Almost Optimal Batch-Regret Tradeoff for Batch Linear Contextual Bandits

Zihan Zhang[*]   **Xiangyang Ji**[†]   **Yuan Zhou**[‡]

Abstract

We study the optimal batch-regret tradeoff for batch linear contextual bandits. For this problem, we design batch learning algorithms and prove that they achieve the optimal regret bounds (up to logarithmic factors) for any batch number M, number of actions K, time horizon T, and dimension d. Therefore, we establish the *full-parameter-range* almost optimal batch-regret tradeoff for the batch linear contextual bandit problem.

Along our analysis, we also prove a new matrix concentration inequality with dependence on their dynamic upper bounds, which, to the best of our knowledge, is the first of its kind in literature and maybe of independent interest.

1 Introduction

Online learning and decision-making is an important aspect of machine learning. In contrast to the traditional batch machine learning where the learner only passively observes the data, an online learner may interact with the data collection process by deciding on which data point to query about. On one hand, sequentially making active queries may fully utilize the power of adaptivity based on the observed data and help to achieve better data efficiency. On the other hand, in many practical scenarios, it is also desirable to limit these queries to a small number of rounds of interaction, which helps to increase the parallelism of the learning process, and reduce the management cost and the total time span. In light of this, the *batch online learning* model, which is a combination of the two major aspects of machine learning, has recently attracted much research attention. It has been shown that for many popular online learning tasks, a very small number of batches may achieve nearly minimax-optimal learning performance, and therefore it is possible to enjoy the benefits of both adaptivity and parallelism.

To understand the impact of the batch constraint to online learning and decision-making problems, in this paper, we study the optimal batch algorithms for the *linear contextual bandit* problem. In a linear contextual problem, the learning algorithm observes a *context* (also referred to as a *context set* as we usually have one context vector for each candidate action) at the beginning of each time period, and the expected reward of each candidate action is determined by a hidden linear function of the context. The learning algorithm has to learn the linear function and maximize its total reward over all time periods. The readers may refer to Section 2 for the detailed problem definition. The linear contextual bandit problem is widely studied due to its simplicity and abstraction (via the context) for the personalized treatment in decision-making, enabling plenty of real-world applications such as advertisement selection, recommendation systems, and clinical trials.

One practical reason that calls for the batch online learning model is due to the *expensive policy deployment and communication cost*. Large-scale online advertisement or recommendation systems (Li et al., 2010) may have to take a long time or pay an expensive overhead cost to update their policies. In many distributed or offline systems (such as designing and updating the policy for autonomous vehicles or robotic arms), such an overhead cost becomes even more significant. During the execution of the policy, we usually may not be able to monitor the process (e.g., the data collected from the environment) in real time, due to the high communication cost and latency in the distributed (or large-scale) system. Similar challenges arise in medicine and clinical trials. The price of implementing a

---

[*]Paul G. Allen School of CSE, University of Washington. Email: `zihanz46@uw.edu`

[†]Department of Automation, Tsinghua University. Email: `xyji@tsinghua.edu.cn`

[‡]Yau Mathematical Sciences Center, Tsinghua University and Beijing Institute of Mathematical Sciences and Applications. Email: `yuan-zhou@tsinghua.edu.cn`

new policy can be costly due to the inherent risks in healthcare, which limits the adaptivity of the deployment method. In such cases, the learner seeks to minimize the number of policy updates under a steady environment, making the batch learning framework with stochastic contexts particularly well-suited.

In the batch online learning model, we refer to the time of each policy update as the beginning of a *batch*. The learning algorithm may decide the policy only based on the data collected from the previous batches, and the data collected from the current batch may only be available at the end of the batch. Due to the non-real-time natural, the learning algorithm also has to decide the length of each batch when the batch begins, without any information from the current batch.

We formalize the above intuition and define the batch complexity of the linear contextual bandit problem as follows.

Definition 1 (Batch complexity). *For a linear contextual bandit problem with time horizon T, we say that the* batch complexity *of a learning algorithm is (at most) M in the, if the learner decides a length T_1 and a policy π_1 (a policy is a mapping from the context space to the set of distributions over all candidate actions) before the learning process starts, and executes π_1 in the first T_1 time steps (which corresponds the first batch). Based on the data (context sets, played actions and the rewards) obtained from the first T_1 steps, the learner then decides T_2 and π_2, and executes π_2 for T_2 time steps (the second batch). The learner repeats the process for M times/batches. In general, at the beginning of the k-th batch, the learner decides T_k (the size of the batch) and π_k based on the data collected from the first $(k-1)$ batches. The batch sizes should satisfy that $\sum_{k=1}^{M} T_k = T$.*

Naturally, there is a tradeoff between the batch complexity and the regret performance in batch online learning. There have been quite a few recent works studying such relationship for multi-armed bandits (Perchet et al., 2016; Gao et al., 2019; Esfandiari et al., 2021). Gao et al. (2019) proved optimal regret bound for every number of batches and show that $M = O(\log \log T)$ batches suffice to achieve the minimax-optimal regret without the batch constraint. Han et al. (2020); Ruan et al. (2021) studied the batch algorithms for linear contextual bandits. While Han et al. (2020) studied a special case of the problem where the contexts follow Gaussian-type distributions, Ruan et al. (2021) provided an algorithm for all context distributions. In particual, Ruan et al. (2021) showed that $M = O(\log \log T)$ batches suffice to achieve the minimax-optimal regret without the batch constraint.[1]

While the exact batch-regret tradeoff for multi-armed bandits is relatively better understood (and arguably easier to study), the optimal tradeoff curve for linear contextual bandits is more challenging and remains open.

In this work, we address this question on the exact batch-regret curve and prove the optimal regret (up to logarithmic factors) for batch linear contextual bandits for the *full range* of the problem parameters Below, we summary our contributions and comparison with the related works.

1.1 Our Contributions

We summarize our contributions and technique ingredients as below. For a more detailed overview of the high-level ideas, we refer the readers to Appendix 3.

Recall that Gao et al. (2019) showed that for multi-armed bandits, the optimal regret using at most M batches is at the order of $T^{\frac{1}{2-2^{-M+1}}}$ (ignoring the polynomial dependence on other problem parameters and the poly-logarithmic dependence on T). For batch linear contextual bandits, we establish a similar (but slightly trickier) tradeoff. More specifically, we design Algorithm 1 (in Section 4), and prove the following regret upper bound.

Theorem 2. *Let d be the dimension of the feature space and K be the number of candidate arms. For any $T \geq d$ and $M \geq 1$, Algorithm 1 may use at most M batches and its regret R_T is bounded by*

$$\tilde{\Theta}\bigg(\min \big\{ T^{\frac{1}{2-2^{-M+2}}} d^{\frac{1-2^{-M+2}}{2-2^{-M+2}}}, T^{\frac{1}{2-2^{-M+1}}} d^{\frac{1-2^{-M+1}}{2-2^{-M+1}}} \min\{K, d\}^{\frac{2^{-M+1}}{2-2^{-M+1}}} \big\} \bigg).$$

---

[1] Both the $O(\log \log T)$ bound in (Ruan et al., 2021) and our work (as well as (Han et al., 2020)) focus on the *stochastic-context case* of linear contextual bandits, which is the most technically interesting and practically useful setting of batch linear contextual bandits. Please refer to Section 1.2.1 for more discussion.

Note that the regret upper bound in Theorem 2 takes the minimum between two terms. When T is comparably large (e.g., $T \geq \tilde{\Omega}(d\min\{K,d\}^{2-2^{-M+2}})$), the second term in our upper bound is smaller, and its asymptotic dependence on T matches the optimal bound for batch multi-armed bandits (Gao et al., 2019).[2] On the other hand, when T is relatively small (e.g., $T \leq \tilde{O}(d\min\{K,d\}^{2-2^{-M+2}})$), a simpler analysis would kick in to give a better regret bound, which results as the first term. We refer the readers to Appendix B for full proof.

When there are no constraints on the number of batches, it is well known that the minimax-optimal regret is $\sqrt{dT\log K} \times \text{poly}\log T$ (see, e.g., (Dani et al., 2008; Chu et al., 2011; Li et al., 2019)). Ruan et al. (2021) showed that with only $M = \lceil \log\log T \rceil + 1$ batches (the logarithms are of base 2), their batch algorithm may match the regret performance (up to logarithmic factors) as the no-batch-constraint setting. The following simple corollary of our Theorem 2 recovers the main result of (Ruan et al., 2021). Moreover, the batch algorithm in (Ruan et al., 2021) only works for $T \geq \tilde{\Omega}(d^{32})$, while in contrast, our Corollary 3 works for every $T \geq d$ (and note that $T < d$ is the trivial scenario).

Corollary 3. *For $T \geq d$ and $M = \lceil \log\log(T) \rceil + 1$, the regret of Algorithm 1 is $R_T \leq \tilde{O}(\sqrt{Td})$.*

While the two-phase regret curve in Theorem 2 may seem completely due to technicality, it surprisingly turns out to be exactly optimal. In Appendix C, we complement Theorem 2 with the following lower bound.

Theorem 4. *Fix any $K \geq 2$, $T \geq d$, and any batch number $M \geq 1$. For any learning algorithm with batch complexity M, there exists a linear contextual bandit problem instance with dimension d and K arms, such that the expected regret R_T is at least*

$$R_T \geq \tilde{\Omega}\bigg(\min\big\{T^{\frac{1}{2-2^{-M+2}}}d^{\frac{1-2^{-M+2}}{2-2^{-M+2}}}, T^{\frac{1}{2-2^{-M+1}}}d^{\frac{1-2^{-M+1}}{2-2^{-M+1}}}\min\{K,d\}^{\frac{2^{-M+1}}{2-2^{-M+1}}}\big\}\bigg).$$

We note that the above upper and lower bounds match (up to factors logarithmic in T, d, and K) for all $M \leq \lceil \log\log T \rceil$ and all non-trivial parameter settings for T, d, and K. When $M > \lceil \log\log T \rceil$, by Corollary 3, our Algorithm 1 already achieves the unconstrained minimax-optimal regret (up to logarithmic factors). Therefore, we achieve near-optimal regret bounds for batch linear contextual bandits under all non-trivial parameter settings.

We also note that the $\Omega(\sqrt{K}T^{\frac{1}{2-2^{-M+1}}})$ regret lower bound established by Gao et al. (2019) for K-arm M-batch multi-armed bandits, and the $\Omega\left(T^{\frac{1}{2-2^{-(M-1)}}}d^{\frac{1-2^{-(M-1)}}{2-2^{-(M-1)}}}\right)$ regret bound by Han et al. (2020) for d-dimensional M batch linear bandits with Gaussian contexts. If we treat multi-armed bandits as a special case of linear contextual bandits with $K = d$ arms with orthogonal features, the lower bounds in Gao et al. (2019) and Han et al. (2020) are weaker than our Theorem 4, especially in the case $d \leq T \leq d\min\{K,d\}^{2-2^{-M+2}}$. This gap demonstrates that intrinsic additional difficulty of the batch linear contextual bandit problem with general context distributions.

In the analysis, we require a stronger matrix concentration inequality to address the challenges posed by a dynamic upper bound. As an independent contribution, we introduce a new matrix concentration inequality (refer to Lemma 20 in Appendix D) to avoid the dynamic upper bound.

1.2 ADDITIONAL RELATED WORKS

1.2.1 LINEAR CONTEXTUAL BANDITS

The linear contextual bandit problem (Abe & Long, 1999; Auer et al., 2002) studies the bandit problem where the actions are associated with (known) features and their mean rewards are defined by an (unknown) linear function of the associated features. Compared with the multi-armed bandit problem, the linear structure on features could help the learner to infer the mean reward of an action given the observation on the other actions, and therefore enables the possibility to achieve regret upper bounds independent from (or weakly dependent on) the number of actions.

---

[2]However, this does not mean that the batch linear contextual bandit problem is easier than batch multi-armed bandits, as the dependence on d and h is worse.

There are generally two types of problem settings studied about linear contextual bandits: *non-adaptive contexts* and *adaptive contexts*. In the non-adaptive-context setting, the context sets are independent from all other randomnesses (including the randomnesses in rewards and used by the algorithm). One can also think of this as that the contexts are fixed (by an adversary) before the learning process starts. In this setting, the optimal minimax regret bound is $\Theta(\sqrt{dT\min\{d, \ln(K)\}})$ up to $\mathrm{poly}\ln(T)$ factors (Auer et al., 2002; Abe et al., 2003; Dani et al., 2008; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Li et al., 2019). In the adaptive-context setting, the context sets are chosen by an adaptive adversarial, where the context sets at any time step may depend on the outcomes and the learner's decisions in previous time steps. In this setting, the problem becomes harder for the learner. To the best of our knowledge, the state-of-the-art regret upper bound for the adaptive-context setting is $O(d\sqrt{T\ln(KT)})$ (Abbasi-Yadkori et al., 2011).

In this work (as well as the most related works (Han et al., 2020; Ruan et al., 2021; Zanette et al., 2021) on batch linear contextual bandits), we focus on a particularly useful case in the non-adaptive-context setting, namely the *stochastic contexts*. In this case, the context sets at each time step are independently generated from a pre-defined (but unknown) distribution D. In many real-world applications such as clinical trial and recommendation system, the patients or customers can often be viewed as independent samples from the population and therefore stochastic contexts are a natural abstraction of these practical scenarios. On the other hand, Han et al. (2020) has shown that even in the non-adaptive-context setting, in the worst case, as many as $\Omega(\sqrt{T})$ batches are needed to achieve any $\sqrt{T}$-type regret, which is less useful in practice.

1.2.2 BANDIT LEARNING WITH LIMITED ADAPTIVITY

Batch learning fits into the broader *learning with limited adaptivity* framework that recently attracts much research attention due to its potentially lower computational cost and close relation to distributed and parallel learning.

The number of batches is a natural measurement of the adaptivity needed by the learner. Besides the above mentioned works (Perchet et al., 2016; Gao et al., 2019) (for batch multi-armed bandits) and Ruan et al. (2021) (for batch linear contextual bandits), Han et al. (2020) studied batch linear contextual bandits with Gaussian-type features, Esfandiari et al. (2019) studied batch adversarial multi-armed bandits and Jin et al. (2021a;b; 2023) focused on the asymptotic regret bounds with batch limits.

Hanna et al. (2023b) designed batch linear bandit algorithms for stochastic contexts via an elegant reduction to the fixed-action-set setting. Hanna et al. (2023a) further proposed an oracle-efficient algorithm for the batched context linear bandit problem with infinite arms at the price of suffering additional $\mathrm{poly}(d)$ factors in the worst case regret.

For the non-contextual linear bandit problem, recent work by Ren et al. (2024) introduced the E^4 algorithm, which achieves a minimax optimal regret bound with $O(\log\log(T))$ batches, and an asymptotic optimal regret bound using just three batches.

2 PRELIMINARIES

Linear Contextual Bandits with Stochastic Context Sets. We consider the linear contextual bandit problem with the hidden linear model described by the d-dimensional vector $\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_\infty \leq 1$. There is also a distribution D over the context sets hidden from the learner. Given the time horizon T, during each time step $t \in \{1, 2, \ldots, T\}$, a stochastic context set of K feature vectors, $X_t = \{\boldsymbol{x}_{t,1}, \boldsymbol{x}_{t,2}, \ldots, \boldsymbol{x}_{t,K}\}$ is drawn from D and revealed to the learner. The feature vectors are in $\mathbb{R}^d$ and D guarantees that $\forall i \in \{1, 2, \ldots, K\} : |\boldsymbol{x}_{t,i}^\top \theta| \leq 1$ almost surely.[3]

The learner has to choose and play an action (defined by its associated feature vector) $\boldsymbol{y}_t \in X_t$ and receives the reward $r_t = \boldsymbol{y}_t^\top \boldsymbol{\theta} + \epsilon_t$, where ϵ_t is an independent sub-Gaussian noise with zero mean and variance proxy bounded by 1. The goal of the learner is to minimize the total (expected) regret $R_T := \mathbb{E}\left[\sum_{t=1}^T \left(\max_i\{\boldsymbol{x}_{t,i}^\top \boldsymbol{\theta}\} - \boldsymbol{y}_t^\top \boldsymbol{\theta}\right)\right].$

---

[3]Note that our formulation is more general than the usual linear contextual bandits setting where $\|\boldsymbol{\theta}\|_2 \leq 1$ and $\|\boldsymbol{x}_{t,i}\|_2 \leq 1$. It also includes K-armed multi-armed bandits as a special case.

Batch Learning. Given the batch complexity M, a batch learning algorithm aims at minimize the regret R_T defined above, subject to the constraints in Definition 1.

Notations. For any non-negative integer N, we let $[N]$ denote $\{1, 2, \ldots, N\}$. We use $\mathbb{E}_{\mathcal{P}}[\cdot]$ and $\Pr_{\mathcal{P}}[\cdot]$ to denote the expectation and probability over the distribution $\mathcal{P}$ respectively. We use $\mathbf{I}$ to denote the d-dimensional identity matrix. We use $\log$ to denote the logarithm base 2, and use $\ln$ to denote the logarithm base e. Throughout the paper, the $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ notations hide the logarithmic factors of T, d, and K. We also define $\mathcal{T}_0 = T_0 = 0$ and $\mathcal{T}_k = \sum_{i=1}^{k} T_i$ for $1 \leq k \leq M$.

3 TECHNICAL OVERVIEW OF OUR ALGORITHMS

In this section, we summarize our techniques contribution and sketch the high-level ideas in the algorithms and regret analysis.

Our algorithms are elimination-based, following (Ruan et al., 2021). At each time step t, give the set of context vectors $X_t = \{\boldsymbol{x}_{t,1}, \boldsymbol{x}_{t,2}, \ldots, \boldsymbol{x}_{t,K}\}$, we maintain a confidence interval $\mathcal{I}_{t,i}$ for $\boldsymbol{x}_{t,i}^{\top}\theta$ for each i. A candidate action $\boldsymbol{x}_{t,i}$ is eliminated when there exists another candidate action $\boldsymbol{x}_{t,i'}$ such that $\mathcal{I}_{t,i'}$ entirely lies above $\mathcal{I}_{t,i}$, meaning that the action $\boldsymbol{x}_{t,i}$ cannot be the optimal action. Then, the clever part of the policy is to decide a distribution over the remaining candidate actions and randomly choose one to commit to according to the distribution.

For the construction of the confidence intervals, we adopt the classical elliptical confidence intervals based on the regularized ordinary least-square (OLS) estimation (Chu et al., 2011). Given a group of context vectors that are played in history $\{\boldsymbol{y}_\tau\}_{\tau=1}^{t}$ and corresponding observed rewards $\{r_\tau\}_{\tau=1}^{t}$ (such that $r_\tau = \boldsymbol{y}_\tau^{\top}\boldsymbol{\theta} + \epsilon_\tau$ where ϵ_τ is an 1-subgaussian noise), we construct the confidence interval for any candidate action with context vector $\boldsymbol{x}$ to be

$$\mathcal{I}(\boldsymbol{x}, \Lambda) = \left[\boldsymbol{x}^{\top}\hat{\boldsymbol{\theta}} - \alpha\sqrt{\boldsymbol{x}^{\top}\Lambda^{-1}\boldsymbol{x}}, \boldsymbol{x}^{\top}\hat{\boldsymbol{\theta}} + \alpha\sqrt{\boldsymbol{x}^{\top}\Lambda^{-1}\boldsymbol{x}}\right] \bigcap [-1, 1],$$

with $\Lambda = \lambda\mathbf{I} + \sum_{\tau=1}^{t}\boldsymbol{y}_\tau\boldsymbol{y}_\tau^{\top}$ is the *regularized information matrix*, $\hat{\boldsymbol{\theta}} = \Lambda^{-1}\sum_{\tau=1}^{t}r_\tau\boldsymbol{y}_\tau$ is the regularized OLS estimation of the hidden vector $\boldsymbol{\theta}$, and α, λ are hyper-parameters satisfying that $\alpha = \Theta(\sqrt{\ln(KdT)} + \lambda\sqrt{d})$.

Define $w(\boldsymbol{x}, \Lambda) = \min\left\{(\sqrt{\ln(KdT)} + \lambda\sqrt{d})\sqrt{\boldsymbol{x}^{\top}\Lambda^{-1}\boldsymbol{x}}, 1\right\}$ to be the *width* of the confidence interval $\mathcal{I}(\boldsymbol{x}, \Lambda)$. To reduce the regret, we would like to design policies to cleverly perform exploration in order to reduce the width of future estimations. Formally, we introduce the following problem which is the key to our optimal batch learning algorithm.

The Problem of Single-Phase Learning for Exploration Policy. *Fix* $m, n \geq 0$. *Let* $\{X_i\}_{i=1}^{m}, \{Y_j\}_{j=1}^{n}$ *be two groups of* i.i.d. *context sets following the same unknown distribution D. After observing $\{X_i\}_{i=1}^{m}$, we are asked to design the parameter λ and an exploration policy π so as to minimize the following expected maximum width* EM-width$(D, \Lambda) := \mathbb{E}_{X \sim D}[\max_{\boldsymbol{x} \in X} w(\boldsymbol{x}, \Lambda)]$, *where $\Lambda = \lambda\mathbf{I} + \sum_{j=1}^{n}\boldsymbol{y}_j\boldsymbol{y}_j^{\top}$ and $\boldsymbol{y}_j \sim \pi(Y_j)$ for all $1 \leq j \leq n$.*

In our batch learning algorithm, we need to solve the above problem once during each batch. During the k-th batch, we let $D = D_k$ be the distribution of the set of the remaining context vectors after the elimination process based on the information learned for the first $(k-1)$ batches. $\{X_i\}_{i=1}^{m}$ is obtained from the $(k-1)$-th batch, and we solve the above problem for π which serves as the exploration policy for the k-th batch. The minimization goal of the above problem helps to reduce the regret starting from the $(k+1)$-th batch.

To facilitate discussion, we use $\mathcal{W}(m, n)$ to denote the minimax optimum of expected maximum width achieved the best learning algorithm $\mathcal{G}$. That is, we let

$$\mathcal{W}(m, n) = \inf_{\mathcal{G}} \sup_{D} \mathbb{E}_{\{X_i, Y_j\} \sim D^{\otimes(m+n)}} \mathbb{E}_{(\pi, \lambda) \sim \mathcal{G}(\{X_i\})} \left[\text{EM-width}(D, \Lambda)\right],$$

where $\mathcal{G}$ is the single-phase learning algorithm to decide π and λ based on $\{X_i\}_{i=1}^{m}$.

In (Ruan et al., 2021), the authors showed the existence of a good policy π and the choice of λ so that the bound of the expected maximum width leads to the desired optimal regret for $M =$

$\Theta(\log \log T)$ batches. Their constructive proof (given the distribution D) is based on a reward-free LinUCB algorithm (Algorithm 2 in their paper). However, to learn such a good policy based on $\{X_i\}_{i=1}^m$, the authors employed more complicated procedures (such as CORELEARNING and COREIDENTIFICATION).

Both (Zanette et al., 2021) and our work are inspired by the reward-free LinUCB and find that one may leverage this algorithmic framework to design learning algorithms as well. The authors of (Zanette et al., 2021) worked on a similar task as the single-phase learning problem defined above and their result implies that $\mathcal{W}(m, n) \leq O\left(\text{poly}\ln(mndT) \cdot \sqrt{\ln(K)} \cdot (\sqrt{d/n} + \sqrt{d/m})\right)$.[4]

In comparison, in Section 4.1 we propose `Exp-Policy` to solve the single-phase learning problem. In Lemma 5 we analyze our algorithm and show that

$$\mathcal{W}(m, n) \leq O\left(\text{poly}\ln(mndT) \cdot \sqrt{\ln(K)} \cdot \left(\sqrt{\frac{d}{n}} + \frac{d}{m}\right)\right). \tag{1}$$

Clearly, the performance of our `Exp-Policy` outperforms that the results in (Zanette et al., 2021) in terms of the dependence on m. Note that in our batch learning algorithm, m represents the number of samples in the previous batch, which is much smaller than the size of the current batch (represented by n). Therefore, the d/m term in our bound usually dominates and it is crucial for us to make this $\sqrt{m}$-factor improvement to achieve the optimal regret in the batch learning model. Indeed, without this improvement, the result of (Zanette et al., 2021) does not even imply the desired optimal regret for $M = \Theta(\log \log T)$ batches (the result of (Ruan et al., 2021)).

The proof of Lemma 5 is based on the analysis of reward-free LinUCB, and involves a scaled-and-clipped update rule and a dynamic concentration inequality for PSD matrices. Below we present the high-level ideas.

Learning the Exploration Policy via Reward-free LinUCB. In (Ruan et al., 2021), the authors showed that given $\{Y_j\}_{j=1}^n \sim D^{\otimes n}$, the reward-free LinUCB algorithm can produce $\{\boldsymbol{y}_j\}_{j=1}^n$ (and therefore also form a policy) such that

$$\text{EM-width}(D, U) \leq O(\sqrt{d \ln(nd/\kappa)/n}), \tag{2}$$

where $U = \kappa \mathbf{I} + \sum_{j=1}^n \boldsymbol{y}_j \boldsymbol{y}_j^\top$, and $\kappa > 0$ is polynomially small (e.g., $\kappa = T^{-2}$) to make sure that U is invertible while we do not lose much in equation 2. In this work, our `Exp-Policy` algorithm cannot direct access $\{Y_j\}_{j=1}^n$ but has to learn the distribution D and construct a policy π based on a much smaller data set $\{X_i\}_{i=1}^m$.[5]

To make learning possible, we first notice that the original reward-free LinUCB in (Ruan et al., 2021) produces $\{\boldsymbol{y}_j\}_{j=1}^n$ by the so-called *argmax policy*: $\boldsymbol{y}_j = \pi_j(Y_j) := \arg\max_{\boldsymbol{y} \in Y_j}\{\boldsymbol{y}^\top W_j \boldsymbol{y}\}$, where $W_j = \kappa \mathbf{I} + \sum_{q=1}^j \boldsymbol{y}_q \boldsymbol{y}_q^\top$ is the regularized information matrix obtained from the samples before j.[6] One may combine $\{\pi_j\}$ via carefully chosen probability weights to form a desired one-shot policy π.

In our algorithm `Exp-Policy` (Algorithm 2), we observe that we may approximately learn the policy π from $\{X_i\}_{i=1}^m$ as long as we are able to approximately construct $\{W_j\}$ based on the $\{X_i\}$ data set. I.e., for any $W = W_j$, we would like to construct $\check{W}$ as long as $\Omega(1) \cdot \check{W} \preccurlyeq W \preccurlyeq \tilde{O}(1) \cdot \check{W}$, and the key here is to lower bound W by $\Omega(1) \cdot \check{W}$.

To illustrate the main technical challenge and our solution, let us consider the following task: let $\boldsymbol{x}_i \sim \pi(X_i)$ and $\check{W} = \frac{n}{m}(\kappa \mathbf{I} + \sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{x}_i^\top)$, we would like to choose appropriate regularization parameter $\lambda > 0$ so that with high probability (over the randomness of $\{X_i\}$ and $\{\boldsymbol{x}_i\}$) it holds that

$$\Omega(1) \cdot \check{W} \preccurlyeq n\left(\lambda \mathbf{I} + \mathbb{E}_{X \sim D, \boldsymbol{x} \sim \pi(X)} \boldsymbol{x} \boldsymbol{x}^\top\right). \tag{3}$$

---

[4]We state this implication by making the "large $|\mathcal{S} \times \mathcal{A}|$" assumption in (Zanette et al., 2021).

[5]Comparing equation 1 and equation 2, we also find that the cost we pay in the expected maximum width for learning is about $\tilde{O}(d/m)$.

[6]We warn the readers that this is an oversimplification of the algorithm by omitting a few important techniques such as the volume-based lazy update of the W_j matrices. However, we choose to the current presentation to better motivate our technical contributions.

We note that this task may seem a bit different from our goal: 1) π is unknown to the learner, and cannot be used to construct $\check{W}$; 2) the upper bound is quite different from $W = \kappa \mathbf{I} + \sum_{j=1}^{n} \boldsymbol{y}_j \boldsymbol{y}_j^\top$. Indeed, these issues may be (quite non-trivially) resolved by observing that 1) π is a mixed policy and can be iteratively updated to its final form and 2) relate the Right-Hand-Side of equation 3 to W by another matrix concentration inequality.

We now focus on the task of equation 3, which is equivalent to

$$\Omega(1) \cdot \left(\kappa \mathbf{I} + \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \preccurlyeq m \left(\lambda \mathbf{I} + \mathbb{E}_{X \sim D, \boldsymbol{x} \sim \pi(X)} \boldsymbol{x} \boldsymbol{x}^\top \right). \tag{4}$$

Note that since $\boldsymbol{x}_i \sim \pi(X_i)$ are *i.i.d.* random variables, standard matrix concentration inequalities would imply equation 4 when $m\lambda \mathbf{I}$ upper bounds $\boldsymbol{x}_i \boldsymbol{x}_i^\top$ (up to logarithmic factors of the inverse of the failure probability) almost surely, i.e., $m\lambda \geq \tilde{\Omega}(1) \Leftrightarrow \lambda \geq \tilde{\Omega}(1/m)$. This choice of λ would lead to a $d/\sqrt{m}$ term instead of the d/m term in equation 1.

The Scaled-and-Clipped Update Rule. While it is not possible to establish equation 1 (with high probability) for a smaller λ (e.g., $\lambda = o(1/m)$), we introduce the *scaled-and-clipped update rule* in the reward-free LinUCB (Line 2 of Algorithm 2) which eventually leads to the improvement of λ.

More concretely, instead of working with $\check{W} = \frac{n}{m}(\kappa \mathbf{I} + \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top)$, we define the *scaled-and-clipped version* of $\boldsymbol{x}_i$ and the *scaled-and-clipped information matrix* U_i,[7]

$$\tilde{\boldsymbol{x}}_i := \min \left\{ \sqrt{\frac{1}{\boldsymbol{x}_i^\top U_{i-1}^{-1} \boldsymbol{x}_i}}, 1 \right\} \boldsymbol{x}_i, \quad U_i = \kappa \mathbf{I} + \sum_{p=1}^{i} \tilde{\boldsymbol{x}}_p \tilde{\boldsymbol{x}}_p^\top.$$

We will use $\{U_i\}$ to construct the mixed argmax policy instead of W's. The downside of this new update rule is that we use shorter feature vectors $\{\tilde{\boldsymbol{x}}_i\}$ instead of the original ones, which leads to the slower growth of the information matrix. However, this slowing effect is not too bad – if we repeat each $\boldsymbol{x}_i$ by $\ln(1/\kappa) = O(\ln T)$ times, scaled-and-clipped information matrix U_i would upper bound the original information matrix. Through a more rigorous analysis, we will see that this effect would only hurt the regret by a logarithmic factor.

On the other hand, the benefit of our scaled-and-clipped update rule is that instead of establishing equation 4, we only need to lower bound the Right-Hand-Side of equation 4 by the scaled-and-clipped information matrix, i.e., to prove that the following inequality holds with high probability.

$$\Omega(1) \cdot \left(\kappa \mathbf{I} + \sum_{i=1}^{m} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top \right) \preccurlyeq m \left(\lambda \mathbf{I} + \mathbb{E}_{X \sim D, \boldsymbol{x} \sim \pi(X)} \boldsymbol{x} \boldsymbol{x}^\top \right), \tag{5}$$

where $\tilde{\boldsymbol{x}}_i = \min \left\{ \sqrt{\frac{1}{\boldsymbol{x}_i^\top U_{i-1}^{-1} \boldsymbol{x}_i}}, 1 \right\} \boldsymbol{x}_i$ and we assume that $\boldsymbol{x}_i \sim \pi(X)$ are *i.i.d.* (as we did in equation 4).

Let $U = \kappa \mathbf{I} + \sum_{i=1}^{m} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top$. To prove equation 5, we only need to show that there exists constants $c_2 > c_1 > 0$ such that

$$c_2 U \preccurlyeq c_1 U + m\lambda \mathbf{I} + m \mathbb{E}_{X \sim D, \boldsymbol{x} \sim \pi(X)} \tilde{\boldsymbol{x}} \tilde{\boldsymbol{x}}^\top, \tag{6}$$

which implies that

$$(c_2 - c_1) U \preccurlyeq m\lambda \mathbf{I} + \mathbb{E}_{X \sim D, \boldsymbol{x} \sim \pi(X)} \tilde{\boldsymbol{x}} \tilde{\boldsymbol{x}}^\top \preccurlyeq m\lambda \mathbf{I} + \mathbb{E}_{X \sim D, \boldsymbol{x} \sim \pi(X)} \boldsymbol{x} \boldsymbol{x}^\top,$$

where the last inequality is due to the clipping operation in the update rule.

Now let us focus on the task of establishing equation 6. Thanks to the definition of the scaled-and-clipped version $\tilde{\boldsymbol{x}}_i$, we have that $\tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top \preccurlyeq U$ almost surely. Therefore, it is possible to establish equation 6 as long as we choose $m\lambda \geq \Omega(\kappa)$ to cover the κI term in U (which only requires that $\lambda \geq \Omega(\kappa/m) = \Omega(m^{-1}T^{-2})$, leading to the better d/m error term in equation 1). On the other hand, however, we note that U, while serving as an upper bound of the random matrices $\tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top$, is also a random variable by itself. We do not find sharp matrix concentration inequalities in literature to fit our need, and we have to resort to the matrix concentration inequality with dynamic upper bounds proved in Lemma 20. The formal version of equation 6 is stated and proved in Lemma 14.

---

[7]The definition of $\tilde{\boldsymbol{x}}_i$ here is slightly different from the real algorithm (up to a logarithmic factor L). We make this simplification only to better explain the main algorithmic ideas.

3.1 LOWER BOUNDS

To construct the lower bound instances, we first construct M mutually independent sub-problems with dimension d/M. At each time step, a uniform random sub-problem (i.e., its context set) is selected and presented to the learner. To analyze the regret performance of any batch learning algorithm, we divide the time horizon T into M consecutive stages with properly chosen lengths $\{T_k\}_{k=1}^M$ for each stage. We will show that, for each $k \in [M]$, during the k-th stage, if the learner does not start a new batch and update its policy, then a large regret would incur for the k-th sub-problem in this stage.

Our construction for the sub-problems is as follows. For each $k \in \{1, 3, 4, 5, \ldots, M\}$, we choose the sub-problem to be the hard instance for the K-armed linear contextual linear bandit problem with burn-in time $\tilde{\Theta}(T_{k-1})$. In other words, we construct the k-th sub-problem so that it is hard for the learner to find a good policy for the k-th problem during the first $\sum_{i \le k-1} T_i$ time steps.

The most interesting sub-problem design is for $k = 2$, where the learner suffer more regret due to lack of context information. In this sub-problem, we consider a linear contextual bandit problem d/M arms (assuming that $K \ge d/M$). However, during each time step, only $(d/(2M) + 1)$ arms are presented to the learner. In particular, we set the first $d/(2M)$ arms to be the *frequent* arms that always appear in the context set. For the rest $d/(2M)$ *infrequent* arms, we choose one of them uniformly randomly and include the chosen one in the context set at each time step. The best exploration strategy for the learner is to play the infrequent arm with a higher probability, e.g., $\frac{1}{2}$ and play the frequent arms with probability $\frac{M}{d}$. However, the learner can not tell whether an arm is frequent or infrequent with insufficient information, and his best strategy is to play the arms with the same probability. As a result, the infrequent arms are insufficiently explored, which leads to an extra $\tilde{\Theta}(\sqrt{d})$ factor in the regret.

4 BATCH LEARNING ALGORITHM

We now present our batch learning algorithm in Algorithm 1, which corresponds to Theorem 2. Let the time schedule $\{T_k\}_{k=1}^M$ such that $\sum_{k=1}^M T_k \ge T$ to be determined later. We can then accordingly calculate the end of each batch by $\mathcal{T}_k = \min\left\{\sum_{\ell=1}^k T_k, T\right\}$.

In preparation for explaining the algorithm, we first introduce a few variables and notations used in the algorithm. During the k-th batch, the algorithm learns an estimation, namely $\hat{\boldsymbol{\theta}}_k$, of the hidden vector $\boldsymbol{\theta}$, as well as an information matrix Λ_k that is used to construct the confidence interval for the estimated rewards based on $\hat{\boldsymbol{\theta}}_k$. More specifically, given the pair $(\Lambda_k, \hat{\boldsymbol{\theta}}_k)$, we set the confidence interval for the expected reward of any feature vector $\boldsymbol{x}$ to be $[\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_k \pm \alpha \sqrt{\boldsymbol{x}^\top \Lambda_k^{-1} \boldsymbol{x}}]$, where $[a \pm b]$ denotes the interval $[a - b, a + b]$ and we set $\alpha := \sqrt{50 \ln(KTd/\delta)}$.

Given the pair $(\Lambda_k, \hat{\boldsymbol{\theta}}_k)$, for any context set $X \subseteq \mathbb{R}^d$, we define the following natural elimination procedure based on the corresponding confidence intervals

$$\mathcal{E}(X; (\Lambda_k, \hat{\boldsymbol{\theta}}_k)) := \left\{ \boldsymbol{x} \in X : \boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_k + \alpha \sqrt{\boldsymbol{x}^\top \Lambda_k^{-1} \boldsymbol{x}} \ge \max_{\boldsymbol{y} \in X} \left(\boldsymbol{y}^\top \hat{\boldsymbol{\theta}}_k - \alpha \sqrt{\boldsymbol{y}^\top \Lambda_k^{-1} \boldsymbol{y}} \right) \right\}. \quad (7)$$

In words, $\mathcal{E}(X; (\Lambda_k, \hat{\boldsymbol{\theta}}_k))$ returns the set of the survived feature vectors, each of which remains possible to hold the highest expected reward when assuming all confidence intervals based on $(\Lambda_k, \hat{\boldsymbol{\theta}}_k)$ contains their estimation targets. By the end of the k-th batch, the algorithm would have learned k pairs $\{(\Lambda_i, \hat{\boldsymbol{\theta}}_i)\}_{i=1}^k$, and we naturally extend our elimination procedure to $\{(\Lambda_i, \hat{\boldsymbol{\theta}}_i)\}_{i=1}^k$ as $\mathcal{E}(X; \{(\Lambda_i, \hat{\boldsymbol{\theta}}_i)\}_{i=1}^k) := \cap_{i=1}^k \mathcal{E}(X; (\Lambda_i, \hat{\boldsymbol{\theta}}_i))$. When $X \sim D$ and given $\{(\Lambda_i, \hat{\boldsymbol{\theta}}_i)\}_{i=1}^k$, we denote the distribution of $\mathcal{E}(X; \{(\Lambda_i, \hat{\boldsymbol{\theta}}_i)\}_{i=1}^k)$ by D_{k+1}.

We now explain the key steps of the algorithm. For the first batch, we take actions according to the local optimal design policy π^{G}, which is defined by the famous *General Equivalence Theorem* in (Kiefer & Wolfowitz, 1960) (see Lemma 6 in Appendix A).

For $k \ge 2$, at any time t during the k-th batch, the algorithm observes the context set $X_t \sim D$, eliminates some of the sub-optimal arms in Line 1, and denote the set of the survived arms by $X_t^{(k)}$.

Algorithm 1 `Batch Learning`

1: **Initialize:** $\lambda \leftarrow 10/T$; $\Lambda_0 \leftarrow \lambda\mathbf{I}$; $\hat{\boldsymbol{\theta}}_0 \leftarrow \mathbf{0}$;
2: **for** $t = 1, 2, \ldots, \mathcal{T}_1$ **do**
3:     Observe X_t;
4:     Play the arm with the feature vector $\boldsymbol{y}_t \sim \pi^{\mathtt{G}}(X_t)$ and receive the reward r_t;
5: **end for**
6: $\Lambda_1 \leftarrow \lambda\mathbf{I} + \sum_{t=1}^{\mathcal{T}_1/2} \boldsymbol{y}_t\boldsymbol{y}_t^\top$; $\hat{\boldsymbol{\theta}}_1 \leftarrow \Lambda_1^{-1} \sum_{t=1}^{\mathcal{T}_1/2} r_t\boldsymbol{y}_t$;
7: $\pi_2 \leftarrow \mathtt{Exp-Policy}\left(\{\mathcal{E}(X_t, \{\Lambda_1, \hat{\theta}_1\})\}_{t=\mathcal{T}_1/2+1}^{\mathcal{T}_1}\right)$;
8: **for** $k = 2, \ldots, M$ **do**
9:     **for** $t = \mathcal{T}_{k-1} + 1, \mathcal{T}_{k-1} + 2, \ldots, \mathcal{T}_k$ **do**
10:       Observe X_t;
11:       $X_t^{(k)} \leftarrow \mathcal{E}(X_t, \{\Lambda_i, \hat{\boldsymbol{\theta}}_i\}_{i=1}^{k-1})$;
12:       Play the arm with the feature vector $\boldsymbol{y}_t \sim \pi_k(X_t^{(k)})$ and receive the reward r_t;
13:     **end for**
14:     $\Lambda_k \leftarrow \lambda\mathbf{I} + \sum_{t=\mathcal{T}_{k-1}+1}^{\mathcal{T}_{k-1}+T_k/2} \boldsymbol{y}_t\boldsymbol{y}_t^\top$; $\hat{\boldsymbol{\theta}}_k \leftarrow \Lambda_k^{-1} \sum_{t=\mathcal{T}_{k-1}+1}^{\mathcal{T}_{k-1}+T_k/2} r_t\boldsymbol{y}_t$;
15:     $X_t^{(k+1)} \leftarrow \mathcal{E}(X_t, \{\Lambda_i, \hat{\boldsymbol{\theta}}_i\}_{i=1}^{k}), \forall \mathcal{T}_{k-1} + 1 \le t \le \mathcal{T}_k$;
16:     $\pi_{k+1} \leftarrow \mathtt{Exp-Policy}(\{X_t^{(k+1)}\}_{t=\mathcal{T}_{k-1}+T_k/2+1}^{\mathcal{T}_k})$;
17: **end for**

By the definition above, the $X_t^{(k)}$ follows the distribution D_k when conditioned on the first $(k-1)$ batches. The algorithm then uses an exploration policy π_{k-1} to stochastically select and play an arm $\boldsymbol{y}_t \in X_t^{(k)}$.

At the end of the k-th batch, our algorithm divides the T_k data points collected in the batch into two parts of the equal sizes. In Line 1, our algorithm calculates Λ_k and $\hat{\boldsymbol{\theta}}_k$ using the standard ridge regression and the first part of the data points. In Line 1, the new exploration policy π_k is learned by the $\mathtt{Exp-Policy}$ procedure using the context sets from the second part of the data points. Note that the context sets fed into $\mathtt{Exp-Policy}$ go through the elimination procedure based on $\{\Lambda_i, \hat{\boldsymbol{\theta}}\}_{i=1}^k$, which depends on the first part of the data points.

4.1 Learning the Exploration Policy

We describe the $\mathtt{Exp-Policy}$ procedure by Algorithm 2. Given a group of context vectors $\{Z_i\}_{i=1}^m$, the algorithm simulates the reward-free linear bandit algorithms. In each time step, the algorithm first clip the context vectors according the current information matrix W, and then chooses the arm with clipped maximal variance. The information matrix is updated with doubling trick, which helps reduce both the number of updates and the complexity of the output policy. With Algorithm 2, we have the results below.

Lemma 5. *Let $\{Z_u\}_{u=1}^m$ be m i.i.d. stochastic context sets following a distribution D. Let π be the output by running $\mathtt{Exp-Policy}$ (Algorithm 2) with the input $\{Z_u\}_{u=1}^m$. Let $\{\tilde{Z}_u\}_{u=1}^n$ be another group of i.i.d. stochastic context sets following the distribution D (which is also independent from $\{Z_u\}_{u=1}^m$). Let $\boldsymbol{y}_u$ be independently sampled from $\pi(\tilde{Z}_u)$ for each $u \in \{1, 2, \ldots, n\}$, and let $\Lambda = \sum_{u=1}^n \boldsymbol{y}_u\boldsymbol{y}_u^\top$. Let $L > 0$ be a constant. With probability $(1 - 3\delta)$, we have that*

$$\mathbb{E}_{X \sim D}\left[\min\left\{\sqrt{\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top\left(\Lambda + \frac{n}{m}\kappa\mathbf{I}\right)^{-1}\boldsymbol{x}}, \sqrt{L}\right\}\right] \le O\left(\sqrt{\frac{d}{n}\ln\left(\frac{md}{\kappa}\right)} + \frac{d\sqrt{L}}{m}\ln\left(\frac{md}{\kappa}\right)\right). \tag{8}$$

Due to space constraints, we postpone the full proof of Theorem 2 and Lemma 5 to Appendix B.

5 Conclusion

In this paper, we study the batch linear contextual linear bandit problem with stochastic context. When the number of batches is limited by M, for any T and d, we achieve matching upper and lower

---

Algorithm 2 `Exp − Policy`

1: **Input:** $\{Z_i\}_{i=1}^m$, L;
2: **Initialization:** $\kappa = \frac{1}{T^2}$ $U_0 \leftarrow \kappa\mathbf{I}$; $\eta \leftarrow 1$, $\tau_\eta \leftarrow \emptyset$, $W_\eta \leftarrow U_0$;
3: **for** $i = 1, 2, \ldots, m$ **do**
4:    $\tau_\eta \leftarrow \tau_\eta \cup \{i\}$;
5:    Choose $\boldsymbol{z}_i \in Z_i$ to maximize $\boldsymbol{z}_i^\top W_\eta^{-1} \boldsymbol{z}_i$;
6:    $\tilde{\boldsymbol{z}}_i \leftarrow \min\left\{\sqrt{L/(\boldsymbol{z}_i^\top W_\eta^{-1} \boldsymbol{z}_i)}, 1\right\} \boldsymbol{z}_i$; $U_i \leftarrow U_{i-1} + \tilde{\boldsymbol{z}}_i\tilde{\boldsymbol{z}}_i^\top$;
7:    **if** $\det(U_i) > 2\det(W_\eta)$ **then**
8:      $\eta \leftarrow \eta + 1$, $\tau_\eta \leftarrow \emptyset$, $W_\eta \leftarrow U_i$;
9:    **end if**
10: **end for**
11: **return** π such that $\pi(X) = \arg\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top W_j^{-1} \boldsymbol{x}$ with probability $|\tau_j|/m$ for $j \in [\eta]$.

---

bounds for the regret (up to logarithmic factors) in both context-blind and context-aware settings. In the algorithm design and analysis, we highlight two key techniques: the scaled-and-clipped update rule and the matrix concentration inequality with dynamic upper bounds. We believe these techniques could help design and analyze batch algorithms for other online learning and decision-making problems with linear reward structures (e.g., the linear Markov Decision Processes).

ACKNOWLEDGMENTS

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24:2312–2320, 2011.

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 3–11, 1999.

Naoki Abe, Alan W Biermann, and Philip M Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

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

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Regret bounds for batched bandits. *arXiv preprint arXiv:1910.04959*, 2019.

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Regret bounds for batched bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7340–7348, May 2021.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32:503–513, 2019.

Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

Osama Hanna, Lin Yang, and Christina Fragouli. Efficient batched algorithm for contextual linear bandits with large action space via soft elimination. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Osama A Hanna, Lin Yang, and Christina Fragouli. Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1791–1821, 2023b.

Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, and Quanquan Gu. Almost optimal anytime algorithm for batched multi-armed bandits. In *International Conference on Machine Learning*, pp. 5065–5073. PMLR, 2021a.

Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Quanquan Gu. Double explore-then-commit: Asymptotic optimality and beyond. In *Conference on Learning Theory*, pp. 2584–2633. PMLR, 2021b.

Tianyuan Jin, Yu Yang, Jing Tang, Xiaokui Xiao, and Pan Xu. Optimal batched best arm identification. *arXiv preprint arXiv:2310.14129*, 2023.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International conference on World Wide Web*, pp. 661–670, 2010.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pp. 2173–2174. PMLR, 2019.

Elliott H Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Les rencontres physiciens-mathématiciens de Strasbourg-RCP25*, 19:0–35, 1973.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, et al. Batched bandit problems. *Annals of Statistics*, 44(2):660–681, 2016.

Dénes Petz. A survey of certain trace inequalities. *Banach Center Publications*, 30(1):287–298, 1994.

Xuanfei Ren, Tianyuan Jin, and Pan Xu. Optimal batched linear bandits. *arXiv preprint arXiv:2406.04137*, 2024.

Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 74–87, 2021.

Joel A Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Andrea Zanette, Kefan Dong, Jonathan N Lee, and Emma Brunskill. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34: 22720–22731, 2021.

A   TECHNICAL LEMMAS

Lemma 6 (General Equivalence Theorem in (Kiefer & Wolfowitz, 1960)). *For any bounded subset $X \subset \mathbb{R}^d$, there exists a distribution $\mathcal{K}(X)$ supported on X, such that for any $\epsilon > 0$, it holds that*

$$\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top \left(\epsilon \mathbf{I} + \mathbb{E}_{\boldsymbol{y} \sim \mathcal{K}(X)}[\boldsymbol{y}\boldsymbol{y}^\top] \right)^{-1} \boldsymbol{x} \le d. \tag{9}$$

Furthermore, there exists a mapping π^{G}, which maps a context X to a distribution over X such that

$$\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top (\epsilon \mathbf{I} + \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)}[\boldsymbol{y}\boldsymbol{y}^\top])^{-1} \boldsymbol{x} \le 2d. \tag{10}$$

In particular, when $\mathrm{supp}(X)$ has a finite size, $\pi^{\mathsf{G}}(X)$ could be implemented within $\mathrm{poly}(|\mathrm{supp}(X)|)$ time.

Lemma 7. *Let $D_{\mathrm{KL}}(P\|Q)$ denote the KL-divergence between P and Q. Let $D_1(P,Q) := |P - Q|_1$ denote the L_1 distance between P and Q. For any two distribution P, Q and a random variable X bounded by $[-N, N]$, we have that*

$$|\mathbb{E}_P[X] - \mathbb{E}_Q[X]| \le N \cdot D_1(P,Q) N \cdot \min\{\sqrt{2D_{\mathrm{KL}}(P\|Q)}, \sqrt{2D_{\mathrm{KL}}(Q\|P)}\} \tag{11}$$

Proof. The second inequality is the famous Pinsker's inequality. For the first inequality, we have that

$$|\mathbb{E}_P[X] - \mathbb{E}_Q[X]| = \left| \int x(dP(x) - dQ(x)) \right| \le N \left| \int |dP(x) - dQ(x)| \right| = ND_1(P,Q). \tag{12}$$

$\square$

Lemma 8. *Let $X_1, X_2, \ldots$ be a sequence of random variables taking value in $[0, l]$. Define $\mathcal{F}_k = \sigma(X_1, X_2, \ldots, X_{k-1})$ and $Y_k = \mathbb{E}[X_k | \mathcal{F}_k]$ for $k \ge 1$. For any $\delta > 0$, we have that*

$$\mathbb{P}\left[\exists n, \sum_{k=1}^n X_k \le 3 \sum_{k=1}^n Y_k + l \ln(1/\delta) \right] \le \delta$$

$$\mathbb{P}\left[\exists n, \sum_{k=1}^n Y_k \ge 3 \sum_{k=1}^n X_k + l \ln(1/\delta) \right] \le \delta.$$

Proof. Let $t \in [0, 1/l]$ be fixed. Consider to bound $Z_k := \exp(t \sum_{k'=1}^k (X_{k'} - 3Y_{k'}))$. By definition, we have that

$$
\begin{aligned}
\mathbb{E}[Z_k | \mathcal{F}_k] &= \exp(t \sum_{k'=1}^{k-1} (X_{k'} - 3Y_{k'})) \mathbb{E}\left[t(X_k - 3Y_k) \right] \\
&\le \exp(t \sum_{k'=1}^{k-1} (X_{k'} - 3Y_{k'})) \exp(-3Y_k) \cdot \mathbb{E}[1 + tX_k + 2t^2 X_k^2] \\
&\le \exp(t \sum_{k'=1}^{k-1} (X_{k'} - 3Y_{k'})) \exp(-3Y_k) \cdot \mathbb{E}[1 + 3tX_k] \\
&= \exp(t \sum_{k'=1}^{k-1} (X_{k'} - 3Y_{k'})) \exp(-3Y_k) \cdot (1 + 3tY_k) \\
&\le \exp(t \sum_{k'=1}^{k-1} (X_{k'} - 3Y_{k'})) \\
&= Z_{k-1},
\end{aligned}
$$

where the second line is by the fact that $e^x \le 1 + x + 2x^2$ for $x \in [0, 1]$. Define $Z_0 = 1$ Then $\{Z_k\}_{k \ge 0}$ is a super-martingale with respect to $\{\mathcal{F}_k\}_{k \ge 1}$. Let τ be the smallest n such that

$\sum_{k=1}^{n} X_k - 3 \sum_{k=1}^{n} Y_k > l \ln(1/\delta)$. It is easy to verify that $Z_{\min\{\tau,n\}} \leq \exp(tl \ln(1/\delta) + tl) < \infty$. Choose $t = 1/l$. By the optimal stopping time theorem, we have that

$$\mathbb{P}\left[\exists n \leq N, \sum_{k=1}^{n} X_k \geq 3 \sum_{k=1}^{n} Y_k + l \ln(1/\delta)\right]$$

$$= \mathbb{P}\left[\tau \leq N\right]$$

$$\leq \mathbb{P}\left[Z_{\min\{\tau,N\}} \geq \exp(tl \ln(1/\delta))\right]$$

$$\leq \frac{\mathbb{E}[Z_{\min\{\tau,N\}}]}{\exp(tl \ln(1/\delta))}$$

$$\leq \delta.$$

Letting $N \to \infty$, we have that

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} X_k \leq 3 \sum_{k=1}^{n} Y_k + l \ln(1/\delta)\right] \leq \delta.$$

Considering $W_k = \mathbb{E}[\exp(t \sum_{k'=1}^{k}(Y_k/3 - X_k))]$, using similar arguments and choosing $t = 1/(3l)$, we have that

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} Y_k \geq 3 \sum_{k=1}^{n} X_k + l \ln(1/\delta)\right] \leq \delta.$$

The proof is completed. □

A.1 STATEMENT AND PROOF OF THE ELLIPTICAL POTENTIAL LEMMA (LEMMA 9)

Lemma 9. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be a sequence of vectors in $\mathbb{R}^d$ such that $\|x_i\|_2 \leq 1$. Let $\Lambda_0 = A$ be a positively definite matrix and $\Lambda_i = \Lambda_0 + \sum_{j=1}^{i} \boldsymbol{x}_j \boldsymbol{x}_j^\top$. If $\boldsymbol{x}_i^\top \Lambda_{i-1}^{-1} \boldsymbol{x}_i \leq 1$ for all $1 \leq i \leq n$, it then holds that*

$$\sum_{i=1}^{n} \boldsymbol{x}_i^\top \Lambda_{i-1}^{-1} \boldsymbol{x}_i \leq 2 \ln\left(\frac{\det(\Lambda_n)}{\det(\Lambda_0)}\right).$$

Proof. Note that $\det(\Lambda_{i+1}) = \det(\Lambda_i)(1 + \boldsymbol{x}_{i+1}^\top \Lambda_i^{-1} \boldsymbol{x}_{i+1})$. Since $\ln(1+x) \geq \frac{x}{2}$ when $0 \leq x \leq 1$, we have that

$$\ln(\det(\Lambda_{i+1})) - \ln(\det(\Lambda_i)) \geq \frac{1}{2} \boldsymbol{x}_{i+1}^\top \Lambda_i^{-1} \boldsymbol{x}_{i+1},$$

which implies that

$$\sum_{i=1}^{n} \boldsymbol{x}_i^\top \Lambda_{i-1}^{-1} \boldsymbol{x}_i \leq 2 \ln\left(\frac{\det(\Lambda_n)}{\det(\Lambda_0)}\right).$$

□

B THE REGRET ANALYSIS: PROOF OF THEOREM 2

Given T and d, we define $\tilde{d} := d \ln(TKd/\delta)$. In the case $d \leq T \leq d \ln(K)$, the regret bound is exactly $\Theta(T)$, and in the case $d \ln(K) < T \leq \tilde{d}$, the regret lower bound is $\Omega(d \ln(K))$ and the upper bound is at most $O(T) \leq O(d \ln(KTd/\delta))$. Therefore, in the case $d \leq T < \tilde{d}$, the trivial upper bound $O(T)$ is optimal up to logarithmic factors in T and d. Below we assume $T \geq \tilde{d}$.

We first define the following desired event where all the confidence intervals contains their estimation targets.

$$E_k := \left\{ \boldsymbol{x}^\top \boldsymbol{\theta} \in [\boldsymbol{x}^\top \hat{\boldsymbol{\theta}}_k \pm \alpha \sqrt{\boldsymbol{x}^\top \Lambda_{k-1}^{-1} \boldsymbol{x}}], \forall \boldsymbol{x} \in X_t, \forall t \in \{1, 2, \ldots, T\} \right\}, \forall k \in \{1, 2, \ldots, M\}$$

We then let $E = \cap_{k=1}^M E_k$ be the whole confidence event.

By the analysis of the ridge regression (Lemma 15, stated and proved in Appendix B.5) and the fact that $\alpha \geq \sqrt{\ln(KTd/\delta)} + \lambda\sqrt{d}$, via a union bound we have that $\Pr[E] \geq 1 - MT\delta$. When E holds, we know that the optimal arm at any time step will never be eliminated by the elimination procedure $\mathcal{E}$. Let $\mathbb{I}[E]$ be the indicator variable which takes value 1 when E holds and value 0 otherwise. We have the following upper bound for the expected regret of the algorithm.

$$R_T \leq \sum_{k=1}^M \sum_{t=\mathcal{T}_{k-1}+1}^{\mathcal{T}_k} \mathbb{E}\left[\mathbb{I}[E]\left(\max_{\boldsymbol{x}\in X_t}\{\boldsymbol{x}^\top\boldsymbol{\theta}\} - \boldsymbol{y}_t^\top\boldsymbol{\theta}\right)\right] + 2T\Pr[\overline{E}]$$

$$\leq \sum_{k=1}^M \sum_{t=\mathcal{T}_{k-1}+1}^{\mathcal{T}_k} \mathbb{E}\left[\mathbb{I}[E]\left(\max_{\boldsymbol{x}\in X_t^{(k)}}\{\boldsymbol{x}^\top\boldsymbol{\theta}\} - \min_{\boldsymbol{x}\in X_t^{(k)}}\{\boldsymbol{x}^\top\boldsymbol{\theta}\}\right)\right] + 2MT^2\delta. \qquad (13)$$

Lemma 10. *For any time step t during any batch k ($k \geq 3$), define $X_t^{(k-1)} := \mathcal{E}(X_t, \{\Lambda_i, \hat{\boldsymbol{\theta}}_i\}_{i=1}^{k-2})$ and we have that*

$$\mathbb{E}\left[\mathbb{I}[E]\left(\max_{\boldsymbol{x}\in X_t^{(k)}}\{\boldsymbol{x}^\top\boldsymbol{\theta}\} - \min_{\boldsymbol{x}\in X_t^{(k)}}\{\boldsymbol{x}^\top\boldsymbol{\theta}\}\right)\right] \leq \mathbb{E}\left[\min\left\{4\alpha\max_{\boldsymbol{x}\in X_t^{(k)}}\sqrt{\boldsymbol{x}^\top\Lambda_{k-1}^{-1}\boldsymbol{x}}, 2\right\}\right]. \qquad (14)$$

Proof. When E happens, for any $\boldsymbol{u}, \boldsymbol{v} \in X_t^{(k)}$, $\boldsymbol{u}^\top\boldsymbol{\theta} \in [\boldsymbol{u}^\top\hat{\boldsymbol{\theta}}_{k-1} \pm \alpha\sqrt{\boldsymbol{u}^\top\Lambda_{k-1}^{-1}\boldsymbol{u}}]$ and $\boldsymbol{v}^\top\boldsymbol{\theta} \in [\boldsymbol{v}^\top\hat{\boldsymbol{\theta}}_{k-1} \pm \alpha\sqrt{\boldsymbol{v}^\top\Lambda_{k-1}^{-1}\boldsymbol{v}}]$. We then have that

$$\mathbb{I}[E]\left(\boldsymbol{u}^\top\boldsymbol{\theta} - \boldsymbol{v}^\top\boldsymbol{\theta}\right)$$

$$\leq \left(\boldsymbol{u}^\top\hat{\boldsymbol{\theta}}_{k-1} + \alpha\sqrt{\boldsymbol{u}^\top\Lambda_{k-1}^{-1}\boldsymbol{u}}\right) - \left(\boldsymbol{v}^\top\hat{\boldsymbol{\theta}}_{k-1} - \alpha\sqrt{\boldsymbol{v}^\top\Lambda_{k-1}^{-1}\boldsymbol{v}}\right)$$

$$= \left(\boldsymbol{u}^\top\hat{\boldsymbol{\theta}}_{k-1} - \alpha\sqrt{\boldsymbol{u}^\top\Lambda_{k-1}^{-1}\boldsymbol{u}}\right) - \left(\boldsymbol{v}^\top\hat{\boldsymbol{\theta}}_{k-1} + \alpha\sqrt{\boldsymbol{v}^\top\Lambda_{k-1}^{-1}\boldsymbol{v}}\right) + 2\alpha\sqrt{\boldsymbol{u}^\top\Lambda_{k-1}^{-1}\boldsymbol{u}} + 2\alpha\sqrt{\boldsymbol{v}^\top\Lambda_{k-1}^{-1}\boldsymbol{v}}$$

$$\leq 2\alpha\sqrt{\boldsymbol{u}^\top\Lambda_{k-1}^{-1}\boldsymbol{u}} + 2\alpha\sqrt{\boldsymbol{v}^\top\Lambda_{k-1}^{-1}\boldsymbol{v}} \qquad (15)$$

$$\leq 4\alpha\max_{\boldsymbol{x}\in X_t^{(k)}}\sqrt{\boldsymbol{x}^\top\Lambda_{k-1}^{-1}\boldsymbol{x}} \leq 4\alpha\max_{\boldsymbol{x}\in X_t^{(k-1)}}\sqrt{\boldsymbol{x}^\top\Lambda_{k-1}^{-1}\boldsymbol{x}},$$

where the equation 15 is because $\boldsymbol{v}$ survived from the elimination based on $(\Lambda_{k-1}, \hat{\boldsymbol{\theta}}_{k-1})$, and the last inequality is because $X_t^{(k)} \subseteq X_t^{(k-1)}$.

Letting $\boldsymbol{u} = \arg\max_{\boldsymbol{x}\in X_t^{(k)}}\{\boldsymbol{x}^\top\boldsymbol{\theta}\}$ and $\boldsymbol{v} = \arg\min_{\boldsymbol{x}\in X_t^{(k)}}\{\boldsymbol{x}^\top\boldsymbol{\theta}\}$ and noting that $\boldsymbol{u}^\top\boldsymbol{\theta}, \boldsymbol{v}^\top\boldsymbol{\theta} \in [-1, 1]$, we have that

$$\mathbb{I}[E]\left(\max_{\boldsymbol{x}\in X_t^{(k)}}\{\boldsymbol{x}^\top\boldsymbol{\theta}\} - \min_{\boldsymbol{x}\in X_t^{(k)}}\{\boldsymbol{x}^\top\boldsymbol{\theta}\}\right) \leq \min\left\{4\alpha\max_{\boldsymbol{x}\in X_t^{(k-1)}}\sqrt{\boldsymbol{x}^\top\Lambda_{k-1}^{-1}\boldsymbol{x}}, 2\right\}. \qquad (16)$$

Taking the expectation over equation 16, we prove the lemma. $\qquad\square$

B.1 REGRET IN THE FIRST AND SECOND BATCHES

The regret in the first batch is bounded by T_1 trivially. For the second batch, we have the lemma below.

Lemma 11. *With probability $1 - \delta$, it holds that*

$$\mathbb{E}_{X\sim D}\left[\max_{\boldsymbol{x}\in\mathcal{E}(X;(\Lambda_1,\hat{\boldsymbol{\theta}}_1))}\sqrt{\min\{\boldsymbol{x}^\top\Lambda_1^{-1}\boldsymbol{x}, 2\}}\right] \leq O\left(\sqrt{\frac{1}{\mathcal{T}_1}\cdot\left(d\ln\left(\frac{\mathcal{T}_1}{\lambda}\right) + \ln(\mathcal{T}_1/\delta)\right)}\right).$$

Proof. Recall the definition of $\boldsymbol{y}_t$ and r_t in line 1, Algorithm 1. Let W_t denote $\lambda \mathbf{I} + \sum_{\tau=1}^{t} \boldsymbol{y}_\tau \boldsymbol{y}_\tau^\top$ for $0 \le t \le \mathcal{T}_1/2$. By the Elliptical Potential Lemma (Lemma 9, stated and proved in Appendix A.1), we have that

$$\sum_{t=1}^{\mathcal{T}_1/2} \min\{\boldsymbol{y}_t^\top W_{t-1}^{-1} \boldsymbol{y}_t, 1\} \le O\left(d \ln\left(\frac{\mathcal{T}_1}{\lambda}\right)\right).$$

By Corollary 22 with $\epsilon = \frac{1}{2}$, we further have that with probability $1 - \delta$, it holds that

$$\sum_{t=1}^{\mathcal{T}_1/2} \mathbb{E}_{X \sim D} \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)} \min\{\boldsymbol{y}^\top W_{t-1}^{-1} \boldsymbol{y}, 1\} \le 2 \sum_{t=1}^{\mathcal{T}_1/2} \min\{\boldsymbol{y}_t^\top W_{t-1}^{-1} \boldsymbol{y}_t, 1\} + 56 \ln(\mathcal{T}_1/\delta)$$

$$\le O\left(d \ln\left(\frac{\mathcal{T}_1}{\lambda}\right) + \ln(\mathcal{T}_1/\delta)\right). \tag{17}$$

Noting that $W_{t-1} \preccurlyeq W_t$ for $1 \le t \le \mathcal{T}_1$ and $\Lambda_1 = W_{\mathcal{T}_1/2}$, by equation 17 we have that

$$\mathcal{T}_1 \mathbb{E}_{X \sim D} \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)} \min\{\boldsymbol{y}^\top \Lambda_1^{-1} \boldsymbol{y}, 1\} \le O\left(d \ln\left(\frac{\mathcal{T}_1}{\lambda}\right) + \ln(\mathcal{T}_1/\delta)\right). \tag{18}$$

Lemma 12. *For any PSD matrix W and context X, we have that*

$$\min\{\max_{\boldsymbol{x} \in X}\{\boldsymbol{x}^\top W^{-1} \boldsymbol{x}\}, 1\} \le \min\{K, d\} \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)} \min\{\boldsymbol{y}^\top W^{-1} \boldsymbol{y}, 1\}. \tag{19}$$

Proof. By Lemma 6, we have that

$$\max_{\boldsymbol{x} \in X}\{\boldsymbol{x}^\top W^{-1} \boldsymbol{x}\} \le \min\{K, d\} \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)} \boldsymbol{y}^\top W^{-1} \boldsymbol{y}. \tag{20}$$

In the case $\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top W^{-1} \boldsymbol{x} \le 1$, we have that $\boldsymbol{y}^\top W^{-1} \boldsymbol{y} \le 1$ for any $\boldsymbol{y} \in X$. It then holds that

$$\max_{\boldsymbol{x} \in X} \min\{\boldsymbol{x}^\top W^{-1} \boldsymbol{x}, 1\} \le \min\{K, d\} \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)} \min\{\boldsymbol{y}^\top W^{-1} \boldsymbol{y}, 1\}.$$

In the case $\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top W^{-1} \boldsymbol{x} > 1$, we analyze as below. When $\min\{K, d\} \mathrm{Pr}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)}[\boldsymbol{y}^\top W^{-1} \boldsymbol{y} > 1] \ge 1$, it is trivial that

$$\min\{K, d\} \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)} \min\{\boldsymbol{y}^\top W^{-1} \boldsymbol{y}, 1\} \ge 1.$$

Otherwise, we have that

$$\min\{K, d\} \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)} \min\{\boldsymbol{y}^\top W^{-1} \boldsymbol{y}, 1\}$$

$$\ge \min\{K, d\} \mathbb{E}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)} \boldsymbol{y}^\top W^{-1} \boldsymbol{y} - \min\{K, d\} \mathrm{Pr}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)}[\boldsymbol{y}^\top W^{-1} \boldsymbol{y} > 1](\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top W^{-1} \boldsymbol{x} - 1)$$

$$\ge \max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top W^{-1} \boldsymbol{x} - \min\{K, d\} \mathrm{Pr}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)}[\boldsymbol{y}^\top W^{-1} \boldsymbol{y} > 1](\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top W^{-1} \boldsymbol{x} - 1) \tag{21}$$

$$= 1 + (1 - \min\{K, d\} \mathrm{Pr}_{\boldsymbol{y} \sim \pi^{\mathsf{G}}(X)}[\boldsymbol{y}^\top W^{-1} \boldsymbol{y} > 1])(\max_{\boldsymbol{x} \in X} \boldsymbol{x}^\top W^{-1} \boldsymbol{x} - 1)$$

$$\ge 1.$$

Here equation 21 holds by Lemma 6. The lemma is proved. $\square$

Setting $W = \Lambda_1$ in Lemma 12, we have that

$$\mathbb{E}_{X \sim D} \max_{\boldsymbol{x} \in X} \min\{\boldsymbol{x}^\top \Lambda_1^{-1} \boldsymbol{x}, 1\} \le \frac{\min\{K, d\}}{\mathcal{T}_1} \cdot O\left(d \ln\left(\frac{\mathcal{T}_1}{\lambda}\right) + \ln(\mathcal{T}_1/\delta)\right). \tag{22}$$

Therefore,

$$\mathbb{E}_{X \sim D} \left[\max_{\boldsymbol{x} \in \mathcal{E}(X; (\Lambda_1, \hat{\boldsymbol{\theta}}_1))} \sqrt{\min\{\boldsymbol{x}^\top \Lambda_1^{-1} \boldsymbol{x}, 1\}} \right]$$

$$\le \sqrt{\mathbb{E}_{X \sim D} \left[\max_{\boldsymbol{x} \in X} \min\{\boldsymbol{x}^\top \Lambda_1^{-1} \boldsymbol{x}, 1\} \right]} \le O\left(\sqrt{\frac{\min\{K, d\}}{\mathcal{T}_1} \cdot \left(d \ln\left(\frac{\mathcal{T}_1}{\lambda}\right) + \ln(\mathcal{T}_1/\delta)\right)} \right).$$

we finish the proof of Lemma 11. $\square$

By Lemma 10 and 11, the regret in the second batch is bounded by

$$O\left(T_2\alpha\cdot\sqrt{\frac{\min\{K,d\}d}{T_1}\ln\left(\frac{T}{\lambda\delta}\right)}\right)+T\delta. \tag{23}$$

B.2 Regret in General Batches

Let

$$L:=\frac{1}{200\ln(Td/\delta)}\qquad\text{and}\qquad\kappa:=\frac{1}{T^2} \tag{24}$$

be the parameters to be used by $\texttt{Exp}-\texttt{Policy}$.

For the k-th batch ($k\geq 3$), we invoke Lemma 5 with $m=T_{k-2}/2$, $\{Z_u\}_{u=1}^m=\{X_t^{(k-1)}\}_{t=\mathcal{T}_{k-3}+T_{k-2}/2+1}^{\mathcal{T}_{k-2}}$, $n=T_{k-1}/2$, $\{\tilde{Z}_u\}_{u=1}^n=\{X_t^{(k-1)}\}_{t=\mathcal{T}_{k-2}+1}^{\mathcal{T}_{k-2}+T_{k-1}/2}$, and $D=D_{k-1}$, we have that with probability $(1-3\delta)$, it holds that

$$\mathbb{E}_{X\sim D_{k-1}}\left[\min\left\{\sqrt{\max_{\boldsymbol{x}\in X}\boldsymbol{x}^\top\left(\sum_{t=\mathcal{T}_{k-2}+1}^{\mathcal{T}_{k-2}+T_{k-1}/2}\boldsymbol{y}_t\boldsymbol{y}_t^\top+\frac{T_{k-1}}{T_{k-2}}\kappa\mathbf{I}\right)^{-1}\boldsymbol{x}},\sqrt{L}\right\}\right]$$
$$\leq\sqrt{\frac{1}{T_{k-1}}}\cdot O\left(\sqrt{d\ln\left(\frac{T_{k-2}d}{\kappa}\right)}\right)+\frac{\sqrt{L}}{T_{k-2}}\cdot O\left(d\ln\left(\frac{T_{k-2}d}{\kappa}\right)\right). \tag{25}$$

Note that $\frac{T_{k-1}}{T_{k-2}}\kappa\leq\lambda$ and therefore $(\sum_{t=\mathcal{T}_{k-2}+1}^{\mathcal{T}_{k-2}+T_{k-1}/2}\boldsymbol{y}_t\boldsymbol{y}_t^\top+\frac{T_{k-1}}{T_{k-2}}\kappa\mathbf{I})^{-1}\succcurlyeq\Lambda_{k-1}^{-1}$. Therefore equation 25 implies that with probability $(1-3\delta)$,

$$\mathbb{E}_{X\sim D_{k-1}}\left[\min\left\{\sqrt{\max_{\boldsymbol{x}\in X}\boldsymbol{x}^\top\Lambda_{k-1}^{-1}\boldsymbol{x}},\sqrt{L}\right\}\right]\leq\sqrt{\frac{1}{T_{k-1}}}\cdot O\left(\sqrt{d\ln\left(Td\right)}\right)+\frac{\sqrt{L}}{T_{k-2}}\cdot O\left(d\ln\left(Td\right)\right). \tag{26}$$

In other words, for each time step t during the k-th batch, we have that

$$\mathbb{E}\left[\min\left\{\sqrt{\max_{\boldsymbol{x}\in X_t^{(k-1)}}\boldsymbol{x}^\top\Lambda_{k-1}^{-1}\boldsymbol{x}},\sqrt{L}\right\}\right]\leq\sqrt{\frac{1}{T_{k-1}}}\cdot O\left(\sqrt{d\ln\left(Td\right)}\right)+\frac{\sqrt{L}}{T_{k-2}}\cdot O\left(d\ln\left(Td\right)\right)+3\delta. \tag{27}$$

B.3 PUTTING ALL TOGETHER

Combining equation 13, Lemma 10 and Lemma 11, the total regret is bounded by

$$R_T \leq 2T_1 + 2T_2 \cdot \min\left\{\alpha\sqrt{\frac{2\min\{K,d\}d\ln\left(\frac{T}{\lambda\delta}\right)}{T_1}}, 1\right\}$$

$$+ \sum_{k=3}^{M} T_k \mathbb{E}\left[\min\left\{4\alpha \max_{\boldsymbol{x}\in X_t^{(k-1)}} \sqrt{\boldsymbol{x}^\top \Lambda_{k-1}^{-1}\boldsymbol{x}}, 2\right\}\right] + 2MT^2\delta$$

$$\leq 2T_1 + 2T_2 \cdot \min\left\{\alpha\sqrt{\frac{2\min\{K,d\}d\ln\left(\frac{T}{\lambda\delta}\right)}{T_1}}, 1\right\}$$

$$+ \sum_{k=3}^{M} 4\alpha T_k \mathbb{E}\left[\min\left\{\max_{\boldsymbol{x}\in X_t^{(k-1)}} \sqrt{\boldsymbol{x}^\top \Lambda_{k-1}^{-1}\boldsymbol{x}}, \sqrt{L}\right\}\right] + 2MT^2\delta \tag{28}$$

$$\leq 2T_1 + 2T_2 \cdot \min\left\{\alpha\sqrt{\frac{2\min\{K,d\}d\ln\left(\frac{T}{\lambda\delta}\right)}{T_1}}, 1\right\}$$

$$+ \sum_{k=3}^{M} 4\alpha T_k \mathbb{E}\left[\min\left\{\max_{\boldsymbol{x}\in X_t^{(k-1)}} \sqrt{\boldsymbol{x}^\top \Lambda_{k-1}^{-1}\boldsymbol{x}}, \sqrt{L}\right\}\right] + 2MT^2\delta, \tag{29}$$

Here equation 28 is because of $2 \leq 4\alpha\sqrt{L}$.

Let $\tilde{d} := d\ln(TdK/\delta) \cdot \ln\left(\frac{T}{\lambda\delta}\right)$. Combining equation 29 and equation 27 (for $k \geq 3$), we have that

$$R_T \leq 2T_1 + 2T_2 \cdot \min\left\{\alpha\sqrt{\frac{2\min\{K,d\}d\ln\left(\frac{T}{\lambda\delta}\right)}{T_1}}, 1\right\}$$

$$+ 4\alpha\sum_{k=3}^{M}\left(\frac{T_k}{\sqrt{T_{k-1}}} \cdot O\left(\sqrt{d\ln(Td)}\right) + \frac{T_k\sqrt{L}}{T_{k-2}} \cdot O\left(d\ln(Td)\right)\right) + 3T\delta + 2MT^2\delta$$

$$\leq 2T_1 + 2T_2 \cdot \min\left\{\alpha\sqrt{\frac{2\min\{K,d\}d\ln\left(\frac{T}{\lambda\delta}\right)}{T_1}}, 1\right\} + O(\sqrt{\ln^2(Td/\delta) + \ln(K)\ln(Td/\delta)}) \times \sum_{k=3}^{M}\frac{T_k\sqrt{d}}{\sqrt{T_{k-1}}}$$

$$+ O\left(\sqrt{\ln(K)\ln^2(Td) + \ln^3(Td)}\right) \times \sum_{k=3}^{M}\frac{T_k d}{T_{k-2}} + O(MT^2\delta)$$

$$\leq 2T_1 + 2T_2 \cdot \min\left\{\sqrt{\frac{2\min\{K,d\}\tilde{d}}{T_1}}, 1\right\} + O(\ln(Td)) \cdot \left(\sum_{k=3}^{M}\frac{T_k\sqrt{\tilde{d}}}{\sqrt{T_{k-1}}} + \sum_{k=3}^{M}\frac{T_k\tilde{d}}{T_{k-2}}\right) + O(MT^2\delta). \tag{30}$$

When T is small (i.e., $d \leq T < \tilde{d}$), Theorem 2 trivially holds because the regret is at most $O(\tilde{d})$, which is further bounded by

$$O\left(\text{poly}\ln(Td)\min\left\{T^{\frac{1}{2-2^{-M+2}}}(d\ln(K))^{\frac{1-2^{-M+2}}{2-2^{-M+2}}}, T^{\frac{1}{2-2^{-M+1}}}(d\ln(K))^{\frac{1-2^{-M+1}}{2-2^{-M+1}}}\min\{K,d\}^{\frac{2^{-M+1}}{2-2^{-M+1}}}\right\}\right).$$

Suppose $T \geq \tilde{d}$. Let $h = \min\{K,d\}$. We discuss the following two cases to design the batch schedule.

Case I: $\tilde{d} \le T \le \tilde{d}h^{2-2^{-M+2}}$. In this case , we define $\gamma := T^{\frac{1}{2-2^{-M+2}}} \tilde{d}^{\frac{1-2^{-M+2}}{2-2^{-M+2}}} \ge \tilde{d}$.

We let

$$T_1 = \gamma, \quad T_2 = \gamma, \quad T_k = \gamma \frac{\sqrt{T_{k-1}}}{\sqrt{\tilde{d}}}, \forall 3 \le k \le M. \tag{31}$$

Then for $3 \le k \le M$, by the iteration rule we have that $T_k = \gamma^{2-2^{-k+2}} \tilde{d}^{-1+2^{-k+2}}$. It is easy to verify that $\sum_{k=1}^M T_k \ge T_M = T$.

Now we verify that the regret for each batch is bounded by $O(\gamma)$. Firstly we have that $T_1, T_2 \le \gamma$. For $k = 3$, we have that $\frac{T_3\sqrt{\tilde{d}}}{\sqrt{T_2}} = \gamma$ and $\frac{T_3\tilde{d}}{T_1} = \gamma^{\frac{1}{2}}\tilde{d}^{\frac{1}{2}} \le \gamma$. For $4 \le k \le M$, noting that $\tilde{d} \le \gamma$, we have that

$$\frac{T_k\sqrt{\tilde{d}}}{\sqrt{T_{k-1}}} = \gamma, \qquad \frac{T_k d}{T_{k-2}} = \gamma^{2^{-k}-2^{-k+2}} \tilde{d}^{1-2^{-k}+2^{-k+2}} \le \gamma. \tag{32}$$

Therefore, the total regret in this case is bounded by

$$O\left(T^{\frac{1}{2-2^{-M+2}}} \tilde{d}^{\frac{1-2^{-M+2}}{2-2^{-M+2}}} \log\log(T) + MT^2\delta\right). \tag{33}$$

Case II: $T > \tilde{d}h^{2-2^{-M+2}}$. In this case, we define $\gamma := T^{\frac{1}{2-2^{-M+1}}} \tilde{d}^{\frac{1-2^{-M+1}}{2-2^{-M+1}}} h^{\frac{2^{-M+1}}{2-2^{-M+1}}} \ge \tilde{d}$. Let

$$T_1 = \gamma, \quad T_2 = \gamma\frac{\sqrt{T_1}}{\sqrt{\tilde{d}h}}, \quad T_k = \gamma\frac{\sqrt{T_{k-1}}}{\sqrt{\tilde{d}}}, \forall 3 \le k \le M. \tag{34}$$

By the iteration rule, we have that $T_k = \gamma^{2-2^{-k+1}} \tilde{d}^{-1+2^{-k+1}} h^{-2^{-k+1}}$. In particular, $\sum_{k=1}^M T_k \ge T_M = \gamma^{2-2^{-M+1}} \tilde{d}^{-1+2^{-M+1}} h^{-2^{-M+1}} = T$.

By definition, we have that $T_1 \le \gamma$ and $T_2\frac{\sqrt{\tilde{d}h}}{\sqrt{T_1}} \le \gamma$. For $k = 3$, we have that

$$\frac{T_3\sqrt{\tilde{d}}}{\sqrt{T_2}} = \gamma, \qquad \frac{T_3\tilde{d}}{T_1} = \gamma^{\frac{3}{4}}\tilde{d}^{\frac{1}{4}}h^{-\frac{1}{4}} \le \gamma.$$

For $4 \le k \le M$, we have that

$$\frac{T_k\sqrt{\tilde{d}}}{\sqrt{T_{k-1}}} = \gamma, \qquad \frac{T_k\tilde{d}}{T_{k-2}} = \gamma^{2^{-k+3}-2^{-k+1}} \tilde{d}^{1+2^{-k+1}-2^{-k+3}} h^{2^{-k+3}-2^{-k+1}} \le \tilde{d}\left(\frac{\gamma h}{\tilde{d}}\right)^{\frac{3}{8}}. \tag{35}$$

So it suffices to verify $\tilde{d}h^{\frac{3}{5}} \le \gamma$. In fact we have that

$$\gamma^{2-2^{-M+1}} = T\tilde{d}^{1-2^{-M+1}} h^{2^{-M+1}} \ge \tilde{d}^{2-2^{-M+1}} h^{2-2^{-M+2}+2^{-M+1}} \ge \tilde{d}^{2-2^{-M+1}} h^{\frac{3}{5}(2-2^{-M+1})}, \tag{36}$$

which implies that $\tilde{d}h^{\frac{3}{5}} \le \gamma$.

Therefore, the total regret in this case is bounded by

$$O\left(T^{\frac{1}{2-2^{-M+1}}} \tilde{d}^{\frac{1-2^{-M+1}}{2-2^{-M+1}}} h^{\frac{2^{-M+1}}{2-2^{-M+1}}} \log\log(T) + MT^2\delta\right). \tag{37}$$

We now finish the discuss about the two cases and combine the two regret upper bounds equation 33 and equation 37. Noting that $T > \tilde{d}h^{2-2^{-M+2}}$ implies that

$$T^{\frac{1}{2-2^{-M+2}}} \tilde{d}^{\frac{1-2^{-M+2}}{2-2^{-M+2}}} > T^{\frac{1}{2-2^{-M+1}}} \tilde{d}^{\frac{1-2^{-M+1}}{2-2^{-M+1}}} h^{\frac{2^{-M+1}}{2-2^{-M+1}}}, \tag{38}$$

18

we have that

$$R_T \leq O\left(\min\left\{T^{\frac{1}{2-2^{-M+2}}}\tilde{d}^{\frac{1-2^{-M+2}}{2-2^{-M+2}}}, T^{\frac{1}{2-2^{-M+1}}}\tilde{d}^{\frac{1-2^{-M+1}}{2-2^{-M+1}}}h^{\frac{2^{-M+1}}{2-2^{-M+1}}}\right\} \cdot \log\log(T) + MT^2\delta\right).$$

Setting $\delta = 1/T^3$, we obtain that

$$R_T \leq O\Bigg(\operatorname{poly}\ln(Td)\min\Big\{T^{\frac{1}{2-2^{-M+2}}}(d\ln(K))^{\frac{1-2^{-M+2}}{2-2^{-M+2}}},$$

$$T^{\frac{1}{2-2^{-M+1}}}(d\ln(K))^{\frac{1-2^{-M+1}}{2-2^{-M+1}}}\min\{K,d\}^{\frac{2^{-M+1}}{2-2^{-M+1}}}\Big\}\Bigg).$$

Theorem 2 is proven.

B.4 ANALYSIS: PROOF OF LEMMA 5

Lemma 5 (restated). *Let $\{Z_u\}_{u=1}^m$ be m i.i.d. stochastic context sets following a distribution D. Let π be the output by running $\mathtt{Exp-Policy}$ with the input $\{Z_u\}_{u=1}^m$. Let $\{\tilde{Z}_u\}_{u=1}^n$ be another group of i.i.d. stochastic context sets following the distribution D (which is also independent from $\{Z_u\}_{u=1}^m$). Let $\boldsymbol{y}_u$ be independently sampled from $\pi(\tilde{Z}_u)$ for each $u \in \{1,2,\ldots,n\}$, and let $\Lambda = \sum_{u=1}^n \boldsymbol{y}_u\boldsymbol{y}_u^\top$. With probability $(1-3\delta)$, we have that*

$$\mathbb{E}_{X\sim D}\left[\min\left\{\sqrt{\max_{\boldsymbol{x}\in X}\boldsymbol{x}^\top(\Lambda + \frac{n}{m}\kappa\mathbf{I})^{-1}\boldsymbol{x}}, \sqrt{L}\right\}\right] \leq \sqrt{\frac{1}{n}}\cdot O\left(\sqrt{d\ln\left(\frac{md}{\kappa}\right)}\right) + \frac{\sqrt{L}}{m}\cdot O\left(d\ln\left(\frac{md}{\kappa}\right)\right). \tag{8}$$

As stated in Section 3, the proof of Lemma 5 utilizes similar ideas in the proof of Theorem 5 in (Ruan et al., 2021). The major difference is that their information matrix starts with $\Omega(1)\cdot\mathbf{I}$ when executing the output policy, while our information matrix could start with $\kappa\mathbf{I}$ with $\kappa = \frac{1}{T^2}$. As a result, it is harder for us to recover the information matrix U_m.

Let D be the distribution defined in the statement of Lemma 5, we first prove the following lemma.

Lemma 13. *With probability $(1-\delta)$, it holds that*

$$m\mathbb{E}_{X\sim D}\left[\min\{\max_{\boldsymbol{x}\in X}\boldsymbol{x}^\top U_m^{-1}\boldsymbol{x}, L\}\right] \leq O(d\ln(md/\kappa)). \tag{39}$$

Proof. Let η_i denote the index η such that $i \in \tau_\eta$. Note that $i \in \tau_\eta$ implies that $\det(U_{i-1}) \leq 2\det(W_\eta)$, which further implies that $\det((W_\eta^{-\frac{1}{2}})^\top U_{i-1}(W_\eta)^{\frac{1}{2}}) \leq 2$. Because $U_{i-1} \succcurlyeq W_\eta$, we have that $(W_\eta^{-\frac{1}{2}})^\top U_{i-1}(W_\eta)^{\frac{1}{2}} \succcurlyeq \mathbf{I}$. Therefore, the maximal eigenvalue of $(W_\eta^{-\frac{1}{2}})^\top U_{i-1}(W_\eta)^{\frac{1}{2}} \succcurlyeq \mathbf{I}$ is at most 2, where it follows that $U_{i-1} \leq 2W_\eta$.

Since $U_m \succcurlyeq W_\eta$ for all $\eta \geq 1$, we have that

$$m\mathbb{E}_{X\sim D}\left[\min\{\max_{\boldsymbol{x}\in X}\boldsymbol{x}^\top U_m^{-1}\boldsymbol{x}, L\}\right] \leq \sum_{i=1}^m \mathbb{E}_{X\sim D}\left[\min\{\max_{\boldsymbol{x}\in X}\boldsymbol{x}^\top W_{\eta_i}^{-1}\boldsymbol{x}, L\}\right]. \tag{40}$$

Invoking Corollary 22 with $\epsilon = 1$ and noting that $\mathbb{E}[\min\{\tilde{\boldsymbol{z}}_i^\top W_{\eta_i}^{-1}\tilde{\boldsymbol{z}}_i, L\}] = \mathbb{E}_{X\sim D}\left[\min\{\max_{\boldsymbol{x}\in X}\boldsymbol{x}^\top W_{\eta_i}^{-1}\boldsymbol{x}, L\}\right]$ when conditioned on the first $(i-1)$ iterations in Algorithm 2, we have with probability $(1-\delta)$, it holds that

$$\sum_{i=1}^m \mathbb{E}_{X\sim D}\left[\min\{\max_{\boldsymbol{x}\in X}\boldsymbol{x}^\top W_{\eta_i}^{-1}\boldsymbol{x}, L\}\right] \leq 2\sum_{i=1}^m \min\{\boldsymbol{z}_i^\top W_{\eta_i}^{-1}\boldsymbol{z}_i, L\} + 20L\ln(1/\delta). \tag{41}$$

By the definition of $\tilde{z}_i$, we further have that

$$\sum_{i=1}^m \min\{z_i^\top W_{\eta_i}^{-1} z_i, L\} = \sum_{i=1}^m \tilde{z}_i^\top W_{\eta_i}^{-1} \tilde{z}_i \le 2 \sum_{i=1}^m \tilde{z}_i^\top U_{i-1}^{-1} \tilde{z}_i \le O(d \ln(md/\kappa)), \quad (42)$$

where the second last inequality is by the fact that $U_{i-1} \preccurlyeq 2W_{\eta_i}$, and the last inequality is by a direct application of the Elliptical Potential Lemma (Lemma 9, stated and proved in Appendix A.1) and the fact that $0 < L < 1$. Combining equation 40, equation 41, equation 42 and the definition of L, we prove the lemma. $\qquad\square$

To proceed, we have the lemma below.

Lemma 14. *Define* $V := \mathbb{E}_{X \sim D, i \sim \pi(X)} \left[\min\left\{ \frac{L}{x_i^\top U_m^{-1} x_i}, 1 \right\} x_i x_i^\top \right]$. *With probability* $1 - \delta$,

$$V \succcurlyeq \frac{1}{6m} U_m - \frac{1}{3m} \kappa \mathbf{I}. \quad (43)$$

Proof. Recall that $\mathcal{F}_{m'} = \sigma(X_1, X_2, \ldots, X_{m'-1})$. By the definition of π, and noting that $U_m \succeq W_{\eta_u}$ implies that $\frac{L}{x_i^\top U_m^{-1} x_i} \ge \frac{L}{x_i^\top W_{\eta_u}^{-1} x_i}$ for $1 \le u \le m$, we have that

$$mV \succcurlyeq \sum_{u=1}^m \mathbb{E}\left[z_{u,i_u} z_{u,i_u}^\top \middle| \mathcal{F}_u \right].$$

Noting that $z_{u,i_u}^\top W_{\eta_u}^{-1} z_{u,i_u} \le L$ implies $z_{u,i_u} z_{u,i_u}^\top \le LW_{\eta_u} \le LU_m$, by Lemma 20 with $Z_m = LU_m$ and $\epsilon = 2$, with probability $(1 - \delta)$, we have that

$$mV \succcurlyeq \mathbb{E}\left[z_{u,i_u} z_{u,i_u}^\top \middle| \mathcal{F}_u \right] \succcurlyeq \frac{1}{3} \sum_{u=1}^m z_{u,i_u} z_{u,i_u}^\top - \frac{100 L \ln(Td/\delta)}{3} U_m$$

$$= \frac{1}{3} \sum_{u=1}^m z_{u,i_u} z_{u,i_u}^\top - \frac{1}{6} U_m = \frac{1}{6} U_m - \frac{1}{3} \kappa \mathbf{I},$$

and the conclusion follows by dividing m on both sides of the inequality. $\qquad\square$

Recall that $\Lambda = \sum_{u=1}^n y_u y_u^\top$. Let

$$\tilde{\Lambda} := \sum_{u=1}^n \min\left\{ \frac{L}{y_u^\top U_m^{-1} y_u}, 1 \right\} y_u y_u^\top.$$

By definition we have that $\Lambda \succcurlyeq \tilde{\Lambda}$.

Noting that for any $1 \le u \le n$,

$$\mathbb{E}\left[\min\left\{ \frac{L}{y_u^\top U_m^{-1} y_u}, 1 \right\} y_u y_u^\top \right] = \mathbb{E}_{X \sim D, i \sim \pi(X)} \left[\min\left\{ \frac{L}{x_i^\top U_m^{-1} x_i}, 1 \right\} x_i x_i^\top \right] = V$$

and $\min\left\{ \frac{L}{y_u^\top U_m^{-1} y_u}, 1 \right\} y_u y_u^\top \preccurlyeq LU_m$, by Corollary 22 with $W = LU_m$ and $\epsilon = \frac{2}{3}$, with probability $1 - \delta$,

$$\tilde{\Lambda} \succcurlyeq \frac{n}{3} V - \frac{68L}{3} U_m.$$

By Lemma 14, we further have that

$$\Lambda \succcurlyeq \tilde{\Lambda} \succcurlyeq \frac{n}{3} V - \frac{68L}{3} U_m \succcurlyeq \frac{n}{3} V - \frac{1}{6} U_m \succcurlyeq \frac{n}{3m} \left(\frac{1}{6} U_m - \frac{1}{3} \kappa \mathbf{I} \right) - \frac{1}{6} U_m = \frac{n}{36m} U_m - \frac{n}{9m} \kappa \mathbf{I}.$$

Therefore, $\Lambda + \frac{n}{m} \kappa \mathbf{I} \ge \frac{n}{36m} U_m$.

In the case $n \geq m$, we have that

$$\mathbb{E}_{X \sim D} \left[\min \left\{ \sqrt{\max_i \boldsymbol{x}_i^\top \left(\Lambda + \frac{n}{m} \kappa \mathbf{I} \right)^{-1} \boldsymbol{x}_i}, \sqrt{L} \right\} \right]$$

$$\leq \sqrt{\frac{36m}{n}} \mathbb{E}_{X \sim D} \left[\min \left\{ \sqrt{\max_i \boldsymbol{x}_i^\top U_m^{-1} \boldsymbol{x}_i}, \sqrt{Ln/m} \right\} \right]$$

$$\leq \sqrt{\frac{36m}{n}} \mathbb{E}_{X \sim D} \left[\min \left\{ \sqrt{\max_i \boldsymbol{x}_i^\top U_m^{-1} \boldsymbol{x}_i}, \sqrt{L} \right\} \right] + \sqrt{\frac{36m}{n}} \cdot \sqrt{\frac{Ln}{m}} \cdot \Pr_{X \sim D} \left[\max_i \boldsymbol{x}_i^\top U_m^{-1} \boldsymbol{x}_i \geq L \right]$$

$$\leq \sqrt{\frac{36m}{n}} \sqrt{\mathbb{E}_{X \sim D} \left[\min \left\{ \max_i \boldsymbol{x}_i^\top U_m^{-1} \boldsymbol{x}_i, L \right\} \right]} + \frac{\sqrt{L}}{m} \cdot O \left(d \ln \left(\frac{md}{\kappa} \right) \right)$$

$$\leq \sqrt{\frac{1}{n}} \cdot O \left(\sqrt{d \ln \left(\frac{md}{\kappa} \right)} \right) + \frac{\sqrt{L}}{m} \cdot O \left(d \ln \left(\frac{md}{\kappa} \right) \right).$$

Here the second last inequality and last inequality are by Lemma 13 and the fact that

$$\Pr_{X \sim D} \left[\max_i \boldsymbol{x}_i^\top U_m^{-1} \boldsymbol{x}_i \geq L \right] \leq \mathbb{E}_{X \sim D} \left[\min \{ \max_i x^\top U_m^{-1} x, L \} \right] \leq \frac{1}{m} \cdot O \left(d \ln \left(\frac{md}{\kappa} \right) \right).$$

In the case $n < m$, with similar arguments we have that

$$\mathbb{E}_{X \sim D} \left[\min \left\{ \sqrt{\max_i \boldsymbol{x}_i^\top \left(\Lambda + \frac{n}{m} \kappa \mathbf{I} \right)^{-1} \boldsymbol{x}_i}, \sqrt{L} \right\} \right]$$

$$\leq \sqrt{\frac{36m}{n}} \mathbb{E}_{X \sim D} \left[\min \left\{ \sqrt{\max_i \boldsymbol{x}_i^\top U_m^{-1} \boldsymbol{x}_i}, \sqrt{Ln/m} \right\} \right]$$

$$\leq \sqrt{\frac{36m}{n}} \mathbb{E}_{X \sim D} \left[\min \left\{ \sqrt{\max_i \boldsymbol{x}_i^\top U_m^{-1} \boldsymbol{x}_i}, \sqrt{L} \right\} \right]$$

$$\leq \sqrt{\frac{36m}{n}} \sqrt{\mathbb{E}_{X \sim D} \left[\min \left\{ \max_i \boldsymbol{x}_i^\top U_m^{-1} \boldsymbol{x}_i, L \right\} \right]}$$

$$\leq \sqrt{\frac{1}{n}} \cdot O \left(\sqrt{d \ln \left(\frac{md}{\kappa} \right)} \right)$$

$$\leq \sqrt{\frac{1}{n}} \cdot O \left(\sqrt{d \ln \left(\frac{md}{\kappa} \right)} \right) + \frac{\sqrt{L}}{m} \cdot O \left(d \ln \left(\frac{md}{\kappa} \right) \right).$$

The proof is completed.

B.5 STATEMENT AND PROOF OF LEMMA 15

The following lemma is a similar version of Lemma 31 in (Ruan et al., 2021) that analyzes the size of the confidence interval by the ridge regression. Note that since we only assume $\|\boldsymbol{\theta}\|_\infty \leq 1$ instead of the upper bound on the Euclidean norm, the calculation is slightly different.

Lemma 15. *Given $\boldsymbol{\theta}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ such that $\|\boldsymbol{\theta}\|_\infty \leq 1$ for all $i \in [n]$, let $r_i = \boldsymbol{x}_i^\top \boldsymbol{\theta} + \epsilon_i$ where $\{\epsilon_i\}_{i=1}^n$ are independent sub-Gaussian random variable with variance proxy 1. Let $\Lambda = \lambda \mathbf{I} + \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$ and $\hat{\boldsymbol{\theta}} = \Lambda^{-1} \sum_{i=1}^n r_i \boldsymbol{x}_i$. For any $\boldsymbol{x} \in \mathbb{R}^d$ and any $\gamma > 0$, we have that*

$$\Pr \left[|\boldsymbol{x}^\top (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})| > (\gamma + \sqrt{d\lambda}) \sqrt{\boldsymbol{x} \Lambda^{-1} \boldsymbol{x}} \right] \leq 2e^{-\frac{\gamma^2}{2}}.$$

Proof. Direct computation gives that

$$
\begin{aligned}
|\boldsymbol{x}^\top(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})| &= \left| \boldsymbol{x}^\top \left(\Lambda^{-1} \sum_{i=1}^n \boldsymbol{x}_i (\boldsymbol{x}_i^\top \boldsymbol{\theta} + \epsilon_i) - \boldsymbol{\theta} \right) \right| \\
&= \left| \boldsymbol{x}^\top \left(\Lambda^{-1} \sum_{i=1}^n \boldsymbol{x}_i \epsilon_i + \Lambda^{-1}(\Lambda - \lambda\mathbf{I})\boldsymbol{\theta} - \boldsymbol{\theta} \right) \right| \\
&= \left| \boldsymbol{x}^\top \Lambda^{-1} \left(\sum_{i=1}^n \boldsymbol{x}_i \epsilon_i - \lambda\boldsymbol{\theta} \right) \right| \\
&\leq \lambda \left| \boldsymbol{x}^\top \Lambda^{-1} \boldsymbol{\theta} \right| + \left| \sum_{i=1}^n \boldsymbol{x}^\top \Lambda^{-1} \boldsymbol{x}_i \epsilon_i \right|.
\end{aligned}
$$

For the first term, by Cauchy-Schwarz and noting that $\Lambda \succcurlyeq \lambda\mathbf{I}$, we have that

$$
\lambda \left| \boldsymbol{x}^\top \Lambda^{-1} \boldsymbol{\theta} \right| \leq \lambda\sqrt{d} \cdot \sqrt{\boldsymbol{x}^\top \Lambda^{-2} \boldsymbol{x}} \leq \sqrt{\lambda d \boldsymbol{x}^\top \Lambda^{-1} \boldsymbol{x}}. \tag{44}
$$

For the second term, by sub-Gaussian concentration inequalities, we have that

$$
\Pr \left[\left| \sum_{i=1}^n \boldsymbol{x}^\top \Lambda^{-1} \boldsymbol{x}_i \epsilon_i \right| > \gamma \sqrt{\boldsymbol{x}^\top \Lambda^{-1} \boldsymbol{x}} \right] \leq 2e^{\frac{-\gamma^2}{2}}. \tag{45}
$$

The conclusion follows by combining equation 44 and equation 45. $\square$

C REGRET LOWER BOUND FOR CONTEXT-BLIND BATCH LEARNING

In this section, we prove Lemma 16, which implies the regret lower bound theorem (Theorem 4). We first sketch the high-level idea below.

High-level idea. To construct the lower bound instances, we first construct M mutually independent sub-problems with dimension d/M. At each time step, a uniform random sub-problem (i.e., its context set) is selected and presented to the learner. To analyze the regret performance of any batch learning algorithm, we divide the time horizon T into M consecutive stages with properly chosen lengths $\{T_k\}_{k=1}^M$ for each stage. We will show that, for each $k \in \{1, 2, \ldots, M\}$, during the k-th stage, if the learner does not start a new batch and update its policy, then a large regret would incur for the k-th sub-problem in this stage.

Our construction for the sub-problems is as follows. For each $k \in \{1, 3, 4, 5, \ldots, M\}$, we choose the sub-problem to be the hard instance for the K-armed linear contextual linear bandit problem with burn-in time $\tilde{\Theta}(T_{k-1})$. In other words, we construct the k-th sub-problem so that it is hard for the learner to find a good policy for the k-th problem during the first $\sum_{i \leq k-1} T_i$ time steps.

The most interesting sub-problem design is for $k = 2$, which takes the advantage of the context-blind setting to force the learner incur more regret. In this sub-problem, we consider a linear contextual bandit problem d/M arms (assuming that $K \geq d/M$). However, during each time step, only $(d/(2M) + 1)$ arms are presented to the learner. In particular, we set the first $d/(2M)$ arms to be the *frequent* arms that always appear in the context set. For the rest $d/(2M)$ *infrequent* arms, we choose one of them uniformly randomly and include the chosen one in the context set at each time step. The best exploration strategy for the learner is to play the infrequent arm with a higher probability, e.g., $\frac{1}{2}$ and play the frequent arms with probability $\frac{M}{d}$. However, due to the context-blind setting, the learner can not tell whether an arm is frequent or infrequent with insufficient information, and his best strategy is to play the arms with the same probability. As a result, the infrequent arms are insufficiently explored, which leads to an extra $\tilde{\Theta}(\sqrt{d})$ factor in the regret.

Lemma 16. *For any algorithm $\mathcal{G}$ with batch complexity M, assuming $T \geq d\log_2(K)$, the minimax regret is at least*

$$
\Omega \left(\frac{1}{\operatorname{poly}\ln(Td)} \min \left\{ \begin{aligned} &T^{\frac{1}{2-2^{-M+1}}} (d\log_2(K))^{\frac{1-2^{-M+1}}{2-2^{-M+1}}} \left(\frac{\min\{K,d\}}{\log_2(K)} \right)^{\frac{2^{-M+1}}{2-2^{-M+1}}} \\ &T^{\frac{1}{2-2^{-M+2}}} (d\log_2(K))^{\frac{1-2^{-M+2}}{2-2^{-M+2}}} \end{aligned} \right\} \right).
$$

Proof. Define $\bar{d} = d/M$ and $\tilde{h} = \frac{1}{2}\min\{K, \bar{d}\}$. Define $\check{d} = \frac{\bar{d}\log_2(K)}{M\ln(dM)} \leq d\log_2(K) \leq d$ and $\check{h} = \frac{\tilde{h}}{\log_2(K)}$. When $\check{d} \leq T \leq \check{d}\check{h}^{2-2^{-M+2}}$, define $\gamma := T^{\frac{1}{2-2^{-M+2}}}\check{d}^{\frac{1-2^{-M+2}}{d2-2^{-M+2}}}$. When $T > \check{h}^{2-2^{-M+2}}$, define $\gamma := T^{\frac{1}{2-2^{-M+1}}}\check{d}^{\frac{1-2^{-M+1}}{2-2^{-M+1}}}\check{h}^{\frac{2^{-M+1}}{2-2^{-M+1}}} \geq \check{d}$. Given γ, we further define $T_1 = \gamma$, $T_2 = \max\{\gamma, \gamma \cdot \sqrt{\frac{T_1\ln(dM)}{\bar{d}h}}\}$ and $T_k = \gamma \cdot \sqrt{\frac{MT_{k-1}\ln(dM)}{\bar{d}\log_2(K)}}$ for any $3 \leq k \leq M$. Let $\epsilon_1 = \frac{1}{100}$, $\epsilon_2 = \min\left\{\frac{1}{100}\sqrt{\frac{\bar{d}^2}{T_1\ln(dM)}}, 1\right\}$ and $\epsilon_k = \frac{1}{100}\sqrt{\frac{\bar{d}\log_2(K)}{MT_{k-1}\ln(dM)}}$ for $k = 3, 4, \ldots, M$.

Fix the algorithm $\mathcal{G}$. Below we assume the randomness of $\mathcal{G}$ is considered in the expectation operator $\mathbb{E}[\cdot]$ and probability operator $\Pr[\cdot]$. Let $\mathrm{Regret}_{\theta,D}(T)$ be the expected regret under $\mathcal{G}$ with hidden parameter as θ and context distribution as D. We aim to design D and θ such that $\mathrm{Regret}_{\theta,D}(T)$ is large enough.

Without loss of generality, we assume d/M is an integer. In the construction below, we divide $\mathbb{R}^d$ into M subspaces, where for each subspace we construct a hard case. We let $\mathbb{R}^d = \otimes_{i=1}^{M}\mathcal{U}^i$, where U^i is the subspace spanned by $\{\mathbf{e}_{(i-1)d/M+j}\}_{j=1}^{d/M}$. In each round the noise is set to be Gaussian with variance 1. The hidden parameter $\theta = \{\theta^i\}_{i=1}^{M}$ is chosen from the space $\mathcal{Y} = \otimes_{i=1}^{M}\mathcal{Y}^i$. The context distribution D is given by the average of $\{D^i\}_{i=1}^{M}$. That is, in the t-th round, the environment first sample $u_t \sim \mathrm{Uniform}(\{1, 2, \ldots, M\})$, and then sample X according to D_{u_t}. In words, we divide the original problem into M independent sub-problems. For the k-th sub-problem is d/M, the dimension is d/M, the parameter space is $\mathcal{Y}^k$ and the context distribution is D_k.

We now define $\mathcal{Y}^k$ and D_k. Let $\bar{d} = d/M$. We have two cases: Case I: $k = 2$; and Case II: $k = 1$ or $3 \leq k \leq M$. Below we respectively define the parameter space $\mathcal{Y}^k$ for θ^k and the context distribution D_k.

Case I: $k = 2$. Define $\mathcal{Y}^2 = \{-\epsilon_2, 0, \epsilon_2\}^{\bar{d}}$. Recall that $\tilde{h} = \frac{1}{2}\min\{K, \bar{d}\}$. For $\mathcal{V} \subset [2\tilde{h}]$ with $|\mathcal{V}| = \tilde{h}$, we define the context distribution $D_2(\mathcal{V})$ by letting $\Pr_{X \sim D_2(\mathcal{V})}[X = \{\mathbf{e}_j\}_{j \in \mathcal{V}} \cup \{\mathbf{e}_i\}] = \frac{1}{\bar{d}-\tilde{h}}$ for any $i \in [\bar{d}]/\mathcal{V}$. In words, the sub-problem for the second batch is a contextual bandit problem with $\tilde{h} + 1$ arms. Among the $h + 1$ arms, there are $\tilde{h}$ arms which appear in each round, and the left arms appears with equal probability as $\frac{1}{\bar{d}-\tilde{h}}$.

In the case $K \gg \tilde{h} + 1$, we simply repeat the first arm for $K - (\tilde{h} + 1)$ times to construct the K-armed linear bandit problem. Without loss of generality, we still use $D_2(\mathcal{V})$ to denote this context. For fixed $\mathcal{V}$ and $\xi \in \{-1, 1\}^{\bar{d}}$, we further define $\theta^2(\mathcal{V}, \xi)$ by setting $\theta_i^2(\mathcal{V}, \xi) = 0$ for $i \in \mathcal{V}$ and $\theta_i^2(\mathcal{V}, \xi) = \xi_i\epsilon_2$ for $i \in [\bar{d}]/\mathcal{V}$.

Since the sub-problem is a contextual bandit problem, we could view the context X as a subset of $[\bar{d}]$. Given a permutation $\sigma \in S_{\bar{d}}$ and a context $X \subset [\bar{d}]$, we define the context vector $\sigma(X) = \{\sigma(i)\}_{i \in X}$. With a slight abuse of notations, we use $D_2(\sigma)$ as the shorthand for $D_2(\sigma([1, 2, \ldots, \tilde{h}]))$.

Let the policy π_1 be the policy for the first batch. Clearly, π_1 is independent of the context distribution. Let the distribution of π_1 be Π_1. We then claim the following lemma.

Lemma 17.

$$\min_{\sigma \in S_{\bar{d}}} \mathbb{E}_{\pi_1 \sim \Pi_1}\mathbb{E}_{X \sim D(\sigma)}\left[\mathbb{I}\left[\pi_1(X) \notin \{\sigma(1), \sigma(2), \ldots, \sigma(\tilde{h})\}\right]\right] \leq \frac{1}{\tilde{h}+1}.$$

Proof. Let $\mathcal{A}$ be the set of subsets of $[\bar{d}]$ with size $\tilde{h} + 1$. Let $Z_i = \{1, 2, \ldots, \tilde{h}, \tilde{h} + i\}$ for $1 \leq i \leq \bar{d} - \tilde{h}$. Note that for any $1 \leq i \leq \bar{d} - \tilde{h}$ and $\sigma \in S_{\bar{d}}$, $\pi(\sigma(Z_i)) \notin \{\sigma(1), \sigma(2), \ldots, \sigma(\tilde{h})\}$

implies that $\pi(X) = \sigma(\tilde{h} + i)$. Then we have that

$$\sum_{\sigma \in S_{\bar{d}}} \frac{1}{\bar{d}!} \mathbb{E}_{\pi_1 \sim \Pi_1} \mathbb{E}_{X \sim D_2(\sigma)} \left[\mathbb{I} \left[\pi_1(X) \notin \{\sigma(1), \sigma(2), \dots, \sigma(\tilde{h})\} \right] \right]$$

$$= \frac{1}{\bar{d}!(\bar{d} - \tilde{h})} \sum_{X \in \mathcal{A}} \sum_{\sigma \in S_{\bar{d}}} \sum_{i=1}^{d_1} \mathbb{I}[\sigma^{-1}(X) = Z_i] \mathbb{E}_{\pi_1 \sim \Pi_1} \left[\mathbb{I} \left[\pi_1(X) = \sigma(\tilde{h} + i) \right] \right]$$

$$= \frac{1}{\bar{d}!(\bar{d} - \tilde{h})} \sum_{X \in \mathcal{A}} \mathbb{E}_{\pi_1 \sim \Pi_1} \left[\sum_{i=1}^{\bar{d} - \tilde{h}} \sum_{\sigma \in S_{\bar{d}}} \mathbb{I}[\sigma^{-1}(X) = Z_i, \pi_1(X) = \sigma(h + i)] \right]$$

$$= \frac{1}{\bar{d}!(\bar{d} - \tilde{h})} \sum_{X \in \mathcal{A}} \mathbb{E}_{\pi_1 \sim \Pi_1} \left[\sum_{i=1}^{d_1} \sum_{\sigma \in S_{\bar{d}}, x \in X} \mathbb{I}[\sigma^{-1}(X) = Z_i, \pi_1(X) = x, x = \sigma(\tilde{h} + i)] \right]$$

$$= \frac{1}{\bar{d}!(\bar{d} - \tilde{h})} \sum_{X \in \mathcal{A}, x \in X} \Pr_{\pi_1 \sim \Pi_1} \left[\pi_1(X) = x \right] \cdot \sum_{\sigma \in S_{\bar{d}}} \sum_{i=1}^{\bar{d} - \tilde{h}} \mathbb{I}[\sigma^{-1}(X) = Z_i, x = \sigma(\tilde{h} + i)]$$

$$= \frac{\tilde{h}!(\bar{d} - \tilde{h} - 1)!}{\bar{d}!} \sum_{X \in \mathcal{A}, x \in X} \Pr_{\pi_1 \sim \Pi_1} \left[\pi_1(X) = x \right]$$

$$= \frac{\tilde{h}!(\bar{d} - \tilde{h} - 1)!}{\bar{d}!} \sum_{X \in \mathcal{A}} 1$$

$$= \frac{\tilde{h}!(\bar{d} - \tilde{h} - 1)!}{\bar{d}!} |\mathcal{A}|$$

$$= \frac{1}{\tilde{h} + 1}. \tag{46}$$

Then the conclusion follows easily. $\qquad\square$

Without loss of generality, we suppose that the identical permutation satisfies the condition in Lemma 17, i.e.,

$$\mathbb{E}_{\pi_1 \sim \Pi_1} \mathbb{E}_{X \sim D_2([\tilde{h}])} \mathbb{I} \left[\pi_1(X) \notin \{1, 2, \dots, \tilde{h}\} \right] \leq \frac{1}{\tilde{h} + 1}. \tag{47}$$

Define $J = \{i \geq \tilde{h} + 1 | \mathbb{E}_{\pi_1 \sim \Pi_1} \mathbb{E}_{X \sim D} \mathbb{I}[\pi(X) = i] \leq \frac{2}{(\tilde{h}+1)(\bar{d}-\tilde{h})}\}$, then by equation 47 we have that $|J| \geq \frac{\bar{d}-\tilde{h}}{2}$. Without loss of generality, we assume $J = \{\tilde{h}+1, \tilde{h}+2, \dots, \tilde{h}+\ell\}$ where $\ell \geq \frac{\bar{d}-\tilde{h}}{2}$ is the size of J. Then the context distribution is fixed as $D_2 = D_2([\tilde{h}])$.

Case II: $k = 1$ **or** $3 \leq k \leq M$. We consider to construct the hard case for K-armed linear contextual bandit problem. We assume $a = \min\{\log_2(K), \bar{d}\}$ and $b = \bar{d}/a$ are both integers. Define $\mathcal{Y}^k = \epsilon_k \cdot \{-1, 1\}^{\bar{d}}$. The context distribution D^k is defined as the uniform distribution over $X^{k,i} = \frac{1}{a} \cdot \{\boldsymbol{x} : \boldsymbol{x}_i = 0, \forall i \notin [(i-1)a+1, ia], \boldsymbol{x}_i \in \{-1, 1\}, \forall i \in [(i-1)a+1, ia]\}$ for $1 \leq i \leq b$.

Now we start to analyze the minimax lower bound over all $\theta \in \mathcal{Y}$ and the context distribution D described above. Let $\{t_i\}_{i=1}^M$ be the time schedule by running the algorithm $\mathcal{G}$. Then there exists $1 \leq i \leq M$ such that $\sum_{j \leq i-1} t_j \leq \sum_{j \leq i-1} T_j < \sum_{j \leq i} T_j \leq \sum_{j \leq i} t_j$, where we define $T_0 = t_0 = 0$. Denote $\mathcal{E}_i$ be the event where i is the smallest number such that $\sum_{j \leq i} t_i \leq \sum_{j \leq i} T_i < \sum_{j \leq i+1} T_{i+1} \leq \sum_{j \leq i+1} t_{i+1}$. Then it holds that $\sum_{i=1}^M \mathbb{I}[\mathcal{E}_i] = 1$.

To proceed, we have the lemma as below.

Lemma 18. *Let π_k be the policy for the k-th batch and $\pi_{(t)}$ be the policy for the t-th step. Fix $j \in [\tilde{h} + 1, \tilde{h} + \ell]$. For $\theta \in \mathcal{Y}$, we define*

$$R_j^k(\theta) = \left\{ \begin{cases} \frac{1}{\bar{d}-\tilde{h}} \sum_{t=T_1+1}^{T_1+T_2} \mathbb{E}_{\theta, \boldsymbol{x} \sim \pi_{(t)}([\tilde{h}] \cap \{\tilde{h}+j\})} \left[\mathbb{I}[\boldsymbol{x} = j, \theta_j^2 < 0] + \mathbb{I}[\boldsymbol{x} \neq j, \theta_j^2 > 0] \right] \cdot \epsilon_2, & k = 2; \\ \frac{1}{\bar{d}} \sum_{t=\mathcal{T}_{k-1}+1}^{\mathcal{T}_k} \mathbb{E}_{\theta, \boldsymbol{x} \sim \pi_{(t)}(X^{k, \lceil j/a \rceil})} \left[\mathbb{I}[\boldsymbol{x}_j \theta_j^k < 0] \right] \epsilon_k, & k = 1 \text{ or } 3 \leq k \leq M. \end{cases} \right. \tag{48}$$

and $R_j(\theta) = \sum_{k=1}^M R_j^k(\theta)$. Then for any $\tilde{h} + 1 \leq j \leq \tilde{h} + \ell$, we have that

$$\frac{1}{|\mathcal{Y}|} \sum_{\theta \in \mathcal{Y}} R_j(\theta) \geq \frac{1}{8(M+1)} \min \left\{ \min_{k \in [M], k \neq 2} \frac{T_k \epsilon_k}{\bar{d}}, \frac{T_2 \epsilon_2}{8\bar{d}} \right\}.$$

With Lemma 18 in hand, noting that $\ell \geq \frac{\bar{d} - \tilde{h}}{4} \geq \frac{\bar{d}}{8}$, we obtain that

$$\frac{1}{|\mathcal{Y}|} \sum_{j=h+1}^{h+\ell} \sum_{\theta \in \mathcal{Y}} R_j(\theta) \geq \frac{1}{32(M+1)} \min \left\{ \min_{k \in [M], k \neq 2} T_k \epsilon_k, \frac{T_2 \epsilon_2}{8} \right\}.$$

Noting that the expected regret under parameter θ is at least $\text{Regret}_{\theta, D}(T) \geq \sum_{j=\tilde{h}+1}^{\tilde{h}+l} R_j(\theta)$, we learn that

$$\frac{1}{|\mathcal{Y}|} \sum_{\theta \in \mathcal{Y}} \text{Regret}_{\theta, D}(T) \geq \frac{1}{|\mathcal{Y}|} \sum_{j=\tilde{h}+1}^{\tilde{h}+\ell} \sum_{\theta \in \mathcal{Y}} R_j(\theta) \geq \Omega\left(\gamma/M\right) \tag{49}$$

Then there exists some $\theta^* \in \mathcal{Y}$ such that $\text{Regret}_{\theta^*, D}(T) \geq \Omega(\gamma/M)$. The proof is completed by definition of γ.

$\square$

It remains to prove Lemma 18.

Proof of Lemma 18. Recall the definition of $\mathcal{E}_k$ for $1 \leq k \leq M$. We further define $\mathcal{H}_k$ as below.

- $\mathcal{H}_1 = \bar{\Omega}$, where $\bar{\Omega}$ is the entire probability space;

- $\mathcal{H}_2 = \{n_j \leq \max\{\frac{6T_1}{(h+1)(\bar{d}-h)M}, 6\ln(dM)\}, \forall h + 1 \leq j \leq h + l\}$, where n_j denotes the number of times the j-th arm is taken in the first batch;

- $\mathcal{H}_k = \{n_i^k \leq \max\{\frac{6MT_{k-1}}{b}, 6\ln(dM)\}, \forall 1 \leq i \leq b\}$ for $3 \leq k \leq M$, where $n_i^k = \sum_{t=1}^{\mathcal{T}_{k-1}} \mathbb{I}[X_t = X^{k,i}]$, i.e., the number of times when the context is $X^{k,i}$.

Since the π^1 is independent of θ, then $\{\mathcal{H}_k\}_{k=1}^M$ is also independent of θ. Using Lemma 8, it is easy to show that $\Pr_\theta[\mathcal{H}_k] \geq 1 - \frac{1}{10M}$ for any $1 \leq k \leq M$ and any $\theta \in \mathcal{Y}$.

Let $p_j^k(\theta) = \mathbb{E}_{\theta, \boldsymbol{x} \sim \pi_k(X^{k, \lceil j/a \rceil})} \left[\mathbb{I}[\boldsymbol{x}_j \theta_j^k < 0] | \mathcal{E}_k \cap \mathcal{H}_k \right]$ for $k = 1$ or $3 \leq k \leq M$. For $k = 2$, we define $p_j^k(\theta) = \mathbb{E}_{\theta, \boldsymbol{x} \sim \pi_2([\tilde{h}] \cap \{\tilde{h}+j\})} \left[\mathbb{I}[\boldsymbol{x} = j, \theta_j^2 < 0] + \mathbb{I}[\boldsymbol{x} \neq j, \theta_j^2 > 0] | \mathcal{E}_2 \cap \mathcal{H}_2 \right]$.

Recall that $\mathcal{F}_t$ denotes the event field over the first t steps. For fixed $\theta \in \mathcal{Y}$, we denote θ_j^k be the vector in $\mathcal{Y}$ by reflecting the j-th dimension of θ^k. Using Pinsker's inequality (Lemma 7), and noting that $\mathcal{E}_k \cap \mathcal{H}_k$ are measurable with respect to $\mathcal{F}_{\mathcal{T}_{k-1}}$.

$$p_j^k(\theta) + p_j^k(\theta_j^k) \geq 1 - \sqrt{\frac{1}{2} D_{\text{KL}} \left(\Pr_\theta^{\mathcal{T}_{k-1}}[\cdot | \mathcal{E}_k \cap \mathcal{H}_k], \Pr_{\theta_j^k}^{\mathcal{T}_{k-1}}[\mathcal{E}_k \cap \mathcal{H}_k] \right)}$$

25

for $k = 1$ or $3 \leq k \leq M$, where $\Pr^t[\cdot]$ denotes probability distribution over the first t steps. By definition of $\mathcal{E}_k$ and $\mathcal{H}_k$, we have that

$$D_{\mathrm{KL}}\left(\Pr_\theta^{\mathcal{T}_{k-1}}[\cdot|\mathcal{E}_k \cap \mathcal{H}_k], \Pr_{\theta_j^k}^{\mathcal{T}_{k-1}}[\mathcal{E}_k \cap \mathcal{H}_k]\right) \leq \max\left\{\frac{6MT_{k-1}}{b}, 6\ln(dM)\right\}\frac{4\epsilon_k^2}{a^2} \leq \frac{1}{8}.$$

As a result, we have that

$$p_j^k(\theta) + p_j^k(\theta_j^k) \geq \frac{3}{4}. \tag{50}$$

Using similar arguments, and noting that $n_j \leq \max\{\frac{6T_1}{(\tilde{h}+1)(\tilde{d}-\tilde{h})M}, 6\ln(dM)\}$ conditioned on $\mathcal{H}_2$, we have that

$$p_j^2(\theta) + p_j^2(\theta_j^k) \geq 1 - \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\Pr_{D_2,\theta}^{T_1}[\cdot|\mathcal{H}_2 \cap \mathcal{E}_2], \Pr_{D_2,\theta_j^2}^{T_1}[\cdot|\mathcal{H}_2 \cap \mathcal{E}_2])} \geq \frac{3}{4}. \tag{51}$$

Let $\zeta_k = \frac{1}{\tilde{d}}T_k\epsilon_k$ for $k = 1$ or $3 \leq k \leq M$ and $\zeta_2 = \frac{1}{8b}T_2\epsilon_2$. By the definition of $R_j(\theta)$, and noting that $\ell \geq \frac{\tilde{d}}{8}$, we have that

$$R_j(\theta) \geq \sum_{k=1}^M \Pr_\theta[\mathcal{E}_k \cap \mathcal{H}_k]p_j^k(\theta)\zeta_k$$

$$R_j(\theta_j^k) = \sum_{k=1}^M \Pr_{\theta_j^k}[\mathcal{E}_k \cap \mathcal{H}_k]p_j^k(\theta_j^k)\zeta_k.$$

using Pinsker's inequality (Lemma 7), for any $1 \leq k \leq M$, it holds that

$$\left|\Pr_\theta[\mathcal{E}_k \cap \mathcal{H}_k] - \Pr_{\theta_k^j}[\mathcal{E}_k \cap \mathcal{H}_k]\right| \leq \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\Pr_\theta^{\mathcal{T}_{k-1}}[\cdot|\mathcal{E}_k \cap \mathcal{H}_k], \Pr_{\theta_j^k}^{\mathcal{T}_{k-1}}[\mathcal{E}_k \cap \mathcal{H}_k]\right)} \leq \frac{1}{2M}.$$

Therefore, we have that

$$R_j(\theta) + \sum_{k=1}^M R_j(\theta_j^k) \geq \sum_{k=1}^M \max\left\{\frac{3}{4}\Pr_\theta[\mathcal{E}_k \cap \mathcal{H}_k] - \frac{1}{2M}, 0\right\}\zeta_k \geq \frac{1}{8}\min_k \zeta_k. \tag{52}$$

Taking sum over $\mathcal{Y}$, we have that

$$\frac{1}{|\mathcal{Y}|}\sum_{\theta \in \mathcal{Y}}\sum_{k=1}^M \Pr_\theta[\mathcal{E}_k \cap \mathcal{H}_k]p_j^k(\theta)T_k\epsilon_k \geq \frac{1}{8(M+1)}\min_k \zeta_k.$$

The proof is completed.

$\square$

D A NEW MATRIX CONCENTRATION INEQUALITY WITH DYNAMIC UPPER BOUNDS

Existing matrix concentration inequalities (see, e.g., (Tropp, 2012)) play an important role in recent works on batch linear contextual bandits (Ruan et al., 2021; Zanette et al., 2021). For example, the proof techniques of Theorem 5.1 in (Tropp, 2012) may yield the following concentration bound in Proposition 19 (and the upper bound on $\sum_k X_k$ may be similarly derived). Special cases of Proposition 19 (taking $W = \mathbf{I}$ and $\epsilon = \text{const.}$) includes Lemma 21 in (Ruan et al., 2021) and Lemmas 11 & 12 in (Zanette et al., 2021).

Proposition 19. *Consider a sequence of independent PSD (positive semi-definite) matrices $X_1, X_2, \ldots, X_n \in \mathbb{R}^{d \times d}$ such that $X_k \preccurlyeq W$ for a fixed PSD matrix W and all $1 \leq k \leq n$. There exists a universal constant $c > 0$ such that for every $\delta > 0$ and $\epsilon \in (0, 1)$, it holds that*

$$\Pr\left[\sum_{k=1}^n X_k \preccurlyeq (1 + \epsilon)\sum_{k=1}^n \mathbb{E}[X_k] + \frac{c\ln(d/\delta)}{\epsilon}W\right] \geq 1 - \delta; \tag{53}$$

$$\Pr\left[\sum_{k=1}^n X_k \succcurlyeq (1 - \epsilon)\sum_{k=1}^n \mathbb{E}[X_k] - \frac{c\ln(d/\delta)}{\epsilon}W\right] \geq 1 - \delta. \tag{54}$$

However, to achieve the optimal batch-regret tradeoff in the context-blind setting, we need a stronger version of Proposition 19 where the uniform upper bound matrix W may be stochastic. In particular, we prove the following lemma as a crucial technical tool in our algorithm analysis (especially for the `Exp-Policy` procedure).

Lemma 20. *Consider a sequence of stochastic PSD matrices $W_1, X_1, W_2, X_2, \ldots, W_n, X_n \in \mathbb{R}^{d \times d}$. Let $\mathcal{F}_k = \sigma(W_1, X_1, W_2, X_2, \ldots, W_{k-1}, X_{k-1})$ and $\mathcal{F}_k^+ = \sigma(W_1, X_1, W_2, X_2, \ldots, W_{k-1}, X_{k-1}, W_k)$ be the natural filtration and $Y_k = \mathbb{E}[X_k | \mathcal{F}_k^+]$ for each $k \in \{1, 2, \ldots, n\}$. Suppose W_k is PSD and increasing in k (with respect to the semidefinite order) and $X_k \preccurlyeq W_k$ for each k. For every $\delta > 0$ and $\epsilon \in (0, 1)$, letting $Z(\delta, \epsilon) = \frac{4(\epsilon^2 + 2\epsilon + 2)}{\epsilon} \ln\left(\frac{(n+1)d}{\delta}\right) W_n$, we have that*

$$\Pr\left[\sum_{k=1}^n X_k \preccurlyeq (1 + \epsilon) \sum_{k=1}^n Y_k + Z(\delta, \epsilon)\right] \geq 1 - \delta; \tag{55}$$

$$\Pr\left[\sum_{k=1}^n X_k \succcurlyeq (1 - \epsilon) \sum_{k=1}^n Y_k - Z(\delta, \epsilon)\right] \geq 1 - \delta. \tag{56}$$

In Lemma 20, the stochastic matrix W_n upper bounds all matrices in $\{X_1, X_2, \ldots, X_n\}$. When it is fixed, the lemma reduces to Proposition 19.[8] There are also a few Freedman's inequalities for matrix martingales (see, e.g., (Tropp, 2011)). However, in these inequalities, while the quadratic variation $\mathbb{E}[X_i^2 | \mathcal{F}_i^+]$ becomes dynamic, the uniform upper bound W_k is still fixed.

We note that if an extra $\text{poly}(d)$ factor were allowed in the $\pm \frac{4(\epsilon^2 + 2\epsilon + 2)}{\epsilon} \ln((n+1)d/\delta) W_n$ terms, the lemma would easily follow from Proposition 19 and an ϵ-net argument. However, reducing these $\text{poly}(d)$ factors is crucial to the *full parameter range* optimality analysis of our batch algorithm.

High-level intuition. One simple approach to prove Lemma 20 is to assume the upper bound matrix W_n (correspondingly U in equation 6) were fixed, apply the ordinary matrix concentration inequality and finally take a union bound over an ϵ-net of W_n. However, such an approach would introduce extra $\text{poly}(d)$ factors in the bound and lead to sub-optimal regret bound for our batch learning algorithm. Our proof of Lemma 20 follows the classical exponential moment method. However, we analyze the exponential moment of a specially chosen matrix, namely $W_n^{-1/2}(\sum_{i=1}^n (X_i - (1+\epsilon)Y_i))W_n^{-1/2}$. To bound the trace of this exponential moment, we resort to some deep analysis about Lieb's theorem on convex trace functions (Theorem 24 and Lemma 23). Besides, we also make the critical observation that the function $\text{Tr}\left(\exp\left(U^\top A U\right)\right)$ is bounded by $\text{Tr}(\exp(A))$ with an additive error at most d for any U such that $U^\top U \preccurlyeq \mathbf{I}$ and any symmetric A (Lemma 28).

Before we start the proof, we introduce some basic properties of PSD matrices as below, whose proof is deferred to Appendix D.4.

Fact 21. *For any two PD matrices A and B, $A \preccurlyeq B$ is equivalent to each of the following inequalities,*

$$B^{-1} \preccurlyeq A^{-1}, \tag{57}$$

$$A^{1/2} B^{-1} A^{1/2} \preccurlyeq \mathbf{I}, \tag{58}$$

$$B^{-1/2} A B^{-1/2} \preccurlyeq \mathbf{I}. \tag{59}$$

We now start to prove Lemma 20, while the two helpful technical lemmas (Lemma 23 and Lemma 26), are deferred to Section D.1 and Section D.2.

Proof of Lemma 20. We first prove Equation equation 55. For each $k \in \{0, 1, 2, \ldots, n\}$, let $Z_k = \frac{4(\epsilon^2 + 2\epsilon + 2)}{\epsilon} W_k$ and

$$E_k := \text{Tr}\left(\exp\left(Z_k^{-1/2}\left(\sum_{i=1}^k (X_i - (1+\epsilon)Y_i)\right) Z_k^{-1/2}\right)\right).$$

---

[8]Indeed, we lose an additional $\ln(n+1)$ term in the $\ln(d/\delta)$ terms in Proposition 19, and we do not know if this compromise is necessary.

By Lemma 23, we have that $\mathbb{E}[E_n] \leq (n+1)d$. Therefore, by Markov inequality,

$$\Pr\left[\lambda_{\max}\left(Z_n^{-1/2}\left(\sum_{i=1}^{n}(X_i - (1+\epsilon)Y_i)\right)Z_n^{-1/2}\right) \geq C\right]$$

$$\leq \Pr\left[\mathrm{Tr}\left(\exp\left(Z_n^{-1/2}\left(\sum_{i=1}^{n}(X_i - (1+\epsilon)Y_i)\right)Z_n^{-1/2}\right)\right) \geq e^C\right] \leq \frac{\mathbb{E}[E_n]}{e^C} \leq (n+1)de^{-C},$$

which (by Fact 21) means that

$$\Pr\left[\sum_{i=1}^{n}(X_i - (1+\epsilon)Y_i) \preccurlyeq CZ_n\right] \geq 1 - (n+1)de^{-C}.$$

Choosing $C = \ln((n+1)d/\delta)$ and recalling that $Z_n = \frac{4(\epsilon^2 + 2\epsilon + 2)}{\epsilon}W_n$, we prove Equation equation 55.

We then prove Equation equation 56. First we define

$$E_k' := \mathrm{Tr}\left(\exp\left(Z_k^{-1/2}\left(\sum_{i=1}^{k}((1-\epsilon)Y_i - X_i)\right)Z_k^{-1/2}\right)\right).$$

Similarly, by Lemma 26, we have that $\mathbb{E}[E_n'] \leq (n+1)d$ and

$$\Pr\left[\lambda_{\max}\left(Z_n^{-1/2}\left(\sum_{i=1}^{n}((1-\epsilon)Y_i - X_i)\right)Z_n^{-1/2}\right) \geq C\right] \tag{60}$$

$$\leq \Pr\left[\mathrm{Tr}\left(\exp\left(Z_n^{-1/2}\left(\sum_{i=1}^{n}((1-\epsilon)Y_i - X_i)\right)Z_n^{-1/2}\right)\right) \geq e^C\right] \leq \frac{\mathbb{E}[E_n']}{e^C} \leq (n+1)de^{-C},$$

which means that

$$\Pr\left[\sum_{i=1}^{n}((1-\epsilon)Y_i - X_i) \preccurlyeq CZ_n\right] \geq 1 - (n+1)de^{-C}.$$

Choosing $C = \ln((n+1)d/\delta)$ we finish the proof. $\qquad\square$

Corollary 22. *Given a sequence of stochastic random variables $X_1, X_2, \ldots, X_n$ such that $0 \leq X_i \leq W$ for any $1 \leq i \leq n$ with probability 1. Let $\mathcal{F}_k = \sigma(X_1, X_2, \ldots, X_{k-1})$ and $Y_k = \mathbb{E}[X_k | \mathcal{F}_k]$ For every $\delta > 0$ and $\epsilon > 0$, we have that*

$$\Pr\left[\sum_{k=1}^{n}X_k \leq (1+\epsilon)\sum_{k=1}^{n}Y_k + \frac{4(\epsilon^2 + 2\epsilon + 2)}{\epsilon}\ln((n+1)/\delta)W\right] \geq 1 - \delta;$$

$$\Pr\left[\sum_{k=1}^{n}X_k \geq (1-\epsilon)\sum_{k=1}^{n}Y_k - \frac{4(\epsilon^2 + 2\epsilon + 2)}{\epsilon}\ln((n+1)/\delta)W\right] \geq 1 - \delta.$$

Proof. Letting $d = 1$ and $W_k = W$ for $1 \leq k \leq n$, by Lemma 20 we finish the proof. $\qquad\square$

D.1 STATEMENT AND PROOF OF LEMMA 23

Lemma 23. *For each $k \in \{1, 2, \ldots, n\}$, we have that*

$$\mathbb{E}\left[E_k | \mathcal{F}_k^+\right] \leq \mathbb{E}\left[E_{k-1} | \mathcal{F}_k^+\right] + d.$$

Proof. Firstly, we introduce a deep theorem from Lieb (Theorem 6, (Lieb, 1973)), which provides theoretical basis for a series of concentration inequalities on self-adjoint matrices.

Theorem 24. *Fix a d-dimensional symmetric matrix H. The function $f(A) := \mathrm{Tr}(\exp(\log(A) + H))$ is concave on the d-dimensional positive definite cone.*

28

Based on Theorem 24, (Tropp, 2012) derived the corollary below.

Corollary 25. *Fix a self-adjoint matrix H. Let X be stochastic symmetric matrix*

$$\mathbb{E}\left[\operatorname{Tr}\left(\exp(X + H)\right)\right] \leq \operatorname{Tr}\left(\exp\left(\log \mathbb{E}[e^X] + H\right)\right). \tag{61}$$

Given Corollary 25 we continue the analysis as below. Note that $\mathbb{E}\left[E_{k-1}|\mathcal{F}_k^+\right]$ is a deterministic value. Throughout this proof, we will condition on $\mathcal{F}_k^+$. We calculate that

$$\mathbb{E}\left[E_k\right]$$

$$= \mathbb{E}\left[\operatorname{Tr}\left(\exp\left(Z_k^{-1/2}\left(\sum_{i=1}^{k}(X_i - (1+\epsilon)Y_i)\right)Z_k^{-1/2}\right)\right)\right]$$

$$= \mathbb{E}\left[\operatorname{Tr}\left(\exp\left(Z_k^{-1/2}\left(\sum_{i=1}^{k-1}(X_i - (1+\epsilon)Y_i)\right)Z_k^{-1/2} + Z_k^{-1/2}(X_k - (1+\epsilon)Y_k)Z_k^{-1/2}\right)\right)\right]$$

$$\leq \operatorname{Tr}\left(\exp\left(Z_k^{-1/2}\left(\sum_{i=1}^{k-1}(X_i - (1+\epsilon)Y_i)\right)Z_k^{-1/2} + \ln\left(\mathbb{E}\left[\exp\left(Z_k^{-1/2}(X_k - (1+\epsilon)Y_k)Z_k^{-1/2}\right)\right]\right)\right)\right)$$

$$\tag{62}$$

$$\leq \operatorname{Tr}\left(\exp\left(Z_k^{-1/2}\left(\sum_{i=1}^{k-1}(X_i - (1+\epsilon)Y_i)\right)Z_k^{-1/2}\right)\right) \tag{63}$$

$$= \operatorname{Tr}\left(\exp\left(Z_k^{-1/2}Z_{k-1}^{1/2}Z_{k-1}^{-1/2}\left(\sum_{i=1}^{k-1}(X_i - (1+\epsilon)Y_i)\right)Z_{k-1}^{1/2}Z_{k-1}^{-1/2}Z_k^{-1/2}\right)\right)$$

$$\leq \operatorname{Tr}\left(\exp\left(Z_{k-1}^{-1/2}\left(\sum_{i=1}^{k}(X_i - (1+\epsilon)Y_i)\right)Z_{k-1}^{-1/2}\right)\right) + d \tag{64}$$

$$= E_{k-1} + d.$$

Here equation 62 is by Corollary 25, equation 63 is by Lemma 27 (stated and proved in Section D.2) and the monotonicity of trace exponential with respect to the semidefinite order (see (Petz, 1994), §2.2), and equation 64 is by Lemma 28 (stated and proved in Section D.3, letting $U = Z_{k-1}^{1/2}Z_k^{-1/2}$ and $A = Z_{k-1}^{-1/2}\left(\sum_{i=1}^{k}(X_i - (1+\epsilon)Y_i)\right)Z_{k-1}^{-1/2}$, and one can verify that $U^\top U = Z_{k-1}^{1/2}Z_k^{-1}Z_{k-1}^{1/2} \preccurlyeq \mathbf{I}$ by Fact 21). $\qquad \square$

Similarly, we may establish the following lemma, whose proof is deferred to Appendix D.5.

Lemma 26. *For each $k \in \{1, 2, \ldots, n\}$, we have that*

$$\mathbb{E}\left[E_k'|\mathcal{F}_k^+\right] \leq \mathbb{E}\left[E_{k-1}'|\mathcal{F}_k^+\right] + d.$$

D.2 STATEMENT AND PROOF OF LEMMA 27

Lemma 27. *For each $k \in \{1, 2, \ldots, n\}$, we have that*

$$\mathbb{E}\left[\exp\left(Z_k^{-1/2}(X_k - (1+\epsilon)Y_k)Z_k^{-1/2}\right)\Big|\mathcal{F}_k^+\right] \preccurlyeq \mathbf{I}.$$

Proof. Throughout this proof, we will condition on $\mathcal{F}_k^+$. Let $U_k := Z_k^{-1/2}X_kZ_k^{-1/2}$ and $V_k := Z_k^{-1/2}Y_kZ_k^{-1/2}$. By the assumption in Lemma 20 and our definition for Z_k, we have that $0 \preccurlyeq U_k \preccurlyeq$

$\frac{\epsilon}{4(\epsilon^2+2\epsilon+2)}\mathbf{I}$ and $\mathbb{E}[U_k] = V_k$. Therefore, $0 \preccurlyeq V_k \preccurlyeq \frac{\epsilon}{4(\epsilon^2+2\epsilon+2)}\mathbf{I}$. We now compute that

$$\mathbb{E}\left[\exp\left(Z_k^{-1/2}(X_k - (1+\epsilon)Y_k)Z_k^{-1/2}\right)\right]$$

$$= \mathbb{E}\left[\exp\left(U_k - (1+\epsilon)V_k\right)\right]$$

$$= \mathbb{E}\left[\mathbf{I} + (U_k - (1+\epsilon)V_k) + \frac{1}{2}(U_k - (1+\epsilon)V_k)^2 + \sum_{i\geq3}\frac{1}{i!}(U_k - (1+\epsilon)V_k)^i\right] \tag{65}$$

$$\leq \mathbb{E}\left[\mathbf{I} + (U_k - (1+\epsilon)V_k) + \frac{1}{2}(U_k - (1+\epsilon)V_k)^2 + \sum_{i\geq3}\frac{1}{i!}(U_k - (1+\epsilon)V_k)^2\right] \tag{66}$$

$$\preccurlyeq \mathbb{E}\left[\mathbf{I} + (U_k - (1+\epsilon)V_k) + 2(U_k - (1+\epsilon)V_k)^2\right]$$

$$\preccurlyeq \mathbb{E}\left[\mathbf{I} + (U_k - (1+\epsilon)V_k) + 4U_k^2 + 4(1+\epsilon)^2V_k^2\right] \tag{67}$$

$$\preccurlyeq \mathbb{E}\left[\mathbf{I} + (U_k - (1+\epsilon)V_k) + \frac{4\epsilon}{4(\epsilon^2+2\epsilon+2)}U_k + \frac{4(1+\epsilon)^2\epsilon}{4(\epsilon^2+2\epsilon+2)}V_k\right] \tag{68}$$

$$= \mathbf{I}.$$

Here, equation 65 is by Taylor series expansion, equation 66 is by the fact that $-\mathbf{I} \preccurlyeq (U_k - (1+\epsilon)V_k) \preccurlyeq \mathbf{I}$ and $X^k - X^2 = X(X^{k-2} - \mathbf{I})X \preccurlyeq 0$ for $k \geq 2$ and $-\mathbf{I} \preccurlyeq X \preccurlyeq \mathbf{I}$. equation 67 is because $4U_k^2 + 4(1+\epsilon)^2V_k^2 - 2(U_k - (1+\epsilon)V_k)^2 = 2(U_k + (1+\epsilon)V_k)^2 \succcurlyeq 0$, and equation 68 is by the fact that $\frac{4(\epsilon^2+2\epsilon+2)}{\epsilon}U_k^2 - U_k = U_k^{1/2}\left(\frac{4(\epsilon^2+2\epsilon+2)}{\epsilon}U_k - \mathbf{I}\right)U_k^{1/2} \preccurlyeq 0$ and $\frac{4(\epsilon^2+2\epsilon+2)}{\epsilon}V_k^2 - V_k = V_k^{1/2}\left(\frac{4(\epsilon^2+2\epsilon+2)}{\epsilon}V_k - \mathbf{I}\right)V_k^{1/2} \preccurlyeq 0$. $\qquad\square$

D.3 STATEMENT AND PROOF OF LEMMA 28

Lemma 28. *Let A be a real symmetric matrix. For any U such that $U^\top U \preccurlyeq \mathbf{I}$, it holds that*

$$\mathrm{Tr}(\exp(U^\top A U)) \leq \mathrm{Tr}(\exp(A)) + d. \tag{69}$$

Proof. We can assume without loss of generality that $A = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ and $\lambda_d \leq \lambda_{d-1} \leq \cdots \leq \lambda_{\ell-1} < 0 \leq \lambda_\ell \leq \cdots \leq \lambda_1$.

Let us write $U^\top = Q^\top L^\top$ to be its QR decomposition where $Q^\top$ is orthogonal and $L^\top$ is an upper triangular matrix (and therefore L is a lower triangular matrix). We claim that

$$U^\top \mathbf{I}_j U \preccurlyeq Q^\top \mathbf{I}_j Q, \tag{70}$$

where we define $\mathbf{I}_j := \mathrm{diag}(0, \ldots, 0, 1, \ldots 1)$ to be a rank j matrix. Note that $U^\top \mathbf{I}_j U = Q^\top L^\top \mathbf{I}_j L Q$. Therefore, to prove the claim, it suffices to show that $L^\top \mathbf{I}_j L \preccurlyeq \mathbf{I}_j$. Since $U^\top U \preccurlyeq \mathbf{I}$, we also have that $L^\top L = QU^\top U Q^\top \preccurlyeq \mathbf{I}$. Note that for any vector x, it holds that $x^\top(L\mathbf{I}_j)^\top(L\mathbf{I}_j)x = (\mathbf{I}_j x)^\top L^\top L(\mathbf{I}_j x) \leq x^\top \mathbf{I}_j x$. Therefore we conclude that $L^\top \mathbf{I}_j L = (L\mathbf{I}_j)^\top(L\mathbf{I}_j) \preccurlyeq \mathbf{I}_j$, where the first equality holds because L is a lower triangular matrix.

Let $\tilde{\lambda}_i := \lambda_i - \lambda_{i-1} \geq 0$ for each $i \in \{1, 2, \ldots, \ell-1\}$, $\tilde{\lambda}_\ell := \lambda_\ell \geq 0$, and $\tilde{\lambda}_j := 0$ for each $j \in \{\ell+1, \ell+2, \ldots, d\}$. By equation 70, we then have that

$$U^\top A U \preccurlyeq \sum_{j=1}^{d}\tilde{\lambda}_j U^\top \mathbf{I}_j U \preccurlyeq \sum_{j=1}^{d}\tilde{\lambda}_j Q^\top \mathbf{I}_j Q.$$

By the monotonicity of trace exponential with respect to the semidefinite order (see, e.g. (Petz, 1994) §2.2), we have that

$$\mathrm{Tr}(\exp(U^\top A U)) \leq \mathrm{Tr}\left(\exp\left(\sum_{j=1}^{d}\tilde{\lambda}_j Q^\top \mathbf{I}_j Q\right)\right)$$

$$= \sum_{j=1}^{d}\exp\left(\sum_{i=j}^{d}\tilde{\lambda}_i\right) \leq \sum_{j=1}^{d}\exp(\lambda_i) + d = \mathrm{Tr}(\exp(A)) + d.$$

$\square$

D.4   Proof of Fact 21

Proof. We first show that $A \preccurlyeq B$ implies equation 58. By $A \preccurlyeq B$, we have that $B - A \succcurlyeq 0$. Therefore, $A^{-1/2}BA^{-1/2} - \mathbf{I} = A^{-1/2}(B - A)A^{-1/2}$ is PSD, which means that $A^{-1/2}BA^{-1/2} \succcurlyeq \mathbf{I}$. We then have $A^{1/2}B^{-1}A^{1/2} \preccurlyeq \mathbf{I}$. This is because if we let $M = A^{-1/2}BA^{-1/2} \succcurlyeq \mathbf{I}$ for convenience, for every vector x, it holds that

$$x^\top A^{1/2}B^{-1}A^{1/2}x = x^\top M^{-1}x = (M^{-1/2}x)^\top(M^{-1/2}x) \leq (M^{-1/2}x)^\top M(M^{-1/2}x) = x^\top x.$$

This proves equation 58, assuming $A \succcurlyeq B$.

We then show that equation 58 implies equation 57. Note that $A^{1/2}B^{-1}A^{1/2} - \mathbf{I} \preccurlyeq 0$ implies that $B^{-1} - A^{-1} = A^{-1/2}(A^{1/2}B^{-1}A^{1/2} - \mathbf{I})A^{-1/2} \preccurlyeq 0$, which leads to $B^{-1} \preccurlyeq A^{-1}$.

By symmetry, we can also prove that equation 57 implies $A \preccurlyeq B$, and therefore establishing the equivalence condition for both equation 57 and equation 58.

Finally, $A - B \preccurlyeq 0$ is equivalent to that $\mathbf{I} - B^{-1/2}AB^{-1/2} = B^{-1/2}(B - A)B^{-1/2}$ is PSD, which is also equivalent to that $B^{-1/2}AB^{-1/2} \preccurlyeq \mathbf{I}$, proving the equivalence for equation 59. $\square$

D.5   Proof of Lemma 26

Proof. Following the arguments in the proof of Lemma 23, it suffices to prove that

$$\mathbb{E}\left[\exp\left(Z_k^{-1/2}((1-\epsilon)Y_k - X_K)Z_k^{-1/2}\right)\Big|\mathcal{F}_k^+\right] \preccurlyeq \mathbf{I}$$

holds for each $1 \leq k \leq n$. Recall the definition of U_k and V_k in Lemma 27. In the analysis below, we will conditioned on $\mathcal{F}_k^+$. By the same arguments in the proof of Lemma 27, and noting that $-\mathbf{I} \preccurlyeq (1-\epsilon)V_k - U_k \preccurlyeq \mathbf{I}$, we have that

$$\mathbb{E}\left[\exp\left(Z_k^{-1/2}((1-\epsilon)Y_k - X_k)Z_k^{-1/2}\right)\right]$$
$$= \mathbb{E}\left[\exp\left((1-\epsilon)V_k - U_k\right)\right]$$
$$= \mathbb{E}\left[I + ((1-\epsilon)V_k - U_k) + \frac{1}{2}((1-\epsilon)V_k - U_k)^2 + \sum_{i\geq 3}\frac{1}{i!}((1-\epsilon)V_k - U_k)^i\right]$$
$$\leq \mathbb{E}\left[I + ((1-\epsilon)V_k - U_k) + \frac{1}{2}((1-\epsilon)V_k - U_k)^2 + \sum_{i\geq 3}\frac{1}{i!}((1-\epsilon)V_k - U_k)^2\right]$$
$$\preccurlyeq \mathbb{E}\left[\mathbf{I} + ((1-\epsilon)V_k - U_k) + 2((1-\epsilon)V_k - U_k)^2\right]$$
$$\preccurlyeq \mathbb{E}\left[\mathbf{I} + ((1-\epsilon)V_k - U_k) + 4U_k^2 + 4(1-\epsilon)^2V_k^2\right]$$
$$\preccurlyeq \mathbb{E}\left[\mathbf{I} + ((1-\epsilon)V_k - U_k) + \frac{4\epsilon}{4(\epsilon^2 + 2\epsilon + 2)}U_k + \frac{4(1-\epsilon)^2\epsilon}{4(\epsilon^2 + 2\epsilon + 2)}V_k\right]$$
$$\preccurlyeq \mathbf{I}.$$

The proof is completed. $\square$