Contextual Semibandits via Supervised Learning Oracles

Akshay Krishnamurthy *, Alekh Agarwal †, and Miroslav Dudík ‡

1Microsoft Research, New York, NY 10011

June 15, 2016

Abstract

We study an online decision making problem where on each round a learner chooses a list of items based on some side information, receives a scalar feedback value for each individual item, and a reward that is linearly related to this feedback. These problems, known as contextual semibandits, arise in crowd-sourcing, recommendation, and many other domains. This paper reduces contextual semibandits to supervised learning, so that we can leverage powerful supervised learning methods in this partial-feedback setting. Our first reduction, which applies when the mapping from feedback to reward is known, leads to a computationally efficient algorithm with a near-optimal regret guarantee. We show that this algorithm outperforms state-of-the-art approaches on real-world learning-to-rank datasets, demonstrating the advantage of oracle-based algorithms. We also develop and analyze a novel algorithm for the setting where the linear transformation is unknown.

1 Introduction

Decision making with partial feedback, motivated by applications including personalized medicine [22] and content recommendation [18], has received increasing attention from the machine learning community. The mathematical abstraction of these problems is bandit optimization, and it involves a learner repeatedly taking an action and observing a reward for the action, with the goal of maximizing reward. While bandit optimization captures many problems of interest, several applications have additional structure: the action is combinatorial in nature and more detailed feedback is provided. For example, in internet applications, we often recommend sets of items and record information about the user’s interaction with each individual item (e.g., click). This additional feedback is unhelpful unless it relates to the overall reward (e.g., number of clicks), and, as in previous work, we assume a linear relationship. This interaction is known as the semibandit feedback model.

Typical bandit and semibandit algorithms achieve reward that is competitive with the single best fixed action, i.e., the best medical treatment or the most popular news article for everyone. This is often inadequate for recommendation applications: while the most popular articles may get some clicks, personalizing content to the users is much more effective. A better strategy is therefore to leverage contextual information to learn a rich policy for selecting actions, and we model this as contextual semibandits. In this setting, the learner

*akshay@cs.umass.edu
†alekha@microsoft.com
‡mdudik@microsoft.com
repeatedly observes a context (user features), chooses a composite action (list of articles), which is an ordered tuple of simple actions, and receives reward for the composite action (number of clicks), but also feedback about each simple action (click). The goal of the learner is to find a policy for mapping contexts to composite actions that achieves high reward.

We typically consider policies in a large but constrained class, for example, linear learners or tree ensembles. Such a class enables us to learn an expressive policy, but introduces a computational challenge of finding a good policy without direct enumeration. We build on the supervised learning literature, which has developed fast algorithms for such policy classes, including logistic regression and SVMs for linear classifiers and boosting for tree ensembles. We access the policy class exclusively through a supervised learning algorithm, viewed as an oracle.

In this paper, we develop and evaluate oracle-based algorithms for the contextual semibandits problem. We make the following contributions:

1. We first study the more common semibandits setting, where the linear function relating the semibandit feedback to the total reward is known. We develop a new algorithm, called VCEE, which extends the oracle-based contextual bandit algorithm of Agarwal et al. [1] to semibandit feedback. We show that VCEE enjoys a regret bound between $\tilde{O}(\sqrt{KLT \log(|\Pi|)})$ and $\tilde{O}(L \sqrt{KT \log(|\Pi|)})$, depending on the combinatorial structure of the problem, when there are $T$ rounds of interaction, $K$ simple actions, a composite action has $L$ simple actions, and $\Pi$ is the policy class.\footnote{We consider finite policy classes. Extensions to infinite classes follow by standard discretization arguments.} The bounds match prior results for both computationally inefficient algorithms and for the non-contextual variant, up to a factor of at most $\sqrt{L}$. Our algorithm admits structural constraints on the action space and makes $\tilde{O}(T^{3/2})$ calls to the supervised learning oracle.

2. We empirically evaluate this algorithm on two large-scale learning-to-rank datasets and compare with other contextual semibandit approaches. These experiments comprehensively demonstrate that effective exploration over a rich policy class can lead to significantly better performance than existing approaches. To our knowledge, this is the first thorough experimental evaluation of not only oracle-based semibandit methods, but of oracle-based contextual bandits as well.

3. We also consider a generalization of the semibandits setting, where the linear function relating the feedback to the reward is unknown. We develop a new algorithm, called EELS, that first explores to learn the weight vector and then, adaptively, switches to act according to an empirically optimal policy. We prove an $\tilde{O} \left( T^{2/3} L^{2/3} (K \log(|\Pi|))^{1/3} \right)$ regret bound via a careful analysis of when to switch. This is the best regret bound achievable by any efficient procedure for this setting.

Related work There is a growing body of work on combinatorial bandit optimization [5, 2] with considerable attention on semibandit feedback [11, 14, 7, 20, 15]. The majority of this research focuses on the non-contextual setting with a known relationship between semibandit feedback and reward, and a typical algorithm here achieves an $\tilde{O}(\sqrt{KLT})$ regret against the best fixed composite action. To our knowledge, only the work of Kale et al. [14] and Qin et al. [20] considers the contextual setting, again with known relationship. The former generalizes the Exp4 bandit algorithm [3] to contextual semibandits, and achieves $\tilde{O}(\sqrt{KLT})$ regret,\footnote{Kale et al. [14] consider the favorable setting where our bounds match, when uniform exploration is allowed.} but requires explicit enumeration of the policy class. The latter instead generalizes the LinUCB algorithm of Chu et al. [8] to semibandits, assuming that the feedback on the simple actions is linearly related to the context. This contrasts with our setting: we make no assumptions about the feedback on the simple actions, but impose that the overall reward is linearly related to this feedback. In our experiments, we compare VCEE against this LinUCB-style algorithm and demonstrate substantial improvements.
We are not aware of any work that attempts to learn a relationship between the reward for the composite action and the feedback on simple actions as we do with EELS. While EELS uses least squares, as done in LinUCB-style approaches, it does so without any modeling assumptions on feedback of simple actions. Crucially, the covariates for its least squares problem are observed after predicting a composite action and not before, unlike in LinUCB.

A supervised learning oracle has been used as a computational primitive in a variety of settings including online learning [12], contextual bandits [21, 25, 1, 10], structured prediction [9], and active learning [13].

2 Preliminaries

Let \( \mathcal{X} \) be a space of contexts and \( A \) a set of \( K \) simple actions. Let \( \Pi \subseteq (\mathcal{X} \to A^L) \) be a set of policies mapping contexts to composite actions (\(|\Pi| = N\)), which are tuples of \( L \) simple actions. Let \( \Delta(\Pi) \) be the set of distributions over policies, and \( \Delta_\leq(\Pi) \) be the set of non-negative weight vectors over policies, summing to at most 1, which we call subdistributions. Let \( 1(A) \) be the indicator of \( A \).

In stochastic contextual semibandits, there is an unknown distribution \( \mathcal{D} \) over triples \( (x, y, \xi) \), where \( x \) is a context, \( y \in \{0, 1\}^K \) is the vector of reward features, with entries indexed by simple actions as \( y(a) \), and \( \xi \in [-1, 1] \) is the reward noise, \( \mathbb{E} [\xi | x, y] = 0 \). Given \( y \in \mathbb{R}^K \) and \( A = (a_1, \ldots, a_L) \in A^L \), we write \( y(A) \in \mathbb{R}^L \) for the vector with entries \( y(a_\ell) \). The learner plays a \( T \) round game where in each round, nature draws \( (x_t, y_t, \xi_t) \sim \mathcal{D} \) and reveals the context \( x_t \). The learner selects a composite action \( A_t = (a_{t,1}, a_{t,2}, \ldots, a_{t,L}) \) and gets reward \( r_t(A_t) = \sum_{\ell=1}^L w_\ast^\ell y_t(a_{t,\ell}) + \xi_t \), where \( w_\ast \in \mathbb{R}^L \) is a possibly unknown but fixed weight vector. The learner is shown the reward \( r_t(A_t) \) and the vector of reward features for the chosen simple actions \( y_t(A_t) \), which is referred to as semibandit feedback.

The goal is to achieve cumulative reward competitive with all \( \pi \in \Pi \). For a policy \( \pi \), let \( \mathcal{R}(\pi) = \mathbb{E}_{(x,y,\xi) \sim \mathcal{D}} [r(\pi(x))] \) denote the expected reward of \( \pi \) and let \( \pi^\ast = \arg\max_{\pi \in \Pi} \mathcal{R}(\pi) \) denote the expected reward maximizer. We measure performance via cumulative empirical regret,

\[
\text{Regret} = \sum_{t=1}^T r_t(\pi^\ast(x_t)) - r_t(A_t).
\]  

Let \( \text{Reg}(\pi) = \mathcal{R}(\pi^\ast) - \mathcal{R}(\pi) \) denote the expected regret for a policy \( \pi \).

Example 1. In personalized search, a learning system repeatedly responds to queries with rankings of search items. This is a contextual semibandit problem where the query and user features form the context, the simple actions are search items, and the composite actions are their lists. The semibandit feedback is whether the user clicked on each item, while the reward may be the click-based discounted cumulative gain (DCG), and achieve a low regret.

We assume that our algorithms have access to a supervised learning oracle, also called an argmax oracle, denoted AMO, that can find a policy with the maximum empirical reward on any appropriate dataset. Specifically, given a dataset \( D = \{x_i, y_i, v_i\}_{i=1}^n \) of contexts \( x_i \), reward-feature vectors \( y_i \in \mathbb{R}^K \) with rewards for all simple actions, and weight vectors \( v_i \in \mathbb{R}^L \), the oracle computes

\[
\text{AMO}(D) := \arg\max_{\pi \in \Pi} \sum_{i=1}^n \langle v_i, y_i(\pi(x_i)) \rangle = \arg\max_{\pi \in \Pi} \sum_{i=1}^n \sum_{\ell=1}^L v_{i,\ell} y_{i}(\pi(x_i)_{\ell}),
\]
We begin with the setting where the weights $\pi(x)$ is the $\ell$th simple action that policy $\pi$ chooses on context $x$. The oracle is supervised as it assumes known features $y_i$ for all simple actions—we only observe them for chosen actions in the semibandit setting. This oracle is the structured generalization of the one considered in contextual bandits [1, 10] and can be any structured prediction approach such as CRFs [16] or SEARN [9].

Both of our algorithms choose composite actions by sampling from a distribution, which allows us to use importance weighting to construct unbiased estimates for the reward features $y$. If on round $t$, a composite action $A_t$ is chosen with probability $Q_t(A_t)$, we define the importance weighted feature vector $\hat{y}_t$ with components $\hat{y}_t(a) = y_t(a)1(a \in A_t)/Q_t(a \in A_t)$, which are unbiased estimators of $y_t(a)$. For a policy $\pi$, we may then define an empirical reward and a regret estimate, resp., as

$$
\eta_t(\pi, w) = \frac{1}{t} \sum_{i=1}^{t} \langle w, \hat{y}_i(\pi(x_i)) \rangle \quad \text{and} \quad \widehat{\text{Reg}}_t(\pi, w) = \max_{\pi'} \eta_t(\pi', w) - \eta_t(\pi, w).
$$

By construction, $\eta_t(\pi, w^*)$ is unbiased for $\mathcal{R}(\pi)$ but $\widehat{\text{Reg}}_t(\pi, w^*)$ is not unbiased for expected policy regret. We use $\mathcal{E}_{z \sim H}$ to denote empirical expectation over contexts in the history of interaction $H$.

Finally, we introduce projections and smoothing of distributions. For any $\mu \in [0, 1/K]$ and any subdistribution $P \in \Delta_\leq(\Pi)$, the smoothed and projected conditional subdistribution $P^\mu(A | x)$ is

$$
P^\mu(A | x) = (1 - K\mu) \sum_{\pi \in \Pi} P(\pi) 1(\pi(x) = A) + K\mu U_x(A).
$$

Here $U_x$ is a uniform distribution over a subset of valid composite actions for context $x$. The set of valid composite actions is defined implicitly through the policy class as $\{\pi(x)\}_{\pi \in \Pi}$ and we design $U_x$ to ensure that the marginal probability of each valid simple action is large. By mixing $U_x$ into our action selection, we ensure that our reward-feature estimates $\hat{y}$ are well-behaved.

The lower bound on the simple action probabilities under $U_x$ appears in our analysis as $p_{\min}$, which is the largest number satisfying $U_x(a \in A) \geq p_{\min}/K$ for all $x$ and all simple actions $a$ valid for $x$. Note that $p_{\min} = L$ when there are no restrictions on the action space, as we can take $U_x$ to be the uniform distribution over all tuples and verify that $U_x(a \in A) = L/K$. In the worst case, $p_{\min} = 1$, since we can always find one composite action for each valid simple action and let $U_x$ be the uniform distribution over this set. This set can be found efficiently by a call to AMO for each simple action $a$, with the dataset of a single point $(x, 1_a \in \mathbb{R}^K, 1 \in \mathbb{R}^L)$, where $1_a(a') = 1(a = a')$. This oracle call is guaranteed to find a policy that chooses the simple action $a$ on context $x$, if one exists.

## 3 Semibandits with known weights

We begin with the setting where the weights $w^*$ are known, and provide an efficient oracle-based algorithm (VCEE, see Algorithm 1) that generalizes the algorithm of Agarwal et al. [1].

VCEE maintains a subdistribution $Q_{t-1} \in \Delta_\leq(\Pi)$, which we turn into a distribution, $\hat{Q}_{t-1}$, by placing the remaining mass on $\pi_{t-1}$, the maximizer of $\eta_{t-1}(\pi, w^*)$. We use $\hat{Q}_{t-1}$ to select a composite action for context $x_t$ in round $t$ via Eq. (3). $Q_{t-1}$ is any solution to the feasibility problem (OP) involving the history of interaction, which aims to balance exploration and exploitation for the policy class via the constraints in Eqs. (4) and (5). Eq. (4) makes sure that the distribution has low empirical regret, which we show implies low expected regret. On the other hand, Eq. (5) ensures that the variance of the reward estimates $\hat{y}$ remains sufficiently small for each policy $\pi$, which helps control the deviation between empirical and expected
Theorem 1. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, VCEE achieves regret $\tilde{O}(\frac{\|w^*\|_1^2}{\min_{\pi} \psi_{\min}} L \sqrt{KT \log(N/\delta) / \min_{\pi} \psi})$. Moreover, VCEE can be efficiently implemented with $\tilde{O}(T^{3/2} \sqrt{K / (\min_{\pi} \log(N/\delta))})$ calls to a supervised learning oracle AMO.

The regret bound scales polynomially with both $K$ and $L$, which represents an exponential improvement in the dependence on $L$, as a classical contextual bandit algorithm would suffer $\Omega(\sqrt{KT \log N})$ regret. This worse bound is achieved by the algorithm of Agarwal et al. [1], which we generalize. We are not aware of other oracle-based algorithms with $O(\sqrt{T})$ regret, but we can use an oracle-based $\epsilon$-greedy scheme to achieve a worse $O(\text{poly}(K, L, \log(N))T^{2/3})$ regret (as verified by our experiments). Before comparing to approaches that do not use oracles, observe that in the most common setting where $w^* = 1$ and all composite actions are valid, VCEE achieves regret $\tilde{O}(\sqrt{KLT \log N})$. This compares to the approaches without oracles.
as follows:
1. Our bound matches the result of Kale et al. [14] for non-stochastic slate bandits. They assume that uniform exploration is possible and that \( w^\star = 1 \), and obtain a regret bound of \( \tilde{O}(\sqrt{KLT} \log N) \) via an algorithm that enumerates the policy space and runs in \( O(NT) \) time. Theorem 1 matches this rate, but our result generalizes to structured action spaces and is computationally efficient. On the other hand, we consider the easier stochastic setting.

2. Apart from Kale et al. [14], the adversarial combinatorial (semi)bandits literature focuses on the non-contextual setting, where the optimal bounds for semibandits are \( \Theta(\sqrt{KLT}) \) [2], and between \( \Omega(L\sqrt{KT}) \) [2] and \( O(L^{3/2}\sqrt{KT}) \) [5] without semibandit feedback. Lifting these results to the contextual setting suggests that our bound may be optimal when \( p_{\min} = \Omega(L) \).

3. When uniform exploration is not allowed, as considered by Kveton et al. [15] in the non-contextual setting, we can set \( p_{\min} = 1 \) and our bound is worse than theirs by a factor of \( \sqrt{L} \). This discrepancy may arise due to the more challenging contextual setting. Note that a UCB-style algorithm, which they use, is no longer suitable. In particular, all \( O(\sqrt{T}) \)-regret contextual bandit algorithms we are aware of involve some degree of uniform exploration, and it seems that \( \tilde{O}(L\sqrt{KT} \ln N) \) is unavoidable if the best exploration distribution has \( p_{\min} = O(1) \).

To summarize, our regret bound agrees with existing results on combinatorial (semi)bandits but represents an exponential improvement over existing computationally efficient approaches.

4 Semibandits with unknown weights

We now consider a generalization of the contextual semibandit problem with a new challenge: the weight vector \( w^\star \) is unknown. This setting is substantially more difficult than the previous one, as it is no longer clear how to use the semibandit feedback to optimize for the overall reward. In this section, we show that the semibandit feedback can still be used effectively, even when the transformation is unknown. Throughout we assume that the true weight vector \( w^\star \) has bounded norm, i.e., \( \|w^\star\| \leq B \).

One restriction required by our analysis is the ability to play any list of \( L \) simple actions without repetitions on every context. Thus, for each context, the set of valid composite actions must include all \( K!/((K-L)! \) ordered lists and we use \( U \) to denote the uniform distribution over these lists. This restriction is natural in domains like ranking and recommendation.

We propose an algorithm that explores first and then, adaptively, switches to exploitation. In the exploration phase, we play composite actions uniformly at random, with the goal of accumulating enough information to learn the weight vector \( w^\star \) for effective policy optimization. The exploration phase lasts for a variable length of time governed by two parameters \( n_\star \) and \( \lambda_\star \). The \( n_\star \) parameter controls the minimum number of rounds of the exploration phase and is \( O(T^{2/3}) \) when \( T \) is large, similar to \( \epsilon \)-greedy style schemes [17]. The adaptivity is implemented by the \( \lambda_\star \) parameter, which imposes a lower bound on the eigenvalues of the reward feature covariance matrix observed during the exploration phase. As a result, we only transition to the exploitation phase after this matrix has suitably large eigenvalues. Since we have no assumptions on the reward features, there is no bound on how many rounds this may take. This is a departure from previous \( \epsilon \)-greedy schemes, and captures the difficulty of learning \( w^\star \) when we observe the regression features only after taking an action.

After the exploration phase of \( t \) rounds, we perform least-squares regression using the observed reward features and the rewards to learn an estimate \( \tilde{w} \) of \( w^\star \). We use \( \tilde{w} \) and importance weighted reward features from the exploration phase to find a policy \( \tilde{\pi} \) with maximum empirical reward, \( \eta_t(\pi, \tilde{w}) \). The remaining rounds comprise the exploitation phase, where we play according to \( \tilde{\pi} \).

The remaining question is how to set \( \lambda_\star \), which governs the length of the exploration phase. The ideal
For any 

2. The deviation bound of Swaminathan et al. [24], which exploits the reward structure but not the semibandit feedback, leads to an algorithm with regret bound that matches ours in the $T, K$, and $\log(N)$ factors and is polynomially worse in its dependence on $L$ and $B$ (see Appendix A). This observation is consistent with non-contextual results, which tend to show that the value of semibandit information is only in $L$ factors [2].

Of course the bound has a sub-optimal dependence on the time horizon $T$, and although this is the best we are aware of for oracle-based procedures in this setting, it is an interesting open question to develop an efficient algorithm with $\text{poly}(K, L)\sqrt{T \log N}$ regret in this setting.
5 Proof Sketches

We next sketch the arguments for our theorems. The regret bound for VCEE is proved in Appendix C, while the oracle complexity bound is in Appendix D. The full proof of Theorem 2 is in Appendix E.

**Proof of Theorem 1:** The result generalizes Agarwal et. al [1], and the proof structure is similar. For the regret bound, we use Eq. (5) to control the deviation of the empirical reward estimates which make up the empirical regret $\hat{\text{Reg}}_t$. A careful inductive argument leads to the following bounds:

$$\text{Reg}(\pi) \leq 2\hat{\text{Reg}}_t(\pi) + c_0 \frac{\|w^*\|^2_2}{\|w^*\|_1} KL\mu_t$$

Here $c_0$ is a universal constant and $\mu_t$ is defined in the pseudocode. Eq. (4) guarantees low empirical regret, which by the above inequalities ensures that the population regret, when playing with $\bar{Q}_t$, is small. The cumulative regret is bounded by $\|w^*\|^2_2 \|w^*\|_1 KL \sum_{t=1}^T \mu_t$ which grows at the rate in Theorem 1.

The number of oracle calls is bounded by the analysis of the number of iterations of coordinate descent used to solve OP, via a potential argument adapted from Agarwal et al. [1].

**Proof of Theorem 2:** For EELS we analyze the exploration and exploitation phases individually, and then optimize $n_\star$ and $\lambda_\star$ to balance these terms. For the exploration phase, the expected per-round regret can be bounded by either $\|w^*\| \sqrt{KV}$ or $\|w^*\| \sqrt{L}$, using the definition of $V$, but the number of rounds depends on $\lambda_{\min}(\Sigma)$ and can be as large as $T$. However, the expected per-round covariance matrix $S = E_{x,y,A}[y(A)y(A)^T]$ also has all eigenvalues at least $V$. Thus, after $t$ rounds of exploration, we expect $\lambda_\star = \lambda_{\min}(\Sigma) \geq tV$, or in other words, the exploration regret is roughly

$$\text{Exploration Regret} \leq t\|w^*\| \min\{\sqrt{KV}, \sqrt{L}\} \leq \frac{\lambda_\star\|w^*\|}{V} \min\{\sqrt{KV}, \sqrt{L}\}.$$ 

Now our choice of $\lambda_\star$ produces a benign dependence on $V$ (Claim 21 in Appendix E) and yields the $T^{2/3}$ bound for the exploration phase.

For the exploitation phase, we bound the error between the empirical reward estimates $\eta_t(\pi, \hat{w})$ and the true reward $R(\pi)$. Since we know $\lambda_{\min}(\Sigma) \geq \lambda_\star$ in this phase, it leads to a bound of the form

$$\text{Exploitation Regret} \leq T\|w^*\| \sqrt{\frac{K\log(N)}{n_\star}} + T \sqrt{\frac{T}{\lambda_\star}} \min\{\sqrt{KV}, \sqrt{L}\}.$$ 

The first term here arises from a deviation bound on the importance weighted reward features, while the second is based on the standard linear regression analysis. Importantly, the second term introduces the $V$ parameter, again producing a mild dependence on $V$, just as with the exploration regret.

This high level argument ignores several important details. First, we must show that using $\hat{V}$ instead of the optimal choice $V$ in the setting of $\lambda_\star$ does not affect the regret. Secondly, since the termination condition for the exploration phase depends on the random variable $\Sigma$, we must derive a high probability bound on the number of exploration rounds to control the regret. Obtaining this bound requires a careful application of the Matrix Bernstein inequality to certify that $\Sigma$ has large eigenvalues.

6 Experimental Results

We conducted experiments to compare VCEE with existing alternatives. As VCEE generalizes the algorithm of Agarwal et al. [1], our experiment also provides insights into oracle-based contextual bandit approaches.
and this is the first detailed empirical study of such algorithms. Since the setting had a known $w^*$, we do not compare EELS. Additional details are in Appendix B.

**Data:** We used two large scale learning-to-rank datasets, MSLR [19] and all folds of the Yahoo! Learning-to-Rank dataset [6]. Both datasets have over 30k unique queries each with a varying number of documents that are annotated with a relevance in $\{0, \ldots, 4\}$. Each query-document pair has a feature vector ($d = 136$ for MSLR and $d = 415$ for Yahoo!) that we use to define our policy class. For MSLR, we choose $K = 10$ documents per query and set $L = 3$, while for Yahoo! we set $K = 6$ and $L = 2$. The goal is to maximize the sum of relevances of shown documents ($w^* = 1$) and the individual relevances are the semibandit feedback. All algorithms make a single pass over the queries.

**Algorithms:** We compare VCEE, implemented with an epoch schedule for solving OP after $2^{i/2}$ rounds (justified by Agarwal et al. [1]), with two baselines. First is the $\epsilon$-GREEDY approach similar to Langford and Zhang [17], with a constant but tuned choice of $\epsilon$. This algorithm explores uniformly with probability $\epsilon$ and exploits with probability $1 - \epsilon$ by following the empirically best policy. The empirically best policy is updated with the same schedule as used for solving OP in VCEE.

We also compare against a semibandit version of LINUCB [20]. This algorithm models the semibandit feedback as linearly related to the query-document features and learns this relationship, while selecting composite actions using an upper-confidence bound strategy. Specifically, the algorithm maintains a weight vector $\theta_t \in \mathbb{R}^d$ formed by solving a ridge regression problem with the semibandit feedback $y_t(a_t, \ell)$ as regression targets. At round $t$, the algorithm is presented document features $\{x_a\}_{a \in A}$ and chooses the $L$ documents with highest $x^T_a \theta_t + \alpha x^T_a \Sigma_t^{-1} x_a$ value. Here, $\Sigma_t$ is the feature covariance and $\alpha$ is a tuning parameter. For computational reasons, we only update $\Sigma_t$ and $\theta_t$ every 100 rounds.

**Oracle implementation:** LINUCB requires a linear policy class, while we instantiate both VCEE and $\epsilon$-GREEDY with three function classes, depth-2 and depth-5 gradient boosted regression trees, and linear functions (abbreviated GB2, GB5 and Lin). Both GB classes use 50 trees. Precise details of how we instantiate the supervised learning oracle can be found in Appendix B.

**Parameter tuning:** Each algorithm has a parameter governing the explore-exploit tradeoff. For VCEE, we set $\mu_t = c \sqrt{1/KL^2}$ with parameter $c$, in $\epsilon$-GREEDY we set $\epsilon$, and in LINUCB we set $\alpha$. We ran each algorithm for 10 repetitions for each of ten logarithmically spaced parameter values.

**Results:** In Figure 1, we plot the average reward (cumulative reward up to round $t$ divided by $t$) on both datasets. For each $t$, we use the parameter that achieves the best average reward across the 10 repetitions at
that $t$. Thus for each $t$, we are showing the performance of each algorithm tuned to maximize reward over $t$ rounds. We found VCEE was fairly stable to parameter tuning, so for VC-GB5 we just use one parameter value ($c = 0.008$) for all $t$ on both datasets. We show confidence bands at twice the standard error for just LINUCB and VC-GB5 to simplify the plot.

Qualitatively, both datasets reveal similar phenomena. First, when using the same policy class, VCEE consistently outperforms $\epsilon$-GREEDY. This agrees with our theory, as VCEE achieves $\sqrt{T}$-type regret, while a tuned $\epsilon$-GREEDY achieves at best a $T^{2/3}$ rate.

Secondly, if we use a rich policy class, VCEE can significantly improve on LINUCB, the empirical state-of-the-art, and one of few practical alternatives to $\epsilon$-GREEDY. Of course, since $\epsilon$-GREEDY does not outperform LINUCB, the tailored exploration of VCEE is critical. Thus, the combination of these two properties is key to improved performance on these datasets. VCEE is the only contextual semibandit algorithm we are aware of that performs adaptive exploration and is agnostic to the policy representation. Note that LINUCB is quite effective and outperforms VCEE with a linear class.

7 Discussion

This paper develops oracle-based algorithms for the contextual semibandits problem both with known and unknown weights. In both cases, our algorithms achieve the best known regret bounds for computationally efficient procedures. Our empirical evaluation of VCEE, clearly demonstrate the advantage of sophisticated oracle-based contextual bandit approaches over both parametric approaches and naive exploration. To our knowledge this is the first detailed empirical evaluation of oracle-based contextual bandit learning.

We close with some promising directions which we look forward to studying in future work:

1. With known weights, can we obtain $\tilde{O}(\sqrt{KLT \log(N)})$ regret even with structured action spaces? This may require a new contextual bandit algorithm that does not use uniform smoothing.

2. With unknown weights, can we achieve a $\sqrt{T}$ dependence while exploiting semibandit feedback?

References

[1] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In ICML, 2014.

[2] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. Math of OR, 2014.

[3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 2002.

[4] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandit algorithms with supervised learning guarantees. In AISTATS, 2011.

[5] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. JCSS, 2012.

[6] Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In Yahoo! Learning to Rank Challenge, 2011.

[7] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In ICML, 2013.
[8] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In AISTATS, 2011.

[9] Hal Daumé III, John Langford, and Daniel Marcu. Search-based structured prediction. MLJ, 2009.

[10] Miroslav Dudík, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In UAI, 2011.

[11] András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. JMLR, 2007.

[12] Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In STOC, 2016.

[13] Daniel J Hsu. Algorithms for active learning. PhD thesis, University of California, San Diego, 2010.

[14] Satyen Kale, Lev Reyzin, and Robert E Schapire. Non-stochastic bandit slate problems. In NIPS, 2010.

[15] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In AISTATS, 2015.

[16] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML, 2001.

[17] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In NIPS, 2008.

[18] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In WWW, 2010.

[19] MSLR. Msrl: Microsoft learning to rank dataset. http://research.microsoft.com/en-us/projects/mslr/.

[20] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In ICDM, 2014.

[21] Alexander Rakhlin and Karthik Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In ICML, 2016.

[22] J. M. Robins. The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In Health Service Research Methodology: A Focus on AIDS, 1989.

[23] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. Electron. Commun. Probab, 18(82):1–9, 2013.

[24] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. arXiv:1605.04812, 2016.

[25] Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert E Schapire. Efficient algorithms for adversarial contextual learning. In ICML, 2016.

[26] Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. FOCM, 2011.
A Comparisons for EELS

In this section we do a detailed comparison of our Theorem 2 to the paper of Swaminathan et al. [24], which is the most directly applicable result.

That paper focuses on off-policy evaluation in a more challenging setting where no semibandit feedback is provided. Specifically, in each round of their setting, the learner observed a context \( x \in X \), chooses a composite action \( A \) (as we do here) and receives reward \( r(A) \in [-1, 1] \) for this choice. They assume that the reward decomposes linearly across the action-position pairs,

\[
\mathbb{E}[r(A)|x, A] = \sum_{\ell=1}^{L} \phi_x(a_\ell, \ell).
\]

In this setting and when exploration is done uniformly, they provide off-policy reward estimation bounds of the form,

\[
|\eta_n(\pi) - R(\pi)| \leq \sqrt{\frac{KL \log(1/\delta)}{n}}.
\]

This bound holds with probability at least \( 1 - \delta \) for any \( \delta \in (0, 1) \). See Theorem 3 and following discussion in Swaminathan et al. [24]. We use notation consistent with our paper.

To do an appropriate comparison, we first need to define the corresponding \( \phi_x(j, a) \) features in their setting. This is easily achieved by setting \( \phi_x(j, a) = w_j^* y(a) \) in our setting. While Swaminathan et al. [24] assume rewards bounded in \([-1, 1]\), we only assume bounded \( y \)'s and noise. Consequently, we need to adjust their bound to incorporate scaling of rewards. If the rewards are scaled to lie in \([-R, R]\), their bound is,

\[
|\eta_n(\pi) - R(\pi)| \leq R \sqrt{\frac{KL \log(1/\delta)}{n}}.
\]

Either deviation bound can be turned into a low-regret algorithm by exploring for the first \( n \) rounds, finding an empirically best policy, and using that policy for the remaining \( T - n \) rounds. Balancing \( n \) lead to a \( T^{2/3} \)-style regret bound.

**Fact 3.** The approach of Swaminathan et al. [24], with rewards scaled in \([-R, R]\) leads to an algorithm with regret bound,

\[
O \left( RT^{2/3}(KL \log(N))^{1/3} \right).
\]

This algorithm can be applied as is to our setting so it is worth comparing to EELS. According to Theorem 2, EELS has a regret bound,

\[
O \left( T^{2/3}(K \log(N))^{1/3} \max \{ B^{1/3} L^{1/2}, BL^{1/6} \} \right).
\]

The dependence on \( T, K, \log(N) \) match between the two algorithms, so we are left with \( L \) and the scale factors \( B, R \). This comparison is somewhat subtle.

Our setting makes no assumptions on the scale of the reward, except for that the noise \( \xi \) is bounded in \([-1, 1]\). Thus, our setting never admits \( R < 1 \). Otherwise if the total rewards are scaled in \([-R, R]\) as in Fact 3, then since the semibandit feedback \( y \in [0, 1]^K \) we get a bound \( B \leq R/\sqrt{L} \). So for any \( R \geq 1 \), we should set \( B = R/\sqrt{L} \), so Theorem 2 gives a dependence of

\[
\max \{ R^{1/3} L^{1/3}, RL^{-1/3} \}
\]

Since \( R \) is always at least 1, this is always better than the bound in Fact 3. Also when \( R \) is large \( (R \geq L) \), the second term dominates our bound, so we improve on Swaminathan et al. [24] by a factor of \( L^{2/3} \).
B Experiment and Implementation Details

B.1 Implementation of VCEE

VCEE is implemented as stated in Algorithm 1 with some modifications, primarily to account for an imperfect oracle. OP is solve using the coordinate descent procedure described in Appendix D.

We set $\psi = 1$ in our implementation and we ignore the log factor in the setting of $\mu_t$. In our parameter tuning, we instead use $\mu_t = c\sqrt{1/KLT}$ and tune $c$ since $p_{\min} = L$, so this can account for the absence of the $\log(t^2N/\delta)$ factor. This additionally means that we ignore the failure probability parameter $\delta$. Otherwise, all other parameters and constants are set as described in Algorithm 1 and OP.

As mentioned in Section 6, we implement AMO via a reduction to squared loss regression. There are many possibilities for this reduction. We specify a dataset with a 5-tuple $D = \{x_i, A_i, y_i, p_i\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $A_i$ is any list of actions, $y_i \in \mathbb{R}^K$ assigns a value to each action, and $\gamma_i \in \mathbb{R}^K$ assigns an importance weight to each action. Since we assume $w^* = 1$ in our experiment, we don’t need to pass in the vectors $v_i \in \mathbb{R}^L$ as described in Eq. 2.

Given such a dataset $D$, we minimize a weighted squared loss objective over a regression class $\mathcal{F}$,

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^n \sum_{a \in A_i} \gamma_i(a)(f(\phi(x_i, a)) - y_i(a))^2$$

(6)

Note that we only include terms corresponding to simple actions in $A_i$ for each $i$. This regression function is associated with the greedy policy that chooses the best valid composite action according to its scoring function.

We access this oracle with two different kinds of examples. When we access OP to find the empirically best policy, we only use the history of the interaction. In this case, we only regress onto the chosen actions in the history and we let $\gamma_i$ be their importance weights. More formally, suppose that at round $t$, we observe context $x_t$, choose composite action $A_t$ with probability $Q_t(A_t)$, and receive feedback $\{y_t(a_t, t)\}_{t=1}^L$. We create a single example $(x_t, A_t, z_t, \gamma_t)$ with $x_t$ is the context, $A_t$ is the chosen composite action, $z_t$ has $z_t(a) = 1\{a \in A_t\}y_t(a)$ and $\gamma_t(a) = Q_t(a \in A_t)^{-1}$. Observe that when this sample is passed into Eq. (6), it leads to a different objective than if we regressed onto the importance weighted reward features $y_t$.

We also create examples to verify the variance constraint within OP. For this, we use the AMO in the more natural way by setting $A_i$ to be a list of all $K$ of the actions, letting $y_t$ be the importance weighted vector, and $\gamma_t = 1 \in \mathbb{R}^K$.

We use this particular implementation because leaving the importance weights inside the square loss term introduces additional variance, which we would like to avoid.

The imperfect oracle introduces one issue that needs to be corrected. Since the oracle is not guaranteed to find the actual maximizing policy on any dataset, we may have that $\hat{\text{Reg}}_t(\pi) < 0$, which can cause the coordinate descent procedure to loop indefinitely. Of course, if we ever find a policy $\pi$ with $\hat{\text{Reg}}_t(\pi) < 0$, it means that we have found an empirically better policy, so we simply switch the leader. We found that with this intuitive change, the coordinate descent procedure always terminates in a just few iterations.

B.2 Implementation of $\varepsilon$-GREEDY

For $\varepsilon$-GREEDY, we also use the oracle defined in Eq. (6). This algorithm only accesses the oracle to find the empirically best policy, and we do this in the same way as VCEE does, i.e., we only regress onto selected actions with importance weights outside of the square loss term. We use all of the data, including the data from exploitation rounds, with importance weighting.
B.3 Implementation of LINUCB

The semibandit version of LINUCB uses ridge regression to predict the semibandit feedback given query-document features $\phi(x, a)$. If the feature vectors are in $d$ dimensions, we start with $\Sigma_1 = I_d$ and $\theta_1 = 0$, the all zeros vector. At round $t$, we receive the query-document feature vectors $\{\phi(x_t, a)\}_{a \in A}$ for query $x_t$ and we choose,

$$A_t = \arg \max_{A \subset A, |A| = L} \left\{ \sum_{a \in A} \theta_t^T \phi(x_t, a) + \alpha \phi(x_t, a)^T \Sigma_t^{-1} \phi(x_t, a) \right\}$$

Since in our experiment we know that $w^* = 1$ the order of the documents is irrelevant. Here $\alpha$ is a parameter of the algorithm that we tune.

After selecting an action, we collect the semibandit feedback $\{y_t(a_t, \ell)\}_{\ell=1}^L$. In the standard implementation we would perform the update,

$$\Sigma_{t+1} \leftarrow \Sigma_t + \sum_{\ell=1}^L \phi(x_t, a_t, \ell) \phi(x_t, a_t, \ell)^T$$

$$\theta_{t+1} \leftarrow \Sigma_{t+1}^{-1} \left( \sum_{i=1}^t \sum_{\ell=1}^L \phi(x_i, a_i, \ell) y_i(a_i, \ell) \right),$$

which are the standard online ridge regression updates. For computational reasons, we only update every 100 iterations, using all of the data. Thus if $\mod (t, 100) \neq 0$ we set $\Sigma_{t+1} \leftarrow \Sigma_t$ and $\theta_{t+1} \leftarrow \theta_t$. If $\mod (t, 100) = 0$, we set,

$$\Sigma_{t+1} \leftarrow I + \sum_{i=1}^t \sum_{\ell=1}^L \phi(x_i, a_i, \ell) \phi(x_i, a_i, \ell)^T$$

$$\theta_{t+1} \leftarrow \Sigma_{t+1}^{-1} \left( \sum_{i=1}^t \sum_{\ell=1}^L \phi(x_i, a_i, \ell) y_i(a_i, \ell) \right),$$

B.4 Policy Classes

For both algorithms, we use the default implementations of function classes in scikit-learn. We instantiate scikit-learn model objects and use the fit() and predict() routines. The model objects we use are

1. sklearn.linear_model.LinearRegression()
2. sklearn.ensemble.GradientBoostingRegressor(n_estimators=50, max_depth=2)
3. sklearn.ensemble.GradientBoostingRegressor(n_estimators=50, max_depth=5)

All three objects accommodate weighted least-squares objectives as required by Eq. (6).

C Proof of Regret Bound in Theorem 1

The proof hinges on two uniform deviation bounds, and then a careful inductive analysis of the regret using the OP. The first deviation bound shows that the variance estimates used in Equation 5 are suitable estimators
Theorem 4. For any \( \delta \in (0, 1) \), if:

\[
\mu_t \geq \sqrt{\frac{\ln(2Nt^2/\delta)}{Ktp_{\min}}} \quad t \geq \frac{4K \ln(2Nt^2/\delta)}{p_{\min}}
\]

then with probability at least \( 1 - \delta \), for all distribution \( P \) over \( \Pi \), all \( \pi \in \Pi \), and all \( t \in \mathbb{N} \), we have:

\[
V(P, \pi, \mu_t) \leq 6.4\hat{V}_t(P, \pi, \mu_t) + 81.3\frac{KL}{p_{\min}}
\]  \( \text{(8)} \)

Proof. The proof of this theorem is similar in spirit to a related theorem in [1]. We first use Freedman’s inequality (Lemma 22) to argue that for a fixed \( P, \pi, \mu, \) and \( t \), the empirical version of the variance is close to the true variance. We use a discretization of the set of all distributions and then take a union bound to extend this deviation inequality to all \( P, \pi, \mu, t \). In particular, we have:

Lemma 5. For fixed \( P, \pi, \mu, t \) and for any \( \lambda \in [0, \frac{\mu p_{\min}}{L}] \), with probability at least \( 1 - \delta \):

\[
V(P, \pi, \mu) - \hat{V}_t(P, \pi, \mu) \leq \frac{(e - 2)\lambda L}{\mu p_{\min}} V(P, \pi, \mu) + \frac{\ln(1/\delta)}{t\lambda}
\]

Proof. Let:

\[
Z_i = \sum_{\ell=1}^{L} \frac{1}{P^{\mu}(\pi(x_i) \mid x_i)} - \mathbb{E}_{x \sim \mathcal{D}_x} \sum_{\ell=1}^{L} \frac{1}{P^{\mu}(\pi(x) \mid x)}
\]

and notice that \( \frac{1}{L} \sum_{i=1}^{L} Z_i = \hat{V}_t(P, \pi, \mu) - V(P, \pi, \mu) \). Clearly, \( \mathbb{E} Z_i = 0 \) for all \( i \) and \( \max_i |Z_i| \leq \frac{L}{\mu p_{\min}} \) since when we smooth by \( \mu \), each action that \( \pi \) could play must appear with probability at least \( \mu p_{\min} \). By the Cauchy-Schwarz and Holder’s Inequalities, the conditional variance is:

\[
\mathbb{E}_{x \sim \mathcal{D}_x} Z_i^2 \leq \mathbb{E}_{x \sim \mathcal{D}_x} \left( \sum_{\ell=1}^{L} \frac{1}{P^{\mu}(\pi(x) \mid x)} \right)^2 \leq L \mathbb{E}_{x \sim \mathcal{D}_x} \sum_{\ell=1}^{L} \frac{1}{P^{\mu}(\pi(x) \mid x)^2}
\]

\[
\leq \frac{L}{\mu p_{\min}} \mathbb{E}_{x \sim \mathcal{D}_x} \sum_{\ell=1}^{L} \frac{1}{P^{\mu}(\pi(x) \mid x)} = \frac{L}{\mu p_{\min}} V(P, \pi, \mu).
\]

The lemma now follows by Freedman’s inequality. \( \square \)

To prove the variance deviation bound, we next use a discretization lemma from [10], which immediately implies that for any \( P \), there exists a distribution \( P' \) supported on at most \( N_t \) policies such that for \( c_t > 0 \), if \( N_t \geq \frac{6}{\gamma_t \mu p_{\min}} \):

\[
V(P, \pi, \mu) - V(P', \pi, \mu_t) + c_t \left( \hat{V}_t(P', \pi, \mu_t) - \hat{V}_t(P, \pi, \mu_t) \right) \leq \gamma_t (V(P, \pi, \mu_t) + c_t \hat{V}_t(P, \pi, \mu_t))
\]
We set $\gamma_t = \sqrt{\frac{1 - K\mu_t}{N_t\mu_t p_{\min}}} + 3 \frac{1 - K\mu_t}{N_t\mu_t p_{\min}}$, $c_t = \frac{1}{1 - 3K\mu_t p_{\min}}$, $N_t = \lceil \frac{12(1 - K\mu_t)}{\mu_t p_{\min}} \rceil$ and $\lambda_t = 0.66\mu_t p_{\min}/L$ and take a union bound over all $t \in \mathbb{N}$, $N_t$-point distributions $P$ over $\Pi$, and all $\pi \in \Pi$ to arrive at:

$$
V(P, \pi, \mu_t) \leq 6.4V_t(P, \pi, \mu_t) + \frac{6.3L \ln(2N^2m^2/\delta)}{\mu_t p_{\min}} + \frac{75L(1 - K\mu_t) \ln N}{\mu_t^2 p_{\min}^2}.
$$

The theorem now follows from the stated bounds on $\mu_t$ and we will set:

$$
\lambda_t = \frac{1}{\|w^*\|_1} \sqrt{\frac{p_{\min}d_t}{2Kt}} \mathbf{1}\{t \leq t_0\} + \frac{\mu_{t-1}p_{\min}}{\|w^*\|_1} \mathbf{1}\{t > t_0\}
$$

The other main deviation bound is a straightforward application of Freedman’s inequality and a union bound. To state the lemma, we must introduce one more definition. Let $V_t(\pi) = \max_{0 \leq \tau \leq t-1} V(\tilde{Q}_\tau, \pi, \mu_\tau)$ where $\tilde{Q}_\tau$ is $Q_\tau$ (the distribution computed at the $\tau$th round of the game) with any additional mass placed on $\pi_\tau$, the empirical regret minimizer at round $\tau$.

**Lemma 6.** For any $\lambda_t-1 \in [0, \mu_t-1p_{\min}/\|w^*\|_1]$ for all $t \in \mathbb{N}$, $\pi \in \Pi$ and for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$
|\eta_t(\pi) - R(\pi)| \leq \|w^*\|_2^2 V_t(\pi) \lambda_{t-1} + \frac{\ln(4t^2 N/\delta)}{t\lambda_{t-1}}
$$

**Proof.** The rewards form a martingale and it is easy to see that the $\tau$th term has range bounded by $\frac{\|w^*\|_1}{\mu_{t-1}p_{\min}} \leq \|w^*\|_1$ since the $\mu$s are non-increasing. Moreover the conditional variance can be bounded by using the Cauchy-Schwarz Inequality:

$$
\mathbb{E}[Z^2|H_{t-1}] \leq \|w^*\|_2^2 \sum_{\ell=1}^L \mathbb{E}_{x \sim D_\pi} \mathbb{E}_{y|x} \frac{y(\pi(x))e^{2}}{\tilde{Q}_{\tau-1}(\pi(x)e|x)} \leq \|w^*\|_2^2 V(\tilde{Q}_{\tau-1}, \pi, \mu_{t-1}) \leq \|w^*\|_2^2 V_t(\pi)
$$

And the claim now follows by Freedman’s inequality.

Equipped with these two deviation bounds we will proceed to prove the main theorem. Define $d_t = \ln(16t^2 N/\delta)$ and let:

$$
t_0 = \min \left\{ \frac{d_t}{t} \leq \frac{p_{\min}}{4K} \right\}.
$$

Note that $t_0 \geq 4$ since $d_t \geq 1$ and $K \geq p_{\min}$. Set $\rho = \min_{t > t_0} \sqrt{t/(t - 1)}$ and note that $\rho \leq \sqrt{2}$. With this definition of $d_t$, we see that with $\mu_t \geq \sqrt{\frac{d_t}{Kp_{\min}}}$ and $t \geq 4Kd_t/p_{\min}$ we have that with probability $1 - \delta/8$ Equation 8 holds for all distributions $P$, policies $\pi$ and $t \in \mathbb{N}$ (provided $t \geq 4Kd_t/p_{\min}$, i.e. $t \geq t_0$). We also have that, for all $t \in \mathbb{N}$, $\pi \in \Pi$, with probability $\geq 1 - \delta/4$:

$$
|\eta_t(\pi) - R(\pi)| \leq \|w^*\|_2^2 V_t(\pi) \lambda_{t-1} + \frac{d_t}{t\lambda_{t-1}},
$$

and we will set:

$$
\lambda_t = \frac{1}{\|w^*\|_1} \sqrt{\frac{p_{\min}d_t}{2Kt}} \mathbf{1}\{t \leq t_0\} + \frac{\mu_{t-1}p_{\min}}{\|w^*\|_1} \mathbf{1}\{t > t_0\}
$$

Notice that to apply Lemma 6 we require $\lambda_t \in [0, \mu_t p_{\min}/\|w^*\|_1]$ so our setting of $\lambda_t$ is only valid for $t \geq t_0$. Let $E$ denote the event that both the variance and reward deviation bounds hold and observe that $\mathbb{P}(E) \geq 1 - \delta/2$.

Using the variance constraint, it is straightforward to prove the following Lemma:
Lemma 7. Assume event $E$ holds, then for any round $t \in \mathbb{N}$ and any policy $\pi \in \Pi$, let $t^*$ be the epoch achieving the max in the definition of $V_t(\pi)$. Then there are universal constants $\theta_1 > 2$ and $\theta_2$ such that:

$$V_t(\pi) \leq \begin{cases} 
2KL/p_{\min} & \text{if } \mu_{t^*} = \frac{1}{2K} \\
\theta_1 KL/p_{\min} + \frac{\|w^*\|_1}{\|w^*\|_2^2} \frac{\tilde{\text{Reg}}_t(\pi)}{\mu_{t^*}} & \text{if } \mu_{t^*} \leq \frac{1}{2K}
\end{cases}$$ (10)

Proof. The first claim follows trivially by the definition of $V_t(\pi)$ and the choice of $\mu_{t^*}$. For the second claim, we use the variance deviation bound and the optimization constraint. In particular, since $\mu_{t^*} = \sqrt{d_{t^*}/(Kt^*p_{\min})}$ we have that $t^* \geq 4Kd_{t^*}/p_{\min}$ so we can apply the variance deviation bound:

$$V(\hat{Q}_{t^*}, \pi, \mu_{t^*}) \leq 6.4V_t(\hat{Q}_{t^*}, \pi, \mu_{t^*}) + 81\frac{KL}{p_{\min}},$$

and we can use the optimization constraint which gives an upper bound on $V_t(\hat{Q}_{t^*}, \pi, \mu_{t^*})$:

$$\hat{V}_t(\hat{Q}_{t^*}, \pi, \mu_{t^*}) \leq \hat{V}_t(Q_{t^*}, \pi, \mu_{t^*}) \leq \frac{2KL}{p_{\min}} + \frac{\|w^*\|_1}{\|w^*\|_2^2} \frac{\tilde{\text{Reg}}_t(\pi)}{\mu_{t^*}}$$

The bound follows by the choice $\theta_1 = 94.1$ and $\theta_2 = \psi/6.4$.

We next compare $\text{Reg}(\pi)$ and $\tilde{\text{Reg}}(\pi)$ using the variance bounds above.

Lemma 8. Assume event $E$ holds and define $c_0 = 4\rho(1 + \theta_1)$. For all $t \geq t_0$ and all policies $\pi \in \Pi$:

$$\text{Reg}(\pi) \leq 2\tilde{\text{Reg}}_t(\pi) + c_0 \frac{\|w^*\|_1^2}{\|w^*\|_2^2} KL\mu_t \quad \text{and} \quad \tilde{\text{Reg}}_t(\pi) \leq 2\text{Reg}(\pi) + c_0 \frac{\|w^*\|_1^2}{\|w^*\|_2^2} KL\mu_t$$ (11)

Proof. The proof is by induction on $t$. As the base case, consider $t = t_0$ where, by definition we have $\mu_{t_0} = 1/(2K)$ for all $t < t_0$ so that $V_t(\pi) \leq 2KL/p_{\min}$ for all $\pi \in \Pi$ by Lemma 7. Using the reward deviation bounds, which hold under $E$ we have:

$$|\eta_t(\pi) - \mathcal{R}(\pi)| \leq \|w^*\|_2^2 V_t(\pi) \lambda_t + \frac{d_t}{t\lambda_t} \leq 2KL\|w^*\|_2^2 \lambda_t/p_{\min} + \frac{d_t}{t\lambda_t},$$

for all $\pi \in \Pi$. Since we are in round $t_0$, we know that $d_{t_0}/t_0 \leq p_{\min}/(4K)$ so we can set $\lambda_t$ as specified above. This gives:

$$|\eta_t(\pi) - \mathcal{R}(\pi)| \leq 2\sqrt{2}\frac{\|w^*\|_2^2}{\|w^*\|_1} KL\mu_{t_0}.$$

Here we use the fact that $\|w^*\|_1 \leq \sqrt{L}\|w^*\|_2$ and the definition of $\mu_{t_0} = \sqrt{d_{t_0}/(Kt_0p_{\min})}$. Now both directions of the bound follow from the triangle inequality and the optimality of $\pi_t$ for $\eta_t(\cdot)$ and $\pi_*$ for $\mathcal{R}(\cdot)$. We also use the fact that $c_0 \geq 4\sqrt{2}$ by definition of $\theta_1$. 

17
For the inductive step, fix some round \( t \) and assume that the claim holds for all for all \( t_0 \leq t' < t \) and all \( \pi \in \Pi \). By the optimality of \( \pi_t \) for \( \eta_t \) and Lemma 6 (with our choice of \( \lambda_t = \mu_{t-1} p_{\min} / \Vert w^* \Vert_1 \)), we have:

\[
\text{Reg}(\pi) - \overline{\text{Reg}}_t(\pi) = (\mathcal{R}(\pi) - \mathcal{R}(\pi_t)) - (\eta_t(\pi_t) - \eta_t(\pi))
\leq (\mathcal{R}(\pi_t) - \mathcal{R}(\pi)) - (\eta_t(\pi_t) - \eta_t(\pi))
\leq (\mathcal{V}_t(\pi_t) + \mathcal{V}_t(\pi)) \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} \mu_{t-1} p_{\min} + 2 \frac{\Vert w^* \Vert_1 d_t}{t \mu_{t-1} p_{\min}}
\]

Now by Lemma 7, there exists rounds \( i, j < t \) such that:

\[
\mathcal{V}_t(\pi) \leq \frac{\theta_1 K L}{p_{\min}} \frac{\Vert w^* \Vert_1}{\Vert w^* \Vert_2} \overline{\text{Reg}}_t(\pi) \mathbf{1}\{\mu_i < 1/(2K)\}
\]

\[
\mathcal{V}_t(\pi_*) \leq \frac{\theta_1 K L}{p_{\min}} \frac{\Vert w^* \Vert_1}{\Vert w^* \Vert_2} \overline{\text{Reg}}_t(\pi_*) \mathbf{1}\{\mu_j < 1/(2K)\}
\]

For the term involving \( \mathcal{V}_t(\pi) \) if \( \mu_i \geq 1/(2K) \) then trivially we have the bound:

\[
\mathcal{V}_t(\pi) \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} \mu_{t-1} p_{\min} \leq \theta_1 \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} K L \mu_{t-1}
\]

On the other hand, if \( \mu_i < 1/(2K) \) then by the applying the inductive hypothesis to \( \overline{\text{Reg}}_t(\pi) \) we have:

\[
\frac{\Vert w^* \Vert_1}{\Vert w^* \Vert_2} \overline{\text{Reg}}_t(\pi) \leq \frac{\Vert w^* \Vert_1}{\Vert w^* \Vert_2} 2 \text{Reg}(\pi) + \frac{c_0 K L}{\theta_2 \mu_i p_{\min}} + \frac{\mu_i}{\theta_2}
\]

\[
\mathcal{V}_t(\pi) \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} \mu_{t-1} p_{\min} \leq (\theta_1 + \frac{c_0}{\theta_2}) K L \mu_{t-1} + \frac{2 \text{Reg}(\pi)}{\theta_2}
\]

Similarly for the \( \mathcal{V}_t(\pi_*) \) term, we have the bound:

\[
\mathcal{V}_t(\pi_*) \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} \mu_{t-1} p_{\min} \leq (\theta_1 + \frac{c_0}{\theta_2}) \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} K L \mu_{t-1} + \frac{2 \text{Reg}(\pi_*)}{\theta_2} \leq (\theta_1 + \frac{c_0}{\theta_2}) \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} K L \mu_{t-1},
\]

since \( \pi_* \) has no regret. Combining these bounds gives:

\[
\text{Reg}(\pi) \leq \frac{1}{1 - \frac{\mu_i}{\theta_2}} \left( \overline{\text{Reg}}_t(\pi) + 2 \left( \theta_1 + \frac{c_0}{\theta_2} \right) \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} K L \mu_{t-1} + \frac{2 \Vert w^* \Vert_1 d_t}{t \mu_{t-1} p_{\min}} \right)
\]

Recall that \( \theta_1 = 94.1, \theta_2 = \psi/6.4, \psi = 100, c_0 = 4 \rho (1 + \theta_1) \) and \( \rho \leq \sqrt{2} \). This means that \( \theta_2/2 \leq 1/2 \) so the pre-multiplier on the \( \overline{\text{Reg}}_t(\pi) \) term is at most 2. For the third term, since \( d_t/t \) is non-increasing, we have the bound \( \frac{d_t}{t \mu_{t-1} p_{\min}} \leq K \mu_{t-1} \) by the definition of \( \mu_{t-1} \). We also use the bound \( \Vert w^* \Vert_2^2 \leq L \Vert w^* \Vert_1^2 \). Since \( \mu_{t-1} \leq \rho \mu_t \) we replace all \( \mu_{t-1} \) terms with \( \rho \mu_t \) above. Lastly, one can verify that if \( \theta_2 \geq 4 \rho \), which it is given our choice of \( \psi = 100 \) and \( \rho \leq \sqrt{2} \), the pre-multiplier to the \( \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} \) term is bounded by \( c_0 \). This gives one direction of the inequality.

The other direction proceeds similarly to before. Under event \( \mathcal{E} \) we have:

\[
\overline{\text{Reg}}_t(\pi) - \text{Reg}(\pi) = \eta_t(\pi_t) - \eta_t(\pi) - \mathcal{R}(\pi) + \mathcal{R}(\pi_t)
\leq \eta_t(\pi_t) - \eta_t(\pi) - \mathcal{R}(\pi_t) + \mathcal{R}(\pi)
\leq (\mathcal{V}_t(\pi_t) + \mathcal{V}_t(\pi)) \frac{\Vert w^* \Vert_2^2}{\Vert w^* \Vert_1} \mu_{t-1} p_{\min} + \frac{2 \Vert w^* \Vert_1 d_t}{t \mu_{t-1} p_{\min}}
\]

18
As before, we have the bound:

\[ V_t(\pi) \| w^* \|_2^2 \mu_{t-1} p_{\min} \leq (\theta_1 + \frac{c_0}{\theta_2}) K L \mu_{t-1} + \frac{2 \text{Reg}(\pi)}{\theta_2} \]

but for the \( V_t(\pi_t) \) term we must use the inductive hypothesis twice. We know there exists a round \( j < t \) for which

\[ V_t(\pi_t) \leq \theta_1 \frac{K L}{p_{\min}} + \frac{\| w^* \|_1}{\theta_2^2 \mu_j p_{\min}} 1\{\mu_j < 1/(2 K)\}. \]

Applying the inductive hypothesis twice gives:

\[
\frac{\| w^* \|_1 \text{Reg}_j(\pi_t)}{\| w^* \|_2^2 \theta_2 \mu_j p_{\min}} \leq \frac{\| w^* \|_1}{\theta_2^2 \mu_j p_{\min}} \left( 2 \text{Reg}(\pi_t) + c_0 \frac{\| w^* \|_1}{\| w^* \|_2} K L \mu_j \right)
\]

\[
\leq \frac{\| w^* \|_1}{\theta_2^2 \mu_j p_{\min}} \left( 2 \text{Reg}(\pi_t) + c_0 \frac{\| w^* \|_1}{\| w^* \|_2} K L \mu_j \right)
\]

\[ \leq \frac{3c_0}{\theta_2^2} \frac{K L}{p_{\min}} \]

Here we use the inductive hypothesis twice, once at round \( j \) and once at round \( t \) and then use the fact that \( \pi_t \) has no regret at round \( t \), i.e. \( \text{Reg}_t(\pi_t) = 0 \). We also use the fact that the \( \mu \)s are non-increasing so that \( \mu_t / \mu_j \leq 1 \). This gives the bound:

\[ V_t(\pi_t) \| w^* \|_2^2 \mu_{t-1} p_{\min} \leq (\theta_1 + \frac{3c_0}{\theta_2^2}) \| w^* \|_1^2 K L \mu_{t-1} \]

Combining the bounds for \( V_t(\pi) \) and \( V_t(\pi_t) \) gives:

\[ \text{Reg}_t(\pi) \leq \left( 1 + \frac{2}{\theta_2} \right) \text{Reg}(\pi) + \left( 2 \theta_1 + \frac{4c_0}{\theta_2} \right) \| w^* \|_1^2 K L \mu_{t-1} + \frac{2 \| w^* \|_1 d_t}{\mu_{t-1} p_{\min}} \]

Since \( \theta_2 \geq 2 \) the pre-multiplier on the first term is at most 2. As before, the third term is bounded by \( 2 \| w^* \|_1^2 K L \mu_{t-1} \) and \( \mu_{t-1} \leq \rho \mu_t \). Then, by definition of \( c_0, \theta_1, \theta_2 \), we have that \( \rho(2 \theta_1 + 4c_0/\theta_2 + 2) \leq c_0 \) which proves the claim.

The last key ingredient of the proof is the following Lemma, which shows that the low-regret constraint in Equation 4, which is based on the regret estimates, actually ensures low regret.

**Lemma 9.** Assume event \( E \) holds. Then for every round \( t \in \mathbb{N} \):

\[
\sum_{\pi \in \Pi} \tilde{Q}_{t-1}(\pi) \text{Reg}(\pi) \leq (4 \psi + c_0) \frac{\| w^* \|_1^2}{\| w^* \|_2^2} K \mu_{t-1}
\]

(12)
Proof. If \( t \leq t_0 \) then \( \mu_{t-1} = 1/(2K) \) in which case (since \( \text{Reg}(\pi) \leq ||w^*||_1 \)):

\[
\sum_{\pi \in \Pi} \hat{Q}_{t-1}(\pi) \text{Reg}(\pi) \leq ||w^*||_1^2 L \leq 2\psi ||w^*||_1^2 KL\mu_{t-1},
\]

so the claim holds. Now for \( t > t_0 \) we have:

\[
\sum_{\pi \in \Pi} \hat{Q}_{t-1} \text{Reg}(\pi) \leq \left( 2 \sum_{\pi \in \Pi} Q_{t-1}(\pi) \text{Reg}_{t-1}(\pi) \right) + c_0 ||w^*||_1^2 KL\mu_{t-1}
\]

\[
\leq (4\psi + c_0) ||w^*||_1^2 KL\mu_t
\]

The first inequality follows by Lemma 8 and the second follows from the fact that \( \hat{Q}_{t-1} \) places its remaining mass on \( \pi_{t-1} \) which suffers no empirical regret at round \( t - 1 \). The last inequality is due to the low regret optimization constraint.

To control the regret, we must first add up the \( \mu_t \)s, which relate to the probability of exploring. Our definition of \( \mu_t \) differs from [1] only in the introduction of \( p_{\min} \), so by a straightforward adaptation we have:

**Lemma 10.** For any \( T \in \mathbb{N} \):

\[
\sum_{t=1}^{T} \mu_t \leq 2\sqrt{Td_T K p_{\min}} \quad \text{and} \quad \sum_{t=1}^{T} \mu_{t-1} \leq \frac{t_0}{2K} + \sqrt{\frac{8Td_T}{Kp_{\min}}}
\]

We are finally ready to prove the theorem by adding up the total regret for the algorithm.

**Lemma 11.** For any \( T \in \mathbb{N} \), with probability at least \( 1 - \delta \), the regret after \( T \) rounds is at most:

\[
||w^*||_1^2 L \left[ 2\sqrt{2T \ln(2/\delta)} + (4\psi + c_0 + 1) \left( \frac{2Kd_{t_0}}{p_{\min}} + \sqrt{\frac{8KTd_T}{p_{\min}}} \right) \right]
\]

Proof. For each round \( t \in \mathbb{N} \) let \( Z_t = r_t(\pi_*(x_t)) - r_t(A_t) - \sum_{\pi \in \Pi} \hat{Q}_{t-1}(\pi) \text{Reg}(\pi) \). Since at round \( t \), we play action \( A_t \) with probability \( \hat{Q}_{t-1}^{\mu_{t-1}} \), this sequence of random variables is clearly centered. Moreover we have \( |Z_t| \leq 2||w^*||_1 \) and it follows by Azuma’s inequality (Lemma 23) that with probability at least \( 1 - \delta/2 \):

\[
\sum_{t=1}^{T} |Z_t| \leq 2||w^*||_1 \sqrt{2T \ln(2/\delta)}
\]

20
There are two types of updates in the algorithm. If the weights
are too large or the regret constraint in

\[ \text{Regret} \leq 2\|w^*\|_1 \sqrt{2T \ln(2/\delta)} + \sum_{t=1}^{T} \sum_{\pi \in \Pi} \tilde{Q}^{\mu_{t-1}}_t(\pi) \text{Reg}(\pi) \]

\[ \leq 2\|w^*\|_1 \sqrt{2T \ln(2/\delta)} + \sum_{t=1}^{T} \sum_{\pi \in \Pi} (1 - K \mu_{t-1}) \tilde{Q}_{t-1}(\pi) \text{Reg}(\pi) + \|w^*\|_1 K \mu_{t-1} \]

\[ \leq 2\|w^*\|_1 \sqrt{2T \ln(2/\delta)} + \sum_{t=1}^{T} (4\psi + c_0 + 1) \|w^*\|_1^2 L K \mu_{t-1} \]

\[ \leq \frac{\|w^*\|_1^2}{\|w^*\|_1} \left[ 2 \sqrt{2T \ln(2/\delta)} + (4\psi + c_0 + 1) \left( \frac{t_0}{2} + \sqrt{\frac{8K T d_T}{p_{\min}}} \right) \right] \]

\[ \leq \frac{\|w^*\|_1^2}{\|w^*\|_1} \left[ 2 \sqrt{2T \ln(2/\delta)} + (4\psi + c_0 + 1) \left( \frac{2K d_{t_0}}{p_{\min}} + \sqrt{\frac{8K T d_T}{p_{\min}}} \right) \right] \]

Here the first inequality is from the application of Azuma’s inequality above. The second one uses the definition of \( \tilde{Q}^{\mu_{t-1}}_t \) to split into rounds where we play as \( Q_{t-1} \) and rounds where we explore. The exploration rounds occur with probability \( K \mu_{t-1} \), and on those rounds we suffer regret at most \( \|w^*\|_1 \). For the other rounds, we use Lemma 9 and then we use Lemma 10. We also use the identity \( \|w^*\|_1 \leq L \|w^*\|_2^2/\|w^*\|_1 \) in order to collect terms. Finally we use the fact that \( t_0 \geq 4K d_{t_0}/p_{\min} \).

\[ \square \]

## D Proof of Oracle Complexity Bound in Theorem 1

In this section we prove the oracle complexity bound in Theorem 1. First we describe how the optimization algorithm (OP), can be solved via a coordinate ascent procedure.

This problem is similar to the one used by Agarwal et. al [1] for classical contextual bandit learning, and following their approach, we provide a coordinate ascent procedure in the policy space (See Algorithm 3). There are two types of updates in the algorithm. If the weights \( Q \) are too large or the regret constraint in
Equation 4 is violated, the algorithm multiplicatively shrinks all of the weights. Otherwise, if there is a policy that is found to violate the variance constraint in Equation 5, the algorithm adds weight to that policy, so that the constraint is no longer violated.

First, if the algorithm halts, then both of the conditions must be satisfied. The regret condition must be satisfied since we know that \( \sum_\pi Q(\pi)(2KL/p_{\min} + b_\pi) \leq 2KL/p_{\min} \) which in particular implies that \( \sum_\pi Q(\pi)b_\pi \leq 2KL/p_{\min} \) as required. Note that this also ensures that \( \sum_\pi Q(\pi) \leq 1 \) so \( Q \in \Delta_{\leq}(\Pi) \). Finally, if we halted, then for each \( \pi \), we must have \( D_\pi(Q) \leq 0 \) which implies \( V_\pi(Q) \leq 2KL/p_{\min} + b_\pi \) so the variance constraint is also satisfied.

The algorithm can be implemented by first accessing the oracle on the importance weighted history \( \hat{H} \) to obtain \( \pi_t \) (so that we can compute \( b_\pi \)). The low regret check in Step 4 of Algorithm 3 can be done efficiently, since each policy in the support of the current distribution \( Q \) was added at a previous iteration of Algorithm 3, and we can store the regret of the policy at that time for no extra computational burden. This allows us to always maintain the expected regret of the current distribution \( Q \) for no added cost. Finding a policy violating the variance check can be done by one call to the oracle \( AMO \). At round \( t \) of the contextual bandit problem, we create a dataset of the form \( (x_i, z_i, v_i) \) of size \( 2t \). The first \( t \) terms come from the variance \( V_\pi(Q) \) and the second \( t \) terms come from the rescaled empirical regret \( \hat{V}_\pi(Q) \).

For any history \( H \) and parameter \( \mu \), Algorithm 3 halts and outputs a set of weights \( Q \in \Delta_{\leq}(\Pi) \) that is feasible for \( OP \). Moreover, Algorithm 3 halts in no more than \( \frac{8\log(1/(K\mu))}{\mu p_{\min}} \) iterations and each iteration can be implemented efficiently, with at most one call to \( AMO \).

As for the iteration complexity, we prove the following theorem.

**Theorem 12.** For any history \( H \) and parameter \( \mu \), Algorithm 3 halts and outputs a set of weights \( Q \in \Delta_{\leq}(\Pi) \) that is feasible for \( OP \). Moreover, Algorithm 3 halts in no more than \( \frac{8\log(1/(K\mu))}{\mu p_{\min}} \) iterations and each iteration can be implemented efficiently, with at most one call to \( AMO \).

Equipped with this theorem, it is easy to see that the total number of calls to the \( AMO \) over the course of the execution of Algorithm 1 can be bounded as \( \tilde{O} \left( T^{3/2} \sqrt{\frac{K}{p_{\min} \log(N/\delta)}} \right) \) by the setting of \( \mu_t \).
We mention in passing that Agarwal et. al [1] also develop two improvements that lead to a more efficient algorithm. They partition the game into epochs and only solve OP once every epoch, rather than in every round as we do here. They also show how to use the weight vector from the previous round to warm-start the next coordinate ascent execution. Both of these optimizations can also be implemented here, and they will lead to a better computational guarantee for the algorithm, although we omit these details to simplify the presentation.

D.1 Proof of Theorem 12

We use the following potential function for the analysis, which is adapted from Agarwal et. al [1],

$$
\Phi(Q) = \left( \frac{\hat{E}[RE(U_x||Q^\mu(\cdot|x))]}{1 - K\mu} + \sum_{\pi} Q(\pi)b_\pi \right) \frac{2K/p_{\text{min}}}{1} - K\mu + \frac{\sum_{\pi} Q(\pi)b_\pi}{2K/p_{\text{min}}}
$$

with:

$$
RE(p||q) = \sum_{a \in A} p_a \ln(p_a/q_a) + q_a - p_a
$$

For intuition, note that the partial derivative of the potential function with respect to a coordinate $Q(\pi)$ relate exactly the variance $V_\pi(Q)$:

$$
\frac{\partial \Phi(Q)}{\partial Q(\pi)} = \left( \frac{\frac{1}{t} \sum_{t=1}^{t} \sum_{a \in \pi(x_t)} \frac{-p_a(1-K\mu)}{Q^\mu(a|x_t)} + (1-K\mu)}{1} + \frac{b_\pi}{2K/p_{\text{min}}} \right)
$$

$$
\leq \left( \frac{-p_{\text{min}}}{K} V_\pi(Q) + L + \frac{p_{\text{min}}b_\pi}{2K} \right)
$$

$$
= \frac{p_{\text{min}}}{2K} \left( -2V_\pi(Q) + \frac{2KL}{p_{\text{min}}} + b_\pi \right)
$$

$$
= \frac{p_{\text{min}}}{2K} \left( -D_\pi(Q) - V_\pi(Q) \right)
$$

This means that if $D_\pi(Q) > 0$, then the partial derivative is very negative, and by increasing the weight on $Q$, we can decrease the potential function $\Phi$.

We establish the five facts:

1. $\Phi(0) \leq L \ln(1/(K\mu))/(1 - K\mu)$. This follows by the fact that the exploration distribution in $Q^\mu$ is exactly $U_x$.
2. $\Phi(Q)$ is convex in $Q$.
3. $\Phi(Q) \geq 0$ for all $Q$.
4. The shrinking update when the regret constraint is violated does not increase the potential. More formally, for any $c < 1$, we have $\Phi(cQ) \leq \Phi(Q)$ whenever $\sum_\pi Q_\pi(2KL/p_{\text{min}} + b_\pi) > 2KL/p_{\text{min}}$.
5. The additive update when $D_\pi > 0$ for some $\pi$ lowers the potential by at least $\frac{L\mu p_{\text{min}}}{(1-K\mu)}$.  

23
The first three are fairly straightforward and the proof of the last two are based on the arguments of Agarwal et. al [1]. For the first claim we have:

$$
\Phi(0) = \sum_{a \in \mathcal{A}} p_a \ln \left( \frac{p_a}{K p_{\pi}} \right) - \frac{(1 - K\mu)p_a}{1 - K\mu} \leq \frac{L \ln(1/(K\mu))}{1 - K\mu}
$$

Since the marginals \( p_a \) sum to at most \( L \). Convexity of this function follows from the fact that the unnormalized relative entropy is convex in the second argument, and the fact that the marginal distribution is linear in the vector \( Q \). The third fact follows by the non-negative of both the empirical regret \( b_\pi \) and of the unnormalized relative entropy \( RE(\cdot || \cdot) \).

For the fourth fact, we prove the following lemma.

**Lemma 13.** Let \( Q \) be a weight vector for which \( \sum_\pi Q(\pi)(2KL/p_{\min} - b_\pi) > 2KL/p_{\min} \) and define \( c = \sum_\pi Q(\pi)(2KL/p_{\min} - b_\pi) < 1 \). Then \( \Phi(cQ) \leq \Phi(Q) \).

**Proof.** Define \( g(c) = B_0 \Phi(cQ) \) and let \( Q^*_\pi(a|x) = (1 - K\mu)cQ_\pi(a|x) + K\mu U_\pi(a|x) \). By the chain rule, using the calculation of the derivative above, we have:

$$
g'(c) = B_0 \sum_\pi Q(\pi) \frac{\partial \Phi(cQ)}{\partial Q(\pi)} \\
\geq \frac{p_{\min} B_0}{2K} \sum_\pi Q(\pi) \left( \frac{2KL}{p_{\min}} + b_\pi - 2E \sum_{a \in \pi(x)} \frac{1}{Q^*_\pi(a|x)} \right)
$$

For the last term, we have:

$$
\sum_\pi Q(\pi) \sum_{a \in \pi(x)} \frac{1}{Q^*_\pi(a|x)} = E \sum_{a \in \mathcal{A}} \sum_{\pi \in \Pi} Q(\pi) \mathbf{1}\{a \in \pi(x)\} \\
= E \sum_{a \in \mathcal{A}} \frac{Q(a|x)}{Q^*_\pi(a|x)} = \frac{1}{c} \sum_{a \in \mathcal{A}} cQ(a|x)
$$

Now define \( q_\pi = cQ(a|x) \) and the inner sum can be upper bounded by:

$$
\leq \sum_{a \in \mathcal{A}} \frac{(1 - K\mu)q_\pi + \mu p_{\min}}{(1 - K\mu)q_\pi + \mu p_{\min}/q_\pi} = K \frac{1}{K} \sum_{a \in \mathcal{A}} \frac{1}{(1 - K\mu)q_\pi + \mu p_{\min}/q_\pi}
$$

$$
\leq K \frac{1}{(1 - K\mu) + K \mu p_{\min}/q_\pi} \leq K \frac{1}{K \mu p_{\min}/q_\pi + 1 - K\mu} \leq KL \frac{1}{p_{\min}(1 - K\mu) + K\mu} \leq KL \frac{1}{p_{\min}}
$$

The first inequality uses the lower bound \( p_{\min}/K \) for exploration distribution, then we use Jensen’s inequality and the fact that \( \sum q_\pi \leq L \) since \( c < 1 \). Finally, we use the fact that \( L/p_{\min} \geq 1 \) and \( K\mu \leq 1 \) so that the first term in the denominator is \( \geq 1 - K\mu \). Plugging this in above we have:

$$
\Phi(c) \geq \frac{p_{\min}B_0}{2K} \left( \sum_\pi Q_\pi \left( \frac{2KL}{p_{\min}} + b_\pi \right) - \frac{2KL}{p_{\min}} \right) > 0
$$
By the condition in the algorithm. Since \( g \) is convex, this means that \( g(c) \) is nondecreasing for all values exceeding \( c \). Since \( c < 1 \), we have:

\[
B_0 \Phi(Q) = g(1) \geq g(c) = B_0 \Phi(cQ)
\]

So any positive \( B_0 \) is fine.

And for the fifth fact, we have:

**Lemma 14.** Let \( Q \) denote a set of weights and suppose, for some policy \( \pi \), that \( D_\pi(Q) > 0 \). Let \( Q' \) be the new set of weights which is identical except that \( Q'(\pi) = Q(\pi) + \alpha \) with \( \alpha = \alpha_\pi(Q) > 0 \). Then

\[
\Phi(Q) - \Phi(Q') \geq \frac{\mu L p_{\text{min}}}{4(1 - K\mu)}
\]

**Proof.** Let \( Q'(\cdot) = Q(\cdot) + \alpha \cdot \pi \), i.e. the update we perform when \( \pi \) is found to violate the inequality in the algorithm. Since \( Q'^\mu(a|x) = Q^\mu(a|x) + (1 - K\mu)\alpha \mathbb{1}\{a \in \pi(x)\} \) only updates a few coordinates of the marginal probabilities, we have by a direct calculation:

\[
2K(\Phi(Q) - \Phi(Q')) = 2K \left( \mathbb{E} \sum_a p_a \ln(p_a/q_a) - p_a \ln(p_a/q_a) + q_a - q_a' \right) \geq \frac{\alpha b_{\text{min}} p_{\text{min}}}{2K}
\]

The term inside the expectation can be bounded using the fact that \( \ln(1 + x) \geq x - x^2/2 \):

\[
\mathbb{E} \sum_a \ln \left( 1 + \frac{(1 - K\mu)}{Q^\mu(a|x)} \right) \geq \mathbb{E} \sum_a \frac{(1 - K\mu)}{Q^\mu(a|x)} - \frac{1}{2} \left( \frac{(1 - K\mu)}{Q^\mu(a|x)} \right)^2 \geq \alpha(1 - K\mu) V_\pi(Q) - \frac{\alpha^2 (1 - K\mu)^2}{2} S_\pi(Q)
\]

Plugging this in above gives a lower bound:

\[
2K(\Phi(Q) - \Phi(Q')) \geq 2p_{\text{min}} \alpha V_\pi(Q) - (1 - K\mu)p_{\text{min}}\alpha^2 S_\pi(Q) - p_{\text{min}}\alpha \frac{2KL}{p_{\text{min}}} + b_{\pi}
\]

Using the definition \( D_\pi(Q) = V_\pi(Q) - \frac{2KL}{p_{\text{min}}} - b_\pi \). Now we set \( \alpha = \frac{(V_\pi(Q) + D_\pi(Q))}{2(1 - K\mu) S_\pi(Q)} \) as in the algorithm and obtain:

\[
2K(\Phi(Q) - \Phi(Q')) \geq \frac{p_{\text{min}}(V_\pi(Q) + D_\pi(Q))^2}{4(1 - K\mu) S_\pi(Q)}
\]
Note that \( S_\pi(Q) \geq \frac{1}{\mu p_{\text{min}}} V_\pi(Q) \) (by bounding one of the terms in the square by the range which is \( \mu p_{\text{min}} \)) and that \( V_\pi(Q) > \frac{2KL}{p_{\text{min}}} \) since \( D_\pi(Q) > 0 \). This gives:

\[
2K(\Phi(Q) - \Phi(Q')) \geq \frac{\mu p_{\text{min}}^2 V_\pi(Q) + D_\pi(Q)}{4(1 - K\mu)} V_\pi(Q)
\]

\[
\geq \frac{\mu p_{\text{min}}^2 V_\pi(Q)}{4(1 - K\mu)} \geq \frac{K L \mu p_{\text{min}}}{2(1 - K\mu)}
\]

Dividing both sides of this inequality by \( 2K \) proves the lemma.

\[\square\]

### E Proof of Theorem 2

The proof of Theorem 2 requires many delicate steps, so we first sketch the overall proof architecture. The first step is to derive a parameter estimation bound for learning in linear models. This is a somewhat standard argument from linear regression analysis, and the important component is that the bound involves the covariance \( \Sigma \) of the feature vectors used in the problem. Combining this with importance weighting on the reward features \( y \) as in VCEE, we prove that the policy used in the exploitation phase has low expected regret, provided that \( \Sigma \) has large eigenvalues.

The next step involves a precise characterization of the mean and deviation of the covariance matrix \( \Sigma \), which relies heavily on the exploration phase employing a uniform exploration strategy. This step involves a careful application of the Matrix Bernstein inequality (Lemma 25). We then bound the expected regret accumulated during the exploration phase; we show, somewhat surprisingly, that the expected regret can be related to the mean of covariance matrix \( \Sigma \). Finally, as the exploitation regret encourages setting \( \lambda^* \) large while the exploitation phase encourages setting small \( \lambda^* \), we optimize this parameter to balance these two terms. Similarly, the exploitation regret encourages setting \( n^* \) large while the exploration phase encourages setting it small, and our choice of \( n^* \) optimizes this tradeoff.

An important definition that will appear throughout the analysis is the expected reward variance, when a single action is chosen uniformly at random.

\[
V = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \frac{1}{K} \sum_{a \in A} y^2(a) - \left( \frac{1}{K} \sum_{a' \in A} y(a') \right)^2 \right]
\]

#### E.1 Estimating \( V \)

The first step is a deviation bound for estimating \( V \).

**Lemma 15.** After \( n^* \) rounds, the estimate \( \hat{V} \) satisfies, with probability at least \( 1 - \delta \),

\[
|\hat{V} - V| \leq \sqrt{\frac{2V \log(2/\delta)}{n^*}} + \frac{4 \log(2/\delta)}{3 n^*}.
\]

**Proof.** Note that our estimator is a sum of identically distributed terms and each term is,

\[
\frac{1}{2K^2} \sum_{a, b \in A} (y(a) - y(b)) \mathbb{1}_{\left\{ a, b \in A \right\}} \frac{1}{U[a, b \in A]}
\]
The mean of this random variable is precisely $V$. 

$$
\mathbb{E}_{x,y,A} \frac{1}{2K^2} \sum_{a,b \in A} (y(a) - y(b))^2 \mathbf{1}_{\{a,b \in A\}} = \frac{1}{2K^2} \mathbb{E}_{x,y} \sum_{a,b \in A} (y(a) - y(b))^2 
$$

$$
= \mathbb{E}_{x,y} \frac{1}{2K^2} \sum_{a,b} (y(a)^2 - 2y(a)y(b) + y(b)^2) 
$$

$$
= \mathbb{E}_{x,y} \left[ \frac{1}{K} \sum_{a} y(a)^2 - \left( \frac{1}{K} \sum_{a} y(a) \right)^2 \right] = V 
$$

The range is bounded by 1, since we choose $L$ actions uniformly at random, the probability for two distinct actions jointly being selected is $U(a,b \in A) = \frac{L(L-1)}{K(K-1)}$ and for a single action it is $U(a \in A) = \frac{L}{K}$. The $(y(a) - y(b))^2$ term is at most one but it’s always zero for $a = b$, so the range is at most 

$$
\frac{1}{2K^2} \sum_{a \neq b \in A} \frac{K(K-1)}{L(L-1)} = \frac{K(K-1)}{K^2} \leq 1. 
$$

As for the second moment, we have a random variable $Z$ such that $Z \geq 0$, $\mathbb{E}[Z] = V$ and range$(Z) \leq 1$. Hence we observe that 

$$
\mathbb{E}[Z^2] \leq \text{range}(Z) \mathbb{E}[Z] \leq V. 
$$

Thus, we may apply Bernstein’s inequality which shows that with probability at least $1 - \delta$, after $n_*$ rounds, we are guaranteed that, 

$$
|\hat{V} - V| \leq \sqrt{\frac{2V \log(2/\delta)}{n_*}} + \frac{4 \log(2/\delta)}{3n_*}. 
$$

Equipped with the deviation bound we can complete the square to find that, 

$$
V - \sqrt{\frac{2V \log(2/\delta)}{n_*}} + \frac{\log(2/\delta)}{2n_*} \leq \hat{V} + \frac{2 \log(2/\delta)}{n_*} 
$$

$$
\Rightarrow (\sqrt{V} - \sqrt{\frac{\log(2/\delta)}{2n_*}})^2 \leq \hat{V} + \frac{2 \log(2/\delta)}{n_*} 
$$

$$
\Rightarrow \sqrt{V} \leq \sqrt{\frac{\log(2/\delta)}{2n_*}} + \sqrt{\hat{V} + \frac{2 \log(2/\delta)}{n_*}} 
$$

$$
\Rightarrow V \leq \left( \sqrt{\frac{\log(2/\delta)}{2n_*}} + \sqrt{\hat{V} + \frac{2 \log(2/\delta)}{n_*}} \right)^2 
$$

$$
\leq 2 \frac{\log(2/\delta)}{2n_*} + 2 \left( \hat{V} + \frac{2 \log(2/\delta)}{n_*} \right) 
$$

$$
\leq 2 \hat{V} + 5 \frac{\log(2/\delta)}{n_*} 
$$

27
Our definition of $\lambda$ uses $\tilde{\lambda}$ which is precisely this value, which we ensure is an upper bound on $V$. Working from the other side of the deviation bound, we know that,

$$
\tilde{\lambda} \leq \left( \sqrt{\lambda} + \sqrt{\frac{\log(2/\delta)}{2n_\star}} \right)^2 + \frac{\log(2/\delta)}{2n_\star} \leq 2\lambda + 2\frac{\log(2/\delta)}{n_\star}
$$

And combining the two, we see that,

$$
V \leq \tilde{\lambda} \leq 4V + \frac{9\log(2/\delta)}{n_\star},
$$

with probability at least $1 - \delta$.

### E.2 Parameter Estimation in Linear Regression

To control the regret associated with the exploitation rounds, we also need to bound $\|\hat{w} - w^*\|_2$ which follows from a standard analysis of linear regression.

At each round $t$, we solve a least squares problem with features $y_t(A_t)$ and response $r_t$ which we know has $E r_t | y_t, A_t = y_t(A_t)^T w^*$. The estimator is,

$$
\hat{w}_t = \arg\min_w \sum_{i=1}^t (y_i(A_i)^T w - r_i)^2
$$

Define the feature covariance,

$$
\Sigma_t = \sum_{i=1}^t y_i(A_i)y_i(A_i)^T
$$

which governs the estimation error of the least squares solution as we show in the next lemma. This analysis is standard.

**Lemma 16.** Let $\Sigma_t$ denote the feature covariance matrix after $t$ rounds of interaction and let $w_t$ be the least-squares solution. There is a universal constant $c > 0$ such that for any $\delta \in (0, 2/e)$, with probability $\geq 1 - \delta$:

$$
\|w_t - w^*\|_{\Sigma_t}^2 \leq cL \ln(2/\delta)
$$

**Proof.** This lemma is just the standard analysis of fixed-design linear regression with bounded noise. By definition of the ordinary least squares estimator, we have $w_t = \Sigma_t^{-1}Y_{1:t}^T r_{1:t}$ where $Y_{1:t} \in \mathbb{R}^{t \times L}$ is the matrix of features, $r_{1:t} \in \mathbb{R}^t$ is the responses and $\Sigma_t$ is the feature covariance defined above. The true weight vector can be written as $w^* = \Sigma_t^{-1}Y_{1:t}^T (r_{1:t} - \xi_{1:t})$ where $\xi_{1:t} \in \mathbb{R}^t$ is the noise vector. Thus:

$$
\|w_t - w^*\|_{\Sigma_t}^2 = \|\Sigma_t^{-1}Y_{1:t}^T \xi_{1:t}\|_{\Sigma_t}^2 = \xi_{1:t}^T Y_{1:t} \Sigma_t^{-1} Y_{1:t}^T \xi_{1:t}.
$$

Since $\Sigma_t^{-1} = (Y_{1:t}^TY_{1:t})^{-1}$ this matrix in the middle is a projection matrix, and it can be written as $UU^T$ where $U \in \mathbb{R}^{t \times d}$ is a matrix with orthonormal columns. We now have to bound the term $\|U^T \xi_{1:t}\|_2^2$. Note that the vector $\xi_{1:t}$ is a subgaussian random vector with independent components, so we can apply subgaussian tail bounds. Specifically, Lemma 24, due to [23], reveals that with probability $\geq 1 - \delta$:

$$
\|U^T \xi_{1:t}\|_2 \leq \sqrt{L} + \sqrt{c \ln(2/\delta)}
$$

for some universal constant $c > 0$. Squaring this inequality and using the upper bound on $\delta$ leads to the claim. \qed
E.3 Analysis of the Reward Covariance $\Sigma$

We now show that the reward covariance matrix has large eigenvalues. This enables us to translate the error in Lemma 16 to the euclidean norm, which we will see plays a roll in bounding the exploitation regret. Interestingly, we will show that the lower bound on the eigenvalues is related to the exploration regret, so that we can continue to explore until the eigenvalues are large, without incurring too much regret.

To prove the bound here, we consider a full sequence of exploration data, which enables us to bypass the data-dependent stopping time. Let $\{x_t, y_t, A_t, \xi_t\}_{t=1}^T$ be a sequence of random variables where $(x, y, \xi) \sim \mathcal{D}$ and $A_t$ is drawn uniformly at random. Let $w_t$ be the least squares solution on the data in this sequence up to round $t$, and let $\Sigma_t$ be the feature covariance for this sequence.

**Lemma 17.** With probability at least $1 - \delta$, for all $t \leq T$,

$$
\Sigma_t \succeq \left( tV - 2\sqrt{L^2tV\log(2LT/\delta)} - 2L\log(2LT/\delta) \right) I_L,
$$

where $I_L$ is the $L \times L$ identity matrix.

**Proof.** There are two main components to the proof, characterizing the mean $\mathbb{E}\Sigma_t$ and the deviation bound.

**Bounding $\mathbb{E}\Sigma_t$:** The first step in the proof is to analyze the expected value of the covariance. Since $y_t, A_t$ are identically distributed, it suffices to consider just one term. Fixing $x$ and $y$, we only reason about the randomness in picking $A$. Let $K$ denote the size of the feasible set of simple actions for this one round and let $S = \mathbb{E}_{A \sim P}[y(A)y(A)^T] \in \mathbb{R}^{L \times L}$ be the mean matrix for that round. We have:

$$
z^T S z = \sum_{t=1}^L z_t^2 z_t y(a)^2 + \sum_{t \neq t'} z_t z_{t'} y(a)y(a') = \frac{\|y\|^2}{K} + \sum_{t \neq t'} z_t z_{t'} y(a)y(a') = \mathbb{E}(y^2) - \frac{1}{K-1} \mathbb{E}y^2.
$$

Define $\bar{y} = \frac{1}{K} \sum_{a \in A} y(a)$, $\mathbb{E}(y^2) = \frac{1}{K} \sum_{a \in A} y(a)^2$ and $\text{Var}(y) = \mathbb{E}(y^2) - \bar{y}^2$. This expression becomes:

$$
z^T S z = \mathbb{E}(y^2)\|z\|^2 + \sum_{t \neq t'} z_t z_{t'} \left( \frac{K}{K-1} \bar{y}^2 - \frac{1}{K-1} \mathbb{E}y^2 \right) = \frac{K}{K-1} \text{Var}(y)\|z\|^2 + \left( \frac{K}{K-1} \bar{y}^2 - \frac{1}{K-1} \mathbb{E}y^2 \right) (z^T 1)^2
$$

The second term is positive for any $z$, as long as the $y$ vector is non-negative, as it is in our case. This means that for any unit vector $z$, we have:

$$
z^T S z \geq \frac{K}{K-1} \text{Var}(y) \geq \text{Var}(y)
$$

Since if $K = 1$ we only have one simple action available, so $\text{Var}(y) = 0$ but $z^T S z \geq 0$ since it is a quadratic form and $S \succeq 0$.

This argument applies for all $x, y$ and $z$, and consequently by linearity of expectation,

$$
\Sigma_t = \mathbb{E}_{x,y,A} S \succeq V I_L
$$
We will obtain the final result with a union bound over all \( t \) where for shorthand we are defining \( \alpha \). This follows since \( u \)\( \mathbb{X} \) Plugging in the definition of \( \mathbb{X} \) and Bernstein’s inequality gives that with probability \( \delta \),

\[
\sum_{i=1}^{t} z_i z_i^T = \sum_{i=1}^{t} \Pi z_i z_i^T \Pi + \sum_{i=1}^{t} \Pi^\perp z_i z_i^T \Pi^\perp.
\]

We will obtain the final result with a union bound over all \( t \leq T \).

In the 1 direction: Recall that we are still considering a fixed context \( x \) and also fix \( y \) for now, considering just the randomness in \( A \) as before. For the first term, it is more convenient to look at the quadratic form \( u^T z_i z_i^T u \), for which each term has mean,

\[
\mathbb{E}_A(u^T z)^2 = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}_A z_\ell^2 + \frac{1}{L} \sum_{\ell \neq \ell'} \mathbb{E}_A z_\ell z_{\ell'} = \mathbb{E}_a y(a)^2 + \frac{L - 1}{K(K - 1)} \sum_{a \neq b} y(a)y(b).
\]

\[
= \frac{K - L}{K - 1} \mathbb{E}_a y(a)^2 + \frac{K}{K - 1} \mathbb{y}^2 - \frac{L}{K - 1} \mathbb{E}_a y(a)^2.
\]

Now taking further expectation over the randomness in \( y \), we see that

\[
\mathbb{E}_{y, A}(u^T z)^2 = \frac{K}{K - 1} - V + \frac{L}{K - 1} (K \mathbb{E}_y \mathbb{y}^2 - \mathbb{E}_{y, A}(y(a)^2)) = \alpha + \beta
\]

where for shorthand we are defining \( \alpha = \frac{K}{K - 1} V \) and \( \beta \) as the residual term.

Let \( X = \sum_{i=1}^{t} u^T z_i z_i^T u = t(\alpha + \beta) \) be the mean over all \( t \) rounds. We now bound the variance and range of the random variable for applying Bernstein’s inequality.

The variance is bounded by the second moment, which by Holder’s inequality is bounded by,

\[
\sigma^2 = \sum_{i=1}^{t} \mathbb{E}(u^T z_i z_i^T u - \overline{u}^T z_i z_i^T u)^2 \leq \sum_{i=1}^{t} \mathbb{E}(u^T z_i z_i^T u)^2 \leq L \sum_{i=1}^{t} \mathbb{E} u^T z_i z_i^T u.
\]

This follows since \( u \) is unit normed and \( \|z\| \leq \sqrt{L} \). The same argument shows that the range is at most \( L \) and Bernstein’s inequality gives that with probability \( 1 - \delta/2 \),

\[
\left| \sum_{i=1}^{t} u^T z_i z_i^T u - X \right| \leq \sqrt{2LX \log(2/\delta)} + \frac{2L}{3} \log(2/\delta).
\]

Plugging in the definition of \( X \) into the deviation term, we can bound as,

\[
\sqrt{2LX \log(2/\delta)} + \frac{2L}{3} \log(2/\delta) \leq \sqrt{2Lt\alpha \log(2/\delta)} + \sqrt{2Lt\beta \log(2/\delta)} + \frac{2L}{3} \log(2/\delta)
\]
This means we can lower bound the empirical sum by,
\[
\sum_{i=1}^{t} u_{i}z_{i}^{T} u \geq t\alpha + t\beta - \sqrt{2Lt\alpha \log(2/\delta)} - \sqrt{2Lt\beta \log(2/\delta)} - \frac{2L}{3} \log(2/\delta)
\]
\[
= t\alpha - \sqrt{2Lt\alpha \log(2/\delta)} + \left( \sqrt{t\beta} - \sqrt{\frac{L}{2} \log(2/\delta)} \right)^2 - \frac{7L}{6} \log(2/\delta)
\]
\[
\geq t\alpha - \sqrt{2Lt\alpha \log(2/\delta)} - \frac{7L}{6} \log(2/\delta)
\]
\[
\geq tV - \sqrt{4LtV \log(2/\delta)} - \frac{7L}{6} \log(2/\delta),
\]
(15)
where the last inequality follows since \( V \leq \alpha \leq 2V \) for all \( K \geq 2 \) and \( V = 0 \) for \( K = 1 \). This suffices to prove the bound in the all-ones direction.

**Orthogonal to 1:** In the \( \Pi_{\perp} \) component we will apply the Matrix Bernstein inequality. This analysis is more straightforward. First, looking at our calculation of the mean matrix \( S \), we can see that,
\[
\mathbb{E}_{\Pi_{\perp}zz^{T}\Pi_{\perp}} = \frac{K}{K-1} V(I - uu^{T})
\]
The range of the random variable is still \( L \). Since we have i.i.d. random variables, the operator norm of the sum of variances is \( t \) times a single term, which is
\[
\| \mathbb{E}(\Pi_{\perp}zz^{T}\Pi_{\perp})^{2} \| \leq \max_{z} \| \Pi_{\perp}zz^{T}\Pi_{\perp} \| \mathbb{E} \text{tr}(\Pi_{\perp}zz^{T}\Pi_{\perp}) \leq \frac{L^{2}K}{K-1} V,
\]
where the inequality follows from Holder’s inequality from matrices and the fact that the matrix is positive semidefinite so the nuclear norm is precisely the trace. The last inequality follows from the definition of the mean matrix. Thus, the Matrix Bernstein inequality (Lemma 25) reveals that with probability at least \( 1 - \delta/2 \),
\[
\| \sum_{i=1}^{t} \Pi_{\perp}z_{i}z_{i}^{T}\Pi_{\perp} - \mathbb{E} \sum_{i=1}^{t} \Pi_{\perp}z_{i}z_{i}^{T}\Pi_{\perp} \| \leq \sqrt{\frac{2L^{2}KtV \log(2L/\delta)}{K-1} + \frac{2L}{3} \log(2L/\delta)}
\]
This implies that for any unit vector \( x \perp 1 \), we have
\[
x^{T} \left( \sum_{i=1}^{t} \Pi_{\perp}z_{i}z_{i}^{T}\Pi_{\perp} \right)x \geq tV - \sqrt{4Lt^{2}V \log(2L/\delta)} - \frac{2L}{3} \log(2L/\delta).
\]
(16)
Combining this bound with the bound in the all-ones direction in Eq. (15) proves the claim. \( \square \)

**E.4 Analysis of the Exploration Regret**

The analysis is made complicated here by the fact that the stopping time of the exploration phase is a random variable. If we let \( \hat{t} \) denote the last round of the exploration phase, this quantity is a random variable that depends on the history of interaction up to and including round \( \hat{t} \). Our proof here will use a high probability non-random bound \( t^{*} \) that satisfies \( \mathbb{P}(\hat{t} \leq t^{*}) \geq 1 - \delta \). We will compute \( t^{*} \) based on our analysis of the reward covariance \( \Sigma \).
A trivial bound on the exploration regret is,

$$
\sum_{t=1}^{t^*} r_t(\pi^*(x_t)) - r_t(A_t) \leq t^* \|w^*\| \sqrt{L}.
$$

This follows from the Cauchy-Schwarz inequality and the fact that the reward features are bounded in $[0, 1]$.

We also need a more precise bound on the exploration regret.

**Lemma 18.** Let $t^*$ be a non-random upper bound on $\hat{t}$ satisfying $\mathbb{P}(\{\hat{t} \leq t^*\}) \geq 1 - \delta$. Then with probability at least $1 - 2\delta$, the expected exploration regret is,

$$
\sum_{t=1}^{\hat{t}} r_t(\pi^*(x_t)) - r_t(A_t) \leq t^* \|w^*\| \sqrt{KV} + \|w^*\| \sqrt{2Lt^* \log(1/\delta)}
$$

**Proof.** Let $\{x_t, y_t, A_t, \xi_t\}_{t=1}^T$ be a sequence of random variables where $(x, y, \xi) \sim \mathcal{D}$ and $A_t$ is drawn uniformly at random. This sequence is adapted to a filtration $\{\mathcal{F}_t\}_{t=1}^T$. The random stopping time $\hat{t}$ is also measurable with respect to this filtration, so that $\{\hat{t} \leq t\} \in \mathcal{F}_t$. We are interested in bounding the probability of the event,

$$
\mathcal{E} = \{\sum_{t=1}^{\hat{t}} (y_t(\pi^*(x_t)) - y_t(A_t))^T w^* \leq \epsilon\}.
$$

This term is exactly the exploration regret, so we want to make sure the probability of this event is large. We will first apply the upper bound,

$$
\sum_{t=1}^{\hat{t}} (y_t(\pi^*(x_t)) - y_t(A_t))^T w^* \leq \sum_{t=1}^{\hat{t}} (y_t(A_t^*) - y_t(A_t))^T w^*,
$$

where $A_t^* = \arg\max_A y_t(A)^T w^*$ is the best possible composite action. This upper bound ensures that every term in the sum is non-negative. Now,

$$
\mathbb{P}(\mathcal{E}) \geq \mathbb{P}\left(\sum_{t=1}^{\hat{t}} (y_t(A_t^*) - y_t(A_t))^T w^* \leq \epsilon\right)
$$

$$
\geq \mathbb{P}\left(\sum_{t=1}^{\hat{t}} (y_t(A_t^*) - y_t(A_t))^T w^* \leq \epsilon \cap \{\hat{t} \leq t^*\}\right)
$$

$$
\geq \mathbb{P}\left(\sum_{t=1}^{t^*} (y_t(A_t^*) - y_t(A_t))^T w^* \leq \epsilon \cap \{\hat{t} \leq t^*\}\right)
$$

$$
\geq 1 - \mathbb{P}\left(\sum_{t=1}^{t^*} (y_t(A_t^*) - y_t(A_t))^T w^* > \epsilon\right) - \mathbb{P}\left(\{\hat{t} > t^*\}\right)
$$

$$
\geq 1 - \delta - \mathbb{P}\left(\sum_{t=1}^{t^*} (y_t(A_t^*) - y_t(A_t))^T w^* > \epsilon\right).
$$
The first line follows from the definition of \( A_t^{*} \) which only increases the sum, so decreases the probability of the event. The second inequality is immediate, while the third inequality holds because all terms in the sequence are non-negative. The fourth inequality is the union bound and the last inequality is by assumption on the event \( \{ \hat{t} \leq t^* \} \). The purpose of this exercise is to remove the dependence on the random stopping time and replace it with a deterministic number of terms in the sum.

Now we can apply a standard concentration analysis. The mean of the random variables is,

\[
E_{x,y,A}(y(A^*) - y(A))T w^* \leq \|w^*\| \|E_{x,y,A}(y(A^*) - y(A))\|
\]

\[
= \|w^*\| \|E_{x,y}(y(A^*) - \bar{y})\|
\]

\[
= \|w^*\| \sqrt{\sum_{\ell=1}^{L} (E_{x,y}(y(A_t^*) - \bar{y}))^2}
\]

\[
\leq \|w^*\| \sqrt{K \sum_{a} (\bar{y} - y(a))^2}
\]

\[
= \|w^*\| \sqrt{KV}
\]

The first inequality is Cauchy-Schwarz while the second inequality is Jensen’s inequality and the third comes from adding non-negative terms. The range of the random variable is easily bounded as,

\[
sup_{x,y,A} \left| (y(A^*) - y(A_t))w - E_{x,y,A}(y(A^*) - y(A_t))w^* \right| \leq \|w^*\| \sqrt{L}
\]

Since the coordinates of \( y \) are bounded in \([0,1]\). Thus Hoeffding’s inequality reveals that with probability at least \( 1 - \delta \),

\[
\sum_{t=1}^{t^*} (y(A^*) - y(A_t))w \leq \sum_{t=1}^{t^*} E_{x,y,A}(y(A^*) - y(A_t))w + \|w^*\| \sqrt{2Lt^* \log(1/\delta)}
\]

\[
\leq t^* \|w^*\| \sqrt{KV} + \|w^*\| \sqrt{2Lt^* \log(1/\delta)}
\]

\( \square \)

**E.5 Analysis of the Exploitation Regret**

In this section we show that after the exploration rounds, we can find a policy that has low expected regret. The technical bulk of this section involves a series of deviation bounds showing that we have good estimates of the expected reward for each policy.

Let \( \{(x_i, y_i)\}_{i=1}^n \) be a set of contexts and reward feature pairs where \( x_i \in X \) and \( y_i \in [0,1]^K \) with \( x_i, y_i \sim D \) i.i.d. Let \( \{A_i\}_{i=1}^n \) be composite actions chosen uniformly at random from the set of all ordered lists of \( L \) actions, so \( A_i \) is uniformy distributed over \( \binom{L}{n} L! L\)-sequences. Finally, let \( \hat{w} \) be any estimate of the true weight vector \( w^* \).
We define per-round versions of the reward variance, which will appear in the argument. We define this quantity as,

\[ v_t = \frac{1}{K} \sum_{a \in A} (y_t(a) - \bar{y}_t)^2, \] where \( \bar{y}_t = \frac{1}{K} \sum_{a' \in A} y_t(a') \).

Finally, we use the importance weighted feature vectors, which we define as,

\[ \hat{y}_t(a) = \frac{1}{\{a \in A_t\}} y_t(a) = \frac{K}{L} \{a \in A_t\} y_t(a). \]

The last identity follows because we are using the uniform exploration distribution.

The empirical reward estimate for a policy \( \pi \) is,

\[ \eta_n(\pi, \hat{w}) = \frac{1}{n} \sum_{i=1}^n \sum_{\ell=1}^L \hat{y}_i(\pi(x_i)_{\ell}) \hat{w}_{\ell}. \]

For shorthand, we write \( y(\pi(x)) \in \mathbb{R}^L \) to have components corresponding to \( y(\pi(x)_{\ell}) \), so that we can write this as an average of \( \hat{y}_i(\pi(x_i)) \hat{w} \).

A natural way to show that we find a policy with low expected regret is to show that for all policies \( \pi \), the empirical reward estimate \( \eta_n(\pi, \hat{w}) \) is close to the true reward, \( \eta(\pi) \), defined as,

\[ \eta(\pi) = \mathbb{E}_{x,y} y(\pi(x))^T w^*. \]

With such a concentration bound, maximizing \( \eta_n(\pi, \hat{w}) \) with via AMO produces a policy with low expected regret.

In our proof, rather than bound the deviation of \( \eta_n \) directly, we instead control a version of \( \eta_n \) that is shifted by \( \bar{y}_t \), defined as,

\[ \psi_n(\pi, \hat{w}) = \frac{1}{n} \sum_{i=1}^n \sum_{\ell=1}^L \hat{y}_i(\pi(x_i)_{\ell}) \hat{w}_{\ell} - \bar{y}_t 1_T \hat{w}, \]

where \( \bar{y} \) is the expected value of the reward feature when an action is chosen uniformly at random and \( 1_L \) is the \( L \)-dimensional all-ones vector. Note in particular that \( \bar{y}_i \) is based on the \( y \) values of all actions, even those that were not chosen at round \( i \). However, this is not an issue since we use \( \bar{y}_i \) only as a proof device, not in the algorithm.

**Lemma 19.** Fix \( \delta \in (0, 1) \) and assume that \( \| \hat{w} - w^* \|_2 \leq \theta \) for some \( \theta > 0 \). For any \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \), we have that for all \( \pi \in \Pi \),

\[ \left| \psi_n(\pi, \hat{w}) - \eta(\pi, w^*) + \mathbb{E}_{x,y} \bar{y} 1_T w^* \right| \]

\[ \leq 4(\theta + \| w^* \|) \sqrt{K} \left( \frac{\log(N/\delta)}{n} + \frac{\sqrt{K} \log(N/\delta)}{L} \right) + \theta \min\{ \sqrt{KV}, 2\sqrt{L} \}. \]

**Proof.** We add and subtract several terms to obtain a decomposition, For shorthand, we subscript by \( \pi \) to
where the last inequality uses $K$. There are two terms to control here. The first term we will control using Bernstein’s inequality, using that fact that $\hat{y}_i$ is coordinate-wise unbiased for $y$. The second term we will control with a purely deterministic analysis, and we will relate this term to $v_i$, the variance of the reward features.

**Term 1:** Note that this term is centered, since $\hat{y}_i$ is an unbiased estimate. Moreover the range of each individual term in the sum can be bounded as,

$$
|\left(\hat{y}_{i,\pi} - \hat{y}_i 1_L - E_{x,y} y_{\pi} - \bar{y} 1_L \right)^T \hat{w}| \leq ||\hat{w}|| ||\hat{y}_{i,\pi} - \hat{y}_i 1_L - E_{x,y} y_{\pi} - \bar{y} 1_L||
$$

To arrive at the second line we are doing two things. First we add and subtract $w^*$ and apply the triangle inequality on the first term. On the second term, notice that these are all $L$-dimensional vectors, and except for the first one, the coordinates are bounded in $[0, 1]$ The first vector has coordinates bounded in $[0, K/L]$, so the norm itself is bounded by $\sqrt{K^2/L}$.

The variance can be bounded by the second moment, which is,

$$
E_{x,y,A} \left(\left(\hat{y}_\pi - \bar{y} 1_L \right)^T \hat{w}\right)^2 \leq ||\hat{w}||^2 E_{x,y,A} \sum_{\ell=1}^L (\hat{y}(\pi(x)\ell) - \bar{y})^2
$$

$$
= ||\hat{w}||^2 E_{x,y,A} \sum_{\ell=1}^L \left(\frac{K}{L} 1\{\pi(x)\ell \in A\} y(\pi(x)\ell) - \bar{y}\right)^2
$$

$$
\leq ||\hat{w}||^2 E_{x,y,A} \sum_{\ell=1}^L \left(\frac{2K^2}{L^2} 1\{\pi(x)\ell \in A\} + 2\right)
$$

$$
\leq 4(\theta + ||w^*||)^2 K,
$$

where the last inequality uses $K \geq L$. Bernstein’s inequality implies that with probability at least $1 - \delta$, for all $\pi \in \Pi$, we have,

$$
\left|\frac{1}{n} \sum_{i=1}^n (\hat{y}_{i,\pi} - \hat{y}_i 1_L)^T \hat{w} - E_{x,y} (y_{\pi} - \bar{y} 1_L)^T \hat{w}\right| \leq \sqrt{\frac{8(\theta + ||w^*||)^2 K \log(N/\delta)}{3n}} + \frac{4(\theta + ||w^*||)K}{3n\sqrt{L}} \log(N/\delta)
$$

**Term 2:** For the second term, we use the Cauchy-Schwarz inequality,

$$
E_{x,y} (y_{\pi} - \bar{y} 1_L)^T (\hat{w} - w^*) \leq ||E_{x,y} (y_{\pi} - \bar{y} 1_L)||_2 ||\hat{w} - w^*||_2
$$
The difference in the weight vectors will be controlled by our analysis of the least squares problem. We need to bound the other quantity here and we will use two different bounds. First,

\[ \|E_{x,y} (y\pi - \bar{y} 1_L)\| \leq E_{x,y} \|y\pi - \bar{y}\| \leq E_{x,y} \|y\pi\| + \|\bar{y}\| \leq 2\sqrt{L}. \]

Second

\[ \|E_{x,y} (y\pi - \bar{y} 1_L)\| = \sqrt{K} \sqrt{\frac{1}{K} \sum_{a \in A} (y(a) - \bar{y})^2} = \sqrt{KV} \]

Here \( V \) is the expected variance of the reward feature when choosing a single action according to the uniform distribution as in Eq. 13.

**Combining everything:** Putting everything together, the bound we have is,

\[ (\theta + \|w^*\|) \left[ \sqrt{\frac{8K \log(N/\delta)}{n}} + \frac{4K \log(N/\delta)}{3n\sqrt{L}} \right] + \theta \min\{\sqrt{KV}, 2\sqrt{L}\} \]

Collecting terms together proves the main result.

Assume that we explore for \( T_1 \) rounds and then call AMO with weight vector \( \hat{w} \) and importance weighted rewards \( \hat{y}_1, \ldots \hat{y}_{T_1} \) to produce a policy \( \hat{\pi} \) that maximizes \( \eta_{T_1}(\pi, \hat{w}) \). With an application of Lemma 19 we know that,

\[ \eta(\pi^*, w^*) - \eta(\hat{\pi}, \hat{w}) = \eta(\pi^*, w^*) - \psi_{T_1}(\pi^*, \hat{w}) + \psi_{T_1}(\hat{\pi}, \hat{w}) - \eta(\hat{\pi}, \hat{w}) + \eta_{T_1}(\pi^*, \hat{w}) - \eta_{T_1}(\hat{\pi}, \hat{w}) \leq 8(\theta + \|w^*\|) \sqrt{K} \left( \sqrt{\frac{\log(N/\delta)}{T_1}} + \sqrt{\frac{K \log(N/\delta)}{L \frac{T_1}{T}}} \right) + 2\theta \min\{\sqrt{KV}, 2\sqrt{L}\} \]

**Proposition 20.** Assume that we explore for \( T_1 \) rounds, we find \( \hat{w} \) satisfying \( \|\hat{w} - w^*\| \leq \theta \), and that the event in Lemma 19 holds. Then the expected exploitation regret is at most,

\[ 8(T - T_1)(\theta + \|w^*\|) \sqrt{K} \left( \sqrt{\frac{\log(N/\delta)}{T_1}} + \sqrt{\frac{K \log(N/\delta)}{L \frac{T_1}{T}}} \right) + 2(T - T_1)\theta \min\{\sqrt{KV}, 2\sqrt{L}\} \]

Of course the exploitation regret minus the expected exploitation regret forms a martingale with bound \( \|w^*\|\sqrt{L} \) by Cauchy-Schwarz. This means that with probability at least \( 1 - \delta \), the exploitation regret is,

\[ \text{Exploitation Regret} \leq 8T(\theta + \|w^*\|) \sqrt{K} \left( \sqrt{\frac{\log(N/\delta)}{T_1}} + \sqrt{\frac{K \log(N/\delta)}{L \frac{T_1}{T}}} \right) + 2T\theta \min\{\sqrt{KV}, 2\sqrt{L}\} + \|w^*\|\sqrt{2LT\log(1/\delta)} \]
E.6 Proving the Final Bound

Choosing the stopping time $t^*$: The first step here is to select the stopping time $t^*$. We set,

$$t^* = \max \left\{ \frac{3\lambda^*}{V}, n^* \right\}$$

This setting of $t^*$ verifies that the stopping time $\hat{t}$ satisfies $P(\hat{t} \leq t^*) \geq 1 - \delta$ through Lemma 17. Specifically, Lemma 17 ensures that after $t^*$ rounds the covariance matrix satisfies,

$$\Sigma_{t^*} \succeq \left( t^*V - 2\sqrt{2t^*V \log(2LT/\delta)} - 2L \log(2LT/\delta) \right) I_L$$

with probability at least $1 - \delta$. If this quantity on the right hand side is larger than $\lambda^*$, then we can be sure that $\hat{t} \leq t^*$. Our setting of $t^*$ verifies this bound. To see why if we complete the square above, we want,

$$\left( \sqrt{t^*V} - \sqrt{L^2 \log(2LT/\delta)} \right)^2 - L^2 \log(2LT/\delta) \geq \lambda^*$$

$$\Rightarrow \left( \sqrt{t^*V} - \sqrt{L^2 \log(2LT/\delta)} \right)^2 \geq \lambda^* + 3L^2 \log(2LT/\delta) \geq \lambda^*$$

$$\Rightarrow t^* \geq \frac{1}{V} \left( \sqrt{\lambda^* + 3L^2 \log(2LT/\delta)} + \sqrt{L^2 \log(2LT/\delta)} \right)^2$$

Our setting is an upper bound on this quantity, using the elementary inequalities $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ and $(a + b)^2 \leq 2a^2 + 2b^2$ and the fact that we set $\lambda^* \geq 9L^2 \log(2LT/\delta)$.

We set,

$$\lambda^* = \max \left\{ 9L^2 \log(2LT/\delta), (T\tilde{V}/B)^{2/3} (L \log(2/\delta))^{1/3} \right\}$$

$$n^* = T^{2/3} (K \log(N/\delta)/L)^{1/3} \max\{1, (B\sqrt{L})^{-2/3}\}$$

Recall that we assume $T > K \log(N/\delta)/L$.

For $\lambda^*$, since we are using an estimate of $V$, we must be extremely careful. By Lemma 15 and Eq. 14, we know $V \leq \tilde{V} \leq 4V + \tau$ where $\tau = \frac{7 \log(2/\delta)}{n^*}$.

The overall exploration regret is bounded via Lemma 18 and Eq. 17,

$$\text{Exploration Regret} \leq \|w^*\| \min\{t^*, T\} \min\{\sqrt{K\tilde{V}}, \sqrt{L} \} + \|w^*\| \sqrt{2LT \log(1/\delta)}$$

$$\leq B \min\left\{ \frac{3\lambda^*}{B}, T \right\} \min\{\sqrt{K\tilde{V}}, \sqrt{L} \} + n^* B\sqrt{L} + B \sqrt{2LT \log(1/\delta)} \quad \text{Term 1, Term 2, Term 3}$$

Meanwhile, for the exploitation regret, we apply Proposition 20 with $\theta = \sqrt{cL \log(2/\delta)/\lambda^*}$, which follows from Lemma 16 and the fact that $\lambda_{\min}(\Sigma) \geq \lambda^*$ during the entire exploitation phase. This bound is therefore
at most,

$$\text{Exploitation Regret} \leq 8TB \left( \sqrt{\frac{K \log(N/\delta)}{n_*}} + \frac{K \log(N/\delta)}{\sqrt{L} n_*} \right) +$$

$$+ 4T \sqrt{\frac{cL \log(2/\delta)}{\lambda_*}} \left( \sqrt{\frac{K \log(N/\delta)}{n_*}} + \frac{K \log(N/\delta)}{\sqrt{L} n_*} \right) +$$

$$+ 4T \sqrt{\frac{cL \log(2/\delta)}{\lambda_*}} \min \{ \sqrt{KV}, 2\sqrt{L} \} + B \sqrt{2LT \log(1/\delta)}$$

This follows because we know that $$n_* \leq T_1 \leq T$$ and $$||w^*|| \leq B$$.

We now use our settings of $$n_*$$ and $$\lambda_*$$ to bound all of the terms.

**Term 1:** For Term 1, we must use a case analysis. First, assume that $$\lambda_*$$ is the second term in its definition and assume that $$V \geq \tau \triangleq 7 \log(2/\delta)/n_*$$.

In this case, we can bound,

$$\text{Term 1} \leq \frac{3\lambda_* B \min \{ \sqrt{KV}, \sqrt{L} \}}{V} \leq \frac{3B(T\tilde{V}/B)^{2/3}(L \log(2/\delta))^{1/3} \min \{ \sqrt{KV}, \sqrt{L} \}}{V} \leq 9B^{1/3}T^{-2/3}V^{-1/3}(L \log(2/\delta))^{1/3} \min \{ \sqrt{KV}, \sqrt{L} \}$$

This argument uses the setting of $$t^* = 3\lambda_* / V$$ and the deviation bound for $$\tilde{V}$$ in Eq. 14. We now show that the term involving $$V$$ and the min is always bounded. We prove,

**Claim 21.**

$$V^{-1/3} \min \{ \sqrt{KV}, \sqrt{L} \} \leq K^{1/3}L^{1/6}$$

**Proof.** We consider two cases. If $$KV \leq L$$ then $$V \leq L / K$$ while the expression is equal to $$V^{1/6} \sqrt{K}$$. On the other hand, if $$L \leq KV$$, then $$V \geq L / K$$ but the expression is equal to $$V^{-1/3} \sqrt{L}$$. Both cases yield the bound. \(\square\)

Thus in this case Term 1 is $$\tilde{O} \left( T^{2/3}L^{1/2}(BK)^{1/3} \right)$$.

Now if $$\lambda_*$$ is still the second term in its definition but $$V \leq \tau$$, then we know that $$V \leq 7\log(2/\delta)/n_*$$.

This means we can bound Term 1 as,

$$\text{Term 1} \leq TB \sqrt{KV} \leq TB \sqrt{\frac{7K \log(2/\delta)}{n_*}} \leq \text{Term 4}.$$

We'll use the bound in Term 4 to control this case.
The next case for Term 1 is when $\lambda_*$ is the first term in its definition. In this case, we know that,

$$9L^2 \log(2LT/\delta) \geq (T\tilde{V}/B)^{2/3}(L \log(2/\delta))^{1/3}$$

$$\Rightarrow V \leq 27L^{5/2} \log(2LT/\delta)^{3/2}(B/T)(\log(2/\delta))^{-1/2} \quad (18)$$

this follows since $\tilde{V} > V$. So we have the bound,

$$= \min\{T, \frac{3\lambda_*}{V}\} \min \left\{ \sqrt{KV}, \sqrt{L} \right\} \leq TB \sqrt{KV}$$

$$= O \left( T^{1/2} B^{3/2} L^{5/4} K^{1/2} \log(LT/\delta) \right).$$

In summary, we have the bound,

$$\text{Term 1} \leq \text{Term 4} + O \left( \left( T^{2/3} L^{1/2} (BK)^{1/3} + T^{1/2} B^{3/2} L^{5/4} K^{1/2} \right) \log(LT/\delta) \right) \quad (19)$$

**Term 2:** Term 2 is fairly straightforward. Our definition of $n_*$ ensures that,

$$n_* B \sqrt{L} \leq T^{2/3} (K \log(N/\delta))^{1/3} \max\{BL^{1/6}, B^{1/3} L^{-1/6}\}$$

Replacing the max by the sum, we have the bound

$$\text{Term 2} \leq O \left( T^{2/3} (K \log(N/\delta))^{1/3} \left( BL^{1/6} + B^{1/3} L^{-1/6} \right) \right) \quad (20)$$

**Term 3 and Term 7:** Term 3 and 7 are identical and we will leave them as is.

$$\text{Term 3} = \text{Term 7} = O \left( B \sqrt{LT \log(1/\delta)} \right) \quad (21)$$

**Term 4 and 5:** We should handle these terms together. First observe that,

$$\text{Term 5} = 4T \sqrt{\frac{cL \log(2/\delta)}{\lambda_*}} \left( \sqrt{\frac{K \log(N/\delta)}{n_*}} + \frac{K \log(N/\delta)}{\sqrt{L} n_*} \right)$$

$$\leq 4T \sqrt{L} \left( \sqrt{\frac{K \log(N/\delta)}{n_*}} + \frac{K \log(N/\delta)}{\sqrt{L} n_*} \right).$$

This is true because $\lambda_* \geq 9L^2 \log(2LT/\delta) \geq L^2 \log(2/\delta)$. Thus,

$$\text{Term 4} + \text{Term 5} \leq O \left( T \left( B + \frac{1}{\sqrt{L}} \right) \left( \sqrt{\frac{K \log(N/\delta)}{n_*}} + \frac{K \log(N/\delta)}{\sqrt{L} n_*} \right) \right)$$

Now if $B \geq \frac{1}{\sqrt{L}}$, the max in the definition of $n_*$ is achieved by the 1 term, and the bound is,

$$O \left( T^{2/3} BL^{1/6} (K \log(N/\delta))^{1/3} + T^{1/3} BL^{-1/6} (K \log(N/\delta))^{2/3} \right)$$
If \( B \leq \frac{1}{\sqrt{L}} \) then the max is achieved by the second term and the bound is,

\[
O \left( T^{2/3}B^{1/3}L^{-1/6}(K \log(N/\delta))^{1/3} + T^{1/3}B^{2/3}L^{-1/3}(K \log(N/\delta))^{2/3} \right)
\]

Either way, the leading order term is,

\[
O \left( T^{2/3}(K \log(N/\delta))^{1/3} \left( BL^{1/6} + B^{1/3}L^{-1/6} \right) \right)
\]

(22)

**Term 6:** The last term we need to address is Term 6,

\[
\text{Term 6} = 2T \sqrt{cL \log(2/\delta)} \min \{ \sqrt{KV}, \sqrt{L} \}
\]

Again here there are two cases, if \( \lambda^* = 9L^2 \log(2LT/\delta) \) then we know that \( V \) must be very small. In particular by Eq. (18),

\[
V \leq 27L^{5/2} \log(2LT/\delta)^{1/2} (B/\delta)(\log(2/\delta))^{-1/2}
\]

We then have,

\[
\text{Term 6} \leq 2T \sqrt{cL \log(2/\delta)} \min \{ \sqrt{KV}, \sqrt{L} \} \leq O \left( \sqrt{BTKL^{3/4} \log(LT/\delta)} \right)
\]

In the other case,

\[
\text{Term 6} \leq 2T \sqrt{cL \log(2/\delta)} \min \{ \sqrt{KV}, \sqrt{L} \}
\]

\[
\leq O \left( T \sqrt{L \log(2/\delta)} \min \{ \sqrt{KV}, \sqrt{L} \} \right)
\]

\[
= \tilde{O} \left( T^{2/3}B^{1/3}L^{1/6} \min \{ \sqrt{KV}, \sqrt{L} \} \right)
\]

\[
= \tilde{O} \left( T^{2/3}(BK)^{1/3}L^{1/2} \right),
\]

(23)

where the last step comes from Claim 21. This is the leading order term for Term 6, and it has the same form as Term 1.

**Putting everything together** Combining Eqs. 19, 20, 21, 22, and 23, we have controlled all terms in the Exploration and Exploitation regret. This yields a regret bound with leading order term,

\[
\text{Regret} = \tilde{O} \left( T^{2/3}(K \log(N/\delta))^{1/3} \max \{ B^{1/3}L^{1/2}, BL^{1/6} \} \right).
\]

Obtaining this bound required a union bound over a constant number of events, so the bound holds with high probability.
F Deviation Bounds

We collect here several deviation bounds that we use in our proofs. All of these results are well known and we point to references rather than provide the proofs. The first inequality, which is a Bernstein-type deviation bound for martingales, is Freedman’s inequality, which is from Beygelzimer et. al [4]

Lemma 22 (Freedman’s Inequality). Let $X_1, X_2, \ldots, X_T$ be a sequence of real-valued random variables. Assume for all $t \in \{1, 2, \ldots, T\}$ that $X_t \leq R$ and $E[X_t|X_1, \ldots, X_{t-1}] = 0$. Define $S = \sum_{t=1}^T X_t$ and $V = \sum_{t=1}^T E[X_t^2|X_1, \ldots, X_{t-1}]$. For any $\delta \in (0, 1)$ and $\lambda \in [0, 1/R]$, with probability at least $1 - \delta$:

$$S \leq (e - 2)\lambda V + \frac{\ln(1/\delta)}{\lambda}$$

We also use Azuma’s inequality, a Hoeffding-type inequality for martingales.

Lemma 23 (Azuma’s Inequality). Let $X_1, X_2, \ldots, X_T$ be a sequence of real-valued random variables. Assume for all $t \in \{1, 2, \ldots, T\}$ that $X_t \leq R$ and $E[X_t|X_1, \ldots, X_{t-1}] = 0$. Define $S = \sum_{t=1}^T X_t$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$S \leq R\sqrt{2T \ln(1/\delta)}$$

We also make use of a vector-valued version of Hoeffding’s inequality, due to Rudelson and Vershynin [23].

Lemma 24 (Vector-valued subgaussian concentration). Let $A \in \mathbb{R}^{m \times n}$ be a fixed matrix, and let $X = (X_1, \ldots, X_n)$ be independent random variables with $E[X] = 0$ and $|X_i| \leq 1$ almost surely. Then there is a universal constant $c > 0$ such that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\|AX\|_2 - \|A\|_F \leq \|A\|_2 \sqrt{c \ln(1/\delta)}$$

Finally, we use a well known matrix-valued version of Bernstein’s inequality, for example from Tropp [26].

Lemma 25 (Matrix-Bernstein). Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint, matrices with dimension $d$. Assume that each random matrix, we have $E[X_k] = 0$ and $\lambda_{\text{max}}(X_k) \leq R$ almost surely. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\lambda_{\text{max}}(\sum_k X_k) \leq \sqrt{2\sigma^2 \ln(d/\delta)} + \frac{2}{3} R \log(d/\delta)$$

with $\sigma^2 = \left\| \sum_k E[X_k^2] \right\|_2$