Design of Experiments for Stochastic Contextual Linear Bandits

Andrea Zanette*
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA
zanette@berkeley.edu

Kefan Dong*
Department of Computer Science
Stanford University
Stanford, CA
kefandong@stanford.edu

Jonathan Lee*
Department of Computer Science
Stanford University
Stanford, CA
jnl@stanford.edu

Emma Brunskill
Department of Computer Science
Stanford University
Stanford, CA
ebrun@cs.stanford.edu

Abstract

In the stochastic linear contextual bandit setting there exist several minimax procedures for exploration with policies that are *reactive* to the data being acquired. In practice, there can be a significant engineering overhead to deploy these algorithms, especially when the dataset is collected in a distributed fashion or when a human in the loop is needed to implement a different policy. Exploring with a single non-reactive policy is beneficial in such cases. Assuming some batch contexts are available, we design a single stochastic policy to collect a good dataset from which a near-optimal policy can be extracted. We present a theoretical analysis as well as numerical experiments on both synthetic and real-world datasets.

1 Introduction

Many settings may substantially benefit from data-driven contextualized decision policies that optimize the desired expected outcome. Online machine learning methods like multi-armed bandits and reinforcement learning, that adaptively change interventions in response to outcomes in a closed loop process (see Figure 1a), may not yet be practical for all domains due to the expertise and infrastructure needed. However running an experiment with a fixed decision policy to identify a good personalized policy is likely to be both simple logistically (since currently such decision policies are often specified by hand) and more easily accepted, since many areas (education, healthcare, social sciences) commonly deploy experiments across a few conditions to find the best approach. For example, education startups, political campaigns, and governmental agencies can use email and

*The first three authors contributed equally. The work was fully completed while Andrea Zanette was a PhD Candidate in the Institute for Computational and Mathematical Engineering at Stanford University. Future revision of this article will be made available at https://arxiv.org/abs/2107.09912.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

text messages to provide targeted information and opportunities, such as information to encourage vaccination, or tips to parents to support their child's developmental stage. Such organizations are familiar with standard experimental design, and but generally lack the infrastructure for continuous online contextual MAB learning. Experiments that involve deploying a nonadaptive policy that fit in with standard workflows but enable data-efficient learning of contextualized policies might offer substantial benefits over AB testing or relying on other segmentation methods that may not directly optimize desired outcomes.

For these reasons, a key opportunity is to design *static* or *nonadaptive* policies that can be used to gather data to identify optimal contextualized decision policies. Indeed a fixed data collection strategy is practically desirable (1) whenever multiple agents collect data asynchronously and communication to update their policy is difficult or impossible, and (2) whenever changing the policy requires a significant overhead either in the engineering infrastructure or in the training of human personnel. Several prior papers limit the number of policy switches with minimal sample complexity impact (Han et al., 2020; Ruan et al., 2020; Ren et al., 2020; Bai et al., 2019; Wang et al., 2021). Motivated by the above settings, we look for a *single*, *nonadaptive* policy for data collection.

Setting and goal We consider the linear stochastic contextual bandit setting where each context $s \in \mathcal{S}$ is sampled from a distribution μ and a context-dependent action set $\mathcal{A}_s$ is made available to the learner. The bandit instance is defined by a feature extractor $\phi(s, a) \in \mathbb{R}^d$ and some unknown parameter $\theta^\star \in \mathbb{R}^d$. Upon choosing an action $a \in \mathcal{A}_s$, the linear reward function $r(s, a) = \phi(s, a)^\top \theta^\star + \eta$ is revealed to the learner corrupted by mean zero 1-subGaussian noise η. Our goal is to construct an exploration policy π_e to gather a dataset, such that after that dataset is gathered, we can extract a near optimal policy $\widehat{\pi}$ from it. In particular, we want to minimize the number of exploration samples.

Perhaps surprisingly, there has been relatively little work on this setting. Prior work on exploration to quickly identify a near-optimal policy focuses on best-arm identification using adaptive policies that react to the observed rewards (Soare et al., 2014; Tao et al., 2018; Jedra and Proutiere, 2020) or design of experiments that produces a nonadaptive policy for data collection (Kiefer and Wolfowitz, 1960; Esfandiari et al., 2019; Lattimore and Szepesvari, 2020); both lines of work assume that a *single*, repeated context with unchanging action set is presented to the learner. In contrast we are interested in identifying near-optimal context-specific decision policies. The closest related work (Ruan et al., 2020) investigates our task as a subtask for online regret learning, but requires an amount of data that scales as $\Omega(d^{16})$, which is impractical in applications with even moderate d.[2]

Without any apriori information, no algorithm can do much better than deploying a random policy, which can require an amount of data that scales exponentially in d, see appendix G.1. However, in many common applications, *prior data about the context distribution* $\mu(s)$ and the state–action feature representation ϕ is available. For example, an organization will often know information about its customers and specify the feature representation used for state–action spaces in advance.[3] The initially available state contexts are referred to as *offline* (state) contexts data.

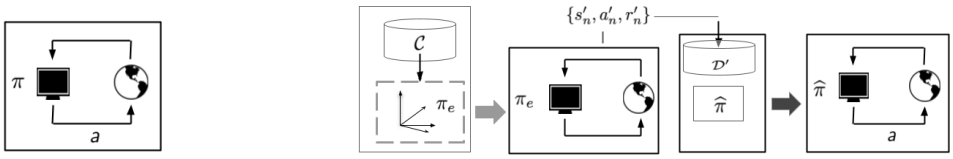
Our algorithm leverages historical context data to enable data efficient design of experiments for stochastic linear contextual bandits. It uses offline context-only data $\mathcal{C}$ to design a nonadaptive policy π_e to collect new, *online* data where reward feedback is observed (see Figure 1b), and uses the resulting dataset $\mathcal{D}'$ to learn a near-optimal on average decision policy $\widehat{\pi}$ for future use. We highlight that the algorithm does not get to adjust the exploratory policy π_e while the online data is being collected.

Contributions We make the following contributions.

- Using past state contexts only, we design a single, nonadaptive policy to acquire online data that can be used to compute a context-dependent decision policy that is near-optimal in expectation across the contexts with high probability, for future use.
- To identify an ϵ-optimal policy, our algorithm achieves the minimax lower bound $\Omega(\min\{d \log \sum_s \mathcal{A}_s, d^2)\}/\epsilon^2)$ on the number of online samples (ignoring log and constants), while keeping the number of offline state contexts required polynomially small ($O(d^2/\epsilon^2)$ or $O(d^3/\epsilon^2)$).

[2]Note that the number of offline data in (Ruan et al., 2020) is independent of $1/\epsilon$. But for most of the practical settings, our bound d^3/ϵ^2 can be much smaller than d^{16}.

[3]In other words, given a set of previously observed states $s_1, \ldots, s_M$, and a known state–action representation ϕ, for any potential action a, we can compute the resulting representation $\phi(s, a)$.



(a) Traditional online RL framework

(b) Design of experiment with an online and offline component

Figure 1: Comparison between the traditional RL setting and design of experiments

- Our experiments on a simulation and on a learning to rank Yahoo dataset show our strategic design of experiments approach can learn better decision policies with less exploration data compared to standard static exploration strategies that fail to exploit structure.

2 Related Work

Most linear bandit papers focus on online regret. Research on sample complexity results has focused primarily on adaptive algorithms for non-contextual bandits with a simple or vectored action space (Soare et al., 2014; Tao et al., 2018; Xu et al., 2018; Fiez et al., 2019; Jedra and Proutiere, 2020; Degenne et al., 2020). Their sample complexity bounds depend on the suboptimality gap of the action set, and are therefore, instance-dependent. The design of experiments literature (Kiefer and Wolfowitz, 1960; Esfandiari et al., 2019; Lattimore and Szepesvari, 2020; Ruan et al., 2020) has also focused on non-contextual bandits, but in designing a non-adaptive policy such that the resulting dataset can be used to identify an optimal arm with high probability.

A few recent papers consider learning the optimal policy for contextual bandits in the pure exploration setting (Deshmukh et al., 2018; Ruan et al., 2020). Deshmukh et al. (2018) propose an algorithm that yields an online reactive policy for when the reward is generated by a RKHS function on the input space: in contrast we focus on nonadaptive policies for linear contextual bandits.

There has been recent interest in reward-free reinforcement learning in Markov decision processes (e.g. (Jin et al., 2020; Wang et al., 2020a)). This work is focused on online adaptive policies. Jin et al. (2020)'s algorithm is similar in structure to our's and could likely be adapted to the offline, bandit setting, but they focus on tabular domains. Note for simple contextual multi-armed bandits (tabular states, no shared structure, Jin et al.'s setting with H=1), a good strategy in the non-adaptive setting is to simply sample actions uniformly; indeed this is optimal given the lack of shared information between state-actions. No sophisticated algorithm is needed in that case.

For linear bandits, the closest reward-free work we are aware of is work on RL with linear function approximationWang et al. (2020a). While their algorithm can be specialized to the linear bandit setting and modified to match our algorithm by using offline data, the theoretical analysis would need to be very different from that in(Wang et al., 2020b)'s paper (and thus it would not yield the results presented in our paper). A key challenge in our case is to compute a non-adaptive policy for gathering data and be able to ensure the covariance matrix of the resulting data will be very similar to the covariance matrix that could result by a simulated adaptive policy. This issue does not arise in (Wang et al., 2020b) because they focus on reactive policies.[4]

The most related paper to this work is by Ruan et al. (2020). The main focus of Ruan et al. (2020) is to minimize the number of policy switches in the online phase while keeping the expected regret optimal. Directly applying Theorem 6 of Ruan et al. (2020) yields a $\widetilde{O}(d^2/\epsilon^2)$ online sample complexity and $\widetilde{O}(d^{16})$ offline sample complexity. Their offline sample complexity is independent of ϵ. As a trade-off, Ruan et al. (2020) have a highly suboptimal dependence on d, which leads to burn-in phase too large to be practically useful; their algorithm also suffers from a high computational complexity. Note their work is focused on a different objective than our's.

[4]In particular, we will shortly introduce a planner-sampler two stage algorithm. We decouple the planner from the sampler to identify a non-reactive policy. Due to the decoupling, the planner needs to accurately predict what will happen when the sampler is run. To ensure this, a key challenge is to ensure that the planner inverse covariance matrix is very similar to the sampler inverse covariance matrix.

3 Setup

We consider the stochastic linear contextual bandit model with stochastic contexts. A bandit instance is characterized by a tuple $\langle \mathcal{S}, \mathcal{A}, \mu, r \rangle$ where $\mathcal{S}$ is the context space and μ is the context distribution. For a context $s \in \mathcal{S}$, the action space is denoted by $\mathcal{A}_s$. The feature map $\phi : (s, a) \mapsto \phi(s, a) \in \mathbb{R}^d$ is assumed to be known and defines the linear reward model $r(s, a) = \phi(s, a)^\top \theta^\star + \eta$ for some $\theta^\star \in \mathbb{R}^d$ parameter and some mean-zero 1-subgaussian random variable η.

We occasionally use the $\widetilde{O}$ notation to suppress polylog factors of the input parameters $d, \lambda, \frac{1}{\delta}$. We write $f \lesssim g$ if $f = O(g)$, and $f \lessapprox g$ if $f = \widetilde{O}(g)$. We say $f = \Omega(g)$ if there exists a positive constant c such that $f \geq cg$. For a positive semi-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$ and a vector $x \in \mathbb{R}^d$, let $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$. For two symmetric matrices A, B, we say $A \preccurlyeq B$ if $B - A$ is positive semi-definite.

Our analysis consists of an offline and online component. We often add $'$ to indicate the online quantities, e.g., we denote with $s_1, s_2, \ldots$ the offline contexts and with $s'_1, s'_2, \ldots$ the online contexts.

3.1 Objective and Error Decomposition

As depicted in Figure 1, our approach is to leverage offline state contexts $\mathcal{C} = \{s_1, \ldots, s_M\}$ where $s_m \sim \mu$ to design a single stochastic policy π_e to acquire data from an online data stream $\mathcal{C}' = \{s'_1, \ldots, s'_N\}$. This generates a dataset $\mathcal{D}' = \{(s'_n, a'_n, r'_n)\}_{n=1,\ldots,N}$ where $s'_n \sim \mu, a'_n \sim \pi_e(s'_n)$ and $r'_n = r(s'_n, a'_n)$. Using $\mathcal{D}'$, the least-square predictor $\widehat{\theta}$ and the corresponding greedy policy $\widehat{\pi}$ can be extracted

$$\widehat{\theta} = \left(\Sigma'_N\right)^{-1} \sum_{i=1}^{N} \phi(s'_n, a'_n) r'_n, \qquad \widehat{\pi}(s) = \arg\max_{a \in \mathcal{A}_s} \phi(s, a)^\top \widehat{\theta} \tag{1}$$

where $\Sigma'_N = \left(\sum_{n=1}^{N} \phi(s'_n, a'_n)\phi(s'_n, a'_n)^\top + \lambda_{reg} I \right)$ is the empirical cumulative covariance matrix with regularization level λ_{reg}. The quality of the dataset $\mathcal{D}'$ (and of the whole two-step procedure) is measured by the suboptimality of the extracted policy $\widehat{\pi}$ obtained after data collection:

$$\mathbb{E}_{s \sim \mu}[\max_a \phi(s, a)^\top \theta^\star - \phi(s, \widehat{\pi}(s))^\top \theta^\star]. \tag{2}$$

Note that Eq. (2) measures the expectation over the contexts of the suboptimality between the resulting decision policy $\widehat{\pi}$ and the optimal policy. This is a looser criteria than a maximum norm bound which evaluates the error over any possible context s: in general this latter error may not be easily reduced if certain directions in feature space are rarely available.

A related objective is to minimize the maximum prediction error on the linear bandit instance. By least square regression analyses (Lattimore and Szepesvári, 2020), with probability at least $1 - \delta$ we have

$$\mathbb{E}_{s \sim \mu} \max_a |\phi(s, a)^\top (\theta^\star - \widehat{\theta})| \leq \sqrt{\beta}\, \mathbb{E}_{s \sim \mu} \max_a \|\phi(s, a)\|_{(\Sigma'_N)^{-1}}, \tag{3}$$

where

$$\sqrt{\beta} = \min\Big\{ \underbrace{\sqrt{2 \ln 2 |\sum_s \mathcal{A}_s| + \ln \frac{1}{\delta}}}_{\text{Small } \sum_s \mathcal{A}_s}, \underbrace{2\sqrt{2d \ln 6 + \ln \frac{1}{\delta}}}_{\text{Large } \sum_s \mathcal{A}_s} \Big\} + \underbrace{\sqrt{\lambda_{reg}} \|\theta^\star\|_2}_{\text{Regularization Effect}}. \tag{4}$$

The above expression assumes that the state-action-rewards (s'_n, a'_n, r'_n) are drawn i.i.d. from a fixed distribution. This is satisfied in our setting as the data-collection policy π_e is nonadaptive to the online data. The parameter β governs the sample complexity as a function of the size of the state-action space and also highlights the impact of the regularization bias.

Small predictive error (Eq. (3)) can be used to bound the suboptimal gap of the greedy policy (Eq. (2)). Therefore to obtain good performance, it is sufficient to bound $\mathbb{E}_{s \sim \mu} \max_a \|\phi(s, a)\|_{(\Sigma'_N)^{-1}}$. This can be achieved by designing an appropriate sampling policy π_e to yield a set of (s'_n, a'_n, r'_n) tuples whose empirical cumulative covariance matrix Σ'_N is as 'large' as possible.

4 Algorithms

Reward-free LINUCB First, assume it is acceptable to have an algorithm that updates its policy reactively. In order to reduce $\mathbb{E}_{s\sim\mu}\max_a \|\phi(s,a)\|_{(\Sigma'_n)^{-1}}$ we could use an algorithm that, every time a context $s \sim \mu$ is observed, chooses the action $\arg\max_{a\in\mathcal{A}_s} \|\phi(s,a)\|_{(\Sigma'_n)^{-1}}$ where the norm $\|\phi(s,a)\|_{(\Sigma'_n)^{-1}}$ that represents the uncertainty is highest (cf. section 3.1). Such procedure is related (although not identical) to the well known globally optimal design of experiment applied to linear bandits (i.e., Lattimore and Szepesvari (2020)). Algorithmically, it corresponds to running the LINUCB algorithm (Abbasi-Yadkori et al., 2011) with the empirical reward function set to zero. In other words, the planner is essentially an *offline, reward-free* version of LINUCB. We emphasize that the associated non-stationary policy is generated offline and is not actively played. One can show that after $\approx d^2/\epsilon^2$ iterations the uncertainty $\sqrt{\beta}\,\mathbb{E}_{s\sim\mu}\max_a \|\phi(s,a)\|_{(\Sigma'_n)^{-1}} \leq \epsilon$; if the algorithm stores the observed reward $r(s,a)$ in every visited context s and chosen action a, the greedy policy that can be extracted from this dataset (of size $\approx d^2/\epsilon^2$) is ϵ-optimal (cf. Eqs. (2),(3)), as desired.

Unfortunately we cannot run this reward-free algorithm online, as its policy is *reactive* to the online stream of observed online contexts s and selected actions a, while we want a nonadaptive policy.

Algorithm 1 PLANNER (Reward-free LINUCB)

Input: Contexts $\mathcal{C} = \{s_1, \dots, s_M\}$, reg. λ_{reg}
$\Sigma_1 = \lambda_{reg} I$
$m = 1$
for $m = 1, 2, \dots M$ **do**
  if $\det(\Sigma_m) > 2\det(\Sigma_{\underline{m}})$ or $m = 1$ **then**
    $\underline{m} \leftarrow m$
    $\Sigma_{\underline{m}} \leftarrow \Sigma_m$
  end if
  Define $\pi_m : s \mapsto \arg\max_{a\in\mathcal{A}_s} \|\phi(s,a)\|_{\Sigma_{\underline{m}}^{-1}}$
  $\Sigma_{m+1} = \Sigma_m + \alpha\phi_m\phi_m^\top;\ \phi_m = \phi(s_m, \pi_m(s_m))$
end for
return policy mixture π_e of $\{\pi_1, \dots, \pi_M\}$

Algorithm 2 SAMPLER

Input: $\pi_e = \{\pi_1, \dots, \pi_M\}$, reg. λ_{reg}
Set $\mathcal{D}' = \emptyset$
for $n = 1, 2, \dots N$ **do**
  Receive context $s'_n \sim \mu$
  Sample $m \in [M]$ uniformly at random
  Select action $a'_n = \pi_m(s'_n)$
  Receive feedback reward r'_n
  Store feedback $\mathcal{D}' = \mathcal{D}' \cup \{s'_n, a'_n, r'_n\}$
end for
return dataset $\mathcal{D}'$

Our algorithm leverages this idea and consists of two subroutines: 1) the *planner* (Alg. 1) which operates on offline contexts and identifies a mixture policy π_e (this is the exploratory policy π_e mentioned in the introduction) 2) the *sampler* (Alg. 2) which runs π_e online to finally gather the dataset. This way, π_e is nonadaptive to the online data.

Planner The purpose of the planner is to use past contexts to compute the exploratory policy. The planner runs the reward-free version of LINUCB on the offline context dataset $\mathcal{C}$ as described earlier in this section. This way, the planner selects the action a in the current offline context s_m that maximizes the uncertainty encoded in $\|\phi(s_m,a)\|_{\Sigma_{\underline{m}}^{-1}}$ where $\Sigma_{\underline{m}}$ is a scaled, regularized, cumulative covariance over the contexts parsed so far and the actions selected. Note this procedure is possible since the state–action function $\phi(s_m,a)$ is assumed to be known for any input (s_m,a) pair, and no actual rewards are observed. The doubling schedule yields a short descriptor for the planner's policies $\{\pi_1, \dots, \pi_M\}$. The variable $\underline{m}$ indicates the last doubling update before iteration m.

A key choice is the parameter $\alpha \in (0,1]$ in the cumulative covariance matrix update. The rationale is that when $\alpha < 1$ each rank-one update $\phi_m\phi_m^\top$ to the cumulative covariance gets *discounted*. The smaller α is, the more offline samples the planner needs to get to a sufficiently positive definite covariance matrix Σ_M. This choice effectively averages the updates and increases the estimation accuracy of the planner's covariance with respect to its conditional expectation.

Sampler Upon termination, the planner identifies a sequence of policies $\pi_1, \dots, \pi_M$. Now consider the average policy π_e: every time an action is needed, π_e samples one index $m \in [M]$ uniformly at random and plays π_m. This is the policy that the sampler (Alg. 2) implements for $N = \alpha M \leq M$ fresh online contexts. Intuitively, executing the mixture policy π_e in the online phase produces the same covariance matrix as Σ_M (in expectation).

Sample Complexity Bounds				
		Small $\sum_s \mathcal{A}_s$	Large $\sum_s \mathcal{A}_s$	Large $\sum_s \mathcal{A}_s$
Offline Data	M	d^3/ϵ^2	d^3/ϵ^2	d^2/ϵ^2
Online Data	N	$(d \ln \sum_s \lvert\mathcal{A}_s\rvert)/\epsilon^2$	d^2/ϵ^2	d^2/ϵ^2
Regularization	λ_{reg}	1	1	d

Table 1: Sample complexity bounds (ignoring constants and log terms) to obtain an ϵ-optimal policy

Upon playing π_m in state s'_n, the corresponding reward is observed and the tuple $(\{s'_n, a'_n, r'_n\})$ is stored. Since π_e is the average policy played by the planner, we expect that running π_e on the online dataset produces a covariance matrix Σ'_N similar to the planner's Σ_M. In a sense, the covariance matrix is a proxy for the amount of information acquired by the algorithm. This means the sampler acquires the same information that the planner would have acquired if it was acting online. Since the planner is the reward-free LINUCB algorithm, we expect its policy to efficiently reduce our uncertainty over the reward parameters and learn a near optimal policy; our analysis will make this intuition precise. Note the sampler's policy is nonadaptive to the online stream of data.

Our work uses mixture policies; mixture policies are not essential, although they arise naturally from the way the algorithms are formulated.

5 Main Result

Theorem 1. *Fix $\epsilon > 0$, and consider running Alg. 1 for $M = \widetilde{\Omega}(\frac{d^2\beta}{\lambda_{reg}\epsilon^2})$ iterations and Alg. 2 for $N = \widetilde{\Omega}(\frac{d\beta}{\epsilon^2})$ iterations with regularization $\lambda_{reg} \in (\Omega(\ln(d/\delta)), d]$. Let $\widehat{\theta}, \widehat{\pi}$ be as in Eq.* (1)*. For any $\epsilon \leq 1$, with probability at least $1-\delta$ the expected maximum uncertainty*

$$\mathbb{E}_{s\sim\mu} \max_{a\in\mathcal{A}_s} |\phi(s,a)^\top(\theta^\star - \widehat{\theta})| \leq \epsilon \tag{5}$$

and the suboptimality of the greedy policy $\widehat{\pi}$ satisfies

$$\mathbb{E}_{s\sim\mu} \max_{a\in\mathcal{A}_s} \left(\phi(s,a) - \phi(s,\widehat{\pi}(s))\right)^\top \theta^\star \leq 2\epsilon. \tag{6}$$

In Table 1 we instantiate the bounds from Theorem 1 in different settings, ignoring constants and log terms. This highlights different tradeoffs between the regularization λ_{reg} and the number of offline contexts ($M = N/\alpha$) needed to achieve the online minimax sample complexity lower bound $\approx \min\{d, \ln \sum_s |\mathcal{A}_s|\} \times d/\epsilon^2$ (Chu et al., 2011; Abbasi-Yadkori et al., 2011). [5] In the large action regime $\ln|\sum_s \mathcal{A}_s| \gtrapprox d$, a regularization level $\lambda_{reg} \approx d$ gives the optimal online sample complexity $N \approx d^2/\epsilon^2$ while requiring a context dataset of the same size ($M \approx d^2/\epsilon^2$, by choosing $\alpha = 1$).

More often, a lower level of regularization can be preferable to introduce less bias. For example $\lambda_{reg} = 1$ is a common choice in linear bandits (Abbasi-Yadkori et al., 2011; Chu et al., 2011). When the cumulative covariance matrix is less regularized, we may need to compensate with additional offline data to ensure the covariance matrices are accurately estimated, which is achieved by setting $M = N/\alpha \approx dN$ (i.e., $\alpha \approx 1/d$). In this way, additional samples are collected by the planner to maintain its covariance estimation accuracy (and hence its planning accuracy) despite the lower regularization.

Note that our algorithm achieves optimal sample complexity for the online data. Even with adaptive policies, the best known minimax upper bound is d^2/ϵ (Abbasi-Yadkori et al., 2011; Chu et al., 2011). Although Chu et al. (2011); Abbasi-Yadkori et al. (2011) consider regret minimization instead of sample complexity. But their results can be easily translated to sample complexity bounds (See Lemma 3 in Appendix A.2, and also Jin et al. (2018, Section 3.1)). However, the optimal rate of the number of offline data is unclear. We believe our results can be improved by new tools in matrix concentration, and we leave this direction for future work. Interestingly, in our experiments the algorithm performed well numerically even when $\lambda_{reg} < 1$ and $\alpha = 1$.

[5] The minimax lower bounds in the literature are stated for the regret setting, but can nonetheless be adapted to derive sample complexity results. See Appendix A.1 for detailed discussion.

More generally, the amount of regularization should be set following the classical bias-variance tradeoff; its choice is mostly a statistical learning question, and its optimal value is normally problem dependent. As a result, the problem dependent tradeoff between different values of λ_{reg} is not reflected in our minimax analyses but can be appreciated in our numerical experiments on the real world dataset.

6 Proof Sketch

We give a brief proof sketch where we ignore constants and log factors; the full analysis can be found in the appendix. First we introduce some notation. We define the scaled cumulative covariance Σ_M at step m for the planner and the cumulative covariance Σ'_n for the sampler at step n:

$$\Sigma_m = \alpha \sum_{j=1}^{m-1} \phi(s_j, a_j)\phi(s_j, a_j)^\top + \lambda_{reg} I, \qquad \Sigma'_n = \sum_{j=1}^{n-1} \phi(s'_j, a'_j)\phi(s'_j, a'_j)^\top + \lambda_{reg} I \tag{7}$$

where the planner's j actions is $a_j = \pi_j(s_j)$ and likewise the sampler's j actions is $a'_j \sim \pi_e(s'_j)$.

Uncertainty Stochastic Process We represent the value of the uncertainty though a stochastic process. Let us define the filtration for the planner $\mathcal{F}_m = \sigma(s_1, \dots, s_{m-1})$ at stage m and for the sampler $\mathcal{F}'_n = \sigma(s_1, \dots, s_M, s'_1, \dots, s'_{n-1})$ at stage n to represent the amount of information available. Let us also define the observed uncertainty U_m for the planner and likewise U'_n for the sampler:

$$U_m \stackrel{def}{=} \max_{a \in \mathcal{A}_{s_m}} \|\phi(s_m, a)\|_{(\Sigma_m)^{-1}}, \qquad U'_n \stackrel{def}{=} \max_{a \in \mathcal{A}_{s_n}} \|\phi(s'_n, a)\|_{(\Sigma'_n)^{-1}}. \tag{8}$$

We ultimately want to bound $\mathbb{E}'_n U'_n = \mathbb{E}[U'_n \mid \mathcal{F}_n] = \mathbb{E}_{s\sim\mu} \max_a \|\phi(s,a)\|_{(\Sigma'_n)^{-1}}$ when $n = N$, i.e., the average uncertainty at the end, see Section 3. Likewise, let us define $\mathbb{E}_m U_m = \mathbb{E}[U_m \mid \mathcal{F}_m] = \mathbb{E}_{s\sim\mu} \max_a \|\phi(s,a)\|_{(\Sigma_m)^{-1}}$.

We show that $\mathbb{E}'_N U'_N$ is minimized in two steps: 1) we show that the planner's uncertainty $\mathbb{E}_M U_M$ can be bounded 2) since the sampler implements the planner's average policy, $\mathbb{E}'_N U'_N$ cannot be too far from $\mathbb{E}_M U_M$.

We highlight that $M \geq N$ in general, i.e., the stochastic processes proceed at different speed. In fact, the planner needs more data than the sampler ($M \geq N$) as the planner's policy is reactive to the offline contexts; this forces us to use more data-intensive concentration inequalities for the planner.

6.1 Offline Uncertainty

The next lemma is formally presented in Lemma 5 and Lemma 7 in the appendix and examines the reduction in the planner's expected uncertainty $\mathbb{E}_M U_M$.

Lemma 1 (Offline Expected Uncertainty)**.** *With high probability we have*

$$\mathbb{E}_M U_M \leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_m U_m \lessapprox \frac{1}{M} \sum_{m=1}^{M} U_m \lessapprox \sqrt{\frac{d}{\alpha M}}.$$

Proof sketch. We start by modifing a result from the classical bandit literature (e.g., Abbasi-Yadkori et al. (2011)) to bound the average realized uncertainty U_m while accounting for the extra α scaling factor contained in the covariance matrix Σ_M:

$$\frac{1}{M} \sum_{m=1}^{M} U_m = \frac{1}{M} \sum_{m=1}^{M} \|\phi(s_m, a_m)\|_{\Sigma_m^{-1}} \lessapprox \sqrt{\frac{d}{\alpha M}}.$$

The parameter α on the right hand side above arises from its inclusion in the definition of cumulative covariance Eq. 7. This justifies the last inequality in the lemma's statement.

While we bounded $\frac{1}{M}\sum_{m=1}^{M} U_m$, we need to bound the average conditional expectations $\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_m U_m$. Bernstein's inequality would bound the actual deviation $\sum_{m=1}^{M} U_m$ by the conditional variances (or expectations $\sum_{m=1}^{M} \mathbb{E}_m U_m$ given our boundness assumptions); here we need the opposite, and so we 'reverse' the inequality in Theorem 3 *(Reverse Bernstein for Martingales)* to claim $\sum_{m=1}^{M} \mathbb{E}_m U_m \lessapprox \sum_{m=1}^{M} U_m$ with high probability to conclude. □

6.2 Online Uncertainty

In the prior section we showed that the planner would be successful in reducing its final uncertainty $\mathbb{E}_M U_M$ if it was collecting reward information; moreover, its sequence of policies $\pi_1, \dots, \pi_M$ only depends on the observed contexts. Because of this, if the sampler runs the planner's average policy π_e, we would expect a similar reduction in the uncertainty $\mathbb{E}'_N U'_N$ after $N = \alpha M$ iterations (the α factor simply arises because the planner's information are discounted by α). The argument is formalized in Lemma 4 *(Relations between Offline and Online Uncertainty)* in the appendix, which we preview here. We let K be the number of policy switches by the sampler which is bounded by $K \lessapprox d$ in Lemma 16 *(Number of Switches)* in appendix.

Lemma 2 (Relations between Offline and Online Uncertainty)**.** *If $\lambda_{reg} = \Omega(\ln \frac{d}{\delta})$ and $M = \Omega\left(\frac{KN}{\lambda_{reg}} \ln \frac{dNK}{\lambda_{reg}\delta}\right)$, upon termination of Alg. 1 and 2 it holds with high probability that*

$$\mathbb{E}'_N U'_N \lesssim \mathbb{E}_M U_M.$$

Proof. Let $d_1, \dots, d_M$ be the conditional distributions of the feature vectors sampled at timesteps $1, \dots, M$ in Alg. 1 after the algorithm has terminated. Conditioned on $\mathcal{F}_M$, the d_i's are non-random. Then the conditional expected covariance matrix given $\mathcal{F}_M$ can be defined as

$$\overline{\Sigma} = \alpha \sum_{i=1}^{M} \mathbb{E}_{\phi \sim d_i} \phi\phi^\top + \lambda_{reg} I \tag{9}$$

Now, our argument relies on some matrix concentration inequalities which hold if

$$\lambda_{reg} = \Omega(\ln \frac{d}{\delta}), \qquad \frac{1}{\alpha} = \Omega\left(\frac{K}{\lambda_{reg}} \ln \frac{dNK}{\lambda_{reg}\delta}\right). \tag{10}$$

Precisely, from Lemma 14 *(Matrix Upper Bound Offline Phase)* and Lemma 15 *(Matrix Upper Bound Online Phase)* with high probability we have that

$$\forall x, \|x\|_2 \le 1: \qquad \|x\|_{(\Sigma'_N)^{-1}} \lesssim \|x\|_{\overline{\Sigma}^{-1}} \qquad \text{and} \qquad \|x\|_{\overline{\Sigma}^{-1}} \lesssim \|x\|_{\Sigma_M^{-1}}. \tag{11}$$

In words, we can relate the planner's (random) scaled covariance Σ_M to its conditional expectation $\overline{\Sigma}$, with the guarantee that it won't be very different from the sampler's covariance Σ'_N (which implements the planner's policy). These concentration inequalities are key to the proof and are proved in the appendix.

Define the policy maximizing the online uncertainty $\pi'_n(s) = \arg\max_{a \in \mathcal{A}_s} \|\phi(s,a)\|_{(\Sigma'_n)^{-1}}$. Under the event in Eq. (11) we can write

$$\begin{aligned}\mathbb{E}'_N U'_N &\stackrel{def}{=} \mathbb{E}_{s\sim\mu} \max_{a\in\mathcal{A}_s} \|\phi(s,a)\|_{(\Sigma'_N)^{-1}} = \mathbb{E}_{s\sim\mu} \|\phi(s,\pi'_N(s))\|_{(\Sigma'_N)^{-1}} \lesssim \mathbb{E}_{s\sim\mu} \|\phi(s,\pi'_N(s))\|_{\overline{\Sigma}^{-1}} \\ &\lesssim \mathbb{E}_{s\sim\mu} \|\phi(s,\pi'_N(s))\|_{\Sigma_M^{-1}} \le \mathbb{E}_{s\sim\mu} \max_a \|\phi(s,a)\|_{\Sigma_M^{-1}} = \mathbb{E}_M U_M.\end{aligned}$$

□

6.3 Conclusion and Tradeoffs

We can now tune the parameters α, λ_{reg}. Starting from Eq. (3), we can write

$$\mathbb{E}_{s\sim\mu} \max_a \phi(s,a)^\top (\theta^\star - \widehat{\theta}) \le \sqrt{\beta}\, \mathbb{E}'_N U'_N \lesssim \sqrt{\beta}\, \mathbb{E}_M U_M \lesssim \sqrt{\frac{\beta d}{\alpha M}} = \sqrt{\frac{\beta d}{N}}. \tag{12}$$

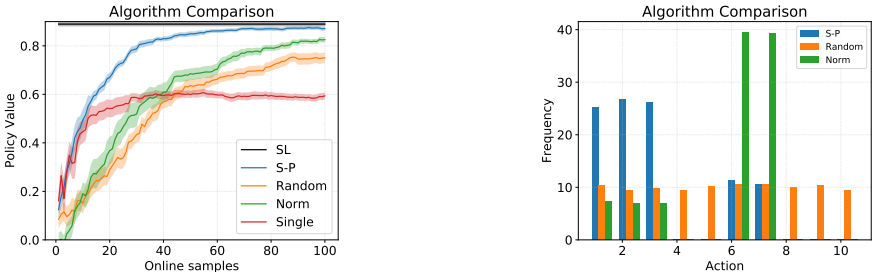
Thus, the rate-optimal online sample complexity $N \approx \frac{\beta d}{\epsilon^2}$ always suffices. However, we need a different number of offline contexts depending on the regularization that we want to adopt. Since $K \lessapprox d$, the reader can verify that if $\lambda_{reg} \approx d$ then the preconditions in Eq. (10) are verified with $M \approx N$, so $\alpha \approx 1$. If a lower value for the regularization $\lambda \approx 1$ is desired, $M \approx dN$ would be required (thus $\alpha \approx \frac{1}{d}$).

The reason why we might need $M \ge N$ (i.e., $\alpha \le 1$) is that the planner's policy is adaptive to the observed offline contexts. This means that the planner's cumulative covariance Σ_M is not a sum of i.i.d. rank-one random matrices; in this case, we need either a more data-hungry concentration inequality or heavier regularization, leading to the the precondtion on α in Eq. (10).

7 Experiments

We now study the empirical properties of the planner and sampler in a synthetic setting and a real-world dataset from the Yahoo! Learning to Rank challenge (Chapelle and Chang, 2011). We also analyze its sensitivity to regularization. Additional details are in the appendix.

7.1 Synthetic setting



(a) (b)

Figure 2: Synthetic Setting. In the left figure, the policy value of the sampler-planner (S-P) is compared with a random algorithm (Random), a largest norm algorithm (Norm), and an algorithm that always chooses the same single action (Single). Supervised learning (SL) represents an approximate upper bound; however, we don't expect the algorithms to be able to reach the SL performance as they all receive bandit feedback compared to the full feedback available to SL (10,000 samples). In the right figure, the empirical distributions over the actions chosen by the algorithms are compared, with the exception of Single, which always chooses action 1.

We first conduct a synthetic validation experiment to verify our algorithms behave consistently with the intuition and theory previously described. We constructed a linear contextual bandit with $d = 20$ and $\mathcal{A} = \{1, \ldots, 10\}$. At each time, features were generated randomly through the following procedure. A category I is sampled uniformly from the set $\{1, 2, 3\}$. For actions $a = 1, 2, 3$ and categories $i \in \{1, 2, 3\}$, we defined diagonal covariance matrices $\Sigma_{a,i} \in \mathbb{R}^{d \times d}$ such that $\Sigma_{a,i}$ has a single coordinate (determined by the pair (a, i)) along the diagonal equal to 1 while others are equal to 10^{-9}. Given the category I, features are distributed as $\phi(s, a) \sim \mathcal{N}(0, \Sigma_{a,I})$. This ensures variance only in certain directions depending on the category sampled and action taken. Independent of the category, for actions $a = 6, 7$, features are distributed as $\phi(s, a) \sim \mathcal{N}(0, \mathrm{diag}(10^{-9}, \ldots, 10^{-9}, 5)$ independently. The last coordinate has large variance regardless of the other actions. Other actions yield features that are zero. The first $d - 1$ coordinates of θ were chosen randomly from $\{-1, 1\}^{d-1}$ and the d-th coordinate is zero.

We considered four algorithms: (1) a random algorithm that chooses actions uniformly from $\mathcal{A}$, (2) a largest norm algorithm that chooses the feature with the largest norm, (3) an algorithm that chooses action 1 and (4) our sampler-planner algorithm. The planner uses an independent context-only dataset of the same size as the one to be used for training. All algorithms are then run in the online phase with reward feedback. The algorithms are evaluated based on their resulting policy values during the online phase on a held-out test set. Policy value is defined as $\mathbb{E}_s \phi(s, \hat{\pi}(s))^\top \theta^\star$ where $\hat{\pi}$ is the learned policy for a given algorithm. All algorithms used $\lambda = 1$ and the planner used $\alpha = 1$ as these worked well. See appendix for more λ settings.

Figure 2 shows the policy values plotted over the number of samples taken and also the action distributions of the algorithms. Each line/bar represents the mean/standard error of 20 trials. Most actions are designed to be uninformative in this setting and it is beneficial to select judiciously from actions $\{1, 2, 3, 6, 7\}$. A random algorithm should and does perform poorly.The largest norm algorithm also fails: while much is learned about the last coordinate of θ by typically taking actions $\{6, 7\}$, little is learned about the others which are necessary for making good decisions. Finally the single action algorithm fails to account for the categories.

7.2 Learning to rank dataset

We evaluate the planner-sampler performance on real-world data from the Yahoo! Learning to Rank challenge (Chapelle and Chang, 2011), which has been used previously (Foster et al., 2018).

The ranking dataset is structured as follows. Each datapoint (row) is a 700-dimensional feature vector associated with a ranking relevance score that takes on values in $\{0, 1, 2, 3, 4\}$. Each datapoint is also associated with a single context, termed a query. Multiple datapoints matched with the same query correspond to actions (documents) that the learner can choose when presented with that query. The dataset is already divided into training, validation, and testing data.

The objective is to choose actions, among those presented for a given query, that maximize the relevance scores. Unlike the previous synthetic dataset, the reward function here is misspecified, i.e., it does not exactly follow a linear model. To reduce the computational burden, we randomly subsampled coordinates of the original 700-dimensional features down to 300-dimensional features. We then normalized them to ensure the norms are at most 1. Similar to prior work, we also limit the number actions (documents) available at any given context to $K = 20$.

As before, we compare the planner-sampler pair to a random algorithm and a supervised learning oracle that observes all the context-action features and relevance scores, which is 200,000 training samples in total. To simulate streaming the data online to the algorithms, we sequentially iterate through the contexts and present the associated features to the learner as actions. During the offline phase, we run the planner on the validation set where none of the relevance scores are observed in order to generate the sampling policies. During the online phase, we run both the sampler and the random algorithm on the training data such that they observe the same contexts but may take different actions. The maximum number of distinct contexts we can iterate over is about 3000. The resulting policies are evaluated on the full test set after every 20 training observations. This experiment was repeated with varying values of the regularization parameter $\lambda \in \{0.1, 1, 10\}$. The planner uses $\alpha = 1$; we did not find significant improvements with $\alpha < 1$. Additional values for λ are reported in the supplementary material.

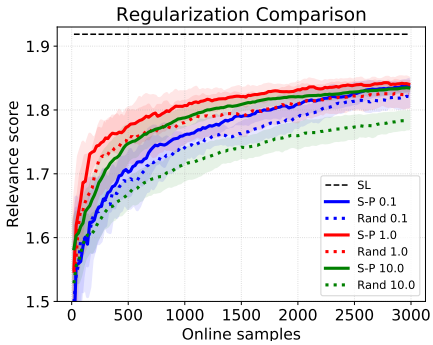


Figure 3: The sampler-planner (S-P) pair is compared with a random sampler (Rand) with varying levels of regularization λ. The supervised learning oracle (SL), which observes state-action pairs in the training set, is shown as baseline.

Each point/band on Figure 3 represents the mean/standard deviation of 10 independent trials. Randomness is due only to internal randomness in the algorithms and the randomized order in which the contexts are revealed. The sampler-planner outperforms the random algorithm in most regimes, including $\lambda = 1$ where the pinnacle performance of both algorithms is achieved. The gap between the algorithms is widened for larger choices of λ. For $\lambda = 0.1$ the curves appear to catch up as more data is gathered; it is possible that they would outperform a larger value of regularization if more samples were available. The optimal choice for λ depends on the bias-variance tradeoff. Despite misspecification, our sampler-planner approach consistently matches or significantly exceeds the performance of the baseline approach, particularly when the number of samples is small.

8 Conclusion

Assuming access to a set of offline contexts, we presents an algorithm to find a single stochastic policy to acquire data strategically. The procedure is validated by a theoretical analysis and by experiments, even when the reward model is not linear. In the future it would be interesting to create an algorithm that works across all values of regularization without impacting the amount of offline context data required, and to develop methods to incorporate prior reward data to reduce the sample complexity of the online data needed to learn a good policy, perhaps through instance-dependent bounds that quantify the benefit of particular prior context-action-reward data in a specific domain.

Our work presents little direct negative potential societal impacts though whether contextualized decision policies are beneficial or harmful depends on the setting and the reward function chosen.

Acknowledgments

This work is supported in part by NSF Grant #2112926 and GFRP (to JL), and a Toyota gift.

References

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*.

Bai, Y., Xie, T., Jiang, N., and Wang, Y. X. (2019). Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32.

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

Chapelle, O. and Chang, Y. (2011). Yahoo! learning to rank challenge overview. In *Proceedings of the learning to rank challenge*, pages 1–24. PMLR.

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

Degenne, R., Ménard, P., Shang, X., and Valko, M. (2020). Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR.

Deshmukh, A. A., Sharma, S., Cutler, J. W., Moldwin, M., and Scott, C. (2018). Simple regret minimization for contextual bandits. *arXiv preprint arXiv:1810.07371*.

Esfandiari, H., Karbasi, A., Mehrabian, A., and Mirrokni, V. (2019). Regret bounds for batched bandits. *arXiv preprint arXiv:1910.04959*.

Fiez, T., Jain, L., Jamieson, K., and Ratliff, L. (2019). Sequential experimental design for transductive linear bandits. *arXiv preprint arXiv:1906.08399*.

Foster, D., Agarwal, A., Dudik, M., Luo, H., and Schapire, R. (2018). Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR.

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

Jedra, Y. and Proutiere, A. (2020). Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020). Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning (ICML)*.

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

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lattimore, T. and Szepesvari, C. (2020). Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning (ICML)*.

Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648.

Ren, Z., Zhou, Z., and Kalagnanam, J. R. (2020). Batched learning in generalized linear contextual bandits with general decision sets. *IEEE Control Systems Letters*.

Ruan, Y., Yang, J., and Zhou, Y. (2020). Linear bandits with limited adaptivity and learning distributional optimal design. *arXiv preprint arXiv:2007.01980*.

Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 828–836.

Tao, C., Blanco, S., and Zhou, Y. (2018). Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pages 4877–4886. PMLR.

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

Wang, R., Du, S. S., Yang, L. F., and Salakhutdinov, R. R. (2020a). On reward-free reinforcement learning with linear function approximation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wang, R., Salakhutdinov, R., and Yang, L. F. (2020b). Provably efficient reinforcement learning with general value function approximation.

Wang, T., Zhou, D., and Gu, Q. (2021). Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *arXiv preprint arXiv:2101.02195*.

Xu, L., Honda, J., and Sugiyama, M. (2018). A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR.

Zanette, A., Cheng, C.-A., and Agarwal, A. (2021). Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*.

Contents

1 Introduction **1**

2 Related Work **3**

3 Setup **4**

3.1 Objective and Error Decomposition 4

4 Algorithms **5**

5 Main Result **6**

6 Proof Sketch **7**

6.1 Offline Uncertainty 7

6.2 Online Uncertainty 8

6.3 Conclusion and Tradeoffs 8

7 Experiments **9**

7.1 Synthetic setting 9

7.2 Learning to rank dataset 9

8 Conclusion **10**

A Sample Complexity Results from Existing Literature **14**

A.1 Sample Complexity Lower Bounds 14

A.2 Sample Complexity Upper Bounds 14

B Proof of Theorem 1 **14**

C Linear Regression **17**

D Bounding the Uncertainty **18**

E Matrix Concentration Inequalities **20**

E.1 Known Matrix Concentration Inequalities 20

E.2 Matrix Concentration Inequalities in All Directions 21

E.3 Relation Between Offline and Online Covariance Matrices 24

F Scalar Concentration Inequalities for Martingales **26**

F.1 Bernstein Inequality for Martingales 26

F.2 Reverse Bernstein Inequality for Martingales 27

G Additional Remarks **28**

G.1 Hard Instance that Requires Exploration 28

G.2 Remarks Regarding Ruan et al. (2020) 29

H Helper Lemmas **30**

I Additional Experiments and Information **30**

I.1 Synthetic Dataset 30

I.2 Yahoo! Learning to Rank Dataset 30

I.3 Additional Plots 30

A Sample Complexity Results from Existing Literature

In this section we discuss how to adapt existing linear bandit results (mostly in the regret setting) to sample complexity results.

A.1 Sample Complexity Lower Bounds

For the large action regime (i.e., $\min\{\ln \sum_s |\mathcal{A}_s|, d\} = d$), the lower bound $\approx d^2/\epsilon^2$ follows by using the same hypercube construction as Lattimore and Szepesvári (2020, Theorem 24.1).

For the small action regime (i.e., $\min\{\ln \sum_s |\mathcal{A}_s|, d\} = \ln \sum_s |\mathcal{A}_s|$). If $O(|A|) = d$, and we ignore the $\ln \sum_s |\mathcal{A}_s|$ term in sample complexity, the lower bound $\approx d/\epsilon^2$ follows by considering a bandit instance with (d) orthogonal arms in $\mathbb{R}^d$. This is equivalent to a multi armed bandit problem, and the sample complexity lower bound then follows from Mannor and Tsitsiklis (2004, Theorem 1).

A.2 Sample Complexity Upper Bounds

The upper bounds come from an instantiations of a general reduction in Jin et al. (2018, Section 3.1).

Lemma 3 (Regret to sample complexity)**.** *Running the algorithm in Abbasi-Yadkori et al. (2011) for $\Omega(\frac{d^2}{\epsilon^2})$ steps induces a ϵ-optimal policy.*

Proof. First of all, Abbasi-Yadkori et al. (2011) achieve a regret after N timesteps of at most $d\sqrt{N}$, meaning that $\sum_{t=1}^{N}(r(s_t, \pi^\star(s_t)) - r(s_t, \pi_t(s_t))) \leq O(d\sqrt{N})$.

Note that π_t is independent of s_t, by the Azuma martingale concentration bound we get $\sum_{t=1}^{N}(\mathbb{E}_s r(s, \pi^\star(s)) - \mathbb{E}_s r(s, \pi_t(s))) \leq O(d\sqrt{N})$. Consider the average policy $\pi(s) = \pi_t(s)$ w.p. $1/N$. We get $\mathbb{E}_\pi \mathbb{E}_s[r(s, \pi(s))] \geq \mathbb{E}_s[r(s, \pi^\star(s))] - O(d/\sqrt{N})$. So when $N \geq \Omega(\frac{d^2}{\epsilon^2})$ the average policy played from timestep 1 to N is at most ϵ-suboptimal. □

B Proof of Theorem 1

Proof. We start by invoking Proposition 2 *(Linear Regression with Union Bound)*. Define a shorthand for the RHS of proposition 2

$$\sqrt{\beta} \stackrel{def}{=} \min\{\alpha_1, \alpha_2\} + \sqrt{\lambda_{reg}}\|\theta^\star\|_2 \tag{13}$$

With a rescaling on δ we write

$$\mathbf{P}\left(\underbrace{\forall (s,a): \quad |\phi(s,a)^\top(\theta^\star - \widehat{\theta})| \lesssim \|\phi(s,a)\|_{(\Sigma'_N)^{-1}} \ln(N)\sqrt{\beta}}_{\mathcal{E}_1}\right) \geq 1 - \frac{\delta}{4}. \tag{14}$$

Notice that under the event $\mathcal{E}_1$ above

$$\mathbb{E}_{s\sim\mu} \max_{a\in\mathcal{A}_s} |\phi(s,a)^\top(\theta^\star - \widehat{\theta})| \leq \ln(N)\sqrt{\beta}\, \mathbb{E}_{s\sim\mu} \max_{a\in\mathcal{A}_s} \|\phi(s,a)\|_{(\Sigma'_N)^{-1}} = \ln(N)\sqrt{\beta} u'_N$$

Recall that K is the number of policy switches by the sampler and Lemma 16 gives $K = \widetilde{O}(d)$. We set

$$\lambda_{reg} = \Omega(\ln(d/\delta)), \tag{15}$$

$$M = \Omega\left(\frac{KN}{\lambda_{reg}} \ln \frac{d^2 N}{\lambda_{reg}\delta}\right), \tag{16}$$

then Lemma 4 *(Relations between Offline and Online Uncertainty)* relates the expected online uncertainty u'_N with the expected offline uncertainty u_M

$$\mathbf{P}\left(\underbrace{u'_N \lesssim u_M}_{\mathcal{E}_2}\right) \geq 1 - \frac{\delta}{2} \tag{17}$$

Finally, the expected offline uncertainty is bounded by in Lemma 5 *(Offline Expected Uncertainty)* chained with Lemma 7 *(Sum of Observed Uncertainties)*

$$\mathbf{P}\left(\underbrace{u_M \lesssim \frac{1}{M}\ln\frac{1}{\delta} + \sqrt{\frac{1}{\alpha M} d \ln\left(\frac{d\lambda_{reg} + M}{d}\right)}}_{\mathcal{E}_3}\right) \geq 1 - \frac{\delta}{4}. \tag{18}$$

A union bound over the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ and chaining the statements now produces with probability at least $1 - \delta$

$$\mathbb{E}_{s\sim\mu} \max_{a\in\mathcal{A}_s} |\phi(s,a)^\top(\theta^\star - \widehat{\theta})| \lesssim \frac{\ln(N)\sqrt{\beta}}{M}\ln\frac{1}{\delta} + \ln(N)\sqrt{\frac{\beta}{\alpha M} d \ln\left(\frac{d\lambda_{reg} + M}{d}\right)} \tag{19}$$

$$\lesssim \frac{\ln(N)\sqrt{\beta}}{M}\ln\frac{1}{\delta} + \ln(N)\sqrt{\frac{\beta}{N} d \ln\left(\frac{d\lambda_{reg} + M}{d}\right)} \tag{20}$$

after recalling $\alpha = N/M$. As a result, we set

$$N = \Omega\left(\frac{d\beta}{\epsilon^2}\ln^3\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right)$$

and

$$M = \Omega\left(\frac{Kd\beta}{\lambda_{reg}\epsilon^2}\ln^4\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right).$$

As a result, we satisfy the preconditions in Eq. (16) because

$$\begin{aligned}
&\frac{KN}{\lambda_{reg}}\ln\frac{d^2N}{\lambda_{reg}\delta} \\
&= \frac{K\frac{d\beta}{\epsilon^2}\ln^3\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)}{\lambda_{reg}} \ln\left(\frac{d^2}{\lambda_{reg}\delta}\frac{d\beta}{\epsilon^2}\ln^3\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right) \\
&\lesssim \frac{Kd\beta}{\lambda_{reg}\epsilon^2}\ln^3\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right) 3\ln\left(\frac{d}{\lambda_{reg}\delta}\frac{d\beta}{\epsilon}\ln\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right) \quad \text{(Note that } \tfrac{d\beta}{\lambda_{reg}\delta} \geq 1\text{)} \\
&\lesssim \frac{Kd\beta}{\lambda_{reg}\epsilon^2}\ln^4\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right).
\end{aligned}$$

Note that $\epsilon \leq Kd\sqrt{\beta}/\lambda_{reg}$. Then

$$\frac{\ln(N)\sqrt{\beta}}{M}\ln\frac{1}{\delta} \lesssim \frac{\epsilon\ln(N)}{\ln^4\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)}\ln\frac{1}{\delta} \tag{21}$$

By the facts $\ln(x\ln(y)) \leq \ln(xy) \leq 2\ln(\max(x,y))$ and $\ln(x^k) = k\ln(x)$, we have

$$\ln(N) = \ln\left(\frac{d\beta}{\epsilon^2}\ln^3\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right) \leq 3\ln\left(\frac{d\beta}{\epsilon}\ln\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right) \tag{22}$$

$$\leq 6\ln\left(\max\left(\frac{d\beta}{\epsilon}, \lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right) \leq 6\ln\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)., \tag{23}$$

since λ is at most d so $\frac{d}{\lambda\delta} \geq 1$. Continuing Eq. (21), we get

$$\frac{\ln(N)\sqrt{\beta}}{M}\ln\frac{1}{\delta} \lesssim \frac{\epsilon\ln(N)}{\ln^4\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)}\ln\frac{1}{\delta} \lesssim \epsilon. \tag{24}$$

Note also that under this choice for N the right most term of Eq. (20) can be upper bounded by

$$\ln(N)\sqrt{\frac{\beta}{N}d\ln\left(\frac{d\lambda_{reg}+M}{d}\right)} \lesssim \epsilon\ln(N)\sqrt{\ln^{-3}\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\ln\left(\frac{d\lambda_{reg}+M}{d}\right)}. \tag{25}$$

Recall that $K = \widetilde{O}(d) \leq O(d^2)$. The logarithmic term can be upper bounded by

$$\ln\left(\frac{d\lambda_{reg}+M}{d}\right) = \ln\left(\lambda_{reg} + \frac{K\beta}{\lambda_{reg}\epsilon^2}\ln^4\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right) \tag{26}$$

$$\lesssim \ln\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon^2}\ln^4\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right) \tag{27}$$

$$\leq 4\ln\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon}\ln\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\right) \tag{28}$$

$$\leq 8\ln\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right). \tag{29}$$

Now we continue with Eq. (25)

$$\begin{aligned}
&\epsilon\ln(N)\sqrt{\ln^{-3}\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\ln\left(\frac{d\lambda_{reg}+M}{d}\right)} \\
\lesssim\ &\epsilon\ln(N)\sqrt{\ln^{-2}\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)} \\
\lesssim\ &\epsilon\ln\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)\sqrt{\ln^{-2}\left(\lambda_{reg} + \frac{d^2\beta}{\lambda_{reg}\epsilon\delta}\right)} \lesssim \epsilon. \qquad \text{(By Eq. (23))}
\end{aligned}$$

Rescaling ϵ by a constant, we ensure

$$\mathbf{P}\left(\mathbb{E}_{s\sim\mu}\max_{a\in\mathcal{A}_s}|\phi(s,a)^\top(\theta^\star - \widehat{\theta})| \leq \epsilon\right) \geq 1-\delta. \tag{30}$$

Now we prove Eq. (6). Let $\pi^\star(s) \stackrel{def}{=} \arg\max_{a\in\mathcal{A}_s}\phi(s,a)^\top\theta^\star$. Define $\Delta(s) \stackrel{def}{=} \max_{a\in\mathcal{A}_s}|\phi(s,a)^\top(\theta^\star - \widehat{\theta})|$ for shorthand. It follows immediately that

$$\phi(s,\widehat{\pi}(a))^\top\theta^\star \geq \phi(s,\widehat{\pi}(a))^\top\widehat{\theta} - \Delta(s), \tag{31}$$

$$\phi(s,\pi^\star(a))^\top\theta^\star \leq \phi(s,\pi^\star(a))^\top\widehat{\theta} + \Delta(s). \tag{32}$$

By the definition of $\widehat{\pi}$, we have

$$\phi(s,\widehat{\pi}(a))^\top\widehat{\theta} \geq \phi(s,\pi^\star(a))^\top\widehat{\theta}. \tag{33}$$

Combining Eqs. (31), (32), and (33) we get

$$\phi(s,\widehat{\pi}(a))^\top\theta^\star \geq \phi(s,\pi^\star(a))^\top\theta^\star - 2\Delta(s). \tag{34}$$

Consequently,

$$\mathbb{E}_{s\sim\mu}\left[(\phi(s,\pi^\star(a)) - \phi(s,\pi(a)))^\top\theta^\star\right] \leq 2\,\mathbb{E}_{s\sim\mu}\left[\Delta(s)\right]. \tag{35}$$

As a result, under the same event indicated by Eq. (30), we get

$$\mathbb{E}_{s\sim\mu}\left[(\phi(s,\pi^\star(s)) - \phi(s,\widehat{\pi}(s)))^\top\theta^\star\right] \leq 2\epsilon, \tag{36}$$

which is exactly Eq. (6). □

C Linear Regression

Proposition 1 (Linear Regression Non-Adaptive Setting)**.** *Consider drawing n i.i.d. copies of ϕ_i from some fixed distribution, and define*

$$\widehat{\theta} = \left(\sum_{i=1}^{n} \phi_i \phi_i^\top + \lambda_{reg} I\right)^{-1} \sum_{i=1}^{n} \phi_i y_i$$

where

$$y_i = \phi_i^\top \theta^\star + \eta_i \tag{37}$$

for some fixed $\theta^\star$ and η_i is mean zero 1-subgaussian conditioned on ϕ_i. Then for any fixed vector x

$$\mathbf{P}\left(x^\top(\theta^\star - \widehat{\theta}) \leq \|x\|_{\Sigma^{-1}} \left(\sqrt{2\ln\frac{2}{\delta}} + \sqrt{\lambda_{reg}}\|\theta^\star\|_2\right)\right) \geq 1-\delta.$$

Proof.

$$x^\top(\theta^\star - \widehat{\theta}) = x^\top\theta^\star - x^\top\left(\sum_{i=1}^{n}\phi_i\phi_i^\top + \lambda_{reg}I\right)^{-1}\left(\sum_{i=1}^{n}\phi_i\left(\phi_i^\top\theta^\star + \eta_n\right) + \lambda_{reg}\theta^\star - \lambda_{reg}\theta^\star\right) \tag{38}$$

$$= -x^\top\Sigma^{-1}\sum_{i=1}^{n}\phi_i\eta_i + \lambda_{reg}x^\top\Sigma^{-1}\theta^\star. \tag{39}$$

Using Cauchy-Schwartz we have

$$\lambda_{reg}x^\top\Sigma^{-1}\theta^\star \leq \lambda_{reg}\|x\|_{\Sigma^{-1}}\|\theta^\star\|_{\Sigma^{-1}}. \tag{40}$$

We have that $\lambda_{reg}\|\theta^\star\|_{\Sigma^{-1}} \leq \sqrt{\lambda_{reg}}\|\theta^\star\|_2$. Since the feature vectors ϕ_i's are sampled from a fixed distribution, conditioned on the sampled state-actions ϕ_i, both the covariance matrix Σ is fixed and the noise η_i is independent and 1 sub-Gaussian. Define $v_i = x^T\Sigma^{-1}\phi_i\eta_i$. Then conditioned on the sampled state-actions ϕ_i, the vs are independent random $(x^T\Sigma^{-1}\phi_i)^2$ sub-Gaussian random variables and we can apply Hoeffding's inequality, conditioned on the observed state-action features ϕ_i:

$$\mathbf{P}\left(|\sum_{i=1}^{n}x^\top\Sigma^{-1}\phi_i\eta_i| \leq \sqrt{2\sum_{i=1}^{n}(x^\top\Sigma^{-1}\phi_i)^2\ln\frac{2}{\delta}} = \|x\|_{\Sigma^{-1}}\sqrt{2\ln\frac{2}{\delta}}\right) \geq 1-\delta. \tag{41}$$

Combining with the regularization part, we conclude. □

Proposition 2 (Linear Regression with Union Bound)**.** *In the same setting as proposition 1 we have that*

$$\mathbf{P}\left(\forall(s,a) : \phi(s,a)^\top(\theta^\star - \widehat{\theta}) \leq \|\phi(s,a)\|_{\Sigma^{-1}}\left(\min\{\alpha_1,\alpha_2\} + \sqrt{\lambda_{reg}}\|\theta^\star\|_2\right)\right) \geq 1-\delta.$$

where

$$\alpha_1 \stackrel{def}{=} \sqrt{2\ln 2|\sum_s \mathcal{A}_s| + \ln\frac{1}{\delta}} \tag{42}$$

$$\alpha_2 \stackrel{def}{=} 2\sqrt{2d\ln 6 + \ln\frac{1}{\delta}} \tag{43}$$

Proof. The proof essentially follows from taking the sharper of two results obtained as follows:

1. Invoke proposition 1 with a union bound over the state-action space (this gives the bound with the α_1 term)
2. This follows from a discretization argument; see e.g., (Lattimore and Szepesvári, 2020), in particular, their equation 20.3.

□

D Bounding the Uncertainty

We use the notation $u_m = \mathbb{E}_m U_m$ and $u'_n = \mathbb{E}'_n U'_n$ for short.

Lemma 4 (Relations between Offline and Online Uncertainty)**.** *If $\lambda_{reg} = \Omega(\ln \frac{d}{\delta})$ and $M = \Omega\left(\frac{KN}{\lambda_{reg}} \ln \frac{dNK}{\lambda_{reg}\delta}\right)$, upon termination of algorithms 1 and 2 it holds that*

$$\mathbf{P}\left(u'_N \lesssim u_M\right) \geq 1 - \frac{\delta}{2}$$

where the probability is over the offline context dataset $\mathcal{C} = \{s_1, \ldots, s_M\}$ and the online context dataset $\mathcal{C}' = \{s'_1, \ldots, s'_N\}$.

Proof. Let $d_1, \ldots, d_M$ be the conditional distributions of the feature vectors sampled at timesteps $1, \ldots, n$ in algorithm 1 after the algorithm has terminated. Conditioned on $\mathcal{F}_M = \sigma(s_1, \ldots, s_M)$, the d_i's are non-random; let

$$\overline{\Sigma} = \alpha \sum_{i=1}^{M} \mathbb{E}_{\phi \sim d_i} \phi\phi^\top + \lambda_{reg} I \tag{44}$$

be the conditional expectation of the cumulative covariance matrix and let

$$\Sigma'_n = \sum_{i=1}^{n} (\phi'_i)(\phi'_i)^\top + \lambda_{reg} I \tag{45}$$

be the cumulative covariance matrix experienced in algorithm 2 where $\phi'_i \stackrel{def}{=} \phi(s'_n, a'_n)$ is the sampled feature during the execution of algorithm 1. If $\lambda_{reg} = \Omega(\ln \frac{d}{\delta})$ and $M = \Omega\left(\frac{KN}{\lambda_{reg}} \ln \frac{dNK}{\lambda_{reg}\delta}\right)$ we obtain from Lemma 14 *(Matrix Upper Bound Offline Phase)* and Lemma 15 *(Matrix Upper Bound Online Phase)*

$$\mathbf{P}\left(\forall x, \|x\|_2 \leq 1 : \|x\|_{\overline{\Sigma}^{-1}} \leq 3 \|x\|_{\Sigma_M^{-1}}\right) \geq 1 - \frac{\delta}{4} \tag{46}$$

$$\mathbf{P}\left(\forall x, \|x\|_2 \leq 1 : \|x\|_{(\Sigma'_N)^{-1}} \leq 9 \|x\|_{\overline{\Sigma}^{-1}}\right) \geq 1 - \frac{\delta}{4}. \tag{47}$$

Define the policy maximizing the online uncertainty

$$\pi'_n(s) = \arg\max_{a \in \mathcal{A}_s} \|\phi(s,a)\|_{(\Sigma'_n)^{-1}}. \tag{48}$$

Under the two above events we can write

$$\begin{aligned}
u'_N &\stackrel{def}{=} \mathbb{E}_{s\sim\mu} \max_{a\in\mathcal{A}_s} \|\phi(s,a)\|_{(\Sigma'_N)^{-1}} && (49)\\
&= \mathbb{E}_{s\sim\mu} \|\phi(s,\pi'_N(s))\|_{(\Sigma'_N)^{-1}} && (50)\\
&\leq 9\, \mathbb{E}_{s\sim\mu} \|\phi(s,\pi'_N(s))\|_{\overline{\Sigma}^{-1}} && (51)\\
&\leq 27\, \mathbb{E}_{s\sim\mu} \|\phi(s,\pi'_N(s))\|_{\Sigma_M^{-1}} && (52)\\
&\leq 27\, \mathbb{E}_{s\sim\mu} \max_a \|\phi(s,a)\|_{\Sigma_M^{-1}} && (53)\\
&\leq 27\, \mathbb{E}_{s\sim\mu} \max_a \|\phi(s,a)\|_{\Sigma_{\underline{M}}^{-1}} && (54)\\
&= 27 u_M. && (55)
\end{aligned}$$

□

Lemma 5 (Offline Expected Uncertainty)**.** *We have that*

$$\mathbf{P}\left(u_M \leq \frac{1}{M}\sum_{m=1}^{M} u_m \lesssim \frac{1}{M}\Big[\ln\frac{1}{\delta} + \sum_{m=1}^{M} U_m\Big] \stackrel{def}{=} \frac{\mathcal{R}(M, \frac{\delta}{4})}{M}\right) \geq 1 - \frac{\delta}{4}.$$

Proof. Consider the event that the sum of the predictable means $\sum_{m=1}^{M} u_m = \sum_{m=1}^{M} \mathbb{E}[U_m \mid \mathcal{F}_m]$ does not deviate significantly from $\sum_{m=1}^{M} U_m$:

$$\mathcal{E}(\delta) \stackrel{def}{=} \left\{ \sum_{n=1}^{M} u_m \leq \frac{1}{4} \left(c_1(\delta) + \sqrt{c_1^2(\delta) + 4 \left(\sum_{m=1}^{M} U_m + c_2(\delta) \right)} \right)^2 \stackrel{def}{=} \mathcal{R}hs(M, \delta) \right\}. \quad (56)$$

Using Theorem 3 *(Reverse Bernstein for Martingales)* (which also defines c_1, c_2) we obtain

$$\mathbf{P}\left(\mathcal{E}(\delta/4)\right) \geq 1 - \frac{\delta}{4}. \quad (57)$$

From Lemma 6 *(Decreasing Uncertainty)* we know that the sequence $\{u_n\}_{n=1}^{N+1}$ is surely decreasing, which means that the last element must be less than the average:

$$u_M \leq \frac{1}{M} \sum_{n=1}^{M} u_m = \frac{\mathcal{R}hs(M, \delta/4)}{M} \quad (58)$$

where in particular the equality holds under the event $\mathcal{E}(\delta)$. Finally using Cauchy-Schwartz we conclude that under the same event

$$\sum_{n=1}^{M} u_m \lesssim c_1^2(\delta) + c_2(\delta) + \sum_{m=1}^{M} U_m \stackrel{def}{=} \mathcal{R}(M, \delta). \quad (59)$$

□

Lemma 6 (Decreasing Uncertainty)**.** *For every n it holds that*

$$u_{n+1} \leq u_n.$$

Proof. By linear algebra, we must have

$$\Sigma_{\underline{n+1}} \succeq \Sigma_{\underline{n}} \qquad \longrightarrow \qquad \Sigma_{\underline{n+1}}^{-1} \preceq \Sigma_{\underline{n}}^{-1}$$

Using the definitions we have:

$$\begin{aligned} u_{n+1} &= \mathbb{E}_{s\sim\mu} \max_{a\sim\mathcal{A}_s} \|\phi(s,a)\|_{\Sigma_{\underline{n+1}}^{-1}} && (60)\\ &\leq \mathbb{E}_{s\sim\mu} \max_{a\sim\mathcal{A}_s} \|\phi(s,a)\|_{\Sigma_{\underline{n}}^{-1}} && (61)\\ &= u_n. && (62) \end{aligned}$$

□

Lemma 7 (Sum of Observed Uncertainties)**.** *If $\lambda_{reg} \geq 1$ and and $\alpha \leq 1$ then*

$$D \stackrel{def}{=} \sum_{n=1}^{M} U_m \lesssim \sqrt{\frac{M}{\alpha} d \ln\left(\frac{d\lambda_{reg} + M}{d}\right)}. \quad (63)$$

Proof. Let $\{\phi_m\}$ be the feature vectors experienced during the offline phase; we have

$$\begin{aligned} \sum_{m=1}^{M} U_m &= \sum_{m=1}^{M} \|\phi_m\|_{\Sigma_{\underline{m}}^{-1}} && (64)\\ &= \sum_{m=1}^{M} \|\phi_m\|_{\left(\alpha \sum_{j=1}^{m} \phi_j \phi_j + \lambda_{reg} I\right)^{-1}} && (65)\\ &\leq \sqrt{M \sum_{m=1}^{M} \|\phi_m\|^2_{\left(\alpha \sum_{j=1}^{m} \phi_j \phi_j + \lambda_{reg} I\right)^{-1}}} && (66) \end{aligned}$$

$$= \sqrt{\frac{M}{\alpha} \sum_{m=1}^{M} \|\sqrt{\alpha}\phi_m\|^2_{\left(\sum_{j=1}^{m} \sqrt{\alpha}\phi_j \sqrt{\alpha}\phi_j + \lambda_{reg} I\right)^{-1}}} \tag{67}$$

$$\leq \sqrt{\frac{M}{\alpha} \times 3 \times \underbrace{\ln \frac{\det\left(\sum_{j=1}^{M} \sqrt{\alpha}\phi_j \sqrt{\alpha}\phi_j^\top + \lambda_{reg} I\right)}{\det\left(\lambda_{reg} I\right)}}_{\stackrel{def}{=} \mathcal{I}}} \tag{68}$$

where the first inequality follows from Cauchy-Schwartz. The second inequality follows from Lemma 8 *(Elliptical Potential Argument Lemma with Doubling)*, where the precondition for Lemma 8 *(Elliptical Potential Argument Lemma with Doubling)* is satisfied by Lemma 9 *(Maximum Determinant Ratio)* since $\alpha \leq 1$. Finally, to bound the information gain $\mathcal{I}$, note $||\sqrt{\alpha}\phi_i||_2 \leq 1$ since $\alpha \leq 1$. Then (Abbasi-Yadkori et al., 2011)'s Lemma 11

$$\begin{aligned} &\ln\det\left(\sum_{j=1}^{M} \sqrt{\alpha}\phi_j \sqrt{\alpha}\phi_j^T + \lambda_{reg} I\right) - \ln\det\left(\lambda_{reg} I\right)) \\ &\leq d\ln((\mathrm{Tr}(\lambda_{reg} I) + M)/d) - \ln\det(\lambda_{reg} I) \end{aligned}$$

Since

$$\ln\det\left(\lambda_{reg} I\right) = d\ln\left(\lambda_{reg}\right) \geq 1. \tag{69}$$

Then

$$\ln\det\left(\sum_{j=1}^{M} \sqrt{\alpha}\phi_j \sqrt{\alpha}\phi_j^\top + \lambda_{reg} I\right) \leq d\ln\left(\frac{d\lambda_{reg} + M}{d}\right) \tag{70}$$

□

Lemma 8 (Elliptical Potential Argument Lemma with Doubling)**.** *(see (Zanette et al., 2021, Lemma 36)) Let $x_1, \cdots, x_M$ be a sequence of vectors such that $\|x_i\|_2 \leq 1$. Define $\Sigma_m = \lambda_{reg} I + \sum_{i=1}^{m-1} x_i x_i^\top$. Suppose $\underline{m} \leq m$ satisfies $\det(\Sigma_m) \leq 4\det(\Sigma_{\underline{m}})$. Then we have*

$$\sum_{m=1}^{M} \|x_i\|_{\Sigma_{\underline{m}}^{-1}} \leq 3\ln\frac{\det(\Sigma_{M+1})}{\det(\lambda_{reg} I)}. \tag{71}$$

Lemma 9 (Maximum Determinant Ratio)**.** *(see (Zanette et al., 2021, Lemma 34)) Let $x_1, \cdots, x_M$ be a sequence of vectors such that $\|x_i\|_2 \leq 1$. and assume $\lambda \geq 1$. Define $\Sigma_m = \lambda_{reg} I + \sum_{i=1}^{m-1} x_i x_i^\top$. Then for $\underline{m} \leq m$ we have $\det(\Sigma_m) \leq 4\det(\Sigma_{\underline{m}})$.*

E Matrix Concentration Inequalities

In this section we present matrix concentration inequalities used in our proof.

E.1 Known Matrix Concentration Inequalities

The following result about eigenvalues lower and upper bounds is well known.

Lemma 10 (Theorem 1.1 of Tropp (2012))**.** *Let $X_1, \cdots, X_n$ be a sequence of independent, positive semi-definite, self-adjoint matrices with dimension d. Suppose $\lambda_{\max}(X_k) \leq R$ almost surely. Define $\mu_{\min} = \lambda_{\min}(\sum_k \mathbb{E}\, X_k)$ and $\mu_{\max} = \lambda_{\max}(\sum_k \mathbb{E}\, X_k)$. Then*

$$\Pr\left(\lambda_{\min}\left(\sum_{k=1}^{n} X_k\right) \leq (1-\delta)\mu_{\min}\right) \leq d\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \textit{ for } \delta \in [0,1), \tag{72}$$

$$\Pr\left(\lambda_{\max}\left(\sum_{k=1}^{n} X_k\right) \geq (1+\delta)\mu_{\max}\right) \leq d\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \textit{ for } \delta \geq 0. \tag{73}$$

We loosen the above inequalities to make them more amenable to a direct use.

Corollary 1. *In the setting of Lemma 10, with $\delta \in [0,1]$*

$$\Pr\left(\lambda_{\min}\left(\sum_{k=1}^{n} X_k\right) \leq (1-\delta)\mu_{\min}\right) \leq d\left(1-\frac{\delta^2}{2}\right)^{\mu_{\min}/R}, \tag{74}$$

$$\Pr\left(\lambda_{\max}\left(\sum_{k=1}^{n} X_k\right) \geq (1+\delta)\mu_{\max}\right) \leq d\left(1-\frac{\delta^2}{4}\right)^{\mu_{\max}/R}. \tag{75}$$

In addition, for any $\mu \geq \mu_{\max}$,

$$\Pr\left(\lambda_{\max}\left(\sum_{k=1}^{n} X_k\right) \geq 2\mu\right) \leq d\exp\left(-\mu/(4R)\right). \tag{76}$$

Proof. The first two inequalities follows from the fact that $\forall \delta \in [0,1]$

$$\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \leq 1-\frac{\delta^2}{4} \tag{77}$$

$$\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \leq 1-\frac{\delta^2}{2}. \tag{78}$$

Now we prove Eq. (76). Let $r = \mu/\mu_{\max}$. Since $r \geq 1$ we have

$$\Pr\left(\lambda_{\max}\left(\sum_{k=1}^{n} X_k\right) \geq (1+\delta)\mu\right) = \Pr\left(\lambda_{\max}\left(\sum_{k=1}^{n} X_k\right) \geq (r+r\delta)\mu_{\max}\right) \tag{79}$$

$$\leq \Pr\left(\lambda_{\max}\left(\sum_{k=1}^{n} X_k\right) \geq (1+r\delta)\mu_{\max}\right) \leq d\left(\frac{e^{r\delta}}{(1+r\delta)^{1+r\delta}}\right)^{\mu_{\max}/R}, \tag{80}$$

where the last inequality is due to Lemma 10.

By basic algebra we have $\frac{e^x}{(1+x)^{1+x}} \leq e^{-x/4}$ for all $x \geq 1$. As a result, let $\delta = 1$ we have

$$d\left(\frac{e^{r}}{(1+r)^{1+r}}\right)^{\mu_{\max}/R} \leq d\exp\left(-\frac{r\mu_{\max}}{4R}\right) = d\exp\left(-\frac{\mu}{4R}\right). \tag{81}$$

Therefore,

$$\Pr\left(\lambda_{\max}\left(\sum_{k=1}^{n} X_k\right) \geq 2\mu\right) \leq d\exp\left(-\frac{\mu}{4R}\right). \tag{82}$$

which completes the proof. □

E.2 Matrix Concentration Inequalities in All Directions

In the following development we need to 'sandwich' the cumulative matrix around its expectation; as a step towards this, we first derive concentration inequalities as a function of the minimum eigenvalue.

For the following lemma, see also Lemma 20 (Ruan et al., 2020).

Lemma 11 (Matrix Upper and Lower Bound with Minimum Eigenvalue)**.** *Let $X_1, \cdots, X_n \sim \mathcal{D}$ be i.i.d. samples from $\mathcal{D}$ where $X_k \in \mathbb{R}^{d\times d}$. Suppose X_k is positive semi-definite for all $k \in [n]$ and $\lambda_{\max}(X_k) \leq 1$ almost surely. Let $\lambda = \lambda_{\min}(\mathbb{E}_{X\sim\mathcal{D}} X) > 0$. Then for $\delta \in [0,1]$ we have*

$$\Pr\left(\frac{1}{n}\sum_{k=1}^{n} X_k \preccurlyeq (1+\delta)\,\mathbb{E}[X]\right) \geq 1-d\left(1-\frac{\delta^2}{4}\right)^{n\lambda} \tag{83}$$

$$\Pr\left(\frac{1}{n}\sum_{k=1}^{n} X_k \succcurlyeq (1-\delta)\,\mathbb{E}[X]\right) \geq 1-d\left(1-\frac{\delta^2}{2}\right)^{n\lambda}. \tag{84}$$

Proof. We prove the first inequality first. Let $\Sigma = \mathbb{E}[X]$ and define $Y = \Sigma^{-1/2} X \Sigma^{-1/2}$. Then using linearity of expectation we have

$$\mathbb{E}[Y_k] = \mathbb{E}[\Sigma^{-1/2} X_k \Sigma^{-1/2}] = \Sigma^{-1/2} \left(\mathbb{E}\, X_k\right) \Sigma^{-1/2} = \Sigma^{-1/2} \Sigma \Sigma^{-1/2} = I. \tag{85}$$

In addition,

$$\|Y_k\|_{\mathrm{op}} \leq \left\|\Sigma^{-1/2}\right\|_{\mathrm{op}} \|X_k\|_{\mathrm{op}} \left\|\Sigma^{-1/2}\right\|_{\mathrm{op}} \leq 1/\lambda$$

almost surely. As a result, $\lambda_{\max}(Y_k^\top) \leq 1/\lambda$. Now corollary 1 gives (with $\mu_{max} = \mu_{min} = n$)

$$\Pr\left(\lambda_{\max}\left(\sum_{k=1}^n Y_k\right) \leq (1-\delta)n\right) \leq d\left(1 - \frac{\delta^2}{2}\right)^{n\lambda}, \tag{86}$$

$$\Pr\left(\lambda_{\max}\left(\sum_{k=1}^n Y_k\right) \geq (1+\delta)n\right) \leq d\left(1 - \frac{\delta^2}{4}\right)^{n\lambda}. \tag{87}$$

Now, to derive the result for $\delta \in [0,1]$ consider

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\sum_{k=1}^n X_k \preccurlyeq (1+\delta)\,\mathbb{E}[X]\right) = \Pr\left(\frac{1}{n}\sum_{k=1}^n Y_k \preccurlyeq (1+\delta) I\right) && \text{(By Eq.(85))}\\
=&\Pr\left(\lambda_{\max}\left(\sum_{k=1}^n Y_k\right) \leq (1+\delta)n\right)\\
=&1 - \Pr\left(\lambda_{\max}\left(\sum_{k=1}^n Y_k\right) > (1+\delta)n\right)\\
\geq&1 - d\left(1 - \frac{\delta^2}{4}\right)^{n\lambda}. && \text{(By Eq. (86))}
\end{aligned}$$

Finally, to derive the other statement we proceed similarly. We have

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\sum_{k=1}^n X_k \succcurlyeq (1-\delta)\,\mathbb{E}[X]\right) = \Pr\left(\frac{1}{n}\sum_{k=1}^n Y_k \succcurlyeq (1-\delta) I\right) && \text{(By Eq.(85))}\\
=&\Pr\left(\lambda_{\max}\left(\sum_{k=1}^n Y_k\right) \geq (1-\delta)n\right)\\
=&1 - \Pr\left(\lambda_{\max}\left(\sum_{k=1}^n Y_k\right) < (1-\delta)n\right)\\
\geq&1 - d\left(1 - \frac{\delta^2}{2}\right)^{n\lambda}. && \text{(By Eq. (87))}
\end{aligned}$$

□

Using the lemma just derived, we can derive a matrix upper bound (in all directions) that does not depend on the minimum eigenvalue.

Lemma 12 (Matrix Upper Bound)**.** *Let $X_1, \cdots, X_n$ be i.i.d. samples from $\mathcal{D}$ where $X_k \in \mathbb{R}^{d\times d}$. Suppose X_k is positive semi-definite and $\|X_k\|_{\mathrm{op}} \leq 1$ for all $k \in [n]$ almost surely. For any fixed $\lambda > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{k=1}^n X_k \preccurlyeq 10\lambda I + 3\,\mathbb{E}_{X\sim\mathcal{D}}[X]\right) \geq 1 - 2d\exp\left(-\frac{n\lambda}{4}\right). \tag{88}$$

Proof. Let $\Sigma = \mathbb{E}_{x\sim\mathcal{D}}[X]$. Consider the spectrum decomposition of Σ, denoted by $\Sigma = \sum_{k=1}^d \lambda_k v_k v_k^\top$. Without loss of generality, we assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Let $R = \sup\{k :$

$\lambda_k \geq \lambda\}$. Define $P_+ = \sum_{k=1}^{R} v_k v_k^\top$, $P_- = \sum_{k=R+1}^{d} v_k v_k^\top$ and $Q = \sum_{k=1}^{R} v_k e_k^\top \in \mathbb{R}^{d\times R}$ where $e_k \in \mathbb{R}^R$ is the k-th basis for R-dimensional space.

Since $\{v_k\}$ is a set of orthogonal basis, we have $P_+ + P_- = I$. By algebraic manipulation we get

$$\frac{1}{n}\sum_{k=1}^{n} X_k = (P_+ + P_-)^\top \left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)(P_+ + P_-) \tag{89}$$

$$= \frac{5}{4n}\sum_{k=1}^{n} P_+^\top X_k P_+ - \frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{2}P_+ - 2P_-\right)^\top X_k \left(\frac{1}{2}P_+ - 2P_-\right) + \frac{5}{n}\sum_{k=1}^{n} P_-^\top X_k P_-. \tag{90}$$

Note that for any $u \in \mathbb{R}^R$ we have $u^\top Q^\top \Sigma Q u = \sum_{k=1}^{R} \lambda_k \langle u, e_k\rangle^2 \geq \lambda \|u\|_2^2$. As a result, $\lambda_{\min}(Q^\top \Sigma Q) = \lambda_{\min}(\mathbb{E}_{x\sim\mathcal{D}} Q^\top X Q) \geq \lambda$. Consequently, applying Lemma 11 we get

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\sum_{k=1}^{n} P_+^\top X_k P_+ \preccurlyeq 2\,\mathbb{E}_{X\sim\mathcal{D}}[P_+^\top X P_+]\right) \\
=\ &\Pr\left(\frac{1}{n}\sum_{k=1}^{n} QQ^\top X_k QQ^\top \preccurlyeq 2\,\mathbb{E}_{X\sim\mathcal{D}}[QQ^\top X QQ^\top]\right) \qquad \text{(By the definition of } Q.\text{)} \\
\geq\ &\Pr\left(\frac{1}{n}\sum_{k=1}^{n} Q^\top X_k Q \preccurlyeq 2\,\mathbb{E}_{X\sim\mathcal{D}}[Q^\top X Q]\right) \\
\geq\ &1 - d\left(1 - \frac{1}{4}\right)^{n\lambda} \geq 1 - d\exp\left(-\frac{n\lambda}{4}\right).
\end{aligned}$$

[Note we cannot directly apply Lemma 11 to the top of the above sequence because that lemma requires a minimum eigenvalue greater than 0 and $P_+^\top X_k P_+$ is not full rank.]

Next, by the linearity of expectation we get $\mathbb{E}_{X\sim\mathcal{D}}[P_+ X P_+^\top] = P_+(\mathbb{E}_{X\sim\mathcal{D}}[X])P_+^\top = P_+\Sigma P_+^\top$. Recall that the spectrum decomposition gives $\Sigma = \sum_{k=1}^{d} \lambda_k v_k v_k^\top$, where $\langle v_k, v_j\rangle = \mathbb{I}[j = k]$. As a result,

$$P_+^\top \Sigma P_+ = \left(\sum_{k=1}^{R} v_k v_k^\top\right)\left(\sum_{k=1}^{d} \lambda_k v_k v_k^\top\right)\left(\sum_{k=1}^{R} v_k v_k^\top\right) \tag{91}$$

$$= \left(\sum_{k=1}^{R} \lambda_k v_k v_k^\top\right) \preccurlyeq \left(\sum_{k=1}^{d} \lambda_k v_k v_k^\top\right) = \Sigma. \tag{92}$$

Consequently we get $\mathbb{E}_{X\sim\mathcal{D}}[P_+ X^\top P_+] = P_+^\top \Sigma P_+ \preccurlyeq \Sigma = \mathbb{E}_{X\sim\mathcal{D}}[X]$. Therefore we have

$$\Pr\left(\frac{1}{n}\sum_{k=1}^{n} P_+^\top X_k P_+ \preccurlyeq 2\,\mathbb{E}_{X\sim\mathcal{D}}[X]\right) \geq 1 - d\exp\left(-\frac{n\lambda}{4}\right). \tag{93}$$

On the other hand, we upper bound the third term in Eq. (90) by Lemma 10. Let $Y_k = P_-^\top X_k P_-$. Then we have $\|Y_k\|_{\text{op}} \leq \|X_k\|_{\text{op}} \leq 1$. In addition,

$$\mathbb{E}[Y_k] = P_- \Sigma P_-^\top. \tag{94}$$

We claim that $\lambda_{\max}(\mathbb{E}[Y_k]) \leq \lambda$. Indeed, for any $v \in S^{d-1}$ we have

$$v^\top \mathbb{E}[Y_k] v = v^\top P_- \Sigma P_-^\top v. \tag{95}$$

Recall that $P_- = \sum_{k=1}^{d} \mathbb{I}[\lambda_k < \lambda]\, v_k v_k^\top$ where $\{\lambda_k\}, \{v_k\}$ are the eigen-vectors and eigen-values of Σ. For any $v \in S^{d-1}$, $P_-^\top v$ lies in the linear space spanned by $\{v_i : \lambda_i < \lambda\}$. As a result, $\left\|\Sigma P_-^\top v\right\|_2 \leq \lambda \|v\|_2$. Consequently,

$$v^\top \mathbb{E}[Y_k] v \leq \left\|P_-^\top v\right\|_2 \left\|\Sigma P_-^\top v\right\|_2 \leq \lambda \|v\|_2^2. \tag{96}$$

for all $v \in S^{d-1}$. Therefore we prove $\lambda_{\max}(\mathbb{E}[Y_k]) \leq \lambda$.

Now apply corollary 1 on Y_k and we get

$$\Pr\left(\frac{1}{n}\sum_{k=1}^{n} Y_k \preccurlyeq 2\lambda I\right) \geq 1 - d\exp\left(-\frac{\lambda n}{4}\right). \tag{97}$$

Under the high probability events described by Eq. (93) and Eq. (97), with probability at least $1 - 2d\exp(-\lambda n/4)$ we get

$$\frac{1}{n}\sum_{k=1}^{n} X_k \tag{98}$$

$$= \frac{5}{4n}\sum_{k=1}^{n} P_+^\top X_k P_+ - \frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{2}P_+ - 2P_-\right)^\top X_k \left(\frac{1}{2}P_+ - 2P_-\right) + \frac{5}{n}\sum_{k=1}^{n} P_-^\top X_k P_- \tag{99}$$

$$\preccurlyeq \frac{5}{4n}\sum_{k=1}^{n} P_+^\top X_k P_+ + \frac{5}{n}\sum_{k=1}^{n} P_-^\top X_k P_- \tag{100}$$

$$\preccurlyeq \frac{5}{2}\,\mathbb{E}_{X\sim\mathcal{D}}[X] + 10\lambda I. \tag{101}$$

□

Likewise, we can easily obtain the following matrix lower bound without any dependence on the minimum eigenvalue (see also Lemma 21 from (Ruan et al., 2020)).

Lemma 13 (Matrix Lower Bound)**.** *Let $X_1, \cdots, X_n$ be i.i.d. samples from $\mathcal{D}$ where $X_k \in \mathbb{R}^{d\times d}$. Suppose X_k is positive semi-definite and rank one and $\|X_k\|_{\text{op}} \leq 1$ for all $k \in [n]$ almost surely. For any fixed $\lambda \geq 0$*

$$\Pr\left(3\lambda I + \frac{1}{n}\sum_{k=1}^{n} X_k \succcurlyeq \frac{1}{8}\,\mathbb{E}_{X\sim\mathcal{D}}[X]\right) \geq 1 - 2d\exp\left(-\frac{n\lambda}{8}\right). \tag{102}$$

Corollary 2. *Let $X_1, \cdots, X_n$ be i.i.d. samples from $\mathcal{D}$ where $X_k \in \mathbb{R}^{d\times d}$. Suppose X_k is positive semi-definite, and $\|X_k\|_{\text{op}} \leq 1$ for all $k \in [n]$ almost surely. For any fixed $t > 0$ we have*

$$\Pr\left(\forall m \leq n, \sum_{k=1}^{m} X_k \preccurlyeq 10tI + 3m\,\mathbb{E}_{X\sim\mathcal{D}}[X]\right) \geq 1 - 2nd\exp\left(-t/4\right). \tag{103}$$

$$\tag{104}$$

Proof. For any fixed $m \leq n$, applying lemma 12 with $\lambda = t/m$ we get

$$\Pr\left(\sum_{k=1}^{m} X_k \preccurlyeq 10tI + 3m\,\mathbb{E}_{X\sim\mathcal{D}}[X]\right) = \Pr\left(\frac{1}{m}\sum_{k=1}^{m} X_k \preccurlyeq 10\frac{t}{m}I + 3\,\mathbb{E}_{X\sim\mathcal{D}}[X]\right) \tag{105}$$

$$\leq 1 - 2d\exp\left(-t/4\right). \tag{106}$$

By union bound over $m \in [n]$ we prove Eq. (103). □

E.3 Relation Between Offline and Online Covariance Matrices

Notation: Let n_k be the expected number of samples in the online phase allocated to policy π^k.

Lemma 14 (Matrix Upper Bound Offline Phase)**.** *algorithm 1 produces a comulative covariance matrix Σ_M that satisfies*

$$\mathbf{P}\left(\Sigma_M \preccurlyeq 3\overline{\Sigma}\right) \geq 1 - \frac{\delta}{4} \tag{107}$$

as long as

$$M \geq \frac{160KN}{\lambda_{reg}} \ln \frac{320dNK}{\lambda_{reg}\delta}. \tag{108}$$

Proof. From Lemma 16 *(Number of Switches)* we know that at most K distinct policies are produced during the execution of algorithm 1 where K is defined in that lemma.

Let $\phi_i^{(k)}$ be the i sampled feature during phase k. Let m_k be the values that m takes on during phase k. Note that although m_k is a random variable, we have $m_k \leq M$ almost surely. As a result, applying corollary 2 with $t = \frac{\lambda_{reg}}{10\alpha K}$ we get

$$\mathbf{P}\left(\sum_{i=1}^{m_k} \phi_i^{(k)} \phi_i^{(k),\top} \preccurlyeq \left(\frac{\lambda_{reg}}{\alpha K} I + 2m_k \, \mathbb{E}_{\phi\sim d^{(k)}}[\phi\phi^\top]\right)\right) \geq 1 - 2dM \exp\left(-\frac{\lambda_{reg}}{40\alpha K}\right). \tag{109}$$

Now multiplying the event inside the probability by $\alpha = \frac{n_k}{m_k} = \frac{N}{M}$ we get

$$\Sigma^{(k)} \stackrel{def}{=} \alpha \sum_{i=1}^{m_k} \phi_i^{(k)} \phi_i^{(k),\top} \preccurlyeq \frac{\lambda_{reg}}{K} I + 2n_k \, \mathbb{E}_{\phi\sim d^{(k)}}[\phi\phi^\top]. \tag{110}$$

After a union bound on the number of phases we can write

$$\mathbf{P}\left(\Sigma_M \stackrel{def}{=} \sum_{k=1}^{K} \Sigma^{(k)} + \lambda_{reg} I \preccurlyeq 2\lambda_{reg} I + 2\sum_{k=1}^{K} n_k \, \mathbb{E}_{\phi\sim d^{(k)}}[\phi\phi^\top] \stackrel{def}{=} 2\overline{\Sigma}\right) \tag{111}$$

$$\geq 1 - 2dMK \exp\left(-\frac{\lambda_{reg}}{40\alpha K}\right). \tag{112}$$

and substituting the value of α on the right hand side now gives

$$\mathbf{P}\left(\Sigma_M \preccurlyeq 3\overline{\Sigma}\right) \geq 1 - 2dMK \exp\left(-\frac{\lambda_{reg} M}{40KN}\right). \tag{113}$$

The final result follows from basic algebra. □

Lemma 15 (Matrix Upper Bound Online Phase)**.** *Recall that*

$$\Sigma'_n = \sum_{j=1}^{n-1} \phi(s'_j, a'_j)\phi(s'_j, a'_j)^\top + \lambda_{reg} I, \qquad \overline{\Sigma} = \alpha \sum_{i=1}^{M} \mathbb{E}_{\phi\sim d_i} \, \phi\phi^\top + \lambda_{reg} I. \tag{114}$$

Algorithm 2 produces a cumulative covariance matrix Σ'_N that satisfies

$$\mathbf{P}\left(9\Sigma'_N \succcurlyeq \overline{\Sigma}\right) \geq 1 - \frac{\delta}{4} \tag{115}$$

as long as $\lambda_{reg} \geq 24 \ln \frac{8d}{\delta}$.

Proof. Notice that conditioned on the run of algorithm 1, the distributions $d^{(1)}, \ldots, d^{(K)}$ of the features ϕ corresponding to the policies $\pi^{(1)}, \ldots, \pi^{(K)}$ are fixed (non-random), hence $\overline{\Sigma} = \sum_{k=1}^{K} \Sigma^{(k)}$ is non-random. Also, let ϕ'_i be the i feature vector collected during the online phase. Notice that conditioned on all the random variables during the offline portion we can write

$$\mathbb{E}\,\Sigma'_N = \lambda_{reg} I + \mathbb{E}\sum_{i=1}^{N} (\phi'_i)(\phi'_i)^\top \tag{116}$$

$$= \lambda_{reg} I + \sum_{k=1}^{K} \mathbb{E} \sum_{i=1}^{n'_k} \mathbb{E}_{\phi\sim d^{(k)}} \, \phi\phi^\top \tag{117}$$

where n'_k is the number of times that policy $\pi^{(k)}$ is sampled during the online phase. Continuing,

$$= \lambda_{reg} I + \sum_{k=1}^{K} \mathbb{E}\, n'_k \, \mathbb{E}_{\phi\sim d^{(k)}} \, \phi\phi^\top \tag{118}$$

$$= \lambda_{reg} I + \sum_{k=1}^{K} n_k \, \mathbb{E}_{\phi\sim d^{(k)}} \, \phi\phi^\top \tag{119}$$

$$\stackrel{def}{=} \overline{\Sigma}. \tag{120}$$

Now, apply Lemma 13 *(Matrix Lower Bound)* with $\lambda = \lambda_{reg}/(3N)$ we get

$$\mathbf{P}\left(8\left(3N\lambda I + \sum_{i=1}^{N}(\phi_i')(\phi_i')^\top\right) \succcurlyeq \sum_{k=1}^{K} n_k \, \mathbb{E}_{\phi\sim d^{(k)}} \, \phi\phi^\top\right) \tag{121}$$

$$\geq 1 - 2d\exp\left(-\frac{N\lambda}{8}\right) = 1 - 2d\exp\left(-\frac{\lambda_{reg}}{24}\right). \tag{122}$$

Recall that $\Sigma_N' = \lambda_{reg} I + \sum_{i=1}^{N}(\phi_i')(\phi_i')^\top$ and $\overline{\Sigma} = \lambda_{reg} I + \sum_{k=1}^{K} n_k \, \mathbb{E}_{\phi\sim d^{(k)}} \, \phi\phi^\top$. Eq. (121) implies that

$$\mathbf{P}\left(3N\lambda I + 8\left(\underbrace{\underbrace{3N\lambda}_{\lambda_{reg}} I + \sum_{i=1}^{N}(\phi_i')(\phi_i')^\top}_{\Sigma_N'}\right) \succcurlyeq \underbrace{3N\lambda I + \sum_{k=1}^{K} n_k \, \mathbb{E}_{\phi\sim d^{(k)}} \, \phi\phi^\top}_{\overline{\Sigma}}\right) \tag{123}$$

$$\geq 1 - 2d\exp\left(-\frac{\lambda_{reg}}{24}\right). \tag{124}$$

By setting $\lambda_{reg} \geq 24 \ln \frac{8d}{\delta}$ we have

$$\mathbf{P}\left(9\Sigma_N' \succcurlyeq \overline{\Sigma}\right) \geq 1 - \delta/4. \tag{125}$$

□

F Scalar Concentration Inequalities for Martingales

F.1 Bernstein Inequality for Martingales

The following lemma is the same as Theorem 1 from Beygelzimer et al. (2011) as is reported here for completeness.

Theorem 2 (Bernstein's Inequality for Martingales)**.** *Consider the stochastic process $\{X_t\}$ adapted to the filtration $\{\mathcal{F}_t\}$. Assume $X_t \leq 1$ almost surely. Then*

$$\forall \lambda \in (0, 1], \qquad \mathbf{P}\left(\sum_{t=1}^{T} X_t \leq \lambda \sum_{t=1}^{T} \mathbb{E}_t \, X_t^2 + \frac{1}{\lambda} \ln \frac{1}{\delta}\right) \geq 1 - \delta, \tag{126}$$

which implies

$$\mathbf{P}\left(\sum_{t=1}^{T} X_t \leq 2\sqrt{\left(\sum_{t=1}^{T} \mathbb{E}_t \, X_t^2\right) \ln \frac{1}{\delta}} + 2 \ln \frac{1}{\delta}\right) \geq 1 - \delta. \tag{127}$$

For completeness, we reprove the theorem below:

Proof. Define the random variable M_t as

$$M_t = M_{t-1} \exp(\lambda X_t - \lambda^2 \, \mathbb{E}_t \, X_t^2) \tag{128}$$

where in particular $M_0 = 1$. Recall the inequality $e^x \leq 1 + x + x^2$ for $x \leq 1$ and $1 + x \leq e^x$:

$$\mathbb{E}_t \, M_t = M_{t-1} \, \mathbb{E}_t \exp(\lambda X_t - \lambda^2 \, \mathbb{E}_t \, X_t^2) \tag{129}$$

$$\leq M_{t-1} \, \mathbb{E}_t \left[(1 + \lambda X_t + \lambda^2 X_t^2)\right] \exp(-\lambda^2 \, \mathbb{E}_t \, X_t^2) \tag{130}$$

$$\leq M_{t-1} \exp(\lambda^2 \, \mathbb{E}_t \, X_t^2) \exp(-\lambda^2 \, \mathbb{E}_t \, X_t^2) \tag{131}$$

$$= M_{t-1}. \tag{132}$$

Thus $\{M_t\}$ is a supermartingale sequence adapted to $\{\mathcal{F}_t\}$. In particular, $\mathbb{E}\, M_t \leq M_0 = 1$ using the tower property. Now by the Markov inequality

$$\mathbf{P}\left(\underbrace{\lambda \sum_{t=1}^{T} X_t - \lambda^2 \sum_{t=1}^{T} \mathbb{E}_t\, X_t^2}_{\ln M_t} > \ln \frac{1}{\delta}\right) = \mathbf{P}\left(M_t > \frac{1}{\delta}\right) \leq \frac{\mathbb{E}\, M_t}{\frac{1}{\delta}} \leq \delta \tag{133}$$

This implies that with probability at least $1-\delta$ the following event holds:

$$\lambda \sum_{t=1}^{T} X_t - \lambda^2 \sum_{t=1}^{T} \mathbb{E}_t\, X_t^2 = \ln M_t \leq \ln \frac{1}{\delta} \tag{134}$$

which is the first part of the thesis. Now, we choose λ. If $\sum_{t=1}^{T} \mathbb{E}_t\, X_t^2 \leq \ln \frac{1}{\delta}$ then under the above event we obtain with $\lambda = 1$ (the largest possible value)

$$\sum_{t=1}^{T} X_t \leq \sum_{t=1}^{T} \mathbb{E}_t\, X_t^2 + \ln \frac{1}{\delta} \leq 2 \ln \frac{1}{\delta}. \tag{135}$$

If conversely $\sum_{t=1}^{T} \mathbb{E}_t\, X_t^2 \geq \ln \frac{1}{\delta}$ then let $\lambda = \sqrt{\frac{\ln \frac{1}{\delta}}{\sum_{t=1}^{T} \mathbb{E}_t\, X_t^2}} \leq 1$ to obtain (still under the same event)

$$\sum_{t=1}^{T} X_t \leq 2\sqrt{\left(\sum_{t=1}^{T} \mathbb{E}_t\, X_t^2\right) \ln \frac{1}{\delta}}. \tag{136}$$

Therefore, summing the rhs of eqs. (135) and (136) to cover both cases we obtain the second part of the thesis. □

F.2 Reverse Bernstein Inequality for Martingales

Bernstein's inequality bounds the sum a random variable $\sum_t X_t$ using second moment information $\sum_t \mathrm{Var}_t\, X_t$; in our case (positive random variables in $[0,1]$), the sum of the conditional variances $\sum_t \mathrm{Var}_t\, X_t$ is upper bounded by the sum of the means $\sum_t \mathbb{E}_t\, X_t$.

This section provides the 'reverse' inequality: Assuming a bound on the sum a random variable $\sum_t X_t$, it bounds the conditional sum of the means $\sum_t \mathbb{E}_t\, X_t$.

Theorem 3 (Reverse Bernstein for Martingales)**.** *Let $(\Sigma, \mathcal{F}, \mathbf{P})$ be a probability space and consider the stochastic process $\{X_t\}$ adapted to the filtration $\{\mathcal{F}_t\}$. Let $\mathbb{E}_t\, X_t \stackrel{def}{=} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ be the conditional expectation of X_t given $\mathcal{F}_t$. If $0 \leq X_t \leq 1$ then it holds that*

$$\mathbf{P}\left(\sum_{t=1}^{T} \mathbb{E}_t\, X_t \geq \frac{1}{4}\left(c_1 + \sqrt{c_1^2 + 4\left(\sum_{t=1}^{T} X_t + c_2\right)}\right)^2\right) \leq \delta, \qquad c_1 = 2\sqrt{\ln \frac{1}{\delta}},\ c_2 = 2 \ln \frac{1}{\delta} \tag{137}$$

Proof. Consider the random 'noise'

$$\xi_t \stackrel{def}{=} \mathbb{E}_t\, X_t - X_t \tag{138}$$

which allows us to write

$$\sum_{t=1}^{T} \mathbb{E}_t\, X_t = \sum_{t}^{T} (\xi_t + X_t) \tag{139}$$

Then Theorem 2 *(Bernstein's Inequality for Martingales)* ensures the following high probability statement for appropriate $c_1 = 2\sqrt{\ln \frac{1}{\delta}}$, $c_2 = 2 \ln \frac{1}{\delta}$:

$$\mathbf{P}\left(\sum_{t=1}^{T} \xi_t \leq c_1 \sqrt{\sum_{t=1}^{T} \mathbb{E}_t \xi_t^2} + c_2\right) \geq 1 - \delta. \tag{140}$$

Notice that since $0 \leq X_t \leq 1$ we have

$$\begin{aligned}
\mathbb{E}_t\, \xi_t^2 &= \mathbb{E}_t (X_t - \mathbb{E}_t\, X_t)^2 && (141)\\
&= \mathbb{E}_t\, X_t^2 - (\mathbb{E}_t\, X_t)^2 && (142)\\
&\leq \mathbb{E}_t\, X_t^2 && (143)\\
&\leq \mathbb{E}_t\, X_t. && (144)
\end{aligned}$$

Plugging back into the above display and using eq. (139) gives

$$\mathbf{P}\left(\sum_{t=1}^{T} \xi_t = \sum_{t=1}^{T} (\mathbb{E}_t\, X_t - X_t) \leq c_1 \sqrt{\sum_{t=1}^{T} \mathbb{E}_t\, X_t} + c_2\right) \geq 1 - \delta \tag{145}$$

or equivalently

$$\mathbf{P}\left(\sum_{t=1}^{T} \mathbb{E}_t\, X_t \leq \sum_{t=1}^{T} X_t + c_1 \sqrt{\sum_{t=1}^{T} \mathbb{E}_t\, X_t} + c_2\right) \geq 1 - \delta. \tag{146}$$

Solving for $\sum_{t=1}^{T} \mathbb{E}_t\, X_t$ gives under such event

$$\sum_{t=1}^{T} \mathbb{E}_t\, X_t \leq \frac{1}{4}\left(c_1 + \sqrt{c_1^2 + 4\left(\sum_{t=1}^{T} X_t + c_2\right)}\right)^2. \tag{147}$$

□

G Additional Remarks

G.1 Hard Instance that Requires Exploration

Even with a fixed context, uniform exploration ignores the structure of the reward function, and is therefore suboptimal. Consider the case where $d = 2$ and $\phi(1,1) = (1,0)$ and $\phi(1,i) = (0,1)$ for $2 \leq i \leq A$. The context is fixed and there are A actions. A uniform exploration requires $\Omega(A/\epsilon^2)$ samples to estimate the reward for action 1. In contrast, the optimal exploration policy is

$$\pi(1) = \begin{cases} 1, & \text{w.p. 1/2,} \\ 2, & \text{w.p. 1/2.} \end{cases}$$

And the corresponding sample complexity is $\widetilde{O}(1/\epsilon^2)$.

On the other hand, the policy that ignores context is also suboptimal. Consider the policy $\pi^G(s)$ that returns the G-optimal design on the action set $\mathcal{A}_s$. The policy π^G explores optimally for a fixed context. Ruan et al. (2020)'s Lemma 4 implies that this policy achieves an online sample complexity $\widetilde{O}(d^3/\epsilon^2)$. For completeness, we also include a hard instance for π^G. Let $\mathcal{S} = \{s_1, \cdots, s_k\}$ and $\mathcal{A}_s = \{a_1, \cdots, a_{k+1}\}$. Assume a uniform distribution over the state space $\mathcal{S}$. The feature vector is defined as follows.

$$\phi(s_i, a_j) = \begin{cases} e_j, & \text{when } j \leq k, \\ e_{i+k}, & \text{when } j = k+1. \end{cases} \tag{148}$$

Note that in this case, the dimension of the feature vector is $d = 2k$. For a fixed $s \in \mathcal{S}$, the G-optimal design returns the uniform exploration policy. As a result, the expected covariance matrix is

$\Sigma = \mathbb{E}_{s\sim\mathcal{S},a\sim\mathcal{A}}\,\phi(s,a)\phi(s,a)^\top = \mathrm{diag}(1/(k+1),\cdots,1/(k+1),1/k(k+1),\cdots,1/k(k+1))$. It follows that

$$\mathbb{E}_{s\sim\mathcal{S}}\max_{a\in\mathcal{A}_s}\phi(s,a)^\top\Sigma^{-1}\phi(s,a) \geq \mathbb{E}_{s\sim\mathcal{S}}\,\phi(s,a_{k+1})^\top\Sigma^{-1}\phi(s,a_{k+1}) = k(k+1) = \Omega(d^2).$$

In contrast, the optimal exploration policy is

$$\pi(s_i) = \begin{cases} U(\{a_1,\cdots,a_k\}), & \text{w.p. } 1/2, \\ a_{k+1}, & \text{w.p. } 1/2, \end{cases} \tag{149}$$

where $U(\cdot)$ denotes the uniform distribution. Correspondingly, we have $\Sigma = \mathrm{diag}(1/2k,\cdots,1/2k)$, and

$$\mathbb{E}_{s\sim\mathcal{S}}\max_{a\in\mathcal{A}_s}\phi(s,a)^\top\Sigma^{-1}\phi(s,a) \leq O(d). \tag{150}$$

G.2 Remarks Regarding Ruan et al. (2020)

Section 6 of Ruan et al. (2020) argues that when $\lambda_{reg} < 1/d$, the covariance matrix doesn't concentrate. Their construction works as follows. Consider a fixed offline dataset $\mathcal{C}$ with size M. The context space is $\mathcal{S} = [d]$. The action space is $\mathcal{A}_1 = \{1\}$ and $\mathcal{A}_s = \{1,2\}$ for $2 \leq s \leq d$. The feature vector is defined as

$$\phi(1,1) = e_1, \quad \phi(s,1) = e_s, \quad \phi(s,2) = \sqrt{1-\frac{d}{M}}e_s + \sqrt{\frac{d}{M}}e_1. \tag{151}$$

The context distribution μ is

$$\mu(1) = \frac{1}{dM}, \quad \mu(s) = \frac{1}{d-1}\left(1-\frac{1}{dM}\right), \forall s \geq 2. \tag{152}$$

Then with probability at least $1/d$, $\mathcal{C}$ contains a single occurrence of context 1. Ruan et al. (2020) argues that in this case, there exits a policy π such that the covariance on $\mathcal{C}$ deviates from the population one when $\lambda_{reg} < \frac{1}{d}$.

The policy is $\pi(s) = 1$. Let $\hat{\Sigma} = \lambda_{reg}I + \sum_{s_i\in\mathcal{C}}\phi(s_i,\pi(s_i))\phi(s_i,\pi(s_i))^\top$. We can compute that

$$\mathbb{E}_{s\sim\mathcal{C}}\max_{a\in\mathcal{A}_s}\phi(s,a)^\top\hat{\Sigma}^{-1}\phi(s,a) \leq O(d/M). \tag{153}$$

Now consider the true distribution μ. Let $\Sigma = \lambda_{reg}I + M\,\mathbb{E}_{s\sim\mu}\,\phi(s,\pi(s))\phi(s,\pi(s))^\top$. By basic algebra we get $\Sigma = \lambda_{reg}I + M\mathrm{diag}(1/(dM),\mu(2),\cdots,\mu(d))$. As a result, we can compute

$$\begin{aligned}
&\mathbb{E}_{s\sim\mu}\max_{a\in\mathcal{A}_s}\phi(s,a)^\top\Sigma^{-1}\phi(s,a) \\
&= \mu(1)\frac{d}{2} + \sum_{s=2}^{d}\mu(s)\max\left(\frac{1}{\lambda_{reg}+M\mu(s)}, \left(1-\frac{d}{M}\right)\frac{1}{\lambda_{reg}+M\mu(s)} + \frac{d}{M}\frac{1}{\lambda_{reg}+1/d}\right) \\
&\geq \Omega(d^2/M).
\end{aligned}$$

Note the inequality follows due to the rightmost term and substituting in $\lambda_{reg} < \frac{1}{d}$. Comparing the estimate of the empirical covariance with the true expectation, we observe the covariance matrix doesn't concentrate. Indeed, in the setting where $\lambda_{reg} < 1$, the concentration events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ in the proof of Theorem 1 fail with constant probability. Since Ruan et al. (2020) focus on small regularization setting, their algorithm first finds a "core" of the contexts. This procedure makes their algorithm much more complex compared with ours, and increases the number of offline samples required.

Now if we set $\lambda_{reg} = 1$, by the same computation we get

$$\mathbb{E}_{s\sim\mu}\max_{a\in\mathcal{A}_s}\phi(s,a)^\top\Sigma^{-1}\phi(s,a) \leq \Omega(d/M).$$

Hence, the concentration events hold and Theorem 1 gives a much tighter offline sample complexity.

H Helper Lemmas

Lemma 16 (Number of Switches)**.** *Algorithm 1 generates at most K distinct policies:*

$$K \leq d \ln_2 \left(1 + \frac{M}{d\lambda_{reg}}\right) = \widetilde{O}(d) \tag{154}$$

Proof. Notice that $\det(\Sigma_1) = \lambda_{reg}^d$ and $\det(\Sigma_M) \leq (\lambda + \frac{M}{d})^d$ (see proof of lemma 11 in (Abbasi-Yadkori et al., 2011)). Every time the policy changes the determinant of Σ_m at least doubles. Let K denote the number of times the policy is updated. By induction,

$$\left(\lambda_{reg} + \frac{M}{d}\right)^d \geq \det(\Sigma_M) \geq 2^K \det(\Sigma_1) \geq 2^K \lambda_{reg}^d \tag{155}$$

Solving for K concludes. □

I Additional Experiments and Information

I.1 Synthetic Dataset

Here we describe how the covariance matrices were defined for the synthetic experiment based on the categories and actions. Recall that this applies for action $a \in \{1, 2, 3\}$ and category $i \in \{1, 2, 3\}$. Recall also that all covariance matrices $\Sigma_{a,i}$ are diagonal of the form $\mathrm{diag}(10^{-9}, \ldots, 10^{-9}, 1, 10^{-9}, \ldots, 10^{-9})$ an differ only by the placement of the coordinate that is equal to 1.

1. For category $i = 1$: $(\Sigma_{1,1})_{11} = 1$, $(\Sigma_{2,1})_{22} = 1$, $(\Sigma_{3,1})_{33} = 1$.
2. For category $i = 2$: $(\Sigma_{1,2})_{44} = 1$, $(\Sigma_{2,2})_{11} = 1$, $(\Sigma_{3,2})_{55} = 1$.
3. For category $i = 3$: $(\Sigma_{1,3})_{66} = 1$, $(\Sigma_{2,3})_{77} = 1$, $(\Sigma_{3,3})_{11} = 1$.

I.2 Yahoo! Learning to Rank Dataset

The Yahoo! dataset is available freely for research purposes through Yahoo! Webscope. Use of the dataset required accepting an agreement not to share the original in a way that the dataset could be reconstructed. The dataset is available at the following link:

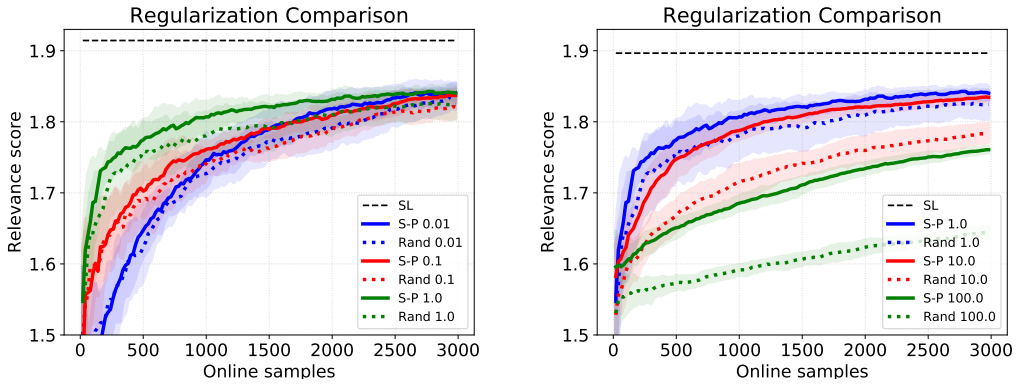
https://webscope.sandbox.yahoo.com/catalog.php?datatype=c

The data consists entirely of numerical feature vectors and does not contain any identifiable or offensive information. To run the experiments, we used a standard Amazon Web Services EC2 c5.xlarge instance with 4 vCPUs and 8gb of memory.

For the subsampling to create 300-dimensional feature vectors, we selected coordinates randomly by sampling with replacement to include out of the full 700.

I.3 Additional Plots

In addition to the three choices of regularization shown in Figure 3 for the real-world dataset, we ran additional more extreme values, but omitted them for clarity in the plot. In Figure 4, we see the same algorithms plotted with both large values of regularization (≥ 1) and small values of regularization (≤ 1).



(a) (b)

Figure 4: The left figure shows the sampler-planner (S-P) compared to the random algorithm with small regularization $\lambda \in \{0.01, 0.1, 1.0\}$. The right shows the same for $\lambda \in \{1.0, 10.0, 100.0\}$.