumption 2.4) and Lemma A.14. The last inequality uses the fact that $S_t = \operatorname{argmax}_{S\in\mathcal{S}}\widetilde{R}_t(S)$ and $\rho_0 \geq 1$. Therefore,

$$\mathbb{E}\left[\sum_{X_{ti}\in\mathcal{X}_t}X_{ti}X_{ti}^\top\mathbb{1}\Big(X_{ti}\in S_t\Big)\right] \succcurlyeq \frac{K}{2C_X\rho_0}\mathbb{E}\left[X_{ti}X_{ti}^\top\right].$$

Now, for $t\in\mathcal{T}_k$, we define

$$\Sigma_{k,t} := \sum_{t'=\tau_{k-1}+1}^{t}\mathbb{E}\left[\sum_{X_{t'i}\in\mathcal{X}_{t'}}X_{t'i}X_{t'i}^\top\mathbb{1}\Big(X_{t'i}\in S_{t'}\Big)\right].$$

Then, since the minimum eigenvalue function $\lambda_{\min}(\cdot)$ is concave over positive semi-definite matrices, we have

$$\begin{aligned}
\lambda_{\min}\left(\Sigma_{k,t}\right) &= \lambda_{\min}\left(\sum_{t'=\tau_{k-1}+1}^{t}\mathbb{E}\left[\sum_{X_{t'i}\in\mathcal{X}_{t'}}X_{t'i}X_{t'i}^\top\mathbb{1}\Big(X_{t'i}\in S_{t'}\Big)\right]\right)\\
&\geq \sum_{s=\tau_{k-1}+1}^{t}\lambda_{\min}\left(\mathbb{E}\left[\sum_{X_{t'i}\in\mathcal{X}_{t'}}X_{t'i}X_{t'i}^\top\mathbb{1}\Big(X_{t'i}\in S_{t'}\Big)\right]\right)\\
&\geq \frac{K(t-\tau_{k-1})\sigma_0}{2\rho_0 C_{\mathcal{X}}} > 0\,. \qquad \text{(A.26)}
\end{aligned}$$

Now, to apply the matrix concentration inequality, we need to show an upper bound on the maximum eigenvalue of $\mathbb{E}\left[\sum_{X_{t'i}\in\mathcal{X}_{t'}}X_{t'i}X_{t'i}^\top\mathbb{1}\Big(X_{t'i}\in S_{t'}\Big)\right]$. We use the fact that

$\|X_{t'i}\| \leq 1$ is bounded. Hence, we have for all τ

$$\lambda_{\max}\left(\mathbb{E}\left[\sum_{X_{t'i}\in\mathcal{X}_{t'}} X_{t'i}X_{t'i}^\top \mathbb{1}\Big(X_{t'i}\in S_{t'}\Big)\right]\right) \leq K\,.$$

Then we can apply Corollary 5.2 in Tropp (2012) to the finite sequence of independent matrices V_t for $t \in \mathcal{T}_k$.

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\min}(V_t) \leq \frac{K(t-\tau_{k-1})\sigma_0}{2\rho_0 C_{\mathcal{X}}}\right) &\leq d\left(\frac{e^{-1/2}}{0.5^{1/2}}\right)^{\frac{(t-\tau_{k-1})\sigma_0}{2\rho_0 C_{\mathcal{X}}}} \\
&= d\exp\left\{\frac{(t-\tau_{k-1})\sigma_0}{2\rho_0 C_{\mathcal{X}}}\log\left(\frac{e^{-1/2}}{0.5^{1/2}}\right)\right\} \\
&\leq d\exp\left\{-\frac{(t-\tau_{k-1})\sigma_0}{20\rho_0 C_{\mathcal{X}}}\right\}
\end{aligned}$$

where the last inequality uses $-\frac{1}{2}-\frac{1}{2}\log\frac{1}{2} \leq -\frac{1}{10}$. Therefore, $\lambda_{\min}(V_t)$ grows linearly as t grows within the episode with probability at least $1 - d\exp\{-(t-\tau_{k-1})\sigma_0/(20\rho_0 C_{\mathcal{X}})\}$. This completes the proof. □

Lemma A.13. *Consider $\{(i_1,...,i_N)\}$, a set of all permutations of $\{1,,,N\}$. Let $\mathbf{V}_{t,\min}(\mathcal{I})$ and $\mathbf{V}_{t,\max}(\mathcal{I})$ be the first and last K sums respectively under given $\mathcal{I}=(i_1,...,i_N)$:*

$$\begin{aligned}
\mathbf{V}_{t,\min}(\mathcal{I}) &= \mathbf{V}_{t,\min}(i_1,...,i_N) := X_{t,i_1}X_{t,i_1}^\top + ... + X_{t,i_K}X_{t,i_K}^\top \\
\mathbf{V}_{t,\max}(\mathcal{I}) &= \mathbf{V}_{t,\max}(i_1,...,i_N) := X_{t,i_{N-K+1}}X_{t,i_{N-K+1}}^\top + ... + X_{t,i_N}X_{t,i_N}^\top
\end{aligned}$$

Then for any fixed vector θ, there is some C_X such that

$$\begin{aligned}
&\sum_{(i_1,...,i_N)} \mathbb{E}\left[(X_{t,i_1}X_{t,i_1}^\top + ... + X_{t,i_N}X_{t,i_N}^\top)\mathbb{1}\{X_{t,i_1}^\top\theta < \cdots < X_{t,i_N}^\top\theta\}\right] \\
&\preccurlyeq \sum_{(i_1,...,i_N)} \frac{N}{K}C_X\mathbb{E}\left[(\mathbf{V}_{t,\min}(\mathcal{I}) + \mathbf{V}_{t,\max}(\mathcal{I}))\mathbb{1}\{X_{t,i_1}^\top\theta < \cdots < X_{t,i_N}^\top\theta\}\right].
\end{aligned}$$

Proof. Here, we use Proposition 4 in Appendix C which shows that there exists some constant C such that for any permutation $(i_1, ..., i_N)$ of $(1, ..., N)$, any integer $n \in \{1, ..., N\}$ and a fixed vector θ,

$$\mathbb{E}\Big[X_{i_n} X_{i_n}^\top \mathbb{1}\{X_{i_1}^\top \theta < ... < X_{i_N}^\top \theta\}\Big] \preccurlyeq C\mathbb{E}\Big[\Big(X_{i_1} X_{i_1}^\top + X_{i_N} X_{i_N}^\top\Big)\mathbb{1}\{X_{i_1}^\top \theta < ... < X_{i_N}^\top \theta\}\Big].$$

One can see that in the case of $K = 1$, then the proposition directly applies. (And, the claim trivially holds with $C_X = 1$ when $K = N$). Now, for $K \in 2, ..., N-1$, it suffices to show that $\mathbf{V}_{t,\min}(\mathcal{I}) + \mathbf{V}_{t,\max}(\mathcal{I})$ dominates any K-sub-sum of $X_{t,i_1} X_{t,i_1}^\top + ... + X_{t,i_N} X_{t,i_N}^\top$. Also, the inequality above immediately implies that

$$\begin{aligned}
&\mathbb{E}\Big[(X_{i_j} X_{i_j}^\top + X_{i_n} X_{i_n}^\top)\mathbb{1}\{X_{i_1}^\top \theta < ... < X_{i_N}^\top \theta\}\Big] \\
&\preccurlyeq 2C\mathbb{E}\Big[\Big(X_{i_1} X_{i_1}^\top + X_{i_N} X_{i_N}^\top\Big)\mathbb{1}\{X_{i_1}^\top \theta < ... < X_{i_N}^\top \theta\}\Big]
\end{aligned}$$

for $j, n \in \{1, ..., N\}$ with $j \neq n$. Then, Consider an arbitrary K-sub-sum over indices $(i'_1, ..., i'_K)$ which is a subset of $(i_1, ..., i_N)$.

$$\mathbb{E}\left[(X_{t,i'_1} X_{t,i'_1}^\top + ... + X_{t,i'_K} X_{t,i'_K}^\top)\mathbb{1}\{X_{t,i_1}^\top \theta < \cdots < X_{t,i_N}^\top \theta\}\right]$$

Without loss of generality, assume $(i'_1, ..., i'_K)$ is sorted in the increasing order with respect the product $X_i^\top \theta$, the same as $(i_1, ..., i_N)$. That is, $X_{t,i'_j}^\top \theta \leq X_{t,i'_n}^\top \theta$ for any $j, n \in \{1, ..., K\}$ with $j < n$. Then, it is easy to see that for the sum of the first and last elements,

$$\begin{aligned}
&\mathbb{E}\left[(X_{t,i'_1} X_{t,i'_1}^\top + X_{t,i'_K} X_{t,i'_K}^\top)\mathbb{1}\{X_{t,i_1}^\top \theta < \cdots < X_{t,i_N}^\top \theta\}\right] \\
&\preccurlyeq 2C\mathbb{E}\left[(X_{t,i_1} X_{t,i_1}^\top + X_{t,i_N} X_{t,i_N}^\top)\mathbb{1}\{X_{t,i_1}^\top \theta < \cdots < X_{t,i_N}^\top \theta\}\right].
\end{aligned}$$

Likewise, for the second and the second to the last elements, we can show

$$\begin{aligned}&\mathbb{E}\left[(X_{t,i_2'}X_{t,i_2'}^\top + X_{t,i_{K-1}'}X_{t,i_{K-1}'}^\top)\mathbb{1}\{X_{t,i_1}^\top\theta < \cdots < X_{t,i_N}^\top\theta\}\right]\\ &\preccurlyeq 2C\mathbb{E}\left[(X_{t,i_2}X_{t,i_2}^\top + X_{t,i_{N-1}}X_{t,i_{N-1}}^\top)\mathbb{1}\{X_{t,i_1}^\top\theta < \cdots < X_{t,i_N}^\top\theta\}\right].\end{aligned}$$

Repeating this procedure $K/2$ times and summing over the inequalities completes the proof since we have shown it for an arbitrary sub-sum. □

Remark A.2. *Since our primary focus in Lemma A.12 is to show $\lambda_{\min}(V_t)$ grows linearly in every episode, we only show a result based on $C_{\mathcal{X}}$ given by Proposition 4. While $C_{\mathcal{X}}$ is a finite value for any i.i.d. distribution, a general bound for $C_{\mathcal{X}}$ can be loose. Note that the exact value of $C_{\mathcal{X}}$ is characterized by the distribution of feature vectors. For example, for multivariate Gaussian and uniform distributions, it can be shown that $C_{\mathcal{X}} = \mathcal{O}(1)$ (see Lemma C.8 and Lemma C.9 in Appendix C.)*

Lemma A.14. *Suppose Assumption 2.4 holds. Then we have*

$$\mathbb{E}\left[\mathbf{V}_{t,\min}(\mathcal{I})\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t} < \cdots < X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right] \preccurlyeq \rho_0\mathbb{E}\left[\mathbf{V}_{t,\max}(\mathcal{I})\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t} < \cdots < X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right].$$

Proof. Let $\mathbf{x}$ be a tuple $(x_{i_1}, ..., x_{i_K})$.

$$\begin{aligned}&\mathbb{E}\left[\mathbf{V}_{t,\min}(\mathcal{I})\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t} < \cdots < X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right]\\ &= \mathbb{E}\left[(X_{t,i_1}X_{t,i_1}^\top + ... + X_{t,i_K}X_{t,i_K}^\top)\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t} < \cdots < X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right]\\ &= \int(x_{i_1}x_{i_1}^\top + ... + x_{i_K}x_{i_K}^\top)\mathbb{1}\left\{x_{i_1}^\top\beta \le ... \le x_{i_N}^\top\beta\right\}p_X(\mathbf{x})d\mathbf{x}\\ &= \int(x_{i_1}x_{i_1}^\top + ... + x_{i_K}x_{i_K}^\top)\mathbb{1}\left\{-x_{i_1}^\top\beta \ge ... \ge -x_{i_N}^\top\beta\right\}p_X(-\mathbf{x})d\mathbf{x}\\ &\preccurlyeq \rho_0\int(x_{i_1}x_{i_1}^\top + ... + x_{i_K}x_{i_K}^\top)\mathbb{1}\left\{x_{i_1}^\top\beta \ge ... \ge x_{i_N}^\top\beta\right\}p_X(\mathbf{x})d\mathbf{x}\\ &= \rho_0\mathbb{E}\left[\mathbf{V}_{t,\max}(\mathcal{I})\mathbb{1}\{X_{t,i_1}^\top\widetilde{\theta}_{k,t} < \cdots < X_{t,i_N}^\top\widetilde{\theta}_{k,t}\}\right]\end{aligned}$$

where the inequality is from the relaxed symmetry in Assumption 2.4. □

A.7 Other Lemmas

Proposition 3. *For each* $\mathcal{E}_t = [\epsilon_{t1}, \epsilon_{t2}, ..., \epsilon_{t|S_t|}]^\top$, $\|\mathcal{E}_t\| \leq \sqrt{2}$.

Proof. Note that by the definition of ϵ_{ti}, we have

$$\epsilon_{t1} + \epsilon_{t2} + ... + \epsilon_{t|S_t|} = 0, \quad \text{and} \quad \epsilon_{ti} \in [-1, 1]. \tag{A.27}$$

Hence the vector $\mathcal{E}_t$ lies within the bounded hyperplane in (A.27). Therefore, the ℓ_2 norm $\|\mathcal{E}_t\| = \sqrt{\epsilon_{t1}^2 + \epsilon_{t2}^2 + ... + \epsilon_{t|S_t|}^2}$ is maximized at the corners of this bounded hyperplane, i.e., for some $i, j \in S_t$, $i \neq j$

$$\epsilon_{ti} = 1, \epsilon_{tj} = -1 \quad \text{and} \quad \epsilon_{tk} = 0, \text{ for all } k \neq i, k \neq j,$$

which gives $\|\mathcal{E}_t\| \leq \sqrt{2}$. □

Lemma A.15 (Hoeffding's inequality)**.** *Let* $X_1, ..., X_n$ *be* n *independent random variables such that* $\mathbb{E}[X_i] = 0$ *and almost surely,* $X_i \in [a_i, b_i]$*, for all* i*. Then for any* $nu > 0$*,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > \nu\right) \leq 2\exp\left(-\frac{2\nu^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Lemma A.16 (Tropp (2012), Corollary 5.2)**.** *Consider a finite sequence* $\{\mathbf{Y}_k\}$ *of independent, random, self-adjoint matrices such that each* $\mathbf{Y}_k$ *is positive semi-definite and* $\lambda_{\max}(\mathbf{Y}_k) \leq R$ *almost surely. Compute the minimum and maximum eigenvalues of the sum of expectations,*

$$\mu_{\min} := \lambda_{\min}\left(\sum_k \mathbb{E}[\mathbf{Y}_k]\right) \quad \textit{and} \quad \mu_{\max} := \lambda_{\max}\left(\sum_k \mathbb{E}[\mathbf{Y}_k]\right).$$

Then

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_k \mathbb{E}[\mathbf{Y}_k]\right) \leq (1-\delta)\mu_{\min}\right\} \leq d \cdot \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \quad \textit{for } \delta \in [0,1] \textit{ and}$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbb{E}[\mathbf{Y}_k]\right) \leq (1+\delta)\mu_{\max}\right\} \leq d \cdot \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \quad \textit{for } \delta \geq 0\,.$$

Appendix B: Thompson Sampling for MNL Contextual Bandits

B.1 Regularized Maximum Likelihood Estimation for MNL Model

We briefly discuss regularized maximum likelihood estimation (MLE) for MNL model – specifically the estimation of the unknown parameter θ^* of the MNL model with the rigde penalty. Recall that $y_t \in \{0,1\}^{|S_t|+1}$ is the user choice where y_{ti} is the i-th component of y_t. Then, the ridge penalized maximum likelihood estimation for MNL model is given by the following minimization problem:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\ell_n(\theta) + \frac{\lambda}{2}\|\theta\|^2 \right] \tag{B.1}$$

where $\ell_n(\theta) = -\sum_{t=1}^{n} \sum_{i \in S_t \cup \{0\}} y_{ti} \log p_{ti}(S_t, \theta)$ with the penalty parameter $\lambda \geq 1$.

Taking the gradient of this penalized log-likelihood function with respect to θ, we obtain

$$\nabla_\theta \left[\ell_n(\theta) + \frac{\lambda}{2}\|\theta\|_2^2 \right] = \sum_{t=1}^{n} \sum_{i \in S_t} (p_{ti}(S_t, \theta) - y_{ti}) x_{ti} + \lambda\theta. \tag{B.2}$$

Instead of using the regularized MLE for the parameter estimation, one could consider using the MLE without regularization. For this, however, one may consider performing a random initialization (random exploration) to ensure that the matrix V_t is invertible.

B.2 Proofs of Lemmas for Theorem 3.1

B.2.1 Proof of Lemma 3.1

Proof. By the mean value theorem, there exists $\bar{u}_{ti} := (1-c)u_{ti} + cu'_{ti}$ for some $c \in (0,1)$ with

$$\begin{aligned}
&\frac{\sum_{i\in S} r_{ti}\exp\left(u_{ti}\right)}{1+\sum_{j\in S}\exp\left(u_{tj}\right)} - \frac{\sum_{i\in S} r_{ti}\exp(u'_{ti})}{1+\sum_{j\in S}\exp(u'_{tj})} \\
&= \sum_{i\in S} r_{ti} p_{ti}(S,\bar{u}_t)(u_{ti}-u'_{ti}) - R_t(S,\bar{u}_t)\cdot\sum_{i\in S} p_{ti}(S,\bar{u}_t)(u_{ti}-u'_{ti}) \\
&= \sum_{i\in S}\Big(r_{ti} - R_t(S,\bar{u}_t)\Big)p_{ti}(S,\bar{u}_t)(u_{ti}-u'_{ti}) \\
&\le \max_{i\in S}|u_{ti}-u'_{ti}|
\end{aligned}$$

where the inequality is from $|r_{ti}| \le 1$, and $p_{ti}(S,\bar{u}_t) \le 1$ is a multinomial probability (and hence $R_t(S,\bar{u}_t) \le 1$). □

B.2.2 Proof of Lemma 3.2

Proof. We first define the function $G_n(\theta)$ which we use throughout the proof:

$$G_n(\theta) = \sum_{t=1}^{n}\sum_{i\in S_t}\left[\left(p_{ti}(S_t,\theta) - p_{ti}(S_t,\theta^*)\right)x_{ti}\right] + \lambda(\theta-\theta^*)$$

$G_n(\theta)$ is the difference in the gradients of the ridge penalized maximum likelihood in (B.2) evaluated at θ and at θ^*. Notice that $G_n(\hat{\theta}) = \sum_{t=1}^{n}\sum_{i\in S_t}\epsilon_{ti}x_{ti} - \lambda\theta^*$ since the choice of $\hat{\theta}$ is given by the ridge penalized maximum likelihood. To see that, first note that $\hat{\theta}$ is the minimizer of (B.1); hence is given by the solution to the following equation:

$$\sum_{t=1}^{n}\sum_{i\in S_t}\Big(p_{ti}(S_t,\hat{\theta}) - y_{ti}\Big)x_{ti} + \lambda\hat{\theta} = 0 \tag{B.3}$$

Therefore, it follows that

$$\begin{aligned} G_n(\hat{\theta}) &= \sum_{t=1}^{n} \sum_{i \in S_t} \left(p_{ti}(S_t, \hat{\theta}) - p_{ti}(S_t, \theta^*) \right) x_{ti} + \lambda(\hat{\theta} - \theta^*) \\ &= \sum_{t=1}^{n} \sum_{i \in S_t} \left(p_{ti}(S_t, \hat{\theta}) - y_{ti} \right) x_{ti} + \lambda \hat{\theta} + \sum_{t=1}^{n} \sum_{i \in S_t} \left(y_{ti} - p_{ti}(S_t, \theta^*) \right) x_{ti} - \lambda \theta^* \\ &= 0 + \sum_{t=1}^{n} \sum_{i \in S_t} \epsilon_{ti} x_{ti} - \lambda \theta^* \end{aligned}$$

where the last equality is from (B.3) and the definition of $\epsilon_{ti} = y_{ti} - p_{ti}(S_t, \theta^*)$. For convenience, we define $Z_n := \sum_{t=1}^{n} \sum_{i \in S_t} \epsilon_{ti} x_{ti}$. Hence, $G_n(\hat{\theta}) = Z_n - \lambda \theta^*$. Also, we will denote $p_{ti}(\theta) := p_{ti}(S_t, \theta)$ when it is clear that S_t is the assortment chosen at round t.

For any $\theta_1, \theta_2 \in \mathbb{R}^d$, the mean value theorem implies that there exists $\bar{\theta} = c\theta_1 + (1-c)\theta_2$ with some $c \in (0, 1)$ such that

$$\begin{aligned} G_n(\theta_1) - G_n(\theta_2) &= \sum_{t=1}^{n} \sum_{i \in S_t} \Big[\left(p_{ti}(\theta_1) - p_{ti}(\theta_2) \right) x_{ti} \Big] + \lambda(\theta_1 - \theta_2) \\ &= \left[\left(\sum_{t=1}^{n} \sum_{i \in S_t} \sum_{j \in S_t} \nabla_j p_{ti}(\bar{\theta}) x_{ti} x_{tj}^\top \right) + \lambda I_d \right] (\theta_1 - \theta_2) \\ &= \left[\sum_{t=1}^{n} \left(\sum_{i \in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i \in S_t} \sum_{j \in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top \right) + \lambda I_d \right] (\theta_1 - \theta_2) \end{aligned}$$

where I_d is a $d \times d$ identitiy matrix. We define the matrix H_t as

$$H_t := \sum_{i \in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i,j \in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top$$

Notice H_t is a Hessian of a negative log-likelihood which is convex. Hence, H_t is positive semidefinite. Also note that

$$(x_i - x_j)(x_i - x_j)^\top = x_i x_i^\top + x_j x_j^\top - x_i x_j^\top - x_j x_i^\top \succeq 0$$

which implies $x_i x_i^\top + x_j x_j^\top \succeq x_i x_j^\top + x_j x_i^\top$. Therefore, it follows that

$$\begin{aligned}
H_t &= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{tj}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \frac{1}{2}\sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) \left(x_{ti} x_{tj}^\top + x_{tj} x_{ti}^\top\right) \\
&\succeq \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \frac{1}{2}\sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) \left(x_{ti} x_{ti}^\top + x_{tj} x_{tj}^\top\right) \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) x_{ti} x_{ti}^\top - \sum_{i\in S_t}\sum_{j\in S_t} p_{ti}(\bar{\theta}) p_{tj}(\bar{\theta}) x_{ti} x_{ti}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) \left(1 - \sum_{j\in S_t} p_{tj}(\bar{\theta})\right) x_{ti} x_{ti}^\top \\
&= \sum_{i\in S_t} p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta}) x_{ti} x_{ti}^\top
\end{aligned}$$

where $p_{t0}(\bar{\theta})$ is the probability of choosing the outside option. Now,

$$\begin{aligned}
G_n(\theta_1) - G_n(\theta_2) &= \left[\sum_{t=1}^{n} H_t + \lambda I_d\right](\theta_1 - \theta_2) \\
&\geq \left[\sum_{t=1}^{n}\sum_{i\in S_t} p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta}) x_{ti} x_{ti}^\top + \lambda I_d\right](\theta_1 - \theta_2) \\
&:= \mathcal{H}(\bar{\theta})(\theta_1 - \theta_2).
\end{aligned}$$

Consider some $\bar{\theta} \in \mathbb{R}^d$. From Assumption 3.3, $p_{ti}(\bar{\theta}) p_{t0}(\bar{\theta})$ is lower-bounded by κ. Then we have

$$(\theta_1 - \theta_2)^\top (G_n(\theta_1) - G_n(\theta_2)) \geq (\theta_1 - \theta_2)^\top (\kappa V_n)(\theta_1 - \theta_2) > 0$$

for any $\theta_1 \neq \theta_2$. By the definition of $G_n(\theta)$, we have $G_n(\theta^*) = 0$. Hence, for any $\theta \in \mathbb{R}^d$,

we have

$$\begin{aligned}
\|G_n(\theta)\|^2_{V_n^{-1}} &= \|G_n(\theta) - G_n(\theta^*)\|^2_{V_n^{-1}} \\
&= \left(G_n(\theta) - G_n(\theta^*)\right)^\top V_n^{-1} \left(G_n(\theta) - G_n(\theta^*)\right) \\
&\geq (\theta - \theta^*)^\top \mathcal{H}(\bar{\theta}) V_n^{-1} \mathcal{H}(\bar{\theta})(\theta - \theta^*) \\
&\geq \kappa^2 (\theta - \theta^*)^\top V_n (\theta - \theta^*) \\
&= \kappa^2 \|\hat{\theta} - \theta^*\|^2_{V_n}
\end{aligned}$$

where the last inequality is from $\mathcal{H}(\bar{\theta}) \succeq \kappa V_n$. Now, recall for $\hat{\theta}$ which is the solution to (B.3), $G_n(\hat{\theta}) = Z_n - \lambda\theta^*$ where $Z_n = \sum_{t=1}^n \sum_{i \in S_t} \epsilon_{ti} x_{ti}$. Hence, we have

$$\kappa \|\hat{\theta} - \theta^*\|_{V_n} \leq \|G_n(\hat{\theta})\|_{V_n^{-1}} \leq \|Z_n\|_{V_n^{-1}} + \lambda \|\theta^*\|_{V_n^{-1}}$$

Then we can use Theorem 1 in Abbasi-Yadkori, Pál, and Szepesvári (2011), which states if the noise ϵ_{ti} is sub-Gaussian with parameter σ (with $\sigma = \frac{1}{2}$ in our problem), then

$$\|Z_n\|^2_{V_n^{-1}} \leq 2\sigma^2 \log\left(\frac{\det(V_n)^{1/2} \det(V)^{-1/2}}{\delta}\right)$$

with probability at least $1 - \delta$. Then we combine with Lemma A.2. So it follows that

$$\|Z_n\|^2_{V_n^{-1}} \leq 2\sigma^2 \left[\frac{d}{2} \log\left(\frac{\text{trace}(V) + nK}{d}\right) - \frac{1}{2} \log\det(V) + \log\frac{1}{\delta}\right].$$

Since $V = \lambda I_d$, it follows that

$$\begin{aligned}\|Z_n\|^2_{V_n^{-1}} &\leq 2\sigma^2 \left[\frac{d}{2}\log\left(\frac{d\lambda + nK}{d}\right) - \frac{1}{2}\log\lambda^d + \log\frac{1}{\delta}\right] \\ &= 2\sigma^2 \left[\frac{d}{2}\log\left(\lambda + \frac{nK}{d}\right) - \frac{d}{2}\log\lambda + \log\frac{1}{\delta}\right] \\ &= 2\sigma^2 \left[\frac{d}{2}\log\left(1 + \frac{nK}{d\lambda}\right) + \log\frac{1}{\delta}\right].\end{aligned}$$

Then for $\|\theta^*\|_{V_n^{-1}}$, we have

$$\|\theta^*\|^2_{V_n^{-1}} \leq \frac{\|\theta^*\|^2}{\lambda_{\min}(V_n)} \leq \frac{\|\theta^*\|^2}{\lambda_{\min}(V)} \leq \frac{\|\theta^*\|^2}{\lambda}.$$

Hence, $\lambda\|\theta^*\|_{V_n^{-1}} \leq \sqrt{\lambda}$ since $\|\theta^*\| \leq 1$. Combining the results and using the fact that $\sigma = \frac{1}{2}$ for our problem, we have that

$$\|\hat{\theta}_n - \theta^*\|_{V_n} \leq \frac{1}{2\kappa}\sqrt{d\log\left(1 + \frac{nK}{d\lambda}\right) + 2\log\frac{1}{\delta}} + \frac{\sqrt{\lambda}}{\kappa}.$$

with probability at least $1 - \delta$.

□

B.2.3 Proof of Lemma 3.3

Proof. First, define event $\hat{\mathcal{E}}_t = \{\|\theta^* - \hat{\theta}_t\|_{V_t} \leq \alpha_t\}$, i.e. the regularized MLE estimate concentrates properly to θ^* in rounds t. From Lemma 3.2, this concentration event holds with probability $1 - \mathcal{O}\left(\frac{1}{t^2}\right)$ for each round t. On $\hat{\mathcal{E}}_t$, we show $x_{ti}^\top\theta^* \leq x_{ti}^\top\hat{\theta}_t + \alpha_t\|x_{ti}\|_{V_t^{-1}}$

for all i.

$$\begin{aligned}
|x_{ti}^\top \hat{\theta}_t - x_{ti}^\top \theta^*| &= \left| \left[V_t^{-1/2} (\hat{\theta}_t - \theta^*) \right]^\top (V_t^{-1/2} x_{ti}) \right| \\
&\leq \left\| V_t^{-1/2} (\hat{\theta}_t - \theta^*) \right\| \left\| V_t^{-1/2} x_{ti} \right\| \\
&= \|\hat{\theta}_t - \theta^*\|_{V_t} \|x_{ti}\|_{V_t^{-1}} \\
&\leq \alpha_t \|x_{ti}\|_{V_t^{-1}}
\end{aligned}$$

where the first inequality is by Hölder's inequality. Hence, it follows that

$$x_{ti}^\top \theta^* - \left(x_{ti}^\top \hat{\theta}_t + \alpha_t \|x_{ti}\|_{V_t^{-1}} \right) \leq 0$$

for all i. Hence, using the restricted monotonicity in Lemma 3.5, if event $\hat{\mathcal{E}}_t$ holds, then we have

$$R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \leq 0.$$

Then we have

$$\begin{aligned}
\mathbb{E}\Big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t \Big] &\leq \mathbb{E}\left[\Big(R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \Big) \mathbb{1}(\hat{\mathcal{E}}_t) \mid \mathcal{F}_t \right] + \mathbb{E}\left[\mathbb{1}(\hat{\mathcal{E}}_t^c) \mid \mathcal{F}_t \right] \\
&\leq 0 + \mathcal{O}(t^{-2}).
\end{aligned}$$

Therefore, summing over all $t \leq T$, we have

$$\sum_{t=1}^{T} \mathbb{E}\Big[R_t(S_t^*, \theta^*) - U_t(S_t^*, \hat{\theta}_t) \mid \mathcal{F}_t \Big] \leq 0 + \sum_{t=1}^{T} \mathcal{O}(t^{-2}) = \mathcal{O}(1).$$

□

B.2.4 Proof of Lemma 3.4

See Section A.1.3. The slight difference between Lemma 3.4 and the proof in Section A.1.3 is the initial Gram matrix V_0. In Section A.1.3, random initialization is used whereas V comes from the regularization in Lemma 3.4. However, these two cases are equivalent; hence slight modification provides for the bound for Lemma 3.4.

B.3 Proofs of Lemmas for Theorem 3.2

B.3.1 Proof of Lemma 3.7

Proof. Given $\mathcal{F}_t$, each of Gaussian random variable $x_{ti}^\top \widetilde{\theta}_t^{(j)}$ has mean $x_{ti}^\top \hat{\theta}_t$ and standard deviation $\alpha_t \|x_{ti}\|_{V_t^{-1}}$.

$$\begin{aligned}
|\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t| &= \alpha_t \|x_{ti}\|_{V_t^{-1}} \frac{\left|\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} - x_{ti}^\top \hat{\theta}_t\right|}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} \\
&\leq \alpha_t \|x_{ti}\|_{V_t^{-1}} \max_j \left| \frac{x_{ti}^\top \widetilde{\theta}_t^{(j)} - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} \right| \\
&= \alpha_t \|x_{ti}\|_{V_t^{-1}} \max_j |Z_j|
\end{aligned}$$

where each Z_j is a standard normal random variable. Using the result from Lemma B.1, we have $\max_j |Z_j| \leq \sqrt{2\log(2M)} + \sqrt{4\log t}$ with probability at least $1 - \frac{1}{t^2}$. Then, for all $i \in [N]$,

$$|\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t| \leq \left(\sqrt{2\log(2M)} + \sqrt{4\log(Nt)} \right) \alpha_t \|x_{ti}\|_{V_t^{-1}}$$

with probability at least $1-\frac{1}{t^2}$. Alternatively, let $m = \text{argmax}_j \, x_{ti}^\top \widetilde{\theta}_t^{(j)}$. Then we can write

$$\begin{aligned}
|\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t| &= \left| \max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} - x_{ti}^\top \hat{\theta}_t \right| \\
&= \left| x_{ti}^\top (\widetilde{\theta}_t^{(m)} - \hat{\theta}_t) \right| \\
&= \left| x_{ti}^\top V_t^{-1/2} V_t^{1/2} (\widetilde{\theta}_t^{(m)} - \hat{\theta}_t) \right| \\
&\leq \alpha_t \|x_{ti}\|_{V_t^{-1}} \left\| \alpha_t^{-1} V_t^{1/2} (\widetilde{\theta}_t^{(m)} - \hat{\theta}_t) \right\| \\
&\leq \alpha_t \|x_{ti}\|_{V_t^{-1}} \max_j \left\| \alpha_t^{-1} V_t^{1/2} (\widetilde{\theta}_t^{(j)} - \hat{\theta}_t) \right\| \\
&= \alpha_t \|x_{ti}\|_{V_t^{-1}} \max_j \|\zeta_j\|
\end{aligned}$$

where each element in $\zeta_j \in \mathbb{R}^d$ is a univariate standard normal variable $\mathcal{N}(0,1)$. Hence, each $\|\zeta_j\| \leq \sqrt{4d \log t}$ with probability at least $1 - \frac{1}{t^2}$. Using the union bound for all $j \in \{1, ..., M\}$, we have with probability at least $1 - \frac{1}{t^2}$

$$|\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t| \leq \sqrt{4d \log(Mt)} \alpha_t \|x_{ti}\|_{V_t^{-1}}.$$

□

Lemma B.1. *Let $Z_i \sim \mathcal{N}(0,1), i = 1, ..., n$ be a standard Gaussian random variable. Then we have*

$$\mathbb{P}\left(\max_i |Z_i| \leq \sqrt{2\log(2n)} + \sqrt{2 \log \frac{1}{\delta}} \right) \geq 1 - \delta.$$

Proof. Using the Chernoff bound, for each Z_i, we have

$$\mathbb{P}(|Z_i| > \epsilon) \leq 2e^{-\epsilon^2/2}.$$

Applying the union bound, we have

$$\begin{aligned}\mathbb{P}\left(\max_i |Z_i| > \sqrt{2\log(2n)} + \epsilon\right) &\leq 2n \exp\left(-(\sqrt{2\log(2n)} + \epsilon)^2/2\right) \\ &= 2n \exp(-\log(2n) - \epsilon\sqrt{2\log(2n)} - \epsilon^2/2) \\ &\leq e^{-\epsilon\sqrt{2\log(2n)}} e^{-\epsilon^2/2} \\ &\leq e^{-\epsilon^2/2}.\end{aligned}$$

Letting $\delta = e^{-\epsilon^2/2}$, we have the result. □

B.3.2 Proof of Lemma 3.6

Proof. Given $\mathcal{F}_t$, each of Gaussian random variable $x_{ti}^\top \widetilde{\theta}_t^{(j)}$ has mean $x_{ti}^\top \hat{\theta}_t$ and standard deviation $\alpha_t \|x_{ti}\|_{V_t^{-1}}$. Hence, for each $i \in S_t^*$, we have

$$\begin{aligned}\mathbb{P}\left(\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} > x_{ti}^\top \theta^* \mid \mathcal{F}_t\right) &= 1 - \mathbb{P}\left(x_{ti}^\top \widetilde{\theta}_t^{(j)} \leq x_{ti}^\top \theta^*, \forall j \in \{1, ..., M\} \mid \mathcal{F}_t\right) \\ &= 1 - \mathbb{P}\left(\frac{x_{ti}^\top \widetilde{\theta}_t^{(j)} - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} \leq \frac{x_{ti}^\top \theta^* - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}}, \forall j \in \{1, ..., M\} \mid \mathcal{F}_t\right) \\ &= 1 - \mathbb{P}\left(Z_j \leq \frac{x_{ti}^\top \theta^* - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}}, \forall j \in \{1, ..., M\} \mid \mathcal{F}_t\right)\end{aligned}$$

where Z_j is a standard normal random variable. By the assumption, we have $|x_{ti}^\top \theta^* - x_{ti}^\top \hat{\theta}_t| \leq \alpha_t \|x_{ti}\|_{V_t^{-1}}$ for all i, Hence, we can bound the RHS term within the probability.

$$\frac{x_{ti}^\top \theta^* - x_{ti}^\top \hat{\theta}_t}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} \leq \frac{\alpha_t \|x_{ti}\|_{V_t^{-1}}}{\alpha_t \|x_{ti}\|_{V_t^{-1}}} = 1$$

Then, it follows that

$$\mathbb{P}\left(\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} > x_{ti}^\top \theta^* \mid \mathcal{F}_t\right) \geq 1 - \left(\mathbb{P}(Z \leq 1)\right)^M. \tag{B.4}$$

Now, since $S_t = \text{argmax}_S \widetilde{R}_t(S)$, we have $\widetilde{R}_t(S_t) \geq \widetilde{R}_t(S_t^*)$. Then combining with Lemma 3.5, we can lower-bound the probability of having an expected revenue optimistic under the sampled parameter (the second inequality below).

$$\begin{aligned}
\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t\right) &\geq \mathbb{P}\left(\widetilde{R}_t(S_t^*) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t\right) \\
&\geq \mathbb{P}\left(\widetilde{u}_{ti} > x_{ti}^\top \theta^*, \forall i \in S_t^* \mid \mathcal{F}_t\right) \\
&= \mathbb{P}\left(\max_j x_{ti}^\top \widetilde{\theta}_t^{(j)} > x_{ti}^\top \theta^*, \forall i \in S_t^* \mid \mathcal{F}_t\right) \\
&\geq 1 - K\left(\mathbb{P}(Z \leq 1)\right)^M
\end{aligned}$$

where the last inequality comes from (B.4) and the union bound. Using the anti-concentration inequality in Lemma B.3, we have $\mathbb{P}(Z \leq 1) \leq 1 - \frac{1}{4\sqrt{e\pi}}$. Hence, it follows that

$$\begin{aligned}
\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t\right) &\geq 1 - K\left(1 - \frac{1}{4\sqrt{e\pi}}\right)^M \\
&\geq 1 - \left(1 - \frac{1}{4\sqrt{e\pi}}\right) \\
&= \frac{1}{4\sqrt{e\pi}}
\end{aligned}$$

where the second inequality comes from our choice of $M = \lceil 1 - \frac{\log K}{\log\left(1 - 1/(4\sqrt{e\pi})\right)} \rceil$ which implies $\left(1 - \frac{1}{4\sqrt{e\pi}}\right)^M \leq \frac{1}{K}\left(1 - \frac{1}{4\sqrt{e\pi}}\right)$.

□

B.3.3 Proof of Lemma 3.8

Proof. The proof is inspired by the techniques used for Theorem 1 in Abeille, Lazaric, et al. (2017). First, we define $\widetilde{\Theta}_t$ the set of parameter samples for which the expected revenue concentrate appropriately to the expected revenue based on the MLE parameter.

Also, we define the set of optimistic parameter samples $\widetilde{\Theta}_t^{\text{opt}}$ which coinciding with $\widetilde{\Theta}_t$.

$$\widetilde{\Theta}_t := \left\{ \{\widetilde{\theta}_t^{(j)}\}_{j=1}^M : \widetilde{R}_t(S_t) - R_t(S_t, \hat{\theta}_t) \le \beta_t \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \right\}$$
$$\widetilde{\Theta}_t^{\text{opt}} := \left\{ \{\widetilde{\theta}_t^{(j)}\}_{j=1}^M : \widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \right\} \cap \widetilde{\Theta}_t$$

Define the event $\mathcal{E}_t$ that both $x_{ti}^\top \hat{\theta}_t$ and $\widetilde{u}_{ti}$ are concentrated around their respective means.

$$\mathcal{E}_t = \{x_{ti}^\top \hat{\theta}_t - x_{ti}^\top \theta^* \le \alpha_t \|x_{ti}\|_{V_t^{-1}}, \forall i\} \cap \{\widetilde{u}_{ti} - x_{ti}^\top \hat{\theta}_t \le \beta_t \|x_{ti}\|_{V_t^{-1}}, \forall i\}.$$

Recall that $S_t = \text{argmax}_{S \in \mathcal{S}} \widetilde{R}(S)$. For any $\widetilde{\theta}_t^{1:M} := \{\widetilde{\theta}_t^{(j)}\}_{j=1}^M \in \widetilde{\Theta}_t^{\text{opt}}$, we have

$$\left(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right) \mathbb{1}(\mathcal{E}_t) \le \left(R_t(S_t^*, \theta_t^*) - \inf_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \max_{S \in \mathcal{S}} \widetilde{R}_t(S, \theta_t^{1:M})\right) \mathbb{1}(\mathcal{E}_t)$$

where $\widetilde{R}_t(S, \theta_t^{1:M})$ is the optimistic expected revenue under the sampled parameters $\theta_t^{1:M}$. Note that we can decompose

$$R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t) = \left(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right) \mathbb{1}(\mathcal{E}_t) + \left(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right) \mathbb{1}(\mathcal{E}_t^c)$$

where we can bound the summation of the second term in the right hand side since event $\mathcal{E}_t$ holds with high probability.

$$\sum_{t=1}^T \mathbb{E}\left[R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right] = \sum_{t=1}^T \mathbb{E}\left[\left(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right) \mathbb{1}(\mathcal{E}_t)\right] + \mathcal{O}(1)$$

Therefore, we are left to bound the first term in the right hand side. Then conditioning on the sample parameter being optimistic, i.e.,$\widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}$, we can further bound with

the expectation over any random choice.

$$
\begin{aligned}
&\mathbb{E}\left[\Big(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\Big)\mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t\right] \\
&\le \mathbb{E}\left[\Big(R_t(S_t^*, \theta_t^*) - \inf_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \max_{S \in \mathcal{S}} \widetilde{R}_t(S, \theta_t^{1:M})\Big)\mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t\right] \\
&\le \mathbb{E}\left[\Big(\widetilde{R}_t(S_t) - \inf_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \max_{S \in \mathcal{S}} \widetilde{R}_t(S, \theta_t^{1:M})\Big)\mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\mathrm{opt}}\right] \\
&\le \mathbb{E}\left[\Big(\widetilde{R}_t(S_t) - \inf_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \widetilde{R}_t(S_t, \theta_t^{1:M})\Big)\mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\mathrm{opt}}\right] \\
&= \mathbb{E}\left[\sup_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \Big(\widetilde{R}_t(S_t) - \widetilde{R}_t(S_t, \theta_t^{1:M})\Big)\mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\mathrm{opt}}\right] \\
&\le \mathbb{E}\left[\sup_{\theta_t^{1:M} \in \widetilde{\Theta}_t} \max_{i \in S_t} \Big|\widetilde{u}_{ti} - x_{ti}^\top \theta_t^{(j)}\Big| \mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\mathrm{opt}}\right] \\
&\le 2\beta_t \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\mathrm{opt}}, \mathcal{E}_t\right] \mathbb{P}(\mathcal{E}_t)
\end{aligned}
$$

where the last inequality is from the definition of the set $\widetilde{\Theta}_t$ and $S_t(\widetilde{\theta}_t^{1:M})$ stands for the optimal assortment under the sampled parameters $\widetilde{\theta}_t^{1:M} = \{\widetilde{\theta}_t^{(j)}\}_{j=1}^M$.

From Lemma 3.6, we have $\mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t, \mathcal{E}_t\right) \ge \frac{1}{4\sqrt{e\pi}} =: \widetilde{p}$. Therefore it follows that

$$
\begin{aligned}
\mathbb{P}\left(\widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\mathrm{opt}} \mid \mathcal{F}_t, \mathcal{E}_t\right) &= \mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \text{ and } \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t, \mathcal{E}_t\right) \\
&\ge \mathbb{P}\left(\widetilde{R}_t(S_t) > R_t(S_t^*, \theta_t^*) \mid \mathcal{F}_t, \mathcal{E}_t\right) - \mathbb{P}\left(\widetilde{\theta}_t^{1:M} \notin \widetilde{\Theta}_t, \mathcal{E}_t\right) \\
&\ge \widetilde{p} - \mathcal{O}(t^{-1}) \\
&\ge \widetilde{p}/2.
\end{aligned}
$$

Now, note that we can write

$$\begin{aligned}
&\mathbb{E}\left[\max_{i\in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \mathcal{E}_t\right] \\
&\geq \mathbb{E}\left[\max_{i\in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t\right] \mathbb{P}\left(\widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}} \mid \mathcal{F}_t, \mathcal{E}_t\right) \\
&\geq \mathbb{E}\left[\max_{i\in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t\right] \cdot \widetilde{p}/2
\end{aligned}$$

Therefore, combining the results, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right)\mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t\right] \\
&\leq 2\beta_t \mathbb{E}\left[\max_{i\in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \widetilde{\theta}_t^{1:M} \in \widetilde{\Theta}_t^{\text{opt}}, \mathcal{E}_t\right] \mathbb{P}(\mathcal{E}_t) \\
&\leq \frac{4\beta_t}{\widetilde{p}} \mathbb{E}\left[\max_{i\in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t, \mathcal{E}_t\right] \mathbb{P}(\mathcal{E}_t) \\
&\leq \frac{4\beta_t}{\widetilde{p}} \mathbb{E}\left[\max_{i\in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right].
\end{aligned}$$

Summing over all t and taking the failure event into consideration, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[\left(R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)\right)\mathbb{1}(\mathcal{E}_t) \mid \mathcal{F}_t\right] \leq \sum_{t=1}^{T} \frac{4\beta_t}{\widetilde{p}} \mathbb{E}\left[\max_{i\in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right].$$

Here, the summation on the RHS contains an expectation, so we cannot directly apply Lemma 3.4. Instead, we use Lemma B.2 to bound the sum of the expectations

$$\sum_{t=1}^{T} \mathbb{E}[R_t(S_t^*, \theta_t^*) - \widetilde{R}_t(S_t)] \leq \sum_{t=1}^{T} \frac{4\beta_t}{\widetilde{p}} \left(\sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda} \log 2T}\right) + \mathcal{O}(1).$$

□

Lemma B.2. *If $\lambda_{\min}(V_t) \geq \lambda$, then with probability $1 - \mathcal{O}(T^{-1})$ we have*

$$\sum_{t=1}^{T} \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] \leq \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} + \sqrt{\frac{8T}{\lambda} \log 2T}.$$

Proof. We rewrite the summation as follows.

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] \\
&= \sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} + \sum_{t=1}^{T} \left(\mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] - \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}\right) \qquad \text{(B.5)}
\end{aligned}$$

The first summation can be bounded by using Lemma 3.4 and the Cauchy-Schwarz inequality.

$$\sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \leq \sqrt{T \sum_{t=1}^{T} \max_{i \in S_t} \|x_{ti}\|^2_{V_t^{-1}}} \leq \sqrt{2dT \log\left(1 + \frac{TK}{d\lambda}\right)} \qquad \text{(B.6)}$$

For the second summation in (B.5), we can apply Azuma-Hoeffding inequality (Lemma B.4). Note that the second summation is a martingale by construction. Also recall that $\max_{i \in S_t} \|x_{ti}\| \leq 1$ for all t, hence we have

$$\mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] - \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}} \leq \frac{2}{\lambda_{\min}(V_t)} \leq \frac{2}{\lambda_{\min}(V)} = \frac{2}{\lambda}.$$

Therefore, $\frac{2}{\lambda}$ is an upper-bound for each element in the second summation. Now applying Azuma-Hoeffding inequality, we have

$$\sum_{t=1}^{T} \left(\mathbb{E}\left[\max_{i \in S_t(\widetilde{\theta}_t^{1:M})} \|x_{ti}\|_{V_t^{-1}} \mid \mathcal{F}_t\right] - \max_{i \in S_t} \|x_{ti}\|_{V_t^{-1}}\right) \leq \sqrt{\frac{8T}{\lambda} \log 2T} \qquad \text{(B.7)}$$

with probability $1 - \mathcal{O}(T^{-1})$. Combining (B.6) and (B.7), we have the result. □

B.3.4 Other Lemmas

The following lemma is used to derive the concentration and anti-concentration inequalities for Gaussian random variables.

Lemma B.3 (Abramowitz and Stegun 1965)**.** *For a Gaussian random variable Z with mean μ and variance σ^2, for any $z \geq 1$,*

$$\frac{1}{2\sqrt{\pi}z} e^{-z^2/2} \leq \mathbb{P}\left(|Z-\mu| > z\sigma\right) \leq \frac{1}{\sqrt{\pi}z} e^{-z^2/2}. \tag{B.8}$$

Lemma B.4 (Azuma-Hoeffding inequality)**.** *If a super-martingale $(Y_t; t \geq 0)$ corresponding to filtration $\mathcal{F}_t$, satisfies $|Y_t - Y_{t-1}| \leq c_t$ for some constant c_t, for all $t = 1, ..., T$, then for any $a \geq 0$,*

$$\mathbb{P}(Y_T - Y_0 \geq a) \leq 2e^{-\frac{a^2}{2\sum_{t=1}^{T} c_t^2}}$$

Appendix C: Sparsity-Agnostic High-Dimensional Bandit Algorithm

C.1 Proofs of Lemmas for Theorem 4.1

C.1.1 Proof of Lemma 4.1

Proof. The proof follows from modifying the proof of the standard Lasso oracle inequality (Bühlmann and Van De Geer, 2011) using martingale theory. Recall from (4.1) that the negative log-likelihood of the GLM is

$$\ell_t(\beta) = -\frac{1}{t}\sum_{\tau=1}^{t}\left[Y_\tau X_\tau^\top \beta - m(X_\tau^\top \beta)\right]$$

where m is a normalizing function with its gradient $\dot{m}(X^\top\beta) = \mu(X^\top\beta)$. Now, we denote the expectation of $\ell_t(\beta)$ over Y by $\bar{\ell}_t(\beta)$:

$$\bar{\ell}_t(\beta) := \mathbb{E}_Y[\ell_t(\beta)] = -\frac{1}{t}\sum_{\tau=1}^{t}\left[\mu(X_\tau^\top\beta^*)X_\tau^\top\beta - m(X_\tau^\top\beta)\right].$$

Note that $\nabla_\beta\bar{\ell}_t(\beta) = -\frac{1}{t}\sum_{\tau=1}^{t}\left[\mu(X_\tau^\top\beta^*) - \mu(X_\tau^\top\beta)\right]X_\tau$. Hence, we have $\nabla_\beta\bar{\ell}_t(\beta^*) = \mathbf{0}_d$ which implies that $\beta^* = \text{argmin}_\beta\,\bar{\ell}_t(\beta)$ given the fact that m is convex in the GLM. Hence, for any parameter $\beta \in \mathbb{R}^d$, the excess risk is defined as

$$\mathcal{E}(\beta) := \bar{\ell}_t(\beta) - \bar{\ell}_t(\beta^*).$$

Note that by definition, $\mathcal{E}(\beta) \geq 0$, for all $\beta \in \mathbb{R}^d$ (with $\mathcal{E}(\beta^*) = 0$). The Lasso estimate $\hat{\beta}_t$ for the GLM is given by the minimization of the penalized negative log-likelihood

$$\hat{\beta}_t := \underset{\beta}{\operatorname{argmin}} \left\{ \ell_t(\beta) + \lambda_t \|\beta\|_1 \right\}$$

where λ is the penalty parameter whose value needs to be chosen to control the noise of the model. Now, we define the empirical process of the problem as

$$v_t(\beta) := \ell_t(\beta) - \bar{\ell}_t(\beta).$$

Note that the randomness in $\{Y_\tau\}$ still plays a role on $\ell_t(\beta)$ and hence on $v_t(\beta)$. Then by the definition of $\hat{\beta}_t$, we have

$$\ell_t(\hat{\beta}_t) + \lambda_t \|\hat{\beta}_t\|_1 \leq \ell_t(\beta^*) + \lambda_t \|\beta^*\|_1.$$

Adding and subtracting terms, we have

$$\ell_t(\hat{\beta}_t) - \bar{\ell}_t(\hat{\beta}_t) + \bar{\ell}_t(\hat{\beta}_t) - \bar{\ell}_t(\beta^*) + \lambda_t \|\hat{\beta}_t\|_1 \leq \ell_t(\beta^*) - \bar{\ell}_t(\beta^*) + \lambda_t \|\beta^*\|_1 .$$

Rearranging terms gives the following "basic inequality" for the GLM

$$\mathcal{E}(\hat{\beta}_t) + \lambda_t \|\hat{\beta}_t\|_1 \leq -[v_t(\hat{\beta}_t) - v_t(\beta^*)] + \lambda_t \|\beta^*\|_1 .$$

The basic inequality implies that in order to provide an upper-bound for the penalized excess risk, we need to control the deviation of the empirical process $[v_t(\hat{\beta}_t) - v_t(\beta^*)]$ (Bühlmann and Van De Geer, 2011). And we bound this deviation of the empirical process in terms of the parameter estimation error $\|\hat{\beta}_t - \beta^*\|_1$. Essentially, $[v_t(\hat{\beta}_t) - v_t(\beta^*)]$ is where the random noise plays a role, and with large enough penalization (suitably large λ) we

can control such randomness in the empirical process. We define the event of the empirical process being controlled by the penalization.

$$\mathcal{T} := \{|v_t(\hat{\beta}_t) - v_t(\beta^*)| \leq \lambda \|\hat{\beta}_t - \beta^*\|_1\}\,. \tag{C.1}$$

Lemma C.1 ensures that we can control this empirical process deviation with high probability. Hence, in the rest of the proof, we restrict ourselves to the case where the empirical process behaves well, i.e., event $\mathcal{T}$ in (C.1) holds.

Lemma C.1. *Assume X_t satisfies $\|X_t\|_2 \leq x_{\max}$ for all t. If $\lambda = \sigma x_{\max}\sqrt{\frac{2[\log(2/\delta)+\log d]}{t}}$, then with probability at least $1-\delta$ we have*

$$|v_t(\hat{\beta}_t) - v_t(\beta^*)| \leq \lambda \|\hat{\beta}_t - \beta^*\|_1\,.$$

On event $\mathcal{T}$, for $\lambda_t \geq 2\lambda$, we have

$$2\mathcal{E}(\hat{\beta}_t) + 2\lambda_t\|\hat{\beta}_t\|_1 \leq \lambda_t\|\hat{\beta}_t - \beta^*\|_1 + 2\lambda_t\|\beta^*\|_1\,. \tag{C.2}$$

Let $\hat{\beta} := \hat{\beta}_t$ for brevity. Using the active set S_0, we can define the following:

$$\beta_{j,S_0} := \beta_j \mathbb{1}\{j \in S_0\} \qquad \beta_{j,S_0^c} := \beta_j \mathbb{1}\{j \notin S_0\}$$

so that $\beta_{S_0} = [\beta_{1,S_0}, ..., \beta_{d,S_0}]^\top$ has zero elements outside the set S_0 and the elements of $\beta_{S_0^c}$ can only be non-zero in the complement of S_0. We can then lower-bound $\|\hat{\beta}\|_1$ using the triangle inequality,

$$\begin{aligned}\|\hat{\beta}\|_1 &= \|\hat{\beta}_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \\ &\geq \|\beta^*_{S_0}\|_1 - \|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1\,.\end{aligned}$$

Also, we can rewrite

$$\begin{aligned}\|\hat{\beta} - \beta^*\|_1 &= \|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 + \|\hat{\beta}_{S_0^c} - \beta^*_{S_0^c}\|_1 \\ &= \|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \,.\end{aligned}$$

Then we continue from (C.2)

$$\begin{aligned}2\mathcal{E}(\hat{\beta}) + 2\lambda_t\|\beta^*_{S_0}\|_1 - 2\lambda_t\|\hat{\beta}_{S_0} - \beta^*_{S_0}\| + 2\lambda_t\|\hat{\beta}_{S_0^c}\|_1 &\leq \lambda_t\|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 + \lambda_t\|\hat{\beta}_{S_0^c}\|_1 + 2\lambda_t\|\beta^*\|_1 \\ &= \lambda_t\|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 + \lambda_t\|\hat{\beta}_{S_0^c}\|_1 + 2\lambda_t\|\beta^*_{S_0}\|_1 \,.\end{aligned}$$

Therefore, we have

$$\begin{aligned}0 \leq 2\mathcal{E}(\hat{\beta}) &\leq 3\lambda_t\|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 - \lambda_t\|\hat{\beta}_{S_0^c}\|_1 \\ &= \lambda_t\left(3\|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 - \|\hat{\beta}_{S_0^c} - \beta^*_{S_0^c}\|_1\right)\end{aligned} \tag{C.3}$$

Then the compatibility condition can be applied to the vector $\hat{\beta} - \beta^*$ which gives

$$\|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1^2 \leq s_0(\hat{\beta} - \beta^*)^\top\hat{\Sigma}(\hat{\beta} - \beta^*)/\phi_t^2 \,. \tag{C.4}$$

From (C.3), we have

$$2\mathcal{E}(\hat{\beta}) + \lambda_t\|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda_t\|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 \,.$$

Therefore, we have

$$\begin{aligned}
2\mathcal{E}(\hat{\beta}) + \lambda_t \|\hat{\beta} - \beta^*\|_1 &= 2\mathcal{E}(\hat{\beta}) + \lambda_t \|\hat{\beta}_{S_0^c}\|_1 + \lambda_t \|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 \\
&\leq 3\lambda_t \|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 + \lambda_t \|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 \\
&= 4\lambda_t \|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 \\
&\leq 4\lambda_t \sqrt{s_0 (\hat{\beta} - \beta^*)^\top \hat{\Sigma} (\hat{\beta} - \beta^*)} / \phi_t \\
&\leq \kappa_0 (\hat{\beta} - \beta^*)^\top \hat{\Sigma} (\hat{\beta} - \beta^*) + \frac{4\lambda_t^2 s_0}{\kappa_0 \phi_t^2} \\
&\leq 2\mathcal{E}(\hat{\beta}) + \frac{4\lambda^2 s_0}{\kappa_0 \phi_t^2}
\end{aligned}$$

where the second inequality is from applying the compatibility condition (C.4) and the third inequality is by using $4uv \leq u^2 + 4v^2$ with $u = \sqrt{\kappa_0 (\hat{\beta} - \beta^*)^\top \hat{\Sigma} (\hat{\beta} - \beta^*)}$ and $v = \frac{\lambda_t \sqrt{s_0}}{\phi_t \sqrt{\kappa_0}}$. The last inequality is from Lemma C.2. Hence, rearranging gives

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{4 s_0 \lambda_t}{\kappa_0 \phi_t^2} \,.$$

This completes the proof. □

C.1.2 Proof of Lemma C.1

Proof. By the definitions of the negative log-likelihood $\ell_t(\beta)$ and its expectation $\bar{\ell}_t(\beta)$, we can rewrite the empirical process $v_t(\beta)$ as

$$\begin{aligned} v_t(\beta) &= \ell_t(\beta) - \bar{\ell}_t(\beta) \\ &= -\frac{1}{t}\sum_{\tau=1}^{t}\left[Y_\tau X_\tau^\top \beta - m(X_\tau^\top \beta)\right] + \frac{1}{t}\sum_{\tau=1}^{t}\left[\mu(X_\tau^\top \beta^*)X_\tau^\top \beta - m(X_\tau^\top \beta)\right] \\ &= -\frac{1}{t}\sum_{\tau=1}^{t}\left[Y_\tau X_\tau^\top \beta - \mu(X_\tau^\top \beta^*)X_\tau^\top \beta\right] \\ &= -\frac{1}{t}\sum_{\tau=1}^{t}\epsilon_\tau X_\tau^\top \beta \end{aligned}$$

where the last equality uses the definition of ϵ_τ. Then, the empirical process deviation is

$$v_t(\hat{\beta}_t) - v_n(\beta^*) = -\frac{1}{t}\sum_{\tau=1}^{t}\epsilon_\tau X_\tau^\top(\hat{\beta}_t - \beta^*).$$

Applying Hölder's inequality, we have

$$|v_t(\hat{\beta}_t) - v_t(\beta^*)| \leq \frac{1}{t}\left\|\sum_{\tau=1}^{t}\epsilon_\tau X_\tau\right\|_\infty \|\hat{\beta}_t - \beta^*\|_1.$$

Then controlling the empirical process reduces to controlling $\frac{1}{t}\left\|\sum_{\tau=1}^{t}\epsilon_\tau X_\tau\right\|_\infty$. Then, using the union bound, it follows that

$$\begin{aligned} \mathbb{P}\left(\frac{1}{t}\left\|\sum_{\tau=1}^{t}\epsilon_\tau X_\tau\right\|_\infty \leq \lambda\right) &= 1 - \mathbb{P}\left(\frac{1}{t}\left\|\sum_{\tau=1}^{t}\epsilon_\tau X_\tau\right\|_\infty > \lambda\right) \\ &\geq 1 - \sum_{j=1}^{d}\mathbb{P}\left(\frac{1}{t}\left|\sum_{\tau=1}^{t}\epsilon_\tau X_\tau^{(j)}\right| > \lambda\right) \end{aligned}$$

where $X_\tau^{(j)}$ is the j-th element of X_τ. For each $j \in [d]$, and $\tau \in [t]$, we let $Z_\tau^{(j)} := \epsilon_\tau X_\tau^{(j)}$. Let $\widetilde{\mathcal{F}}_{t-1}$ denote the sigma-field that contains all observed information prior to taking an

action in round t, i.e., $\widetilde{\mathcal{F}}_{t-1}$ is generated by random variables of previously chosen actions $\{a_1, ..., a_{t-1}\}$, their features $\{X_1, ..., X_{t-1}\}$, the corresponding rewards $\{Y_1, ..., Y_{t-1}\}$ and the set of feature vectors $\mathcal{X}_t = \{X_{t,1}, ..., X_{t,K}\}$ in round t.

Then, each $\{Z_\tau^{(j)}\}_{\tau=1}^t$ for $j \in [d]$ is a martingale difference sequence adapted to the filtration $\widetilde{\mathcal{F}}_1 \subset ... \subset \widetilde{\mathcal{F}}_\tau$ since $\mathbb{E}[\epsilon_\tau X_\tau^{(j)} | \widetilde{\mathcal{F}}_{\tau-1}] = X_\tau^{(j)} \mathbb{E}[\epsilon_\tau | \widetilde{\mathcal{F}}_{\tau-1}] = 0$ for each j. Note that each $X_\tau^{(j)}$ is a bounded random variable with $|X_\tau^{(j)}| \leq \|X_\tau\|_\infty \leq \|X_\tau\|_2 \leq x_{\max}$. Then from the fact that ϵ_τ is σ^2-sub-Gaussian, it follows that $Z_\tau^{(j)}$ is also σ^2-sub-Gaussian. That is,

$$\begin{aligned}
\mathbb{E}\left[\exp(\alpha Z_\tau^{(j)}) \mid \widetilde{\mathcal{F}}_{\tau-1}\right] &= \mathbb{E}\left[\exp\left\{\left(\alpha X_\tau^{(j)}\right)\epsilon_\tau\right\} \mid \widetilde{\mathcal{F}}_{\tau-1}\right] \\
&\leq \mathbb{E}\left[\exp(\alpha x_{\max}\epsilon_\tau) \mid \widetilde{\mathcal{F}}_{\tau-1}\right] \\
&\leq \exp\left(\frac{\alpha^2 x_{\max}^2 \sigma^2}{2}\right)
\end{aligned}$$

for any $\alpha \in \mathbb{R}$. Then, using the concentration result in Lemma C.11, we have

$$\mathbb{P}\left(\left|\sum_{\tau=1}^t \epsilon_{ti} X_\tau^{(j)}\right| > t\lambda\right) \leq 2\exp\left(-\frac{t^2\lambda^2}{2t\sigma^2 x_{\max}^2}\right) \leq 2\exp\left(-\frac{t\lambda^2}{2\sigma^2 x_{\max}^2}\right).$$

So, with $\lambda = \sigma x_{\max}\sqrt{\frac{2[\log(2/\delta)+\log d]}{t}}$, we have

$$\mathbb{P}\left(\frac{1}{t}\left\|\sum_{\tau=1}^t \epsilon_\tau X_\tau\right\|_\infty \leq \lambda\right) \geq 1 - 2d\exp\left(\log\frac{\delta}{2} - \log d\right) = 1 - \delta\,.$$

□

Lemma C.2. *The excess risk is lower-bounded by*

$$\mathcal{E}(\hat{\beta}_t) \geq \frac{\kappa_0}{2}(\hat{\beta}_t - \beta^*)^\top \hat{\Sigma}(\hat{\beta}_t - \beta^*)\,.$$

Proof. By the definition of the excess risk $\mathcal{E}(\beta)$, we have

$$\begin{aligned}\mathcal{E}(\beta) &= \bar{\ell}_t(\beta) - \bar{\ell}_t(\beta^*) \\ &= -\frac{1}{t}\sum_{\tau=1}^{t}\left[\mu(X_\tau^\top\beta^*)X_\tau^\top\beta - m(X_\tau^\top\beta)\right] + \frac{1}{t}\sum_{\tau=1}^{t}\left[\mu(X_\tau^\top\beta^*)X_\tau^\top\beta^* - m(X_\tau^\top\beta^*)\right]\,.\end{aligned}$$

Since $\dot{m}(\cdot) = \mu(\cdot)$, we have $\nabla_\beta\bar{\ell}_t(\beta^*) = \mathbf{0}_d$. Hence, the gradient of the excess risk $\nabla_\beta\mathcal{E}(\beta)$ and the Hessian are given as

$$\begin{aligned}\nabla_\beta\mathcal{E}(\beta) &= -\frac{1}{t}\sum_{\tau=1}^{t}\left[\mu(X_\tau^\top\beta^*)X_\tau - \mu(X_\tau^\top\beta)X_\tau\right], \\ H_{\mathcal{E}}(\beta) &:= \nabla_\beta^2\mathcal{E}(\beta) = \frac{1}{t}\sum_{\tau=1}^{t}\dot{\mu}(X_\tau^\top\beta)X_\tau X_\tau^\top\,.\end{aligned}$$

Using the Taylor expansion, with $\bar{\beta} = c\beta^* + (1-c)\hat{\beta}$ for some $c \in (0,1)$

$$\mathcal{E}(\hat{\beta}_t) = \mathcal{E}(\beta^*) + \nabla_\beta\mathcal{E}(\beta^*)^\top(\hat{\beta}_t - \beta^*) + \frac{1}{2}(\hat{\beta}_t - \beta^*)^\top H_{\mathcal{E}}(\bar{\beta})(\hat{\beta}_t - \beta^*)\,. \tag{C.5}$$

Note that by the definition of β^*, we have $\mathcal{E}(\beta^*) = 0$ and $\nabla_\beta\mathcal{E}(\beta^*) = \nabla_\beta\ell(\beta^*) = \mathbf{0}_d$. Hence, combining with the definition of the Hessian, we have

$$\begin{aligned}\mathcal{E}(\hat{\beta}_t) &= \frac{1}{2}(\hat{\beta}_t - \beta^*)^\top\left[\frac{1}{t}\sum_{\tau=1}^{t}\dot{\mu}(X_\tau^\top\bar{\beta})X_\tau X_\tau^\top\right](\hat{\beta}_t - \beta^*) \\ &\geq \frac{\kappa_0}{2}(\hat{\beta}_t - \beta^*)^\top\hat{\Sigma}(\hat{\beta}_t - \beta^*)\end{aligned}$$

where the last inequality is from Assumption 4.2 and $\hat{\Sigma} = \frac{1}{t}\sum_{\tau=1}^{t}X_\tau X_\tau^\top$. □

C.1.3 Proof of Lemma 4.2

Proof. Consider $\mathcal{X} = \{X_1, X_2\}$. Let the joint density function of x_1, x_2 as $p_{\mathcal{X}}(x_1, x_2)$. Then we have

$$\begin{aligned}
\mathbb{E}[\mathbf{X}^\top \mathbf{X}] &= \int (x_1 x_1^\top + x_2 x_2^\top) p_{\mathcal{X}}(x_1, x_2) dx_1, x_2 \\
&= \int x_1 x_1^\top \left[\mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} + \mathbb{1}\left\{ (x_1 - x_2)^\top \beta \leq 0 \right\} \right] p_{\mathcal{X}}(x_1, x_2) dx_1, x_2 \\
&\quad + \int x_2 x_2^\top \left[\mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} + \mathbb{1}\left\{ (x_1 - x_2)^\top \beta \leq 0 \right\} \right] p_{\mathcal{X}}(x_1, x_2) dx_1, x_2
\end{aligned}$$

Let's first look at the first integral.

$$\begin{aligned}
&\int x_1 x_1^\top \left[\mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} + \mathbb{1}\left\{ (x_1 - x_2)^\top \beta \leq 0 \right\} \right] p_{\mathcal{X}}(x_1, x_2) dx_1, x_2 \\
&= \int x_1 x_1^\top \left[\mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} p_{\mathcal{X}}(x_1, x_2) + \mathbb{1}\left\{ -(x_1 - x_2)^\top \beta \geq 0 \right\} p_{\mathcal{X}}(x_1, x_2) \right] dx_1, x_2 \\
&\preccurlyeq \int x_1 x_1^\top \mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} p_{\mathcal{X}}(x_1, x_2) dx_1, x_2 \\
&\quad + \nu \int x_1 x_1^\top \mathbb{1}\left\{ -(x_1 - x_2)^\top \beta \geq 0 \right\} p_{\mathcal{X}}(-x_1, -x_2) dx_1, x_2 \\
&= \int x_1 x_1^\top \mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} p_{\mathcal{X}}(x_1, x_2) dx_1, x_2 \\
&\quad + \nu \int x_1 x_1^\top \mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} p_{\mathcal{X}}(x_1, x_2) dx_1, x_2 \\
&= (1 + \nu) \int x_1 x_1^\top \mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} p_{\mathcal{X}}(x_1, x_2) dx_1, x_2 \\
&= (1 + \nu) \mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{ X_1 = \operatorname*{argmax}_{X \in \mathcal{X}} X^\top \beta \} \right]
\end{aligned}$$

where the inequality follows from Assumption 4.4. Likewise, we can show for the second integral that

$$\begin{aligned}
&\int x_2 x_2^\top \left[\mathbb{1}\left\{ (x_1 - x_2)^\top \beta \geq 0 \right\} + \mathbb{1}\left\{ (x_1 - x_2)^\top \beta \leq 0 \right\} \right] p_{\mathcal{X}}(x_1, x_2) dx_1, x_2 \\
&= (1 + \nu) \mathbb{E}\left[X_2 X_2^\top \mathbb{1}\{ X_1 = \operatorname*{argmax}_{X \in \mathcal{X}} X^\top \beta \} \right].
\end{aligned}$$

Hence,

$$\mathbb{E}[\mathbf{X}^\top \mathbf{X}] = (1+\nu)\left(\mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_1 = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top \beta\}\right] + \mathbb{E}\left[X_2 X_2^\top \mathbb{1}\{X_2 = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top \beta\}\right]\right).$$

Therefore, with the fact that $\nu \geq 1$, we have

$$\sum_{i=1}^{2} \mathbb{E}\left[X_i X_i^\top \mathbb{1}\{X_i = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top \beta\}\right] \succcurlyeq \frac{2}{1+\nu} \cdot \frac{1}{2}\mathbb{E}[\mathbf{X}^\top \mathbf{X}] \succcurlyeq \nu^{-1}\Sigma\,.$$

□

C.1.4 Bernstein-type Inequality for Adapted Samples

In this section, we derive a Bernstein-type inequality for adapted samples which is shown in Lemma C.5. We first define the following function of a random variable X_t which is used throughout this section.

Definition C.1. *For all i, j with $1 \leq i \leq j \leq d$, we define $\gamma_t^{ij}(X_t)$ to be a real-value function which take random variable $X_t \in \mathbb{R}^d$ as input:*

$$\gamma_t^{ij}(X_t) := \frac{1}{2x_{\max}^2}\left(X_t^{(i)} X_t^{(j)} - \mathbb{E}[X_t^{(i)} X_t^{(j)} \mid \mathcal{F}_{t-1}]\right) \tag{C.6}$$

where $X_t^{(i)}$ is the i-th element of X_t.

It is easy to see that $\mathbb{E}\left[\gamma_t^{ij}(X_t) \mid \mathcal{F}_{t-1}\right] = 0$ and $\mathbb{E}\left[|\gamma_t^{ij}(X_t)|^m \mid \mathcal{F}_{t-1}\right] \leq 1$ for all integer $m \geq 2$. While we introduce this specific function $\gamma_t^{ij}(X_t)$ in order to connect to the matrix concentration $\|\Sigma_\tau - \hat{\Sigma}_\tau\|_\infty$, Lemma C.4 and Lemma C.5 can be applied to any function $\gamma_t^{ij}(X_t)$ that satisfies the zero mean and the bounded m-th moment conditions.

Lemma C.3 (Bühlmann and Van De Geer (2011), Lemma 14.1)**.** *Let $Z_t \in \mathbb{R}$ be a random variable with $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$. Then it holds that*

$$\log \mathbb{E}\left[e^{Z_t} \mid \mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[e^{|Z_t|} \mid \mathcal{F}_{t-1}\right] - 1 - \mathbb{E}\left[|Z| \mid \mathcal{F}_{t-1}\right].$$

Proof. The proof follows directly from the proof of Lemma 14.1 in Bühlmann and Van De Geer (2011), applying their result to a conditional expectation. For any $c > 0$,

$$\begin{aligned}\exp(Z_t - c) - 1 &\leq \frac{\exp(Z_t)}{1+c} - 1 \\ &= \frac{e^{Z_t} - 1 - Z_t + Z_t - c}{1+c} \\ &\leq \frac{e^{|Z_t|} - 1 - |Z_t| + Z_t - c}{1+c}.\end{aligned}$$

Let $c = \mathbb{E}\left[e^{|Z_t|} \mid \mathcal{F}_{t-1}\right] - 1 - \mathbb{E}\left[|Z| \mid \mathcal{F}_{t-1}\right]$. Hence, since $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$,

$$\mathbb{E}\left[\exp(Z_t - c) \mid \mathcal{F}_{t-1}\right] - 1 \leq \frac{\mathbb{E}\left[e^{|Z_t|} \mid \mathcal{F}_{t-1}\right] - 1 - \mathbb{E}\left[|Z_t| \mid \mathcal{F}_{t-1}\right] - c}{1+c} = 0.$$

□

Lemma C.4. *Suppose $\mathbb{E}\left[\gamma_t^{ij}(X_t) \mid \mathcal{F}_{t-1}\right] = 0$ and $\mathbb{E}\left[|\gamma_t^{ij}(X_t)|^m \mid \mathcal{F}_{t-1}\right] \leq m!$ for all integer $m \geq 2$, all $t \geq 1$ and all $1 \leq i \leq j \leq d$. Then, for $L > 1$ we have*

$$\mathbb{E}\left[\exp\left(\frac{1}{L}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right)\right] \leq \exp\left(\frac{\tau}{L(L-1)}\right).$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{1}{L}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{1}{L}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right) \mid \mathcal{F}_{\tau-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\gamma_\tau^{ij}(X_\tau)}{L}\right) \mid \mathcal{F}_{\tau-1}\right]\exp\left(\frac{1}{L}\sum_{t=1}^{\tau-1}\gamma_t^{ij}(X_t)\right)\right] \\ &\le e^{\frac{1}{L(L-1)}}\mathbb{E}\left[\exp\left(\frac{1}{L}\sum_{t=1}^{\tau-1}\gamma_t^{ij}(X_t)\right)\right]\end{aligned}$$

where the inequality is from Lemma C.3 and noting that

$$\begin{aligned}\log\mathbb{E}\left[\exp\left(\frac{\gamma_\tau^{ij}(X_\tau)}{L}\right) \mid \mathcal{F}_{\tau-1}\right] &\le \mathbb{E}\left[e^{|\gamma_\tau^{ij}(X_\tau)|/\tau} - 1 - \frac{|\gamma_\tau^{ij}(X_\tau)|}{L} \mid \mathcal{F}_{\tau-1}\right] \\ &= \mathbb{E}\left[\sum_{m=2}^{\infty}\frac{|\gamma_\tau^{ij}(X_\tau)|^m}{L^m m!} \mid \mathcal{F}_{\tau-1}\right] \\ &= \sum_{m=2}^{\infty}\frac{\mathbb{E}\left[|\gamma_\tau^{ij}(X_\tau)|^m \mid \mathcal{F}_{\tau-1}\right]}{L^m m!} \\ &\le \frac{1}{L(L-1)}.\end{aligned}$$

Then, repeatedly applying this to the rest of the sum $\frac{1}{L}\sum_{t=1}^{\tau-1}\gamma_t^{ij}(X_t)$, we have

$$\mathbb{E}\left[\exp\left(\frac{1}{L}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right)\right] \le \exp\left(\frac{\tau}{L(L-1)}\right).$$

□

Lemma C.5 (Bernstein-type inequality for adapted samples)**.** *Suppose* $\mathbb{E}\left[\gamma_t^{ij}(X_t) \mid \mathcal{F}_{t-1}\right] = 0$ *and* $\mathbb{E}\left[|\gamma_t^{ij}(X_t)|^m \mid \mathcal{F}_{t-1}\right] \le m!$ *for all integer* $m \ge 2$*, all* $t \ge 1$ *and all* $1 \le i \le j \le d$*. Then for all* $w > 0$*, we have*

$$\mathbb{P}\left(\max_{1\le i\le j\le d}\left|\frac{1}{\tau}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right| \ge w + \sqrt{2w} + \sqrt{\frac{4\log(2d^2)}{\tau}} + \frac{2\log(2d^2)}{\tau}\right) \le \exp\left(-\frac{\tau w}{2}\right).$$

Proof. Using the Chernoff bound and Lemma C.4, for any $L > 1$ we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t) \geq a\right) &= \mathbb{P}\left(\exp\left(\frac{1}{L}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right) \geq \exp\Big(\frac{a}{L}\Big)\right) \\
&\leq \frac{\mathbb{E}\left[\exp\Big(\frac{1}{L}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\Big)\right]}{\exp\Big(\frac{a}{L}\Big)} \\
&\leq \exp\Big(-\frac{a}{L}\Big)\exp\left(\frac{\tau}{L(L-1)}\right) \\
&= \exp\left(-\frac{a}{L} + \frac{\tau}{L(L-1)}\right).
\end{aligned}$$

Here, $L = \frac{\tau+a+\sqrt{\tau^2+\tau a}}{a}$ minimizes the right hand side above for $L > 1$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t) \geq a\right) &\leq \exp\left\{-\frac{a^2}{\tau+a+\sqrt{\tau^2+\tau a}} + \frac{\tau a^2}{(\tau+a+\sqrt{\tau^2+\tau a})(\tau+\sqrt{\tau^2+\tau a})}\right\} \\
&= \exp\left\{-\left(\frac{\sqrt{1+a/\tau}}{1+\sqrt{1+a/\tau}}\right)\frac{a^2}{\tau+a+\sqrt{\tau^2+\tau a}}\right\} \\
&\leq \exp\left\{-\frac{a^2}{2\left(\tau+a+\sqrt{\tau^2+\tau a}\right)}\right\} \\
&\leq \exp\left\{-\frac{a^2}{2\left(\tau+a+\sqrt{\tau^2+2\tau a}\right)}\right\}.
\end{aligned}$$

Choosing $a = \tau\left(w + \sqrt{2w}\right)$ gives

$$\mathbb{P}\left(\frac{1}{\tau}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t) \geq w + \sqrt{2w}\right) \leq \exp\Big(-\frac{\tau w}{2}\Big). \tag{C.7}$$

Then for the maximal inequality, we first apply the union bound to (C.7).

$$\begin{aligned}\mathbb{P}\left(\max_{1\le i\le j\le d}\frac{1}{\tau}\left|\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right|\ge w+\sqrt{2w}\right) &\le \sum_{1\le i\le j\le d} 2\mathbb{P}\left(\frac{1}{\tau}\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\ge w+\sqrt{2w}\right)\\ &\le 2d^2\exp\Big(-\frac{\tau w}{2}\Big)\\ &= \exp\Big(-\frac{\tau w}{2}+\log(2d^2)\Big)\ .\end{aligned}$$

Then,

$$\begin{aligned}&\mathbb{P}\left(\max_{1\le i\le j\le d}\frac{1}{\tau}\left|\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right|\ge w+\sqrt{2w}+\sqrt{\frac{4\log(2d^2)}{\tau}}+\frac{2\log(2d^2)}{\tau}\right)\\ &\le \mathbb{P}\left(\max_{1\le i\le j\le d}\frac{1}{\tau}\left|\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right|\ge \left(w+\frac{2\log(2d^2)}{\tau}\right)+\sqrt{2\left(w+\frac{2\log(2d^2)}{\tau}\right)}\right)\\ &= \mathbb{P}\left(\max_{1\le i\le j\le d}\frac{1}{\tau}\left|\sum_{t=1}^{\tau}\gamma_t^{ij}(X_t)\right|\ge w'+\sqrt{2w'}\right)\\ &\le \exp\Big(-\frac{\tau w'}{2}+\log(2d^2)\Big)\\ &= \exp\Big(-\frac{\tau w}{2}\Big)\end{aligned}$$

where $w' = w + \frac{2\log(2d^2)}{\tau}$. □

C.1.5 Proof of Lemma 4.3

Proof. Notice the difference between the unconditional theoretical Gram matrix Σ and its adapted version $\mathbb{E}[X_t X_t^\top|\mathcal{F}_{t-1}]$ which is a conditional covariance matrix conditioned on the history $\mathcal{F}_{t-1}$. Recall that from Algorithm 8, in each round t we choose X_t given the history $\mathcal{F}_{t-1}$. More precisely, we compute β_t based on $\mathcal{F}_{t-1}$ and choose X_t which maximizes the product $X_t^\top\hat{\beta}_t$, i.e., $\operatorname{argmax}_{X\in\mathcal{X}_t} X^\top\hat{\beta}_t$ where $\mathcal{X}_t = \{X_{t,1}, X_{t,2}\}$. Hence, we

can write $\mathbb{E}[X_t X_t^\top | \mathcal{F}_{t-1}]$ as the following:

$$\mathbb{E}[X_t X_t^\top | \mathcal{F}_{t-1}] = \sum_{i=1}^{2} \mathbb{E}_{\mathcal{X}_t}\Big[X_{t,i} X_{t,i}^\top \mathbb{1}\{X_{ti} = \operatorname*{argmax}_{X \in \mathcal{X}_t} X^\top \hat{\beta}_t\} \mid \hat{\beta}_t \Big] .$$

From Lemma 4.2, it follows that

$$\mathbb{E}[X_t X_t^\top | \mathcal{F}_{t-1}] \succcurlyeq \nu^{-1} \Sigma .$$

Now, taking an average over t gives,

$$\Sigma_\tau = \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}[X_t X_t^\top | \mathcal{F}_{t-1}] \succcurlyeq \nu^{-1} \Sigma .$$

Then, we define $\widetilde{\beta}$ corresponding to compatibility constant $\phi^2(\Sigma_\tau, S_0)$, that is,

$$\widetilde{\beta} := \operatorname*{argmin}_{\beta} \left\{ \frac{\beta^\top \Sigma_\tau \beta}{\|\beta_{S_0}\|_1^2} : \|\beta_{S_0^c}\|_1 \le 3 \|\beta_{S_0}\|_1 \neq 0 \right\} .$$

Therefore, it follows that

$$\frac{\widetilde{\beta}^\top \Sigma_\tau \widetilde{\beta}}{\|\widetilde{\beta}_{S_0}\|_1^2} \ge \frac{\widetilde{\beta}^\top \Sigma \widetilde{\beta}}{\nu \|\widetilde{\beta}_{S_0}\|_1^2} \ge \frac{\phi_0^2}{\nu} \tag{C.8}$$

where the second inequality is by the compatibility condition on Σ. Thus, Σ_τ satisfies the compatibility condition with compatibility constant $\phi^2(\Sigma_\tau, S_0) = \frac{\phi_0^2}{\nu}$.
Now, noting that $\frac{1}{2x_{\max}^2} \|\Sigma_\tau - \hat{\Sigma}_\tau\|_\infty = \max_{1 \le i \le j \le d} \frac{1}{\tau} \left| \sum_{t=1}^{\tau} \gamma_t^{ij}(X_t) \right|$ for $\gamma_t^{ij}(\cdot)$ defined in (C.6), we can use a Bernstein-type inequality for adapted samples in Lemma C.5 to get

$$\mathbb{P}\left(\frac{\|\Sigma_\tau - \hat{\Sigma}_\tau\|_\infty}{2x_{\max}^2} \ge w + \sqrt{2w} + \sqrt{\frac{4 \log(2d^2)}{\tau}} + \frac{2 \log(2d^2)}{\tau} \right) \le \exp\Big(-\frac{\tau w}{2} \Big) .$$

For $\tau \geq \frac{2\log(2d^2)}{C_0(s_0)^2}$ where $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 \nu x_{\max}^2}\right)$, letting $w = C_0(s_0)^2$ gives

$$\begin{aligned} w + \sqrt{2w} + \sqrt{\frac{4\log(2d^2)}{\tau}} + \frac{2\log(2d^2)}{\tau} &\leq 2\left(C_0(s_0)^2 + \sqrt{2} C_0(s_0)\right) \\ &\leq 4 C_0(s_0) \\ &\leq \frac{\phi_0^2}{64 s_0 \nu x_{\max}^2} \\ &= \frac{\phi^2(\Sigma_\tau, S_0)}{64 s_0 x_{\max}^2}\,. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{P}\left(\frac{\|\Sigma_\tau - \hat{\Sigma}_\tau\|_\infty}{2x_{\max}^2} \geq \frac{\phi^2(\Sigma_\tau, S_0)}{64 s_0 x_{\max}^2}\right) &\leq \mathbb{P}\left(\frac{\|\Sigma_\tau - \hat{\Sigma}_\tau\|_\infty}{2x_{\max}^2} \geq w + \sqrt{2w} + \sqrt{\frac{4\log(2d^2)}{\tau}} + \frac{2\log(2d^2)}{\tau}\right) \\ &\leq \exp\left(-\frac{\tau w}{2}\right) \\ &= \exp\left(-\frac{\tau C_0(s_0)^2}{2}\right)\,. \end{aligned}$$

□

Corollary C.1. *For $t \geq \frac{2\log(2d^2)}{C_0(s_0)^2}$ where $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 \nu x_{\max}^2}\right)$, the empirical Gram matrix $\hat{\Sigma}_t$ satisfies the compatibility condition with compatibility constant $\phi_t \geq \frac{\phi_0^2}{2\nu} > 0$ with probability at least $1 - \exp\{-tC_0(s_0)^2/2\}$.*

Proof. We can use Corollary 4.1 (Bühlmann and Van De Geer (2011), Corollary 6.8) to show that the empirical Gram matrix $\hat{\Sigma}_\tau$ satisfies the compatibility condition as long as Σ_τ satisfies the compatibility condition. From (C.8), we know Σ_τ satisfies the compatibility condition with compatibility constant $\frac{\phi_0^2}{\nu}$. Then, combining Lemma 4.3 and Corollary 4.1, it follows that given $\|\Sigma_t - \hat{\Sigma}_t\|_\infty \leq \frac{\phi_0^2}{32 s_0 \nu}$ for $t \geq \lceil T_0 \rceil$, we have

$$\phi^2(\hat{\Sigma}_t, S_0) \geq \frac{\phi^2(\Sigma_t, S_0)}{2} \geq \frac{\phi_0^2}{2\nu} > 0\,.$$

That is, $\hat{\Sigma}_\tau$ satisfies the compatibility condition with compatibility constant which is at least $\frac{\phi_0^2}{2\nu} > 0$. □

C.2 Proof of Theorem 4.1

Proof. First, let $T_0 := \frac{2\log(2d^2)}{C_0(s_0)^2}$ where $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 \nu x_{\max}^2}\right)$. Also, we define the high probability event $\mathcal{E}_t$:

$$\mathcal{E}_t := \left\{ \|\Sigma_t - \hat{\Sigma}_t\|_\infty \geq \frac{\phi_0^2}{32 s_0 \nu} \right\}.$$

Hence, on this event $\mathcal{E}_t$, if $t \geq T_0$, then from Corollary C.1 we have $\phi_t^2 \geq \frac{\phi_0^2}{2\nu}$, i.e., the compatibility condition holds in round t. Slightly overloading the subscript for brevity, let $X_t := X_{t,a_t}$ be a feature of the arm chosen in round t and $X_{a_t^*} := X_{t,a_t^*}$ be the feature of the optimal arm in round t. First, we look at the (non-expected) immediate regret $\mathrm{Reg}(t)$ with $\mathcal{R}(t) = \mathbb{E}[\mathrm{Reg}(t)]$ in round t. Notice that by Assumptions 4.1 and 4.2 and by the mean value theorem, $\mathrm{Reg}(t)$ is bounded by

$$\mathrm{Reg}(t) \leq \kappa_1 \left(X_{a_t^*}^\top \beta^* - X_t^\top \beta^* \right) \leq \kappa_1 \|X_{a_t^*} - X_t\|_2 \|\beta^*\|_2 \leq 2\kappa_1 x_{\max} b$$

Then we can decompose the immediate regret as follows.

$$\begin{aligned}
\mathrm{Reg}(t) &= \mathrm{Reg}(t)\mathbb{1}(t \leq T_0) + \mathrm{Reg}(t)\mathbb{1}(t > T_0, \mathcal{E}_t) + \mathrm{Reg}(t)\mathbb{1}(t > T_0, \mathcal{E}_t^c) \\
&\leq 2\kappa_1 x_{\max} b \mathbb{1}(t \leq T_0) + \mathrm{Reg}(t)\mathbb{1}(t > T_0, \mathcal{E}_t) + 2\kappa_1 x_{\max} b \mathbb{1}(t > T_0, \mathcal{E}_t^c) \\
&= 2\kappa_1 x_{\max} b \mathbb{1}(t \leq T_0) + \mathrm{Reg}(t)\mathbb{1}\left(\mu(X_t^\top \hat{\beta}_t) \geq \mu(X_{a_t^*}^\top \hat{\beta}_t), t > T_0, \mathcal{E}_t\right) \\
&\quad + 2\kappa_1 x_{\max} b \mathbb{1}(t > T_0, \mathcal{E}_t^c)
\end{aligned}$$

where the last equality follows from the optimality of X_t with respect to parameter $\hat{\beta}_t$, i.e., $X_t = \text{argmax}_{X \in \mathcal{X}_t} \mu(X^\top \hat{\beta}_t)$. For the second term, we have

$$\begin{aligned}
\mathbb{P}\left(\mu(X_t^\top \hat{\beta}_t) \geq \mu(X_{a_t^*}^\top \hat{\beta}_t)\right) &= \mathbb{P}\left(\mu(X_t^\top \hat{\beta}_t) - \mu(X_{a_t^*}^\top \hat{\beta}_t) + \text{Reg}(t) \geq \text{Reg}(t)\right) \\
&= \mathbb{P}\left((\mu(X_t^\top \hat{\beta}_t) - \mu(X_t^\top \beta^*)) - (\mu(X_{a_t^*}^\top \hat{\beta}_t) - \mu(X_{a_t^*}^\top \beta^*)) \geq \text{Reg}(t)\right) \\
&\leq \mathbb{P}\left(|\mu(X_t^\top \hat{\beta}_t) - \mu(X_t^\top \beta^*)| + |\mu(X_{a_t^*}^\top \hat{\beta}_t) - \mu(X_{a_t^*}^\top \beta^*)| \geq \text{Reg}(t)\right) \\
&\leq \mathbb{P}\left(\kappa_1 \|\hat{\beta}_t - \beta^*\|_1 \|X_t\|_\infty + \kappa_1 \|\hat{\beta}_t - \beta^*\|_1 \|X_{a_t^*}\|_\infty \geq \text{Reg}(t)\right) \\
&\leq \mathbb{P}\left(2\kappa_1 \|\hat{\beta}_t - \beta^*\|_1 \geq \text{Reg}(t)\right)
\end{aligned}$$

where the last inequality is from the fact that each $X_{t,i}$ is bounded. For an arbitrary constant $g_t > 0$, we continue with expected regret $\mathcal{R}(t) = \mathbb{E}[\text{Reg}(t)]$ for $t > T_0$.

$$\begin{aligned}
\mathcal{R}(t) &\leq \mathbb{E}\left[\text{Reg}(t) \mathbb{1}\left(2\kappa_1 \|\hat{\beta}_t - \beta^*\|_1 \geq \text{Reg}(t), \mathcal{E}_t\right)\right] + 2\kappa_1 x_{\max} b \mathbb{P}(\mathcal{E}_t^c) \\
&= \mathbb{E}\left[\text{Reg}(t) \mathbb{1}\left(2\kappa_1 \|\hat{\beta}_t - \beta^*\|_1 \geq \text{Reg}(t), \text{Reg}(t) \leq \kappa_1 g_t, \mathcal{E}_t\right)\right] \\
&\quad + \mathbb{E}\left[\text{Reg}(t) \mathbb{1}\left(2\kappa_1 \|\hat{\beta}_t - \beta^*\|_1 \geq \text{Reg}(t), \text{Reg}(t) > \kappa_1 g_t, \mathcal{E}_t, \right)\right] + 2\kappa_1 x_{\max} b \mathbb{P}(\mathcal{E}_t^c) \\
&\leq \kappa_1 g_t + \kappa_1 \mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_1 \geq g_t, \mathcal{E}_t\right) + 2\kappa_1 x_{\max} b \mathbb{P}(\mathcal{E}_t^c)\,.
\end{aligned}$$

Summing over all rounds after the initial T_0 rounds, we have

$$\sum_{t=\lceil T_0 \rceil}^{T} \mathcal{R}(t) \leq \underbrace{\kappa_1 \sum_{t=\lceil T_0 \rceil}^{T} g_t}_{(a)} + \underbrace{\kappa_1 \sum_{t=\lceil T_0 \rceil}^{T} \mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_1 \geq g_t, \mathcal{E}_t\right)}_{(b)} + \underbrace{2\kappa_1 x_{\max} b \sum_{t=\lceil T_0 \rceil}^{T} \mathbb{P}(\mathcal{E}_t^c)}_{(c)}\,. \tag{C.9}$$

We first bound the term (b) in (C.9). We choose $g_t := \frac{2s_0\lambda_t}{\kappa_0 \phi_t^2} = \frac{4\sigma x_{\max} s_0}{\kappa_0 \phi_t^2} \sqrt{\frac{4\log t + 2\log d}{t}}$. Then using Lemma 4.1, we have

$$\mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_1 \geq g_t, \mathcal{E}_t\right) \leq \frac{2}{t^2}\,.$$

for all $t \geq T_0$. Therefore, it follows that

$$\sum_{t=\lceil T_0 \rceil}^{T} \mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_1 \geq g_t, \mathcal{E}_t\right) \leq \sum_{t=\lceil T_0 \rceil}^{T} \frac{2}{t^2} \leq \sum_{t=1}^{\infty} \frac{2}{t^2} \leq \frac{\pi^2}{3} < 4\,.$$

For the term (a) in (C.9), we have $\phi_t^2 \geq \frac{\phi_0^2}{2\nu}$ provided that event $\mathcal{E}_t$ holds. Hence, we have

$$\begin{aligned}
\sum_{t=\lceil T_0 \rceil}^{T} g_t &= \sum_{t=\lceil T_0 \rceil}^{T} \frac{4\sigma x_{\max} s_0}{\kappa_0 \phi_t^2} \sqrt{\frac{4\log t + 2\log d}{t}} \\
&\leq \sum_{t=\lceil T_0 \rceil}^{T} \frac{8\nu\sigma x_{\max} s_0}{\kappa_0 \phi_0^2} \sqrt{\frac{4\log t + 2\log d}{t}} \\
&\leq \frac{8\nu\sigma x_{\max} s_0 \sqrt{4\log T + 2\log d}}{\kappa_0 \phi_0^2} \sum_{t=\lceil T_0 \rceil}^{T} \frac{1}{\sqrt{t}} \\
&\leq \frac{8\nu\sigma x_{\max} s_0 \sqrt{4\log T + 2\log d}}{\kappa_0 \phi_0^2} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \\
&\leq \frac{16\nu\sigma x_{\max} s_0 \sqrt{4\log T + 2\log d}}{\kappa_0 \phi_0^2} \sqrt{T}
\end{aligned}$$

where the last inequality is from the fact that $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq \int_{t=0}^{T} \frac{1}{\sqrt{t}} = 2\sqrt{T}$.

Finally, for the term (c) in (C.9), we have from Lemma 4.3:

$$\begin{aligned}
\sum_{t=\lceil T_0 \rceil}^{T} \mathbb{P}(\mathcal{E}_t^c) &\leq \sum_{t=\lceil T_0 \rceil}^{T} \mathbb{P}\left(\|\Sigma_t - \hat{\Sigma}_t\|_\infty \geq \frac{\phi_0^2}{32 s_0 \nu}\right) \\
&\leq \sum_{t=\lceil T_0 \rceil}^{T} \exp\left(-\frac{t C_0(s_0)^2}{2}\right) \\
&\leq \sum_{t=1}^{\infty} \exp\left(-\frac{t C_0(s_0)^2}{2}\right) \\
&\leq \frac{2}{C_0(s_0)^2}\,.
\end{aligned}$$

□

C.3 Proof of Theorem 4.2

The proof follows similar arguments as the proof of Theorem 4.1. The key difference is that the RE condition involves ℓ_2 norm and therefore the analysis requires the Lasso oracle inequality of the GLM in ℓ_2 norm, which we provide as an extension of Lemma 4.1.

Corollary C.2. *Assume that the RE condition holds for $\hat{\Sigma}_t$ with active set S_0 and restricted eigenvalue ϕ_t. For some $\delta \in (0,1)$, let the regularization parameter λ_t be*

$$\lambda_t := 2\sigma x_{\max}\sqrt{\frac{2[\log(2/\delta) + \log d]}{t}}\,.$$

Then with probability at least $1-\delta$, we have

$$\|\hat{\beta}_t - \beta^*\|_2 \le \frac{3\sqrt{s_0}\lambda_t}{\kappa_0 \phi_t^2}\,.$$

Proof. Continuing from (C.3) in Lemma 4.1, the RE condition can be applied to the vector $\hat{\beta} - \beta^*$ which gives

$$\|\hat{\beta} - \beta^*\|_2^2 \le \frac{(\hat{\beta} - \beta^*)^\top \hat{\Sigma}_t (\hat{\beta} - \beta^*)}{\phi_t^2}\,. \tag{C.10}$$

Again from (C.3), we can use the margin condition in Lemma C.2

$$\begin{aligned} 3\lambda_t \|\hat{\beta}_{S_0} - \beta^*_{S_0}\|_1 &\ge 2\mathcal{E}(\hat{\beta}_n) \\ &\ge \kappa_0 (\hat{\beta} - \beta^*)^\top \hat{\Sigma}_t (\hat{\beta} - \beta^*) \\ &\ge \kappa_0 \phi_t^2 \|\hat{\beta} - \beta^*\|_2^2 \end{aligned}$$

where the last inequality is from (C.10) applying the RE condition. Then, it follows that

$$\begin{aligned}\kappa_0\phi_t^2\|\hat{\beta}-\beta^*\|_2^2 &\leq 3\lambda_t\|\hat{\beta}_{S_0}-\beta^*_{S_0}\|_1\\ &\leq 3\lambda_t\sqrt{s_0}\|\hat{\beta}_{S_0}-\beta^*_{S_0}\|_2\\ &\leq 3\lambda_t\sqrt{s_0}\|\hat{\beta}-\beta^*\|_2\,.\end{aligned}$$

Hence, dividing the both sides by $\|\hat{\beta}-\beta^*\|_2$ and rearranging gives

$$\|\hat{\beta}-\beta^*\|_2 \leq \frac{3\sqrt{s_0}\lambda_t}{\kappa_0\phi_t^2}\,.$$

This complete the proof. □

C.3.1 Ensuring the RE Condition for the Empirical Gram Matrix

To distinguish from the compatibility constant, we introduce the definition of a generic restricted eigenvalue of matrix M over active set S_0.

Definition C.2. *The restricted eigenvalue of M over S_0 is*

$$\phi_{RE}^2(M,S_0) := \min_{\beta}\left\{\frac{\beta^\top M\beta}{\|\beta\|_2^2} : \|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1 \neq 0\right\}.$$

Note that Assumption 4.5 only provides the RE condition for the theoretical Gram matrix Σ. Then, we follow the same arguments as in the analysis under the compatibility condition to show that $\phi^2_{\mathrm{RE}}(\Sigma_t,S_0) \geq \frac{\phi^2_{\mathrm{RE}}(\Sigma,S_0)}{\nu} > 0$, i.e., Σ_t satisfies the RE condition. Then using Lemma 4.3, we can show that $\hat{\Sigma}_t$ concentrates to Σ_t with high probability. The following lemma (similar to Corollary 4.1) ensures the RE condition of $\hat{\Sigma}_t$ conditioned on the matrix concentration of the empirical Gram matrix $\hat{\Sigma}_t$.

Lemma C.6. *Suppose that the RE condition holds for Σ_0 and the index set S with cardinality $s = |S|$, with restricted eigenvalue $\phi^2_{RE}(\Sigma_0, S) > 0$, and that $\|\Sigma_1 - \Sigma_0\|_\infty \leq \Delta$, where $32s\Delta \leq \phi^2_{RE}(\Sigma_0, S)$. Then, for the set S, the RE condition holds as well for Σ_1, with $\phi^2_{RE}(\Sigma_1, S) \geq \phi^2_{RE}(\Sigma_0, S)/2$.*

Proof. The proof is an adaptation of Lemma 6.17 in Bühlmann and Van De Geer, 2011 to the RE condition.

$$\begin{aligned}\left|\beta^\top \Sigma_1 \beta - \beta^\top \Sigma_0 \beta\right| &= \left|\beta^\top (\Sigma_1 - \Sigma_0)\beta\right| \\ &\leq \|\Sigma_1 - \Sigma_0\|_\infty \|\beta\|_1^2 \\ &\leq \Delta \|\beta\|_1^2\end{aligned}$$

For β such that $\|\beta_{S^c}\| \leq 3\|\beta_S\|$, we have the RE condition satisfied for Σ_0. Hence, we have

$$\|\beta\|_1 \leq 4\|\beta_S\|_1 \leq 4\sqrt{s}\|\beta_S\|_2 \leq 4\sqrt{s}\|\beta\|_2 \leq \frac{4\sqrt{s_0 \beta^\top \Sigma_0 \beta}}{\phi_{\mathrm{RE}}(\Sigma_0, S)}\,.$$

Therefore, it follows that

$$\left|\beta^\top \Sigma_1 \beta - \beta^\top \Sigma_0 \beta\right| \leq \frac{16 s \Delta \beta^\top \Sigma_0 \beta}{\phi^2_{\mathrm{RE}}(\Sigma_0, S)}\,.$$

Since $\beta^\top \Sigma_0 \beta > 0$, dividing the both sides by $\beta^\top \Sigma_0 \beta$ gives

$$\left|\frac{\beta^\top \Sigma_1 \beta}{\beta^\top \Sigma_0 \beta} - 1\right| \leq \frac{16 s \Delta}{\phi^2_{\mathrm{RE}}(\Sigma_0, S)}$$

Now, since $32s\Delta \leq \phi^2_{\mathrm{RE}}(\Sigma_0, S)$, it follows that

$$\frac{1}{2} \cdot \frac{\beta^\top \Sigma_0 \beta}{\|\beta\|_2^2} \leq \frac{\beta^\top \Sigma_1 \beta}{\|\beta\|_2^2} \leq \frac{3}{2} \cdot \frac{\beta^\top \Sigma_0 \beta}{\|\beta\|_2^2} .$$

Hence,

$$\phi^2_{\mathrm{RE}}(\Sigma_1, S) \geq \frac{\phi^2_{\mathrm{RE}}(\Sigma_0, S)}{2} .$$

□

C.3.2 Proof of Theorem 4.2

Proof. The proof of Theorem 4.2 follows the similar arguments as the proof of Theorem 4.1. The only difference is that we use ℓ_2 error bound $\|\hat{\beta}_t - \beta^*\|_2$ instead of $\|\hat{\beta}_t - \beta^*\|_1$. First, note that

$$\begin{aligned}
\mathbb{P}\left(\mu(X_t^\top \hat{\beta}_t) \geq \mu(X_{a_t^*}^\top \hat{\beta}_t)\right) &\leq \mathbb{P}\left(|\mu(X_t^\top \hat{\beta}_t) - \mu(X_t^\top \beta^*)| + |\mu(X_{a_t^*}^\top \hat{\beta}_t) - \mu(X_{a_t^*}^\top \beta^*)| \geq \mathrm{Reg}(t)\right) \\
&\leq \mathbb{P}\left(\kappa_1 \|\hat{\beta}_t - \beta^*\|_2 \|X_t\|_2 + \kappa_1 \|\hat{\beta}_t - \beta^*\|_2 \|X_t^*\|_2 \geq \mathrm{Reg}(t)\right) \\
&\leq \mathbb{P}\left(2\kappa_1 \|\hat{\beta}_t - \beta^*\|_2 \geq \mathrm{Reg}(t)\right) .
\end{aligned}$$

For an arbitrary constant $g_t > 0$, we continue with expected regret $\mathbb{E}[\mathrm{Reg}(t)]$ for $t > T_0$.

$$\mathcal{R}(t) \leq \kappa_1 g_t + \kappa_1 \mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_2 \geq g_t, \mathcal{E}_t\right) + 2\kappa_1 x_{\max} b \mathbb{P}(\mathcal{E}_t^c) .$$

Hence, the cumulative regret is bounded by

$$\sum_{t=1}^{T} \mathcal{R}(t) \leq 2\kappa_1 x_{\max} b T_0 + \kappa_1 \sum_{t=\lceil T_0 \rceil}^{T} g_t + \kappa_1 \sum_{t=\lceil T_0 \rceil}^{T} \mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_2 \geq g_t, \mathcal{E}_t\right) + 2\kappa_1 x_{\max} b \sum_{t=\lceil T_0 \rceil}^{T} \mathbb{P}(\mathcal{E}_t^c) .$$

Let $g_t := \frac{3\sqrt{s_0}\lambda_t}{2\kappa_0\phi_t^2} = \frac{6\sigma x_{\max}}{\kappa_0\phi_t^2}\sqrt{\frac{s_0(4\log t+2\log d)}{t}}$. From Lemma 4.1, we have

$$\mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_2 \geq g_t, \mathcal{E}_t\right) \leq \frac{2}{t^2}$$

for all t. Therefore, it follows that

$$\sum_{t=\lceil T_0\rceil}^{T} \mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_2 \geq g_t, \mathcal{E}_t\right) \leq \sum_{t=1}^{T} \mathbb{P}\left(2\|\hat{\beta}_t - \beta^*\|_2 \geq g_t, \mathcal{E}_t\right) \leq \frac{\pi^2}{3} < 4\,.$$

For $t \geq T_0$, we have $\phi_t^2 \geq \frac{\phi_1^2}{2\nu}$ provided that event $\mathcal{E}_t$ holds. Hence, we have

$$\begin{aligned}
\sum_{t=\lceil T_0\rceil}^{T} g_t &= \sum_{t=\lceil T_0\rceil}^{T} \frac{6\sigma x_{\max}}{\kappa_0\phi_t^2}\sqrt{\frac{s_0(4\log t+2\log d)}{t}} \\
&\leq \sum_{t=\lceil T_0\rceil}^{T} \frac{12\nu\sigma x_{\max}}{\kappa_0\phi_1^2}\sqrt{\frac{s_0(4\log t+2\log d)}{t}} \\
&\leq \frac{12\nu\sigma x_{\max}\sqrt{s_0(4\log T+2\log d)}}{\kappa_0\phi_1^2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \\
&\leq \frac{24\nu\sigma x_{\max}\sqrt{s_0(4\log T+2\log d)}}{\kappa_0\phi_1^2}\sqrt{T}
\end{aligned}$$

where the last inequality is from the fact that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq \int_{t=0}^{T}\frac{1}{\sqrt{t}} = 2\sqrt{T}$. Combining all the results with the bounds on T_0 and $\sum_{t=\lceil T_0\rceil}^{T}\mathbb{P}(\mathcal{E}_t^c)$ from the proof of Theorem 4.1, the expected regret under the RE condition is bounded by

$$\mathcal{R}^{\pi}(T) \leq 4\kappa_1 + \frac{4\kappa_1 x_{\max} b(\log(2d^2)+1)}{C_2(\phi_1, s_0)^2} + \frac{48\kappa_1\nu\sigma x_{\max}\sqrt{s_0 T\log(dT)}}{\kappa_0\phi_1^2}$$

where $C_2(\phi_1, s_0) = \min\left(\frac{1}{2}, \frac{\phi_1^2}{256 s_0 \nu x_{\max}^2}\right)$. □

C.4 Regret Analysis for K-Armed Case

C.4.1 Proof Outline of Theorem 4.3

As discussed in Section 4.6, the analysis for the K-armed bandit mostly follows the proof of the two-armed bandit analysis in Section 4.4. Assuming the compatibility condition of the empirical Gram matrix $\hat{\Sigma}_t$, the Lasso oracle inequality for adapted samples in Lemma 4.1 can be directly applied. Hence, what we have left is ensuring the compatibility condition of $\hat{\Sigma}_t$. As before, for each $\mathbb{E}[X_\tau X_\tau^\top | \mathcal{F}_\tau]$ in Σ_t, the history $\mathcal{F}_\tau$ affects how feature vector X_τ is chosen. Similar to the two-armed bandit case, we rewrite Σ_t as

$$\Sigma_t = \frac{1}{t} \sum_{\tau=1}^{t} \sum_{i=1}^{K} \mathbb{E}_{\mathcal{X}_t} \Big[X_{\tau,i} X_{\tau,i}^\top \mathbb{1}\{X_{\tau,i} = \operatorname*{argmax}_{X \in \mathcal{X}_\tau} X^\top \hat{\beta}_\tau\} \mid \hat{\beta}_\tau \Big] .$$

Recall that the compatibility condition is only assumed for the theoretical Gram matrix Σ (Assumption 4.3). Again, the adapted Gram matrix Σ_t is used to bridge Σ and $\hat{\Sigma}_t$ to ensure the compatibility of $\hat{\Sigma}_t$. The key difference between the two-armed bandit analysis and the K-armed bandit analysis lies in how Σ_t is controlled by Σ. In particular, under the balanced covariance condition in Assumption 4.6, we show the following lemma which is a generalization of Lemma 4.2.

Lemma C.7. *Suppose Assumption 4.6 holds. For a fixed vector $\beta \in \mathbb{R}^d$, we have*

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{X}_t} \Big[X_{t,i} X_{t,i}^\top \mathbb{1}\{X_i = \operatorname*{argmax}_{X \in \mathcal{X}_t} X^\top \beta\} \Big] \succcurlyeq (2\nu C_{\mathcal{X}})^{-1} \Sigma .$$

With this result, we can lower-bound the compatibility constant $\phi^2(\Sigma_t, S_0)$ of the adapted Gram matrix in terms of the compatibility constant $\phi^2(\Sigma, S_0)$ for the theoretical

Gram matrix. That is, we have $\Sigma_t \succcurlyeq (2\nu C_{\mathcal{X}})^{-1}\Sigma$ which implies that

$$\phi^2(\Sigma_t, S_0) \geq \frac{\phi^2(\Sigma, S_0)}{2\nu C_{\mathcal{X}}} > 0\,.$$

Hence, Σ_t satisfies the compatibility condition. Then, we can show that $\hat{\Sigma}_t$ concentrates to Σ_t with high probability which directly follows from applying Lemma 4.2, which is formally stated as follows.

Corollary C.3. *For* $t \geq \frac{2\log(2d^2)}{C_1(s_0)^2}$ *where* $C_1(s_0) = \min\left(\frac{1}{2}, \frac{\phi_0^2}{256 s_0 \nu C_{\mathcal{X}} x_{\max}^2}\right)$, *we have*

$$\mathbb{P}\left(\|\Sigma_t - \hat{\Sigma}_t\|_\infty \geq \frac{\phi_0^2}{32 s_0 \nu C_{\mathcal{X}}}\right) \leq \exp\left\{-\frac{t C_1(s_0)^2}{2}\right\}.$$

Now, we can invoke Corollary 4.1 to connect this matrix concentration result to guaranteeing the compatibility condition of $\hat{\Sigma}_t$. Therefore, $\hat{\Sigma}_t$ satisfies the compatibility condition with compatibility constant $\phi_t^2 = \frac{\phi_0^2}{4\nu C_{\mathcal{X}}} > 0$. The rest of the proof of Theorem 4.3 directly follows the proof of Theorem 4.1 using this compatibility constant.

C.4.2 Proof of Lemma C.7

Proof. Since the distribution of $\mathcal{X}_t = \{X_{t,1}, ..., X_{t,K}\}$ is time-invariant, we suppress the subscript on t and write $\mathcal{X} = \{X_1, ..., X_K\}$. Let joint distribution of $\mathcal{X}$ as $p_{\mathcal{X}}(x_1, ..., x_K) = p_{\mathcal{X}}(\mathbf{x})$ where we let $\mathbf{x} = (x_1, ..., x_K)$. All expectations in this proof is taken with respect to the tuple $\mathcal{X}$. Then the theoretical Gram matrix is defined as

$$\begin{aligned}\mathbb{E}[\mathbf{X}^\top \mathbf{X}] &= \mathbb{E}\left[\sum_{i=1}^K X_i X_i^\top\right] \\ &= \int (x_1 x_1^\top + ... + x_K x_K^\top) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x}\end{aligned}$$

Let's first focus on $\int x_1 x_1^\top p_{\mathcal{X}}(\mathbf{x})d\mathbf{x}$.

$$\begin{aligned}\int x_1 x_1^\top p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} = &\int x_1 x_1^\top \mathbb{1}\Big\{x_1 = \operatorname*{argmax}_{x_i\in\mathcal{X}} x_i^\top \beta\Big\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} \\ &+ \int x_1 x_1^\top \mathbb{1}\Big\{x_1 = \operatorname*{argmin}_{x_i\in\mathcal{X}} x_i^\top \beta\Big\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} \\ &+ \int x_1 x_1^\top \mathbb{1}\Big\{x_1 \neq \operatorname*{argmax}_{x_i\in\mathcal{X}} x_i^\top \beta, x_1 \neq \operatorname*{argmin}_{x_i\in\mathcal{X}} x_i^\top \beta\Big\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x}\,.\end{aligned}$$

We define three disjoint sets of possible orderings for $\{1, ..., K\}$ as follows.

Definition C.3. *We define the following sets of permutations of* $(1, ..., K)$*.*

$$\begin{aligned}\mathcal{I}_1^{\max} &:= \{\textit{indices } (i_1, ..., i_K) \textit{ such that } i_K = 1\} \\ \mathcal{I}_1^{\min} &:= \{\textit{indices } (i_1, ..., i_K) \textit{ such that } i_1 = 1\} \\ \mathcal{I}_1^{mid} &:= \{\textit{indices } (i_1, ..., i_K) \textit{ such that } i_1 \neq 1 \textit{ and } i_K \neq 1\}.\end{aligned}$$

Then, for $\int x_1 x_1^\top \mathbb{1}\{x_1 = \operatorname{argmin}_{x_i\in\mathcal{X}} x_i^\top \beta\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x}$, we can write

$$\int x_1 x_1^\top \mathbb{1}\Big\{x_1 = \operatorname*{argmin}_{x_i\in\mathcal{X}} x_i^\top \beta\Big\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} = \sum_{(i_1,...,i_K)\in\mathcal{I}_1^{\min}} \int x_1 x_1^\top \mathbb{1}\Big\{x_{i_1}^\top \beta \leq ... \leq x_{i_K}^\top \beta\Big\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x}$$

Then for any $(i_1, ..., i_K) \in \mathcal{I}_1^{\min}$,

$$\begin{aligned}\int x_1 x_1^\top \mathbb{1}\Big\{x_{i_1}^\top \beta \leq ... \leq x_{i_K}^\top \beta\Big\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} &= \int x_1 x_1^\top \mathbb{1}\Big\{ -x_{i_1}^\top \beta \geq ... \geq -x_{i_K}^\top \beta\Big\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} \\ &\preccurlyeq \nu \int x_1 x_1^\top \mathbb{1}\Big\{ -x_{i_1}^\top \beta \geq ... \geq -x_{i_K}^\top \beta\Big\}p_{\mathcal{X}}(-\mathbf{x})d\mathbf{x} \\ &= \nu \int x_1 x_1^\top \mathbb{1}\Big\{x_{i_1}^\top \beta \geq ... \geq x_{i_K}^\top \beta\Big\}p_{\mathcal{X}}(\mathbf{x})d\mathbf{x}\end{aligned}$$

where the inequality is again from Assumption 4.4. Since the elements in $\mathcal{I}_1^{\min}$ can be

considered as reversed orderings of elements in $\mathcal{I}_1^{\max}$ (and obviously $|\mathcal{I}_1^{\min}| = |\mathcal{I}_1^{\max}|$),

$$\begin{aligned}
\mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_1 = \operatorname*{argmin}_{X\in\mathcal{X}} X^\top\beta\}\right] &= \int x_1 x_1^\top \mathbb{1}\Big\{x_1 = \operatorname*{argmin}_{x_i\in\mathcal{X}} x_i^\top\beta\Big\} p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} \\
&= \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\min}} \int x_1 x_1^\top \mathbb{1}\Big\{x_{i_1}^\top\beta \le \ldots \le x_{i_K}^\top\beta\Big\} p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} \\
&\preccurlyeq \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\min}} \nu \int x_1 x_1^\top \mathbb{1}\Big\{x_{i_1}^\top\beta \ge \ldots \ge x_{i_K}^\top\beta\Big\} p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} \\
&= \nu \int x_1 x_1^\top \mathbb{1}\Big\{x_1 = \operatorname*{argmax}_{x_i\in\mathcal{X}} x_i^\top\beta\Big\} p_{\mathcal{X}}(\mathbf{x})d\mathbf{x} \\
&= \nu\mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_1 = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top\beta\}\right].
\end{aligned}$$

Also, using the definitions of $\mathcal{I}_1^{\min}, \mathcal{I}_1^{\text{mid}}$ and $\mathcal{I}_1^{\max}$, we can rewrite $\mathbb{E}\left[X_1 X_1^\top\right]$.

$$\begin{aligned}
\mathbb{E}\left[X_1 X_1^\top\right] &= \mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_1 = \operatorname*{argmin}_{X\in\mathcal{X}} X^\top\beta\}\right] + \mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_1 = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top\beta\}\right] \\
&\quad + \mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_1 \ne \operatorname*{argmin}_{X\in\mathcal{X}} X^\top\beta, X_1 \ne \operatorname*{argmax}_{X\in\mathcal{X}} X^\top\beta\}\right] \\
&= \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\min}} \mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\max}} \mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\text{mid}}} \mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&= \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\min}} \mathbb{E}\left[X_{i_1} X_{i_1}^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\max}} \mathbb{E}\left[X_{i_K} X_{i_K}^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\text{mid}}} \mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right].
\end{aligned}$$

sFrom Assumption 4.6, we have

$$\mathbb{E}\left[X_1X_1^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \preccurlyeq C_{\mathcal{X}}\mathbb{E}\left[(X_{i_1}X_{i_1}^\top + X_{i_K}X_{i_K}^\top)\mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right].$$

Then it follows that

$$\begin{aligned}
\mathbb{E}\left[X_1X_1^\top\right] &\preccurlyeq \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\min}} \mathbb{E}\left[X_{i_1}X_{i_1}^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\max}} \mathbb{E}\left[X_{i_K}X_{i_K}^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\mathrm{mid}}} C_{\mathcal{X}}\mathbb{E}\left[\left(X_{i_1}X_{i_1}^\top + X_{i_K}X_{i_K}^\top\right)\mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\preccurlyeq \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\min}} C_{\mathcal{X}}\mathbb{E}\left[\left(X_{i_1}X_{i_1}^\top + X_{i_K}X_{i_K}^\top\right)\mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\max}} C_{\mathcal{X}}\mathbb{E}\left[\left(X_{i_1}X_{i_1}^\top + X_{i_K}X_{i_K}^\top\right)\mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\mathrm{mid}}} C_{\mathcal{X}}\mathbb{E}\left[\left(X_{i_1}X_{i_1}^\top + X_{i_K}X_{i_K}^\top\right)\mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right].
\end{aligned}$$

Since $\mathcal{I}_1^{\min}, \mathcal{I}_1^{\mathrm{mid}}$ and $\mathcal{I}_1^{\max}$ are disjoint sets, we can write

$$\begin{aligned}
\mathbb{E}\left[X_iX_i^\top \mathbb{1}\{X_i = \operatorname*{argmin}_{X\in\mathcal{X}} X^\top\beta\}\right] &= \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\min}} \mathbb{E}\left[X_{i_1}X_{i_1}^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\max}} \mathbb{E}\left[X_{i_1}X_{i_1}^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right] \\
&\quad + \sum_{(i_1,\ldots,i_K)\in\mathcal{I}_1^{\mathrm{mid}}} \mathbb{E}\left[X_{i_1}X_{i_1}^\top \mathbb{1}\{X_{i_1}^\top\beta < \cdots < X_{i_K}^\top\beta\}\right].
\end{aligned}$$

We can also express $\mathbb{E}\left[X_i X_i^\top \mathbb{1}\{X_i = \operatorname{argmax}_{X\in\mathcal{X}} X^\top\beta\}\right]$ similarly. Therefore, we have

$$\begin{aligned}\mathbb{E}\left[X_1 X_1^\top\right] &\preccurlyeq C_{\mathcal{X}} \sum_{i=1}^{K}\left(\mathbb{E}\left[X_i X_i^\top \mathbb{1}\{X_i = \operatorname*{argmin}_{X\in\mathcal{X}} X^\top\beta\}\right] + \mathbb{E}\left[X_i X_i^\top \mathbb{1}\{X_i = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top\beta\}\right]\right)\\ &\preccurlyeq C_{\mathcal{X}}(1+\nu)\sum_{i=1}^{K}\mathbb{E}\left[X_i X_i^\top \mathbb{1}\{X_i = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top\beta\}\right].\end{aligned}$$

Then, summing $\mathbb{E}\left[X_j X_j^\top\right]$ over all $j = 1, ..., K$ gives

$$\mathbb{E}[\mathbf{X}^\top\mathbf{X}] = \sum_{j=1}^{K}\mathbb{E}\left[X_j X_j^\top\right] \preccurlyeq KC_{\mathcal{X}}(1+\nu)\sum_{i=1}^{K}\mathbb{E}\left[X_i X_i^\top \mathbb{1}\{X_i = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top\beta\}\right].$$

Hence,

$$\sum_{i=1}^{K}\mathbb{E}\left[X_i X_i^\top \mathbb{1}\{X_i = \operatorname*{argmax}_{X\in\mathcal{X}} X^\top\beta\}\right] \succcurlyeq \frac{1}{C_{\mathcal{X}}(1+\nu)}\cdot\frac{1}{K}\mathbb{E}[\mathbf{X}^\top\mathbf{X}] \succcurlyeq (2C_{\mathcal{X}}\nu)^{-1}\Sigma\,.$$

□

C.4.3 Proposition 4

Proposition 4. *In the case of independent arms, both a multivariate Gaussian distribution and a uniform distribution on a unit sphere satisfy Assumption 4.6 with $C_{\mathcal{X}} = \mathcal{O}(1)$. For an arbitrary distribution, it holds with $C_{\mathcal{X}} = \binom{K-1}{K_0}$ where $K_0 = \lceil (K-1)/2 \rceil$.*

The proof of Proposition 4 involves the following few technical lemmas.

Lemma C.8. *Suppose each $X_i \in \mathbb{R}^d$ is i.i.d. Gaussian with mean μ and covariance matrix Γ. For any permutation $(i_1, ..., i_K)$ of $(1, ..., K)$, any integer $k \in \{2, ..., K-1\}$ and*

fixed β,

$$\mathbb{E}\Big[X_{i_k}X_{i_k}^\top \mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\Big] \preccurlyeq \mathbb{E}\Big[X_{i_1}X_{i_1}^\top \mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\Big] + \mathbb{E}\Big[X_{i_K}X_{i_K}^\top \mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\Big].$$

Proof. It suffices to show that for any $y \in \mathbb{R}^d$

$$\begin{aligned}&\mathbb{E}\left[(X_{i_k}^\top y)^2 \mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right]\\ &\leq \mathbb{E}\left[(X_{i_1}^\top y)^2 \mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right] + \mathbb{E}\left[(X_{i_K}^\top y)^2 \mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right].\end{aligned}$$

Now, we can write

$$y = \widetilde{\beta}(\widetilde{\beta}^\top y) + \sum_{j=1}^{d-1} g_j g_j^\top y := \widetilde{\beta} w_0 + \sum_{j=1}^{d-1} g_j g_j^\top y\,.$$

where $w_0 = \widetilde{\beta}^\top y$ and $\widetilde{\beta} = \frac{\beta}{\|\beta\|}$ and $\left[\widetilde{\beta}, g_i, ..., g_{d-1}\right]$ form an orthonormal basis. For $i \in [N]$, we can write

$$\begin{aligned}X_i^\top y &= (X_i^\top \widetilde{\beta}) w_0 + X_i^\top \left(\sum_{j=1}^{d-1} g_j g_j^\top\right) y\\ &= (X_i^\top \widetilde{\beta}) w_0 + \left[\left(\sum_{j=1}^{d-1} g_j g_j^\top\right) X_i\right]^\top y\,.\end{aligned}$$

Then we define the following two random variables

$$U_i := X_i^\top \widetilde{\beta}, \qquad V_i := G X_i$$

where $G = \sum_{j=1}^{d-1} g_j g_j^\top$. Then we have

$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu^\top \widetilde{\beta} \\ G\mu \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right)$$

where

$$\begin{aligned} A_{11} &= \widetilde{\beta}^\top \Gamma \widetilde{\beta} \in \mathbb{R} \\ A_{12} &= A_{21}^\top = \widetilde{\beta}^\top \Gamma G^\top \in \mathbb{R}^{1\times d} \\ A_{22} &= G\Gamma G^\top \in \mathbb{R}^{d\times d}\,. \end{aligned}$$

Then, we know from Lemma C.12 that the conditional distribution $V_i \mid U_i$ of a multivariate normal distribution is also a multivariate normal distribution. In particular,

$$V_i \mid U_i = u_i \sim \mathcal{N}\left(G\mu + A_{21}A_{11}^{-1}(u_i - \mu^\top \widetilde{\beta}), B\right)$$

where $B = A_{22} - A_{21}A_{11}^{-1}A_{12}$. Therefore, given $U_{i_k} = u_{i_k}$, we can write

$$\begin{aligned} X_{i_k}^\top y &= u_{i_k} w_0 + V_{i_k}^\top y \\ &= u_{i_k} w_0 + \left(G\mu + A_{21}A_{11}^{-1}(u_{i_k} - \mu^\top \widetilde{\beta}) + B^{1/2} Z\right)^\top y\,. \end{aligned}$$

where $Z \sim \mathcal{N}(0, I_d)$ and $Z \perp\!\!\!\perp U_{i_k}$. Rearranging gives

$$X_{i_k}^\top y = u_{i_k}\left(w_0 + A_{11}^{-1}A_{12}y\right) + \left(G\mu - A_{21}A_{11}^{-1}\mu^\top \widetilde{\beta}\right)^\top y + Z^\top B^{1/2} y\,.$$

Hence, $X_{i_k}^\top y$ is a linear function of u_{i_k}. Then it follows that

$$\begin{aligned}
\left(X_{i_k}^\top y\right)^2 &= \left[u_{i_k}\left(w_0 + A_{11}^{-1}A_{12}y\right) + \left(G\mu - A_{21}A_{11}^{-1}\mu^\top\widetilde{\beta}\right)^\top y + Z^\top B^{1/2}y\right]^2 \\
&\le \max\left\{\left[u_{i_1}\left(w_0 + A_{11}^{-1}A_{12}y\right) + \left(G\mu - A_{21}A_{11}^{-1}\mu^\top\widetilde{\beta}\right)^\top y + Z^\top B^{1/2}y\right]^2,\right. \\
&\qquad \left.\left[u_{i_K}\left(w_0 + A_{11}^{-1}A_{12}y\right) + \left(G\mu - A_{21}A_{11}^{-1}\mu^\top\widetilde{\beta}\right)^\top y + Z^\top B^{1/2}y\right]^2\right\} \\
&\le \left[u_{i_1}\left(w_0 + A_{11}^{-1}A_{12}y\right) + \left(G\mu - A_{21}A_{11}^{-1}\mu^\top\widetilde{\beta}\right)^\top y + Z^\top B^{1/2}y\right]^2 \\
&\quad + \left[u_{i_K}\left(w_0 + A_{11}^{-1}A_{12}y\right) + \left(G\mu - A_{21}A_{11}^{-1}\mu^\top\widetilde{\beta}\right)^\top y + Z^\top B^{1/2}y\right]^2.
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
&\mathbb{E}\left[(X_{i_k}^\top y)^2\mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right] \\
&\le \mathbb{E}\left[(X_{i_1}^\top y)^2\mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right] + \mathbb{E}\left[(X_{i_K}^\top y)^2\mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right].
\end{aligned}$$

Hence,

$$\mathbb{E}\left[X_{i_k}X_{i_k}^\top\mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right] \preccurlyeq \mathbb{E}\left[(X_{i_1}X_{i_1}^\top + X_{i_K}X_{i_K}^\top)\mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right].$$

□

Lemma C.9. *Suppose $X \in \mathbb{R}^d$ is uniformly distributed on the unit sphere $\mathcal{S}^{d-1}$ and $K = o(d)$. For fixed vector $\beta \in \mathbb{R}^d$ and a given integer $k \in \{2, ..., K-1\}$,*

$$\mathbb{E}\left[X_{i_k}X_{i_k}^\top\mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right] \preccurlyeq C_{\mathcal{X}}\mathbb{E}\left[(X_{i_1}X_{i_1}^\top + X_{i_K}X_{i_K}^\top)\mathbb{1}\{X_{i_1}^\top\beta < ... < X_{i_K}^\top\beta\}\right].$$

where $C_{\mathcal{X}} = \mathcal{O}(1)$.

Proof. Here, we instead show directly

$$\mathbb{E}[XX^\top] \preccurlyeq C\left(\mathbb{E}\Big[XX^\top \mathbb{1}\{X = \underset{X_i \in \{X_1,\dots,X_K\}}{\operatorname{argmax}} X_i^\top \widetilde{\beta}\}\Big] + \mathbb{E}\Big[XX^\top \mathbb{1}\{X = \underset{X_i \in \{X_1,\dots,X_K\}}{\operatorname{argmin}} X_i^\top \widetilde{\beta}\}\Big]\right)$$

for some constant C. It can be shown that if $C = \mathcal{O}(1)$, then the claim holds with $C_{\mathcal{X}} = \mathcal{O}(1)$. Suppose $X \in \mathbb{R}^d$ is uniformly distributed on the unit sphere $\mathcal{S}^{d-1} := \{s \in \mathbb{R}^d : \|s\|_2 = 1\}$. Then by Lemma 2 in Cambanis, Huang, and Simons, 1981, we can write for each X_i,

$$X_i \sim \Big(B_i U_{i,1}, (1 - B_i^2)^{1/2} U_{i,2}\Big)$$

where $B_i \sim \text{beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$, $U_{i,1} = \pm 1$ with probability $\frac{1}{2}$, $U_{i,2} \sim \text{unif}(\mathcal{S}^{d-2})$. $U_{i,1}$, $U_{i,2}$ and B_i are independent of each other. Similar to the analysis of the Gaussian case, we can normalize β so that $\widetilde{\beta} = \frac{\beta}{\|\beta\|}$. Without loss of generality, assume that $\widetilde{\beta} = [1, 0, ..., 0]^\top$. That is, only the first element is non-zero. We can do this since X is spherical and rotation invariant. Then we can write

$$\mathbb{E}\Big[XX^\top \mathbb{1}\{X = \underset{X_i \in \{X_1,\dots,X_K\}}{\operatorname{argmax}} X_i^\top \widetilde{\beta}\}\Big] = \mathbb{E}\Big[XX^\top \mathbb{1}\{X = \underset{X_i \in \{X_1,\dots,X_K\}}{\operatorname{argmax}} X_i^{(1)}\}\Big]$$

where $X_i^{(1)}$ is the first element of X_i. Similarly,

$$\mathbb{E}\Big[XX^\top \mathbb{1}\{X = \underset{X_i \in \{X_1,\dots,X_K\}}{\operatorname{argmin}} X_i^\top \widetilde{\beta}\}\Big] = \mathbb{E}\Big[XX^\top \mathbb{1}\{X = \underset{X_i \in \{X_1,\dots,X_K\}}{\operatorname{argmin}} X_i^{(1)}\}\Big].$$

Now, from the definition of X, for $B \sim \text{beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$ we have

$$X_i X_i^\top = \begin{bmatrix} B_i^2 & B_i\sqrt{1 - B_i^2} U_{i,1} U_{i,2}^\top \\ B_i\sqrt{1 - B_i^2} U_{i,1} U_{i,2} & (1 - B_i^2) U_{i,2} U_{i,2}^\top \end{bmatrix}.$$

By the independence of U_1, U_2, and B, we have

$$\mathbb{E}\left[XX^\top\right] = \mathbb{E}\begin{bmatrix} B^2 & 0 \\ 0 & \frac{1}{d-1}(1-B^2)I_{d-1} \end{bmatrix}.$$

By the definitions of B_i and $U_{i,1}$, it follows that

$$\mathbb{E}\Big[XX^\top \mathbb{1}\{B = \max_{B_i \in \{B_1,\ldots,B_K\}} B_i\}\Big] \preccurlyeq \mathbb{E}\Big[XX^\top \mathbb{1}\{X = \operatorname*{argmax}_{X_i \in \{X_1,\ldots,X_K\}} X_i^{(1)}\}\Big] + \mathbb{E}\Big[XX^\top \mathbb{1}\{X = \operatorname*{argmin}_{X_i \in \{X_1,\ldots,X_K\}} X_i^{(1)}\}\Big].$$

Since $\mathbb{E}[B^2] = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$ for $B \sim \text{beta}(\alpha,\beta)$, we have $\mathbb{E}[B^2] = \frac{3}{d(d+2)}$ and $\frac{1-\mathbb{E}[B^2]}{d-1} = \frac{d+3}{d(d+2)}$ using $\alpha = \frac{1}{2}$ and $\beta = \frac{d-1}{2}$. Clearly, $\lambda_{\min}(\mathbb{E}\left[XX^\top\right]) = \frac{3}{d(d+2)}$. Similarly, for the matrix $\mathbb{E}\left[XX^\top \mathbb{1}\{B = \max_i B_i\}\right]$, we have

$$\mathbb{E}\left[XX^\top \mathbb{1}\{B = \max_i B_i\}\right] = \mathbb{E}\begin{bmatrix} B^2\mathbb{1}\{B = \max_i B_i\} & 0 \\ 0 & \frac{1}{d-1}(1-B^2)\mathbb{1}\{B = \max_i B_i\}I_{d-1} \end{bmatrix}.$$

Note that $\mathbb{E}[B^2\mathbb{1}\{B = \max_i B_i\}] = \sum_{j=1}^K \mathbb{E}[B_j^2\mathbb{1}\{B_j = \max_i B_i\}] \geq \mathbb{E}[B^2]$. Then, we need to show

$$C(1 - \mathbb{E}[B^2\mathbb{1}\{B = \max_i B_i\}]) \geq 1 - \mathbb{E}[B^2]$$

for some C. Note that $\mathbb{E}[B^2\mathbb{1}\{B = \max_i B_i\}] \leq N\mathbb{E}[B^2]$. Hence, we can show

$$C \geq \frac{1-\mathbb{E}[B^2]}{1-N\mathbb{E}[B^2]} = \frac{1-\frac{3}{d(d+2)}}{1-\frac{3K}{d(d+2)}} = \frac{d^2+d-3}{d^2+d-3K}.$$

Since $K = o(d)$, we have $C = \mathcal{O}(1)$. Hence,

$$\begin{aligned}
\mathbb{E}[XX^\top] &\preccurlyeq C\mathbb{E}\Big[XX^\top \mathbb{1}\{B = \max_{B_i \in \{B_1,...,B_K\}} B_i\}\Big] \\
&\preccurlyeq C\left(\mathbb{E}\Big[XX^\top \mathbb{1}\{X = \operatorname*{argmax}_{X_i \in \{X_1,...,X_K\}} X_i^{(1)}\}\Big] + \mathbb{E}\Big[XX^\top \mathbb{1}\{X = \operatorname*{argmin}_{X_i \in \{X_1,...,X_K\}} X_i^{(1)}\}\Big]\right) \\
&= C\left(\mathbb{E}\Big[XX^\top \mathbb{1}\{X = \operatorname*{argmax}_{X_i \in \{X_1,...,X_K\}} X_i^\top \widetilde{\beta}\}\Big] + \mathbb{E}\Big[XX^\top \mathbb{1}\{X = \operatorname*{argmin}_{X_i \in \{X_1,...,X_K\}} X_i^\top \widetilde{\beta}\}\Big]\right)
\end{aligned}$$

which implies $C_{\mathcal{X}} = \mathcal{O}(1)$. □

Lemma C.10. *Consider i.i.d. arbitrary distribution $p_{\mathcal{X}}$. Fix some vector $\beta \in \mathbb{R}^d$. For a given integer $k \in \{2, ..., K-1\}$,*

$$\begin{aligned}
&\mathbb{E}\left[X_k X_k^\top \mathbb{1}\{X_1^\top \beta < ... < X_k^\top \beta < ... < X_K^\top \beta\}\right] \\
&\preccurlyeq C_{K,k}\mathbb{E}\left[(X_1 X_1^\top + X_K X_K^\top)\mathbb{1}\{X_1^\top \beta < ... < X_K^\top \beta\}\right]
\end{aligned}$$

where $C_{\mathcal{X}} = \binom{K-1}{(K-1)/2}$ assuming K is odd — if K is even, we can use $\lceil (K-1)/2 \rceil$.

Proof. First notice that

$$\begin{aligned}
&\mathbb{E}\left[X_k X_k^\top \mathbb{1}\{X_1^\top \beta < \cdots < X_k^\top \beta < \cdots < X_K^\top \beta\}\right] \\
&= \mathbb{E}_V\left[VV^\top \mathbb{E}_{X_{1:K}/X_k}\left[\mathbb{1}\{X_1^\top \beta < \cdots < X_{k-1}^\top \beta < V^\top \beta < X_{k+1}^\top \beta < \cdots < X_K^\top \beta\} \mid V\right]\right]
\end{aligned}$$

where $X_{1:K}/X_k$ denotes $X_1, ..., X_{k-1}, X_{k+1}, ..., X_K$. Also,

$$\begin{aligned}
&\mathbb{E}\left[X_1 X_1^\top \mathbb{1}\{X_1^\top \beta < \cdots < X_K^\top \beta\}\right] \\
&= \mathbb{E}_V\left[VV^\top \mathbb{E}_{X_{2:K}}\left[\mathbb{1}\{V^\top \beta < X_2^\top \beta < \cdots < X_K^\top \beta\} \mid V\right]\right] \\
&\mathbb{E}\left[X_K X_K^\top \mathbb{1}\{X_1^\top \beta < \cdots < X_K^\top \beta\}\right] \\
&= \mathbb{E}_V\left[VV^\top \mathbb{E}_{X_{1:K-1}}\left[\mathbb{1}\{X_1^\top \beta < \cdots < X_{K-1}^\top \beta < V^\top \beta\} \mid V\right]\right]
\end{aligned}$$

Let $\psi(y) := \mathbb{P}(X^\top \beta \le y)$ denote the CDF of $X^\top \beta$. Then

$$\begin{aligned}
&\mathbb{P}\left(X_1^\top \beta < \cdots < X_{k-1}^\top \beta < V^\top \beta < X_{k+1}^\top \beta < \cdots < X_K^\top \beta\right) \\
&= \prod_{i=1}^{k-1} \mathbb{P}\left(X_i^\top \beta \le V^\top \beta\right) \frac{1}{(k-1)!} \prod_{i=k+1}^{N} \mathbb{P}\left(X_i^\top \beta \ge V^\top \beta\right) \frac{1}{(K-k)!} \\
&= \frac{1}{(k-1)!(K-k)!} \psi(V^\top \beta)^{k-1} \left(1 - \psi(V^\top \beta)\right)^{K-k}.
\end{aligned}$$

Likewise

$$\begin{aligned}
\mathbb{P}\left(V^\top \beta < X_2^\top \beta < \cdots < X_K^\top \beta\right) &= \frac{1}{(K-1)!} \left(1 - \psi(V^\top \beta)\right)^{K-1}, \\
\mathbb{P}\left(X_1^\top \beta < \cdots < X_{K-1}^\top \beta < V^\top \beta\right) &= \frac{1}{(K-1)!} \psi(V^\top \beta)^{K-1}.
\end{aligned}$$

Then, we need to show there exists $C_{K,k}$ such that

$$\begin{aligned}
&\mathbb{P}\left(X_1^\top \beta < \cdots < X_{k-1}^\top \beta < V^\top \beta < X_{k+1}^\top \beta < \cdots < X_K^\top \beta\right) \\
&\le C_{K,k} \left[\mathbb{P}\left(V^\top \beta < X_2^\top \beta < \cdots < X_K^\top \beta\right) + \mathbb{P}\left(X_1^\top \beta < \cdots < X_{K-1}^\top \beta < V^\top \beta\right)\right].
\end{aligned}$$

That is,

$$\frac{\psi(V^\top \beta)^{k-1} \left(1 - \psi(V^\top \beta)\right)^{K-k}}{(k-1)!(K-k)!} \le \frac{C_{K,k}}{(K-1)!} \left[\left(1 - \psi(V^\top \beta)\right)^{K-1} + \psi(V^\top \beta)^{K-1}\right].$$

Hence,

$$C_{K,k} \ge \binom{K-1}{k-1} \frac{\psi(V^\top \beta)^{k-1} \left(1 - \psi(V^\top \beta)\right)^{K-k}}{\left(1 - \psi(V^\top \beta)\right)^{K-1} + \psi(V^\top \beta)^{K-1}}.$$

Since $\psi(V^\top\beta) \in [0,1]$, we have

$$\frac{\psi(V^\top\beta)^{k-1}\left(1-\psi(V^\top\beta)\right)^{K-k}}{\left(1-\psi(V^\top\beta)\right)^{K-1}+\psi(V^\top\beta)^{K-1}} \leq 1$$

for all K and k. Hence, for $C_{K,k} = \binom{K-1}{k-1}$,

$$\begin{aligned}
&\mathbb{E}\left[X_k X_k^\top \mathbb{1}\{X_1^\top\beta < \cdots < X_k^\top\beta < \cdots < X_K^\top\beta\}\right] \\
&\preccurlyeq C_{K,k}\mathbb{E}\left[(X_1X_1^\top + X_K X_K^\top)\mathbb{1}\{X_1^\top\beta < \cdots < X_K^\top\beta\}\right].
\end{aligned}$$

□

C.5 Other lemmas

Lemma C.11 (Wainwright (2019), Theorem 2.19)**.** *Let $\{Z_\tau, \mathcal{F}_\tau\}_\tau^\infty$ be a martingale difference sequence, and suppose that Z_τ is σ^2-sub-Gaussian in an adapted sense, i.e., for all $\alpha \in \mathbb{R}$, $\mathbb{E}[e^{\alpha Z_\tau}|\mathcal{F}_{\tau-1}] \leq e^{\alpha^2\sigma^2/2}$ almost surely. Then for all $\gamma \geq 0$, $\mathbb{P}\left[|\sum_{\tau=1}^n Z_\tau| \geq \gamma\right] \leq 2\exp[-\gamma^2/(2n\sigma^2)]$.*

Note that Lemma C.12 is a well-known result, but for the sake of completeness, we present its formal statment and proof.

Lemma C.12. *Let $X \in \mathbb{R}^d$ follow a multivariate Gaussian distribution with mean μ and covarance matrix Σ and consider the partition of X with*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right).$$

Then the conditional distribution of X_1 given X_2 is also a multivariate Gaussian distri-

bution. In particular

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

Proof. Define $Z = X_1 + \mathbf{A}X_2$ where $\mathbf{A} = -\Sigma_{12}\Sigma_{22}^{-1}$. Now we can write

$$\begin{aligned}
\text{cov}(Z, X_2) &= \text{cov}(X_1, X_2) + \text{cov}(\mathbf{A}X_2, X_2) \\
&= \Sigma_{12} + \mathbf{A}\text{var}(X_2) \\
&= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} \\
&= 0
\end{aligned}$$

Therefore Z and X_2 are not correlated and, since they are jointly normal, they are independent[1]. Now, clearly we have $\mathbb{E}(Z) = \mu_1 + \mathbf{A}\mu_2$. Then

$$\begin{aligned}
\mathbb{E}[X_1|X_2] &= \mathbb{E}[Z - \mathbf{A}X_2|X_2] \\
&= \mathbb{E}[Z|X_2] - \mathbb{E}[\mathbf{A}X_2|X_2] \\
&= \mathbb{E}[Z] - \mathbf{A}X_2 \\
&= \mu_1 + \mathbf{A}(\mu_2 - X_2) \\
&= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2).
\end{aligned}$$

[1]If a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent.

For the covariance matrix, note that

$$\begin{aligned}
\mathrm{var}(X_1|X_2) &= \mathrm{var}(Z - \mathbf{A}X_2|X_2) \\
&= \mathrm{var}(Z|X_2) + \mathrm{var}(\mathbf{A}X_2|X_2) - \mathbf{A}\mathrm{cov}(Z, -X_2) - \mathrm{cov}(Z, -X_2)\mathbf{A}^\top \\
&= \mathrm{var}(Z|X_2) \\
&= \mathrm{var}(Z)
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
\mathrm{var}(X_1|X_2) &= \mathrm{var}(Z) \\
&= \mathrm{var}(X_1 + \mathbf{A}X_2) \\
&= \mathrm{var}(X_1) + \mathbf{A}\mathrm{var}(X_2)\mathbf{A}^\top + \mathbf{A}\mathrm{cov}(X_1, X_2) + \mathrm{cov}(X_2, X_1)\mathbf{A}^\top \\
&= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}
\end{aligned}$$

□

C.6 Additional Experiment Results

C.6.1 Details on Experimental Setup

For feature vectors drawn from the uniform distribution, we sample each feature vector X independently from a d-dimensional hypercube $[-1, 1]^d$. For elliptically distributed feature vectors, we construct each feature vector $X \in \mathbb{R}^d$ following the definition in Theorem 1

of Cambanis, Huang, and Simons (1981):

$$X = \mu + RAU^{(k)}$$

where $\mu \in \mathbb{R}^d$ is a mean vector, $U^{(k)} \in \mathbb{R}^k$ is uniformly distributed on the unit sphere in $\mathbb{R}^k$, $R \in \mathbb{R}$ is a random variable independent of $U^{(k)}$, and A is a $d \times k$-dimensional matrix with rank k. We sample R from Gaussian distribution $\mathcal{N}(0, 1)$, and sample each element of A uniformly in $[0, 1]$. We use zero mean $\mu = \mathbf{0}_d$.

C.6.2 Additional Results for Two-Armed Bandits

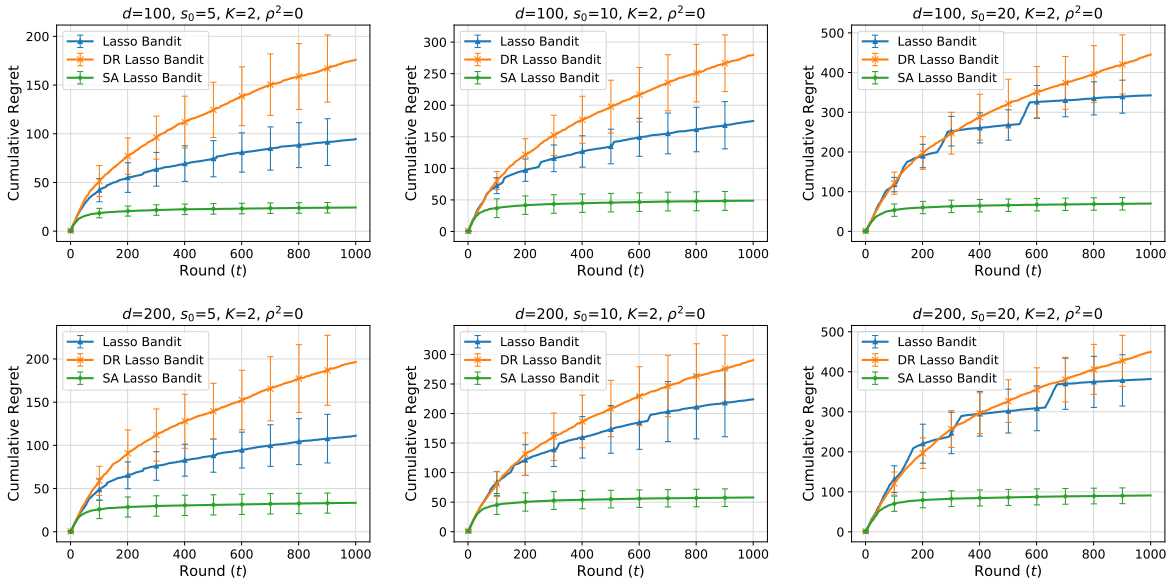

Figure C.1: The plots show the t-round cumulative regret of SA Lasso Bandit (Algorithm 8), DR Lasso Bandit (Kim and Paik, 2019), and Lasso Bandit (Bastani and Bayati, 2020) for $K = 2$, $d \in \{100, 200\}$ and varying sparsity $s_0 \in \{5, 10, 20\}$ under no correlation between arms, $\rho^2 = 0$.

Figure C.1 shows the evaluations in two-armed bandits with independent arms whose features are drawn from a multivariate Gaussian distribution. Comparing the numerical results in Figure C.1 with those in Figure 4.1 and Figure 4.2, we observe that the per-

formance of DR Lasso Bandit substantially deteriorates as correlation between arms decreases whereas the performances of SA Lasso Bandit and Lasso Bandit decrease more gracefully with a decrease in arm correlation. Throughout these experiments, our proposed algorithm, SA Lasso Bandit, consistently exhibits the fastest convergence to the optimal action and robust performances under various instances.

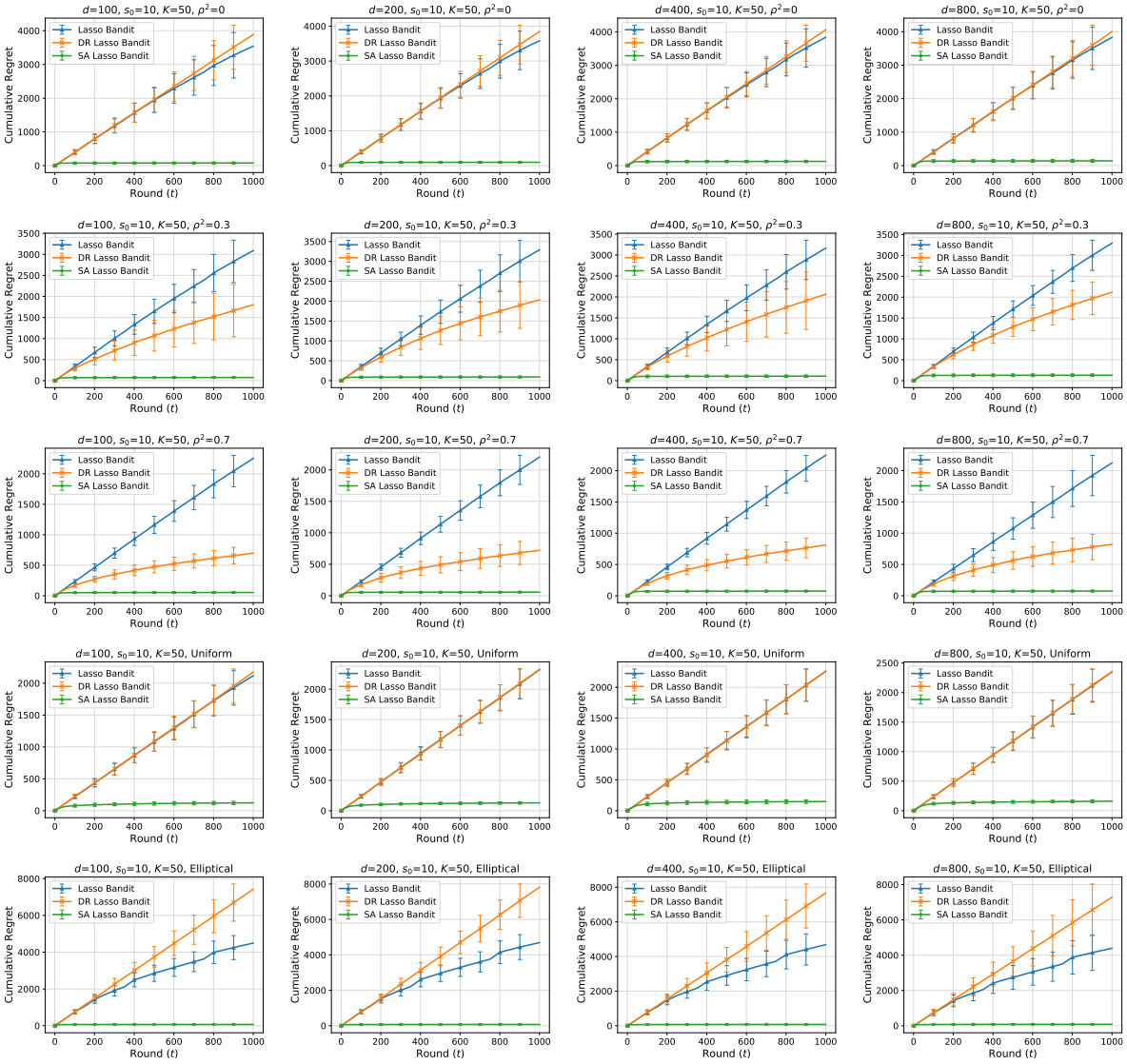

Figure C.2: The plots show the t-round regret of SA Lasso Bandit (Algorithm 8), DR Lasso Bandit (Kim and Paik, 2019), and Lasso Bandit (Bastani and Bayati, 2020) for $K = 50$ and $s_0 = 10$. The first three rows are the results with features drawn from multivariate Gaussian distributions with varying levels of correlation between arms $\rho^2 \in \{0, 0.3, 0.7\}$. In the fourth row, features are drawn from a multi-dimensional uniform distribution. In the fourth row, features are drawn from a non-Gaussian elliptical distribution. For each row, we present evaluations for varying feature dimensions, $d \in \{100, 200, 400, 800\}$.