Offline-to-Online Learning in Linear Bandits

Kushagra Chandak, Toshinori Kitamura, Xiaoqi Tan

Keywords: Linear bandits, offline-to-online, regret minimization

Summary

We study online learning with an additional offline dataset in the stochastic linear bandit setting. Although this problem arises frequently in practice, the offline-to-online tradeoff remains poorly understood in structured environments. We propose a linear bandit algorithm that balances this tradeoff: it relies on offline data during early rounds, and increasingly favors exploration as the horizon grows. We establish regret bounds showing that our method is simultaneously competitive with both purely online and purely offline solutions. In particular, it achieves sublinear regret relative to the optimal action in the number of online interactions, while its regret relative to an offline reference decreases as the number of offline samples grows. Empirical results further demonstrate its effectiveness across various problem parameters.

Contribution(s)

1. This paper presents an algorithm for online learning with an additional offline dataset in the stochastic linear bandit setting with regret bounds.
   Context: Prior work provides an algorithm for the multi-armed bandit case (Sentenac et al., 2025).
2. Empirical results showing competitiveness of our algorithm against an online and an offline solution simultaneously
   Context: None

Offline-to-Online Learning in Linear Bandits

Kushagra Chandak, Toshinori Kitamura, Xiaoqi Tan
{kchandak,toshinor,xiaoqi.tan}@ualberta.ca

Department of Computing Science, University of Alberta, Canada

Abstract

We study online learning with an additional offline dataset in the stochastic linear bandit setting. Although this problem arises frequently in practice, the offline-to-online tradeoff remains poorly understood in structured environments. We propose a linear bandit algorithm that balances this tradeoff: it relies on offline data during early rounds, and increasingly favors exploration as the horizon grows. We establish regret bounds showing that our method is simultaneously competitive with both purely online and purely offline solutions. In particular, it achieves sublinear regret relative to the optimal action in the number of online interactions, while its regret relative to an offline reference decreases as the number of offline samples grows. Empirical results further demonstrate its effectiveness across various problem parameters.

1 Introduction

We study the *offline-to-online learning* problem in the stochastic linear bandit model (Lattimore & Szepesvári, 2020), where a learner sequentially selects actions and receives noisy linear rewards while having access to an offline dataset. Such offline-to-online learning problems arise frequently in practice, for example, a recommendation system typically has access to historical user-preference data which can be used to improve the next version of the system. Despite its practical importance, the offline-to-online learning problem remains poorly understood in structured environments, even in the simple linear bandit setting.

In offline learning, a common approach is *pessimism in the face of uncertainty* (Li et al., 2022; Xiao et al., 2021), which yields solutions that are competitive with respect to the offline data. However, such methods do not explore and may incur large regret relative to the optimal policy when the learning horizon is long. In contrast, online learning algorithms typically rely on *optimism in the face of uncertainty* (Li et al., 2010; Abbasi-Yadkori et al., 2011), which encourages exploration and achieves sublinear online regret. Nevertheless, when the learning horizon is short, optimistic strategies may explore excessively and consequently suffer large regret relative to the policy derived from the offline data. The central challenge in the offline-to-online setting is therefore to balance these two principles, effectively trading off between exploiting the offline solution and exploring to improve long-term performance (Sentenac et al., 2025).

We achieve this balance by managing an *exploration budget* in the linear bandit setup. Intuitively, by selecting the pessimistic action computed from the offline data, the learner can ensure temporarily low regret relative to the offline dataset, thereby accumulating budget that can later be used for exploration. Once sufficient budget has been accumulated, the learner can select optimistic actions to explore potentially better alternatives. Building on this idea, in Section 4, we propose the algorithm LinOtO. In each online round, LinOtO carefully balances between exploration using upper confidence bounds (UCB) and budget accumulation using lower confidence bounds (LCB). Proposition 2 formally characterizes the budget accumulation by using LCB in the stochastic linear bandit model.

We show that, with high probability, the performance of our algorithm is simultaneously close to both the optimism-based online solution (Theorem 1) and the pessimism-based offline solution (Theorem 2). Moreover, in Section 6, we conduct numerical experiments across a range of problem settings and demonstrate that LinOtO performs nearly as well as the better of the offline and online solutions. Overall, our work contributes to a growing line of research that leverages prior knowledge or hints to guide the exploration–exploitation tradeoff in online learning (Cutkosky et al., 2022; Wei et al., 2020; Bhaskara et al., 2023; Alamdari et al., 2024).

2 Related Work

There is an extensive body of literature on both online and offline learning. In this section, we focus on three areas closely related to our offline-to-online setting.

Warm starting online learning with offline data. The problem of leveraging offline data to warm start online learning has been studied across various settings. For the multi-armed bandit (MAB) case, Shivaswamy & Joachims (2012) proposed the HUCB1 algorithm, which improves the regret of the standard UCB1 (Auer et al., 2002) by constructing tighter confidence intervals. Cheung & Lyu (2024) generalized this result to scenarios with distribution shifts between the offline and the online data, establishing tight upper and lower bounds. The work of He et al. (2024) considers the case where offline and online data have different feedback structures for MAB, also proposing algorithms based on UCB. Hao et al. (2023) proposed a Thompson Sampling (TS) algorithm for linear bandits and derived Bayesian regret bounds when the offline data is generated by an expert. The recent work of Vijayan et al. (2025) proposes an elimination-based algorithm for linear bandits using an extended D-optimal design for exploration, providing near-optimal regret bounds for warm-starting with offline data. In contrast to our work, they solely focus on online regret and do not consider the notion of regret relative to an offline reference. We refer readers to Vijayan et al. (2025) for an extensive discussion on warm starting online learning using offline data.

Online learning with hints. Our problem can also be viewed through the lens of online learning with hints, prior knowledge, or side information. Cutkosky et al. (2022) considered the linear bandit problem in which a hint provides a (possibly imperfect) estimate of the optimal action. They proposed an algorithm that improves upon the standard regret guarantees when the hint is accurate while remaining robust to inaccuracies. More recently, Bhaskara et al. (2023) investigated a similar problem in the adversarial linear bandit optimization setting. In a different direction, Wei et al. (2020) studied the contextual bandit problem in both stochastic and adversarial settings where hints are given as predictions of losses, showing that the algorithm can achieve improved regret bounds when the prediction error is small. In all of these works, hints are typically provided as *black-box* side information. This contrasts with our setting, where prior knowledge is available in the structured form of an offline dataset.

Conservative exploration. In conservative exploration (Wu et al., 2016; Kazerouni et al., 2017), algorithms balance exploration and safety by assuming access to a known safe baseline action. Our offline-to-online learning setting can be cast within this framework by treating the pessimistic action computed from the offline dataset as the safe baseline. This perspective was adopted by Sentenac et al. (2025) in the multi-armed bandit (MAB) setting, whose work is most closely related to ours. They propose an algorithm that balances the UCB1 (Auer et al., 2002) and LCB algorithms, and provide a detailed analysis of the tradeoffs between UCB1 and LCB. However, their analysis is restricted to the unstructured MAB setting, in which the regret scales with the number of actions. This dependence can become prohibitive when the action space is large. In contrast, in linear bandits the regret scales with the dimension of the action vectors, which allows for substantially improved scalability in high-dimensional action spaces.

3 Problem Setting

Notation. The set $\{1, \dots, n\}$ is denoted by $[n]$ for any $n \in \mathbb{N}$. The set of probability distributions over any set S is denoted by $\mathcal{M}_1(S)$ and the cardinality of S is denoted by $|S|$. We write the inner product between two vectors $x, y \in \mathbb{R}^d$ as $\langle x, y \rangle = x^\top y$. For a symmetric positive definite matrix A, we denote $\|x\|_A = \sqrt{x^\top A x} = \sqrt{\langle Ax, x \rangle}$ for any vector x. For two positive semidefinite matrices A and B, we write $A \preceq B$ if $B - A$ is positive semidefinite. The trace of a matrix A (sum of its diagonal entries) is denoted by $\mathrm{tr}(A)$ and the determinant of A is denoted by $\det(A)$. Throughout the paper, $\log$ is used to denote the natural logarithm. We use $\tilde{O}(\cdot)$ to hide logarithmic factors in the problem parameters.

3.1 Linear Bandits with Offline Data

Let $\mathcal{A} = \{a_1, \dots, a_k\} \subset \mathbb{R}^d$ be the set of actions (vectors) that the learner can choose and $\mathcal{D} \coloneqq (\mathcal{D}_a)_{a \in \mathcal{A}}$ be the offline dataset. Similar to previous works (Sentenac et al., 2025; Vijayan et al., 2025; Cheung & Lyu, 2024), we consider the setting in which the offline data is collected by a fixed design, i.e., the actions are chosen by a non-adaptive policy with deterministic action counts. More concretely, for each $a \in \mathcal{A}$, $\mathcal{D}_a \coloneqq (X_1^{(a)}, \dots, X_{m_a}^{(a)})$ is the dataset corresponding to action a and m_a denotes the number of rewards sampled from action a. The reward samples $X_i^{(a)} \sim P_a$ for all $i \in [m_a]$, where $\nu \coloneqq (P_a)_{a \in \mathcal{A}}$ are the reward distributions for each action constituting the *unknown* environment. Given the *offline* dataset $\mathcal{D}$, the learner interacts *online* with the environment ν over a sequence of $n \in \mathbb{N}$ rounds. At every round $t \in [n]$, the learner chooses action $A_t \in \mathcal{A}$, and observes a reward $X_t \sim P_{A_t}$. The goal of the learner is to choose actions to maximize the total reward collected over these n rounds while also leveraging the offline dataset $\mathcal{D}$ to improve the performance. We make the following standard assumptions on the reward distributions and the actions.

Assumption 1 (Linear reward function)**.** *There exists an unknown parameter $\theta_* \in \mathbb{R}^d$ such that $X_t = \langle \theta_*, A_t \rangle + \eta_t$ and $X_i^{(a)} = \langle \theta_*, a \rangle + \eta_i^{(a)}$ for all $t \in [n]$, $i \in [m_a]$ and $a \in \mathcal{A}$, where η_t and $\eta_i^{(a)}$ are the noise terms.*

Assumption 2 (Subgaussian noise)**.** *The noise terms $(\eta_t)_{t \in [n]}$ are conditionally 1-subgaussian given the past online and the offline data. That is, if $\mathcal{F}_t = \sigma(\mathcal{D}, A_1, X_1, \dots, A_{t-1}, X_{t-1}, A_t)$ is the σ-algebra generated by the past online data up to round t and the offline dataset $\mathcal{D}$, then for all $\lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp(\lambda \eta_t) \mid \mathcal{F}_t] \le \exp\left(\lambda^2/2\right)$ almost surely. Further, $(\eta_i^{(a)})_{i \in [m_a], a \in \mathcal{A}}$ are independent 1-subgaussian, i.e., $\mathbb{E}\left[\exp(\lambda \eta_i^{(a)})\right] \le \exp(\lambda^2/2)$ for all $\lambda \in \mathbb{R}$.*

Assumption 3 (Bounded actions, parameters, and rewards)**.** *There exists constants $B, L > 0$ such that $\|\theta_*\|_2 \le B$, $\|a\|_2 \le L$ and $\langle \theta_*, a \rangle \in [0, 1]$ for all $a \in \mathcal{A}$.*

3.2 Performance Metrics

We denote the mean reward of action $a \in \mathcal{A}$ by $\mu_a \coloneqq \langle \theta_*, a \rangle$. We measure the performance of any policy π, which may depend on the history and the offline data, relative to two baselines: (1) the highest mean reward of any arm $\mu_* \coloneqq \max_{a \in \mathcal{A}} \mu_a$ and (2) the mean reward of the offline dataset $\mu_{\text{off}} \coloneqq \sum_{a \in \mathcal{A}} m_a \mu_a / m$, where $m = \sum_{a \in \mathcal{A}} m_a$ is the total number of offline reward samples. The pseudo-regret relative to μ_* is

$$R_n(\pi, \nu) \coloneqq \sum_{t=1}^{n} \left(\mu_* - \langle \theta_*, A_t \rangle\right) ,$$

and the pseudo-regret relative to μ_{off} is

$$R_n^{\text{off}}(\pi, \nu) \coloneqq \sum_{t=1}^{n} \left(\mu_{\text{off}} - \langle \theta_*, A_t \rangle\right) .$$

Note that both $R_n(\pi, \nu)$ and $R_n^{\text{off}}(\pi, \nu)$ are random quantities because the sequence of actions $(A_t)_{t\in[n]}$ may be randomized. The pseudo-regret relative to μ_* is the standard notion of regret while the pseudo-regret $R_n^{\text{off}}(\pi, \nu)$ captures the loss of performance of π compared to our offline reference μ_{off}, induced by the offline sample counts $(m_a)_{a\in\mathcal{A}}$. Together, the two metrics characterize different aspects of the offline-to-online tradeoff: $R_n(\pi, \nu)$ measures how far the algorithm is from optimal performance, while $R_n^{\text{off}}(\pi, \nu)$ quantifies the improvement (or degradation) relative to the historical action distribution represented in the offline dataset. When the dependence on ν is clear from the context, we write $R_n(\pi)$ and $R_n^{\text{off}}(\pi)$ for brevity.

3.3 Parameter Estimation

As is common in the linear bandit literature (Abbasi-Yadkori et al., 2011), we estimate the true parameter θ_* using the least squares estimator $\hat{\theta}_t$ constructed at the *end* of round t using both the offline and the online data:

$$\hat{\theta}_t = V_t^{-1} \left(\sum_{a\in\mathcal{A}} \sum_{i=1}^{m_a} X_i^{(a)} a + \sum_{s=1}^{t} X_s A_s \right) . \tag{1}$$

Here, $V_t = V_0 + \sum_{s=1}^{t} A_s A_s^\top$ is the covariance matrix of both the offline and the online data and $V_0 = \sum_{a\in\mathcal{A}} m_a a a^\top$ is the covariance matrix corresponding to only the offline data.[1]

4 Method

In the offline-to-online setting, the goal is to minimize both the regret against the optimal action, R_n, and the regret against the offline reference, R_n^{off}. Upper Confidence Bound (UCB) and Lower Confidence Bound (LCB) methods have been shown to be effective in achieving strong online and offline performance, respectively (Abbasi-Yadkori et al., 2011; Li et al., 2022; Auer et al., 2002; Xiao et al., 2021). This section presents our algorithm LinOtO, which carefully switches between UCB and LCB strategies to achieve competitive performance with respect to both R_n and R_n^{off}.

To build confidence bounds, we construct confidence sets $\mathcal{C}_t$ around the least squares parameter estimate $\hat{\theta}_t$ given by the following proposition. The proof is deferred to Appendix A.

Proposition 1. *Let $\delta \in (0, 1)$. Then with probability at least $1 - \delta$, it holds that for all $t \in [n]$*

$$\left\| \hat{\theta}_{t-1} - \theta_* \right\|_{V_{t-1}} \le \sqrt{\beta_{\text{off}}} + \sqrt{\beta_t} ,$$

where $\sqrt{\beta_{\text{off}}} = \sqrt{8d \log 6 + 8 \log(2/\delta)}$ is the confidence radius corresponding to the offline data and $\sqrt{\beta_t} = \sqrt{2 \log(2/\delta) + \log\left(\det V_{t-1} / \det V_0\right)}$ with $\beta_1 = 0$ is the confidence radius from both the offline and the online data. Furthermore, for all $t \in [n]$, let

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d \colon \left\| \theta - \hat{\theta}_{t-1} \right\|_{V_{t-1}} \le \sqrt{\rho_t} \right\} , \tag{2}$$

where $\sqrt{\rho_t} \coloneqq \sqrt{\beta_{\text{off}}} + \sqrt{\beta_t}$. Then $\mathbb{P}(\text{exists } t \in [n] \colon \theta_* \notin \mathcal{C}_t) \le \delta$.

Using the confidence set $\mathcal{C}_t$, LinOtO switches between an optimistic action U_t and a pessimistic action L_t given by

$$\begin{aligned} &U_t = \arg\max_{a\in\mathcal{A}} \text{UCB}_t(a) , \quad L_t = \arg\max_{a\in\mathcal{A}} \text{LCB}_t(a), \\ &\text{where } \text{UCB}_t(a) = \sup_{\theta\in\mathcal{C}_t} \langle \theta, a \rangle , \quad \text{LCB}_t(a) = \max_{t'\le t} \inf_{\theta\in\mathcal{C}_{t'}} \langle \theta, a \rangle . \end{aligned} \tag{3}$$

Essentially, the UCB action U_t can reduce the regret R_n by optimism but may increase the regret R_n^{off}, while the LCB action L_t can reduce the regret R_n^{off} by pessimism but may increase the regret R_n. We balance between these two actions to achieve a good performance for both R_n and R_n^{off}.

[1] For analysis purposes, we assume that V_0 is invertible; if not, we can add a small regularization term to ensure invertibility.

4.1 Algorithm

We control the balance between R_n and R_n^{off} using an *exploration budget* accumulated by choosing the pessimistic action L_t. Recall that R_n^{off} represents the reward deficit of an algorithm relative to μ_{off}, and the LCB strategy can avoid large budget deficits in multi-armed bandits (Sentenac et al., 2025). Based on this intuition, we record the algorithm's excess reward over μ_{off} and use it as an exploration budget to decide when to choose the optimistic action U_t for exploration. If choosing U_t would not significantly deplete the budget, the algorithm proceeds with exploration; otherwise, it selects L_t to accumulate additional budget before exploring further.

This switching mechanism is inspired by conservative bandits (Wu et al., 2016; Kazerouni et al., 2017), and a similar technique was used by Sentenac et al. (2025) in the multi-armed offline-to-online bandit setting. This budget-based exploration approach crucially relies on constructing a *safe* baseline action—namely, the LCB action in our setting—that is guaranteed to achieve a reward not too far from μ_{off}. However, the budget gained from the safe LCB action was previously unknown in the linear bandit setting. The following proposition provides the key to constructing this baseline and relies on a technical result that bounds the average uncertainty of actions over the offline dataset. The proof can be found in Appendix B.1.

Proposition 2. *Fix $\delta_1 > 0$. Let $\mu_{off} = (1/m)\sum_{a\in\mathcal{A}} m_a \langle \theta_*, a\rangle$ and $L_{off} = \arg\max_{a\in\mathcal{A}} LCB_1(a)$ be the LCB action computed from the offline dataset. Then with probability at least $1-\delta_1$,*

$$\mu_{off} - \mu_{L_{off}} \le 2\sqrt{\frac{d\beta_{off}}{m}},$$

where $\beta_{off} = 8d\log 6 + 8\log(1/\delta_1)$.

Proposition 2 implies that the mean reward of L_{off} is not too far from the mean reward of the offline dataset. Since L_{off} can be computed directly from the offline data, we construct our baseline reward using a lower bound on $\mu_{L_{\text{off}}}$ (with some slack). That is, we define our baseline safe reward as

$$r_b \coloneqq \text{LCB}_1(L_{\text{off}}) - \alpha S\,, \tag{4}$$

where α is a tuning parameter and $S \coloneqq \sqrt{d\beta_{\text{off}}/m}$ represents the slack allowed from the lower bound on the offline LCB action's reward. Now we are ready to define the exploration budget $B(t)$ at the beginning of round t:

$$B(t) \coloneqq \sum_{s=1}^{t-1} \text{LCB}_t(A_s) + \text{LCB}_t(U_t) - tr_b\,. \tag{5}$$

Our algorithm LinOtO computes $B(t)$ at the beginning of each round t using the LCB values derived from the confidence set $\mathcal{C}_t$ (Proposition 1) and the baseline reward r_b (eq. (4)). If $B(t) > 0$, choosing U_t will not cause the total reward to drop below r_b; therefore, LinOtO selects U_t. Otherwise, it chooses L_t and accumulates an exploration budget of $\text{LCB}_t(A_t) - r_b \ge \text{LCB}_1(L_{\text{off}}) - r_b = \alpha S$. After executing the chosen action $A_t \in \{U_t, L_t\}$, the algorithm observes a reward sample $X_t \sim P_{A_t}$. Finally, it updates the estimate $\hat{\theta}_t$ using eq. (1) and the covariance matrix V_t, and constructs the confidence set $\mathcal{C}_{t+1}$ for the next round using eq. (2). The pseudocode for LinOtO is presented in Algorithm 1.

Remark 1 (Computational complexity)**.** *We make the following remarks on the computational complexity of LinOtO.*

1. *Computing U_t and L_t on line 4 in Algorithm 1 requires solving a bilinear optimization program, which is known to be intractable in general. However, for the finite action set considered in our setting, it can be solved easily. Computation can further be improved by using the Sherman-Morrison formula to incrementally update V_t^{-1}. Therefore the per-round computational complexity of computing the actions becomes $O(|\mathcal{A}| + |\mathcal{A}|d^2)$ (Lattimore & Szepesvári, 2020, Chapter 19).*

Algorithm 1: Offline-to-Online Algorithm for Linear Bandits (LinOtO)

Input: Offline dataset $\mathcal{D}$, horizon n, baseline tuning parameter α

Compute $\hat{\theta}_0$ and V_0 using $\mathcal{D}$ and construct the confidence set $\mathcal{C}_1$ using eq. (2)
Compute baseline reward r_b using eq. (4)
for $t = 1$ **to** n **do**
    Compute the UCB action U_t and the LCB action L_t using $\mathcal{C}_t$ as in eq. (3)
    Compute budget $B(t)$ using eq. (5)
    if $B(t) > 0$ **then**
        Choose action $A_t = U_t$
    else
        Choose action $A_t = L_t$
    Observe reward X_t
    Update V_t and $\hat{\theta}_t$ using eq. (1), and construct confidence set $\mathcal{C}_{t+1}$ using eq. (2)

2. *Recall that the LCB value is* $\text{LCB}_t(a) = \max_{t' \le t} \inf_{\theta \in \mathcal{C}_{t'}} \langle \theta, a \rangle$. *We can reduce the computational overhead of evaluating* $\text{LCB}_t(a)$ *by maintaining a running maximum at each round* t.
3. *Instead of computing the sum of individual LCBs in eq. (5), our implementation computes the LCB of the aggregate action vector* $\sum_{s=1}^{t-1} A_s$. *This heuristic substantially reduces the computational cost, since the algorithm can incrementally maintain the cumulative action vector* $\sum_{s=1}^{t-1} A_s$ *without re-evaluating* $\text{LCB}_t(A_s)$ *for every past round* s. *Empirically, we observe that both implementations exhibit very similar performance.*

Remark 2 (Unknown horizon)**.** *Algorithm 1 can be adapted to the unknown horizon setting using the doubling trick: run the algorithm in epochs of length* $n = 2^k$ *for* $k = 0, 1, 2, \ldots$, *resetting all historical computations (except the offline estimates) at the end of each epoch. This adapted algorithm enjoys almost identical regret guarantees to Algorithm 1. For further details, see Sentenac et al. (2025).*

5 Regret Analysis of LinOtO

In this section, we present regret bounds for LinOtO with respect to both R_n and R_n^{off}, establishing its competitiveness against both LinUCB and LinLCB (the algorithm that chooses the LCB action in every round). The full proofs of all the results in this section are long and therefore deferred to the supplementary section.

5.1 Regret relative to μ_*

We bound the regret of LinOtO against the optimal action by decomposing R_n into rounds where the UCB (U_t) and LCB (L_t) actions are chosen:

$$\begin{aligned} R_n(\text{LinOtO}) &= \sum_{t \in \mathcal{U}_n} \left(\langle \theta_*, a_* \rangle - \langle \theta_*, U_t \rangle \right) + \sum_{t \in \mathcal{L}_n} \left(\langle \theta_*, a_* \rangle - \langle \theta_*, L_t \rangle \right) \\ &\le \sum_{t \in \mathcal{U}_n} \left(\langle \theta_*, a_* \rangle - \langle \theta_*, U_t \rangle \right) + |\mathcal{L}_n| \max_{t \in \mathcal{L}_n} \left(\langle \theta_*, a_* \rangle - \langle \theta_*, L_t \rangle \right) \\ &\le \sum_{t \in \mathcal{U}_n} \left(\langle \theta_*, a_* \rangle - \langle \theta_*, U_t \rangle \right) + 2\sqrt{\beta_{\text{off}}} \, |\mathcal{L}_n| \, \|a_*\|_{V_0^{-1}} \ , \end{aligned}$$

where $\mathcal{U}_n$ and $\mathcal{L}_n$ denote the rounds where the learner chooses UCB and LCB actions, respectively. In the last line, we used the following suboptimality bound for LCB, which is a special case of Li et al. (2022, Theorem 1).

Lemma 1 (LCB suboptimality w.r.t optimal action)**.** *Suppose that the event $\mathcal{E} = \{\theta_* \in \mathcal{C}_t \text{ for all } t \in [n]\}$ holds. Then for any $t \in [n]$,*

$$\langle \theta_*, a_* \rangle - \langle \theta_*, L_t \rangle \leq 2\sqrt{\beta_{\text{off}}} \|a_*\|_{V_0^{-1}} .$$

The regret during $\mathcal{U}_n$ is bounded via a LinUCB analysis adapted for offline data. Thus, the key is to control the regret during $\mathcal{L}_n$ by bounding the total number of times the LCB action is chosen, namely, $|\mathcal{L}_n|$.

Bounding $|\mathcal{L}_n|$. By eq. (5), LinOtO chooses L_t only to cover budget deficits ($B(t) < 0$), generating an αS surplus that is then consumed by U_t. As the confidence bounds shrink over time (from eq. (2)), the lower confidence bound of U_t will eventually exceed the baseline r_b. Once this occurs, playing U_t becomes self-funding, and L_t will no longer be chosen. By equating the budget deficit incurred by U_t with the αS increments, we can bound the total number of LCB actions by:

$$|\mathcal{L}_n| = \tilde{O}\left(\frac{d^2}{\alpha S(\Delta_{\text{off}} + \alpha S)}\right) ,$$

where $\Delta_{\text{off}} = \langle \theta_*, a_* - L_{\text{off}} \rangle$. For the full proof, see Appendix C.

Finally, to state the regret bound, we define an "effective dimension" (Valko et al., 2014):

$$d_{\text{eff}} = \max\left\{ i \in [d] \,:\, (i-1)\lambda_i(V_0) \leq \frac{nL^2}{\log(1 + nL^2/\lambda_1(V_0))} \right\} ,$$

where $\lambda_1(V_0) \leq \cdots \leq \lambda_d(V_0)$ are the eigenvalues of the offline covariance matrix V_0. The effective dimension d_{eff} represents the maximum number of directions $i \in [d]$ for which the corresponding eigenvalues $\lambda_j(V_0)$ (for $j \leq i$) are small. It captures the maximum number of directions where the offline data lacks sufficient coverage. Note that $d_{\text{eff}} \leq d$.

By combining the above discussions, we have the following regret bound on R_n.

Theorem 1. *With probability at least $1 - \delta$, the pseudo-regret of LinOtO is bounded as*

$$R_n(\textit{LinOtO}) = \tilde{O}\left(\sqrt{n\rho_n d_{\textit{eff}} \log\left(1 + \frac{nL^2}{\lambda_{\min}(V_0)}\right)} + \frac{d^{2.5}\|a_*\|_{V_0^{-1}}}{\alpha S(\Delta_{\textit{off}} + \alpha S)}\right) .$$

When $\lambda_{\min}(V_0) = \Omega(mL^2/d)$ and $\Delta_{\textit{off}} + \alpha S \gtrsim \sqrt{\rho_n d/m}$, then with probability $1 - \delta$, $R_n(\textit{LinOtO}) = O\left(n\sqrt{\rho_n d d_{\textit{eff}}/m}\right)$.

Remark 3 (Comparison with LinUCB)**.** *The regret bound of LinOtO consists of two terms: the first corresponds to the regret due to UCB actions, and the second corresponds to the regret from LCB actions. The first term is sublinear in the horizon n and depends on the offline data through $d_{\textit{eff}}$. If the offline data is not well-covered ($d_{\textit{eff}} = d$), we recover the standard $\tilde{O}(d\sqrt{n})$ bound for LinUCB (Abbasi-Yadkori et al., 2011). However, when there is good offline coverage with large m, the standard LinUCB bound improves to $O\left(n\sqrt{\rho_n d d_{\textit{eff}}/m}\right)$. Moreover, the second term is independent of n (up to log factors), and hence not growing with online data. Therefore, overall, the regret of LinOtO is comparable to LinUCB up to an additive constant.*

Remark 4 (Analysis of α)**.** *When α is large, the budget is more likely to be positive and LinOtO behaves closer to LinUCB. From Theorem 1, the regret bound in this case is similar to that of LinUCB. When α is small, then LinOtO mostly plays the LCB actions and its regret corresponding to the LCB plays increases as shown in Theorem 1. A more refined analysis of how to tune α is left for future work.*

5.2 Regret relative to μ_{off}

Now we turn to R_n^{off}, the regret of LinOtO relative to the offline reference μ_{off}. The intuition here is that choosing the LCB action incurs a constant regret per round (from Proposition 2 and the

monotonicity of LCB values), and the confidence widths decrease whenever the UCB action is chosen (from eq. (2)). Therefore, the cumulative regret relative to μ_{off} is of the same order as the offline optimal LinLCB.

Theorem 2. *With probability at least* $1-\delta$*, we have*

$$R_n^{\text{off}}(\textit{LinOtO}) = O\left((\alpha+2)\sqrt{\frac{d\rho_1}{m}}n\right).$$

To compare the performance of LinOtO against LinLCB, we next state the regret of LinLCB relative to μ_{off}.

Proposition 3. *With probability* $1-\delta$*, we have*

$$R_n^{\text{off}}(\textit{LinLCB}) = O\left(\sqrt{\frac{d\rho_n}{m}}n\right).$$

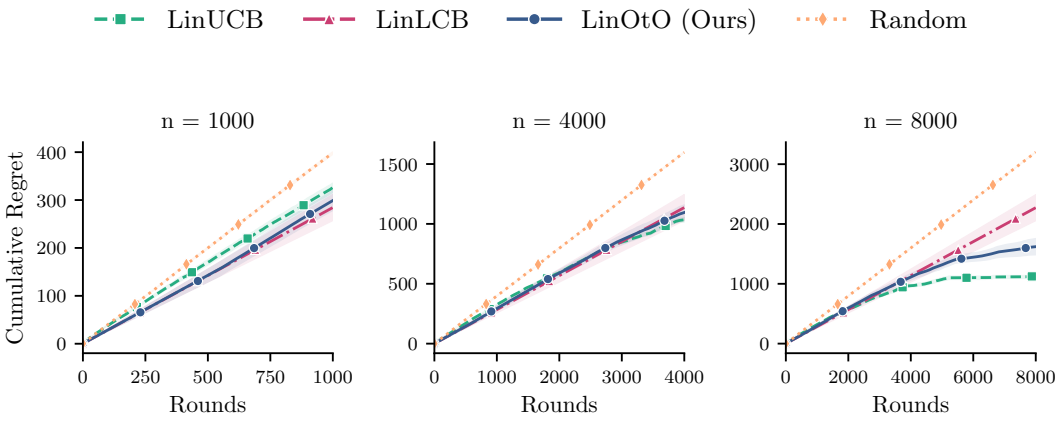
From Theorem 2 and Proposition 3, we observe that the regret of LinOtO is of the same order as LinLCB as ρ_1 and ρ_n differ only by a factor of $\log n$. Note that R_n^{off} scales linearly with the online horizon n because LinLCB does not explore. However, the per-round regret vanishes as $O(1/\sqrt{m})$. Moreover, there is a Pareto tradeoff between $R_n(\text{LinOtO})$ and $R_n^{\text{off}}(\text{LinOtO})$ in terms of α. For large α, say $O(\sqrt{m})$, the second term in Theorem 1 becomes smaller improving the online regret R_n. However, R_n^{off} becomes a constant in m. On the other hand, setting $\alpha = O(1)$ preserves the $1/\sqrt{m}$ decay of R_n^{off} at the cost of a larger R_n bound.

6 Experimental Results

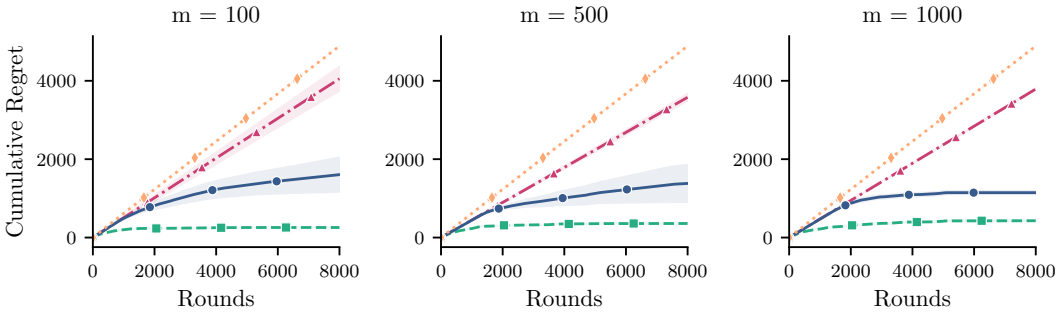
In this section, we evaluate the empirical performance of LinOtO on synthetic linear bandit tasks to demonstrate its competitiveness against warm-started LinUCB, LinLCB, and a random policy. For the default setting, we generate $|\mathcal{A}| = 100$ actions uniformly over unit sphere in dimension $d = 20$. The true parameter is also generated randomly over the unit sphere and α is set to 10^{-4}. Both the offline and the online reward noises are sampled from standard normal distribution. The offline dataset is constructed with $m = 100$ samples evenly distributed over 10 suboptimal actions. We add a small regularization term of 10^{-4} to the offline covariance matrix to ensure invertibility.

Varying the Horizon Length. In Figure 1a, we fix the offline dataset size to $m = 100$ and vary the online horizon length $n \in \{1000, 4000, 8000\}$ which is an input to Algorithm 1, and plot the cumulative pseudo-regret R_n over rounds. We observe that in the early rounds, LinOtO successfully leverages the offline data to maintain a low initial regret, closely tracking the safe baseline of LinLCB. As n increases and the exploration budget permits, LinOtO transitions to online exploration. It eventually achieves a sublinear regret similar to LinUCB. In contrast, LinLCB continues to follow the conservative baseline and therefore incurs linear regret.

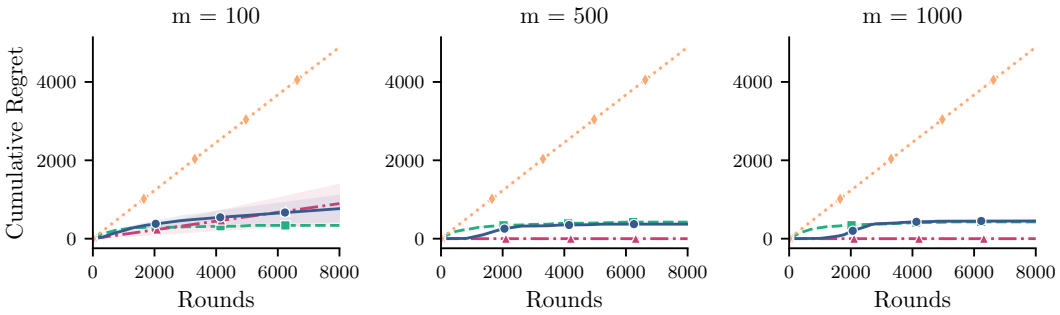
Varying Offline Data Size (Excluding Optimal Action). Next, we examine the impact of the offline data size $m \in \{100, 500, 1000\}$ in a setting where the dataset excludes the optimal action. In this regime, the offline reference action L_{off} can be strictly suboptimal. Figure 1b plots the cumulative regret for $n = 8000$ with different offline sample sizes m. Because LinLCB does not explore, it commits to a suboptimal safe action and therefore incurs linear regret. In contrast, LinOtO uses its exploration budget to safely deviate from the suboptimal offline prior. Although its performance initially resembles that of LinLCB, it eventually explores and achieves sublinear regret. When $m = 1000$, the performance of LinOtO improves further and becomes closer to that of LinUCB, consistent with the theoretical insights discussed in Section 5.



(a) Cumulative regret across varying horizon lengths. For a smaller horizon ($n = 1000$), the performance of LinOtO closely tracks LinLCB due to limited initial exploration. However, for a larger horizon ($n = 8000$), LinOtO successfully explores and achieves a sublinear regret profile matching LinUCB.



(b) Cumulative regret for varying offline dataset sizes with no optimal actions in the offline dataset. We vary the number of offline samples m (excluding samples of the optimal action) while fixing the horizon to $n = 8000$. For $m = 1000$, we set $\alpha = 0.01$. LinOtO achieves a regret bounded between LinUCB and LinLCB, converging to the performance of LinUCB for large m



(c) Cumulative regret for varying offline sample size m with optimal actions in the offline dataset. LinLCB performs best by immediately exploiting the high-quality offline data. Both LinUCB and LinOtO incur some exploration penalty but quickly converge.

Figure 1: Cumulative regret of LinUCB, LinLCB, LinOtO, and a random policy (Random) as a function of number of online rounds. The shaded area is the standard error measured over 5 independent runs for all the algorithms.

Varying Offline Data Size (Including Optimal Action). In Figure 1c, we evaluate the performance of LinOtO when the offline dataset contains samples from the optimal action. The performance of all algorithms (except Random) improves as the number of offline samples m increases.

As expected, LinLCB quickly identifies the optimal action and achieves (near) zero regret. Both LinOtO and LinUCB perform a small amount of exploration but rapidly converge to the optimal action. Additional experimental results are provided in Appendix E.

7 Conclusion

In this work, we studied the problem of offline-to-online learning in the stochastic linear bandit model. By carefully switching between the LinUCB and LinLCB strategies, we developed the algorithm LinOtO, which is designed to achieve strong performance relative to both online learning and the offline baseline. Theoretically, we established regret bounds showing that LinOtO is simultaneously competitive with both LinUCB and LinLCB. We also validated these theoretical guarantees through empirical evaluations across a range of environmental parameters.

A primary limitation of the current framework is the assumption of fixed-design offline data, rather than data collected through an adaptive logging policy. Moreover, the offline reference does not account for the coverage properties of the offline dataset. Consequently, the offline regret bounds scale with the total number of offline samples rather than with a more refined coverage-dependent quantity. Addressing these limitations remains an important open problem. Additional future directions include studying distribution shift between offline data and online interactions, as well as extending the framework to nonlinear function approximation settings.

Acknowledgments

KC and XT acknowledge support from Alberta Machine Intelligence Institute (Amii), Alberta Major Innovation Fund, and NSERC Discovery Grant RGPIN-2022-03646.

A Confidence Sets Construction

We start by constructing the confidence set $\mathcal{C}_1$ from the offline data. The least squares estimate $\hat{\theta}_0$ of θ_* is given by

$$\hat{\theta}_0 = V_0^{-1} \sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} X_i^{(a)} a \,,$$

where $V_0 = \sum_{a \in \mathcal{A}} m_a a a^\top$ is the offline covariance matrix. To get the confidence set around $\hat{\theta}_0$, we use the following lemma, which can be obtained using concentration of subgaussian random variables and a covering argument (Lattimore & Szepesvári, 2020, Equation 20.3).

Lemma 2 (Lattimore & Szepesvári (2020))**.** *For all $\delta_0 \in (0,1)$, we have*

$$\mathbb{P}\left(\left\| \sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \eta_i^{(a)} a \right\|_{V_0^{-1}} \geq \sqrt{\beta_{\mathit{off}}(\delta_0)} \right) \leq \delta_0,$$

where $\sqrt{\beta_{\text{off}}(\delta_0)} := \sqrt{8d \log 6 + 8 \log(1/\delta_0)}$ is the offline confidence radius. The next lemma gives a uniform concentration guarantee for all $t \in [n]$ based on a martingale argument (Lattimore & Szepesvári, 2020, Theorem 20.4).

Lemma 3 (Lattimore & Szepesvári (2020))**.** *For all $\delta_1 \in (0,1)$,*

$$\mathbb{P}\left(\textit{exists } t > 1 \,:\, \left\| \sum_{s=1}^{t} \eta_s A_s \right\|_{V_t^{-1}} \geq \sqrt{\beta_{t+1}(\delta_1)} \right) \leq \delta_1,$$

where $\sqrt{\beta_t(\delta_1)} := \sqrt{2 \log(2/\delta_1) + \log(\det V_{t-1} / \det V_0)}$ with $V_t = V_0 + \sum_{s=1}^{t} A_s A_s^\top$.

Combining Lemmas 2 and 3, we get

Lemma 4. *For all $\delta \in (0,1)$, there exists $\delta_0, \delta_1 \in (0,1)$ such that*

$$\mathbb{P}\left(\textit{exists } t > 1 \,:\, \left\| \sum_{s=1}^{t} \eta_s A_s \right\|_{V_t^{-1}} \geq \beta_{t+1}(\delta_1) \;\; \textit{or} \;\; \left\| \sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \eta_i^{(a)} a \right\|_{V_0^{-1}} \geq \beta_{\textit{off}}(\delta_0) \right) \leq \delta \,.$$

Proof. The proof follows from Lemmas 2 and 3 by choosing $\delta_0 = \delta_1 = \delta/2$ and a union bound. □

A.1 Proof of Proposition 1

Proof. Using the definition of $X_i^{(a)}$ and X_s, we have

$$\begin{aligned}
\hat{\theta}_t - \theta_* &= V_t^{-1} \left(\sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} X_i^{(a)} a + \sum_{s=1}^{t} X_s A_s \right) - \theta_* \\
&= V_t^{-1} \left(\sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \left(\langle \theta_*, a \rangle + \eta_i^{(a)} \right) a + \sum_{s=1}^{t} (\langle \theta_*, A_s \rangle + \eta_s) A_s \right) - \theta_* \\
&= V_t^{-1} \left(\sum_{a \in \mathcal{A}} m_a a a^\top \theta_* + \sum_{s=1}^{t} A_s A_s^\top \theta_* + \sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \eta_i^{(a)} a + \sum_{s=1}^{t} \eta_s A_s \right) - \theta_* \\
&= V_t^{-1} \left(V_t \theta_* + \sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \eta_i^{(a)} a + \sum_{s=1}^{t} \eta_s A_s \right) - \theta_* \\
&= V_t^{-1} \left(\sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \eta_i^{(a)} a + \sum_{s=1}^{t} \eta_s A_s \right) .
\end{aligned}$$

Noting that $\left\| B^{-1} x \right\|_B = \|x\|_{B^{-1}}$ for any positive definite matrix B, we have

$$\begin{aligned}
\left\| \hat{\theta}_t - \theta_* \right\|_{V_t} &= \left\| \sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \eta_i^{(a)} a + \sum_{s=1}^{t} \eta_s A_s \right\|_{V_t^{-1}} \\
&\leq \left\| \sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \eta_i^{(a)} a \right\|_{V_t^{-1}} + \left\| \sum_{s=1}^{t} \eta_s A_s \right\|_{V_t^{-1}} \\
&\leq \left\| \sum_{a \in \mathcal{A}} \sum_{i=1}^{m_a} \eta_i^{(a)} a \right\|_{V_0^{-1}} + \left\| \sum_{s=1}^{t} \eta_s A_s \right\|_{V_t^{-1}} && \text{(since } V_0 \preceq V_t\text{)} \\
&\leq \sqrt{\rho_t} \,.
\end{aligned}$$

where the last step holds with probability at least $1 - \delta$ from Lemma 4. □

B Baseline Construction

Our baseline construction using Proposition 2 requires the following technical lemma.

Lemma 5. *Let $V_0 = \sum_{a \in \mathcal{A}} m_a a a^\top$ and $m = \sum_{a \in \mathcal{A}} m_a$. Then*

$$\frac{1}{m} \sum_{a \in \mathcal{A}} m_a \|a\|_{V_0^{-1}} \leq \sqrt{\frac{d}{m}}.$$

Proof. First, using Cauchy-Schwarz inequality and $\sum_{a\in\mathcal{A}} m_a/m = 1$, we have

$$\sum_{a\in\mathcal{A}} \frac{m_a}{m} \|a\|_{V_0^{-1}} \le \sqrt{\sum_{a\in\mathcal{A}} \frac{m_a}{m}} \sqrt{\sum_{a\in\mathcal{A}} \frac{m_a}{m} \|a\|^2_{V_0^{-1}}} = \sqrt{\sum_{a\in\mathcal{A}} \frac{m_a}{m} \|a\|^2_{V_0^{-1}}}\,.$$

Using the cyclicity of trace, the last expression inside the square root can be rewritten as

$$\operatorname{tr}\left(\sum_{a\in\mathcal{A}} \frac{m_a}{m} a^\top V_0^{-1} a\right) = \operatorname{tr}\left(V_0^{-1} \frac{1}{m} \sum_{a\in\mathcal{A}} m_a a a^\top\right) = \frac{1}{m} \operatorname{tr}\left(V_0^{-1} V_0\right) = \frac{d}{m}\,.$$

Therefore $\sum_{a\in\mathcal{A}} (m_a/m) \|a\|_{V_0^{-1}} \le \sqrt{d/m}$. □

B.1 Proof of Proposition 2

Proof. Since $\theta_* \in \mathcal{C}_1$ with high probability, we have

$$\begin{aligned}
\mu_{L_{\text{off}}} &\ge \text{LCB}_1(L_{\text{off}}) \\
&\ge \frac{1}{m} \sum_{a\in\mathcal{A}} m_a \inf_{\theta\in\mathcal{C}_1} \langle \theta, a\rangle && \text{(since max} \ge \text{avg)} \\
&= \frac{1}{m} \sum_{a\in\mathcal{A}} m_a \inf_{\theta\in\mathcal{C}_1} \left(\langle \theta_*, a\rangle - \langle \theta_* - \theta, a\rangle\right) \\
&= \mu_{\text{off}} - \frac{1}{m} \sum_{a\in\mathcal{A}} m_a \sup_{\theta\in\mathcal{C}_1} \langle \theta_* - \theta, a\rangle \\
&\ge \mu_{\text{off}} - \frac{2\sqrt{\beta_{\text{off}}}}{m} \sum_{a\in\mathcal{A}} m_a \|a\|_{V_0^{-1}} && \text{(by Cauchy-Schwarz)} \\
&\ge \mu_{\text{off}} - 2\sqrt{\frac{d\beta_{\text{off}}}{m}}. && \text{(by Lemma 5)}
\end{aligned}$$

Rearranging the above proves the result. □

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Neural Information Processing Systems*, 2011.

Parand A Alamdari, Yanshuai Cao, and Kevin H Wilson. Jump starting bandits with LLM-generated prior knowledge. *Conference on Empirical Methods in Natural Language Processing*, 2024.

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

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Bandit online linear optimization with hints and queries. *International Conference on Machine Learning*, 2023.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. *International Conference on Machine Learning*, 2024.

Ashok Cutkosky, Chris Dann, Abhimanyu Das, and Qiuyi Zhang. Leveraging initial hints for free in stochastic linear bandits. *International Conference on Algorithmic Learning Theory*, 2022.

Botao Hao, Rahul Jain, Tor Lattimore, Benjamin Van Roy, and Zheng Wen. Leveraging demonstrations to improve online learning: Quality matters. *International Conference on Machine Learning*, 2023.

Qijia He, Minghan Wang, Xutong Liu, Zhiyong Wang, and Fang Kong. Learning across the gap: Hybrid multi-armed bandits with heterogeneous offline and online data. *Neural Information Processing Systems*, 2024.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Neural Information Processing Systems*, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *Neural Information Processing Systems*, 2022.

Lihong Li, Wei Gold Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. *International Conference on World Wide Web*, 2010.

Flore Sentenac, Ilbin Lee, and Csaba Szepesvári. Balancing optimism and pessimism in offline-to-online learning. *arXiv Preprint arXiv:2502.08259*, 2025.

Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. *Artificial Intelligence and Statistics*, 2012.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. *International Conference on Machine Learning*, 2014.

Sushant Vijayan, Arun Suggala, Karthikeyan Shanmugam, and Soumyabrata Pal. Regret minimization in linear bandits with offline data via extended D-optimal exploration. *arXiv Preprint arXiv:2508.08420*, 2025.

Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? *Conference on Learning Theory*, 2020.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. *International Conference on Machine Learning*, 2016.

Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. *International Conference on Machine Learning*, 2021.

Supplementary Materials

The following content was not necessarily subject to peer review.

C Regret Analysis of LinOtO

Our analysis is based on the proof of linear conservative bandits (Kazerouni et al., 2017) with some key modifications since our baseline is chosen based on the offline data. We will assume that the event $\mathcal{E} = \{\theta_* \in \mathcal{C}_t \text{ for all } t \in [n]\}$ holds throughout, which happens with probability at least $1-\delta$ by Proposition 1.

C.1 Proof of Theorem 1

Proof. Let $\mathcal{U}_t$ be the set of rounds till round t when the budget was non-negative, that is, when LinOtO chooses the UCB action. Similarly, let $\mathcal{L}_t$ be the set of rounds when LinOtO chooses the LCB action. Then the regret of LinOtO can be decomposed as

$$\begin{aligned} R_n(\text{LinOtO}) &= \sum_{t=1}^{n} \langle \theta_*, a_* \rangle - \langle \theta_*, A_t \rangle \\ &= \sum_{t \in \mathcal{U}_n} \left(\langle \theta_*, a_* \rangle - \langle \theta_*, U_t \rangle \right) + \sum_{t \in \mathcal{L}_n} \left(\langle \theta_*, a_* \rangle - \langle \theta_*, L_t \rangle \right) \\ &\leq \sum_{t \in \mathcal{U}_n} \left(\langle \theta_*, a_* \rangle - \langle \theta_*, U_t \rangle \right) + 2\sqrt{\beta_{\text{off}}} \, \|a_*\|_{V_0^{-1}} \, |\mathcal{L}_n|. \qquad \text{(by Lemma 1)} \end{aligned}$$

Now we bound $|\mathcal{L}_n|$. Let $\tau \coloneqq \max\{t \in [n] \,:\, A_t = L_t\}$ be the last round when the LCB action is chosen. Then $\mathcal{L}_\tau = \mathcal{L}_n$. Since L_τ was chosen in round τ, we also have $B(\tau) < 0$. That is,

$$\sum_{t=1}^{\tau-1} \text{LCB}_\tau(A_t) + \text{LCB}_\tau(U_\tau) - \tau r_b < 0 \,.$$

Decomposing the sum over $\tau - 1$ into rounds when UCB and LCB actions are chosen, and distributing τr_b over these two sets of rounds, the inequality in the last display can be rewritten as

$$\sum_{t \in \mathcal{U}_{\tau-1}} (\text{LCB}_\tau(A_t) - r_b) + \sum_{t \in \mathcal{L}_{\tau-1}} (\text{LCB}_\tau(A_t) - r_b) + \text{LCB}_\tau(U_\tau) - r_b < 0 \,. \tag{6}$$

For all $t \in \mathcal{L}_{\tau-1}$, the second term above can be lower bounded as follows:

$$\begin{aligned} \text{LCB}_\tau(A_t) - r_b &= \text{LCB}_\tau(L_t) - r_b && \text{(since } A_t = L_t \text{ for all } t \in \mathcal{L}_{\tau-1}\text{)} \\ &\geq \text{LCB}_t(L_t) - r_b && \text{(since LCB values are non-decreasing)} \\ &\geq \text{LCB}_t(L_{\text{off}}) - rb && \text{(since } L_t = \arg\max_{a \in \mathcal{A}} \text{LCB}_\tau(a)\text{)} \\ &\geq \text{LCB}_1(L_{\text{off}}) - r_b && \text{(since LCB values are non-decreasing)} \\ &= \alpha S \,. && \text{(by the definition of } r_b \text{ in eq. (4))} \end{aligned}$$

Therefore, eq. (6) becomes

$$\sum_{t \in \mathcal{U}_{\tau-1}} (\text{LCB}_\tau(A_t) - r_b) + \sum_{t \in \mathcal{L}_{\tau-1}} \alpha S + \text{LCB}_\tau(U_\tau) - r_b < 0 \,.$$

Since $\sum_{t \in \mathcal{L}_{\tau-1}} \alpha S = |\mathcal{L}_{\tau-1}| \alpha S$, we have

$$|\mathcal{L}_{\tau-1}| \alpha S < r_b - \text{LCB}_\tau(U_\tau) + \sum_{t \in \mathcal{U}_{\tau-1}} \left(r_b - \text{LCB}_\tau(A_t) \right) . \tag{7}$$

Let $\tilde{\theta}_t \coloneqq \arg\max_{\theta \in \mathcal{C}_t} \langle \theta, U_t \rangle$ be the optimistic estimate of θ_* in round t. Then for all $t \in \mathcal{U}_{\tau-1}$, we have

$$\begin{aligned}
\mathrm{LCB}_\tau(A_t) &= \mathrm{LCB}_\tau(U_t) \\
&= \max_{t' \le \tau} \inf_{\theta \in \mathcal{C}_{t'}} \langle \theta, U_t \rangle \\
&= \max_{t' \le \tau} \inf_{\theta \in \mathcal{C}_{t'}} -\langle \theta_* - \theta, U_t \rangle + \langle \theta_*, U_t \rangle \\
&= \max_{t' \le \tau} \inf_{\theta \in \mathcal{C}_{t'}} -\langle \theta_* - \theta, U_t \rangle - \left\langle \tilde{\theta}_t - \theta_*, U_t \right\rangle + \left\langle \tilde{\theta}_t, U_t \right\rangle \qquad (8) \\
&\ge \max_{t' \le \tau} \inf_{\theta \in \mathcal{C}_{t'}} -\|\theta_* - \theta\|_{V_{t-1}} \|U_t\|_{V_{t-1}^{-1}} - \left\|\tilde{\theta}_t - \theta_*\right\|_{V_{t-1}} \|U_t\|_{V_{t-1}^{-1}} + \left\langle \tilde{\theta}_t, U_t \right\rangle \\
&\ge \inf_{\theta \in \mathcal{C}_t} -\|\theta_* - \theta\|_{V_{t-1}} \|U_t\|_{V_{t-1}^{-1}} - \left\|\tilde{\theta}_t - \theta_*\right\|_{V_{t-1}} \|U_t\|_{V_{t-1}^{-1}} + \left\langle \tilde{\theta}_t, U_t \right\rangle \\
&\ge -2\sqrt{\rho_t} \|U_t\|_{V_{t-1}^{-1}} - 2\sqrt{\rho_t} \|U_t\|_{V_{t-1}^{-1}} + \left\langle \tilde{\theta}_t, U_t \right\rangle \\
&\ge -4\sqrt{\rho_\tau} \|U_t\|_{V_{t-1}^{-1}} + \left\langle \tilde{\theta}_t, U_t \right\rangle . \qquad \text{(since } \rho_\tau \ge \rho_t \text{ for all } t \le \tau\text{)}
\end{aligned}$$

Since the mean rewards are between $[0, 1]$, we can also bound $\mathrm{LCB}_\tau(A_t)$ from Equation (8) as $\mathrm{LCB}_\tau(A_t) \ge -4\sqrt{\rho_\tau} + \left\langle \tilde{\theta}_t, U_t \right\rangle$ to get

$$\mathrm{LCB}_\tau(A_t) \ge -4\sqrt{\rho_\tau} \left(1 \wedge \|U_t\|_{V_{t-1}^{-1}}\right) + \left\langle \tilde{\theta}_t, U_t \right\rangle ,$$

where $a \wedge b \coloneqq \min\{a, b\}$. Similarly, we can also bound $\mathrm{LCB}_\tau(U_\tau)$ as

$$\mathrm{LCB}_\tau(U_\tau) = \max_{t' \le \tau} \inf_{\theta \in \mathcal{C}_{t'}} \langle \theta, U_\tau \rangle \ge -4\sqrt{\rho_\tau} \left(1 \wedge \|U_\tau\|_{V_{\tau-1}^{-1}}\right) + \left\langle \tilde{\theta}_\tau, U_\tau \right\rangle .$$

Substituting the previous two displays in eq. (7) and rearranging, we get

$$|\mathcal{L}_{\tau-1}|\alpha S < \sum_{t \in \mathcal{U}_{\tau-1} \cup \{\tau\}} \left(r_b - \left\langle \tilde{\theta}_t, U_t \right\rangle + 4\sqrt{\rho_\tau} \left(1 \wedge \|U_t\|_{V_{t-1}^{-1}}\right)\right) . \qquad (9)$$

Note that for any $t \in [n]$,

$$\begin{aligned}
r_b - \left\langle \tilde{\theta}_t, U_t \right\rangle &= \mathrm{LCB}_1(L_{\text{off}}) - \alpha S - \left\langle \tilde{\theta}_t, U_t \right\rangle \\
&\le \langle \theta_*, L_{\text{off}} \rangle - \langle \theta_*, a_* \rangle - \alpha S \\
&= -\Delta_{\text{off}} - \alpha S ,
\end{aligned}$$

where $\Delta_{\text{off}} = \langle \theta_*, a_* - L_{\text{off}} \rangle$ is the suboptimality gap of the offline LCB action L_{off}. So eq. (9) can be rewritten as

$$|\mathcal{L}_{\tau-1}|\alpha S < -(\Delta_{\text{off}} + \alpha S)\left(|\mathcal{U}_{\tau-1}| + 1\right) + 4\sqrt{\rho_\tau} \sum_{t \in \mathcal{U}_{\tau-1} \cup \{\tau\}} \left(1 \wedge \|U_t\|_{V_{t-1}^{-1}}\right) .$$

Using Cauchy-Schwarz and the elliptical potential lemma (Abbasi-Yadkori et al., 2011; Kazerouni et al., 2017), we have

$$\sum_{t \in \mathcal{U}_{\tau-1} \cup \{\tau\}} \left(1 \wedge \|U_t\|_{V_{t-1}^{-1}}\right) \le \sqrt{2\left(|\mathcal{U}_{\tau-1}| + 1\right) d \log\left(1 + \frac{\left(|\mathcal{U}_{\tau-1}| + 1\right) L^2}{d \det(V_0)^{1/d}}\right)} ,$$

which can be further bounded using $\det V_0 \ge (\lambda_{\min}(V_0))^d$ as

$$\sum_{t \in \mathcal{U}_{\tau-1} \cup \{\tau\}} \left(1 \wedge \|U_t\|_{V_{t-1}^{-1}}\right) \le \sqrt{2\left(|\mathcal{U}_{\tau-1}| + 1\right) d \log\left(1 + \frac{\left(|\mathcal{U}_{\tau-1}| + 1\right) L^2}{d\lambda_{\min}(V_0)}\right)} .$$

Therefore

$$\begin{aligned}|\mathcal{L}_{\tau-1}|\alpha S &< -(\Delta_{\text{off}}+\alpha S)\left(|\mathcal{U}_{\tau-1}|+1\right) \\ &\quad + 4\sqrt{\rho_\tau}\sqrt{2d\left(|\mathcal{U}_{\tau-1}|+1\right)\log\left(1+\frac{\left(|\mathcal{U}_{\tau-1}|+1\right)L^2}{d\lambda_{\min}(V_0)}\right)} \\ &\leq -(\Delta_{\text{off}}+\alpha S)\left(|\mathcal{U}_{\tau-1}|+1\right) \\ &\quad + 4\sqrt{2\rho_n\left(|\mathcal{U}_{\tau-1}|+1\right)d\log\left(1+\frac{\left(|\mathcal{U}_{\tau-1}|+1\right)L^2}{d\lambda_{\min}(V_0)}\right)}. \end{aligned} \tag{10}$$

Since $|\mathcal{U}_{\tau-1}|+1 \leq n$, the RHS of Equation (10) becomes a quadratic function in $\sqrt{|\mathcal{U}_{\tau-1}|+1}$ and can be bounded by its maximum value. Therefore,

$$|\mathcal{L}_{\tau-1}| = \tilde{O}\left(\frac{d^2}{\alpha S(\Delta_{\text{off}}+\alpha S)}\right).$$

Finally, since $|\mathcal{L}_n| = |\mathcal{L}_\tau| = |\mathcal{L}_{\tau-1}|+1$, the regret corresponding to the LCB plays for Case 1 can be bounded as

$$\begin{aligned}\sum_{t\in\mathcal{L}_n}\left(\langle\theta_*,a_*\rangle-\langle\theta_*,L_t\rangle\right) &\leq 2\sqrt{\beta_{\text{off}}}\,\|a_*\|_{V_0^{-1}}\,|\mathcal{L}_{\tau-1}| \\ &= \tilde{O}\left(\frac{d^{2.5}\,\|a_*\|_{V_0^{-1}}}{\alpha S(\Delta_{\text{off}}+\alpha S)}\right).\end{aligned}$$

To bound the regret corresponding to the UCB plays, we can adapt the standard LinUCB analysis (Abbasi-Yadkori et al., 2011; Lattimore & Szepesvári, 2020) and of (Valko et al., 2014) to get

$$\sum_{t\in\mathcal{U}_n}\left(\langle\theta_*,a_*\rangle-\langle\theta_*,U_t\rangle\right) \leq \sqrt{8n\rho_n d_{\text{eff}}\log\left(1+\frac{nL^2}{\lambda_{\min}(V_0)}\right)}, \tag{11}$$

Therefore, with probability $1-\delta$, the total regret of LinOtO can be bounded as

$$R_n(\text{LinOtO}) = \tilde{O}\left(\sqrt{n\rho_n d_{\text{eff}}\log\left(1+\frac{nL^2}{\lambda_{\min}(V_0)}\right)}+\frac{d^{2.5}\,\|a_*\|_{V_0^{-1}}}{\alpha S(\Delta_{\text{off}}+\alpha S)}\right).$$

When $\lambda_{\min}(V_0) = \Omega(mL^2/d)$, using $\log(1+x)\leq x$, we can upper bound Equation (10) as

$$\begin{aligned}|\mathcal{L}_{\tau-1}|\alpha S &\leq -(\Delta_{\text{off}}+\alpha S)\left(|\mathcal{U}_{\tau-1}|+1\right) \\ &\quad + 4\left(|\mathcal{U}_{\tau-1}|+1\right)\sqrt{2d\rho_n/m},\end{aligned}$$

which can be bounded by 0 for large enough α satisfying $\Delta_{\text{off}}+\alpha S \geq 4\sqrt{2d\rho_n/m}$. In this case, the regret for the UCB plays can be upper bounded using Equation (11) as

$$\sum_{t\in\mathcal{U}_n}\left(\langle\theta_*,a_*\rangle-\langle\theta_*,U_t\rangle\right) = O\left(n\sqrt{\frac{\rho_n d d_{\text{eff}}}{m}}\right).$$

We thus complete the proof of Theorem 1. □

D Offline-Reference Regret Analysis of LinOtO

We start with the following lemma.

Lemma 6. *On the event $\mathcal{E}$, it holds that $r_b \geq \mu_{off} - (\alpha+2)S$.*

Proof. We have

$$\begin{aligned}
r_b = \max_{a\in\mathcal{A}} \inf_{\theta\in\mathcal{C}_1} \langle\theta, a\rangle - \alpha S &\geq \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \inf_{\theta\in\mathcal{C}_1} \langle\theta, a\rangle - \alpha S \\
&= \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \left(\inf_{\theta\in\mathcal{C}_1} -\langle\theta_* - \theta, a\rangle + \langle\theta_*, a\rangle\right) - \alpha S \\
&= \frac{1}{m}\sum_{a\in\mathcal{A}} -m_a \sup_{\theta\in\mathcal{C}_1} \langle\theta_* - \theta, a\rangle + \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \langle\theta_*, a\rangle - \alpha S \\
&\geq -\frac{1}{m}\sum_{a\in\mathcal{A}} m_a \sup_{\theta\in\mathcal{C}_1} \|\theta - \theta_*\|_{V_0} \|a\|_{V_0^{-1}} + \mu_{\text{off}} - \alpha S \\
&\geq -\frac{2\sqrt{\beta_{\text{off}}}}{m}\sum_{a\in\mathcal{A}} m_a \|a\|_{V_0^{-1}} + \mu_{\text{off}} - \alpha S \\
&\geq -2S + \mu_{\text{off}} - \alpha S \\
&= \mu_{\text{off}} - (\alpha + 2)S\,.
\end{aligned}$$

This establishes Lemma 6. □

D.1 Proof of Theorem 2

Proof. Let $\mathcal{U}_t \coloneqq \{s \in [t] \,:\, A_s = U_s\}$ and $\mathcal{L}_t \coloneqq \{s \in [t] \,:\, A_s = L_s\}$. Then

$$\begin{aligned}
R_n^{\text{off}}(\text{LinOtO}) &= \sum_{t=1}^{n} (\mu_{\text{off}} - \langle\theta_*, A_t\rangle) \\
&= \sum_{t\in\mathcal{U}_n} (\mu_{\text{off}} - \langle\theta_*, A_t\rangle) + \sum_{t\in\mathcal{L}_n} (\mu_{\text{off}} - \langle\theta_*, A_t\rangle) \\
&= \sum_{t\in\mathcal{U}_n} (\mu_{\text{off}} - \langle\theta_*, U_t\rangle) + \sum_{t\in\mathcal{L}_n} (\mu_{\text{off}} - \langle\theta_*, L_t\rangle)\,.
\end{aligned}$$

To bound the regret corresponding to the rounds when the UCB action was chosen, consider the "realized" budget at the end of round n, that is, the excess reward of the actions chosen by LinOtO till round n over the baseline r_b:

$$\begin{aligned}
B_+(n) &\coloneqq \sum_{t=1}^{n} \text{LCB}_n(A_t) - nr_b \\
&= \sum_{t=1}^{n-1} \text{LCB}_n(A_t) + \text{LCB}_n(A_n) - nr_b.
\end{aligned}$$

If $A_n = U_n$, then $B_+(n) = B(n) \geq 0$. If $A_n = L_n$, then

$$\begin{aligned}
B_+(n) &= \sum_{t=1}^{n-1} \text{LCB}_n(A_t) + \text{LCB}_n(L_n) - nr_b \\
&\geq \sum_{t=1}^{n-1} \text{LCB}_{n-1}(A_t) - (n-1)r_b + \text{LCB}_n(L_n) - r_b \\
&\geq B_+(n-1) + \text{LCB}_n(L_{\text{off}}) - r_b && \text{(since } L_n = \arg\max_{a\in\mathcal{A}} \text{LCB}_n(a)\text{)} \\
&\geq B_+(n-1) + \text{LCB}_1(L_{\text{off}}) - r_b && \text{(since LCB values are non-decreasing)} \\
&= B_+(n-1) + \alpha S\,.
\end{aligned}$$

Similarly, $B_+(n-1) \geq 0$ or $B_+(n-1) \geq B_+(n-2) + \alpha S$. Therefore, by iterating the above argument, we have $B_+(n) \geq 0$ or $B_+(n) \geq |\mathcal{L}_n|\alpha S$. In either case, we have $B_+(n) \geq 0$. This

implies

$$\sum_{t\in\mathcal{U}_n} \mathrm{LCB}_n(A_t) + \sum_{t\in\mathcal{L}_n} \mathrm{LCB}_n(A_t) - nr_b \geq 0\,.$$

Rearranging and using the fact that $A_t = U_t$ if $t \in \mathcal{U}_n$ and $A_t = L_t$ if $t \in \mathcal{L}_n$, we get

$$\begin{aligned}
\sum_{t\in\mathcal{U}_n} \mathrm{LCB}_n(U_t) &\geq \sum_{t\in\mathcal{L}_n} (r_b - \mathrm{LCB}_n(L_t)) + \sum_{t\in\mathcal{U}_n} r_b \\
&\geq \sum_{t\in\mathcal{L}_n} (\mu_{\text{off}} - \mathrm{LCB}_n(L_t) - (\alpha+2)S) + \sum_{t\in\mathcal{U}_n} r_b \\
&\geq \sum_{t\in\mathcal{L}_n} (\mu_{\text{off}} - \langle\theta_*, L_t\rangle) + \sum_{t\in\mathcal{U}_n} r_b - (\alpha+2)Sn\,.
\end{aligned}$$

Using the previous display with the regret during UCB iterations, we have

$$\begin{aligned}
\sum_{t\in\mathcal{U}_n} (\mu_{\text{off}} - \langle\theta_*, U_t\rangle) &\leq \sum_{t\in\mathcal{U}_n} (\mu_{\text{off}} - \mathrm{LCB}_n(U_t)) \\
&\leq \sum_{t\in\mathcal{U}_n} (\mu_{\text{off}} - r_b) + (\alpha+2)Sn - \sum_{t\in\mathcal{L}_n} (\mu_{\text{off}} - \langle\theta_*, L_t\rangle) \\
&\leq 2(\alpha+2)Sn - \sum_{t\in\mathcal{L}_n} (\mu_{\text{off}} - \langle\theta_*, L_t\rangle)\,. \qquad \text{(by Lemma 6)}
\end{aligned}$$

Finally, the total $R_n^{\text{off}}(\text{LinOtO})$ can be bounded as

$$\begin{aligned}
R_n^{\text{off}}(\text{LinOtO}) &= \sum_{t\in\mathcal{U}_n} (\mu_{\text{off}} - \langle\theta_*, U_t\rangle) + \sum_{t\in\mathcal{L}_n} (\mu_{\text{off}} - \langle\theta_*, L_t\rangle) \\
&\leq 2(\alpha+2)Sn \\
&= 2(\alpha+2)\sqrt{\frac{d\beta_{\text{off}}}{m}}n\,.
\end{aligned}$$

This completes the proof of Theorem 2. □

D.2 Proof of Proposition 3

The proof follows a similar structure to the proof of Proposition 2.

Proof. When the event $\theta_* \in \mathcal{C}_t$ for all $t \in [n]$ holds, we have that with probability $1-\delta$ and for all $t \in [n]$,

$$\mu_{L_t} = \langle\theta_*, L_t\rangle \geq \max_{t'\leq t} \inf_{\theta\in\mathcal{C}_{t'}} \langle\theta, L_t\rangle = \mathrm{LCB}_t(L_t)\,.$$

Since $\max$ is always greater than the average, we have

$$\mathrm{LCB}_t(L_t) = \max_{a\in\mathcal{A}} \mathrm{LCB}_t(a) \geq \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \mathrm{LCB}_t(a) = \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \max_{t'\leq t} \inf_{\theta\in\mathcal{C}_{t'}} \langle\theta, a\rangle\,.$$

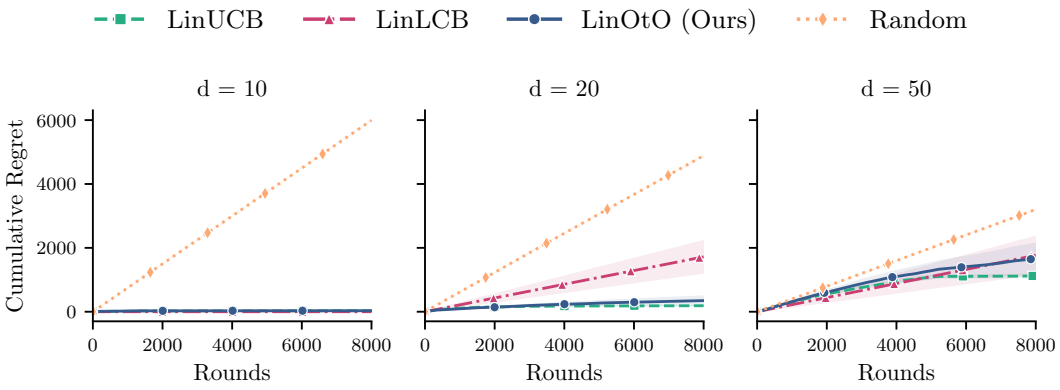


Figure 2: Cumulative regret with varying dimension d.

From the above two displays, we have

$$
\begin{aligned}
\mu_{L_t} &\geq \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \max_{t'\leq t} \inf_{\theta\in\mathcal{C}_{t'}} \langle \theta, a\rangle \\
&= \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \max_{t'\leq t} \inf_{\theta\in\mathcal{C}_{t'}} \left(\langle \theta_*, a\rangle - \langle \theta_* - \theta, a\rangle\right) \\
&= \frac{1}{m}\sum_{a\in\mathcal{A}} m_a\mu_a + \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \max_{t'\leq t}\left(-\sup_{\theta\in\mathcal{C}_{t'}} \langle \theta_* - \theta, a\rangle\right) \\
&\geq \mu_{\text{off}} + \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \max_{t'\leq t}\left(-\sup_{\theta\in\mathcal{C}_{t'}} \|\theta-\theta_*\|_{V_{t'-1}} \|a\|_{V_{t'-1}^{-1}}\right) && \text{(Cauchy-Schwarz)} \\
&\geq \mu_{\text{off}} + \frac{1}{m}\sum_{a\in\mathcal{A}} m_a \max_{t'\leq t}\left(-2\sqrt{\rho_{t'}}\,\|a\|_{V_{t'-1}^{-1}}\right) && \text{(since } \sup_{\theta\in\mathcal{C}_{t'}} \|\theta-\theta_*\|_{V_{t'-1}} \leq 2\sqrt{\rho_{t'}}\text{)} \\
&> \mu_{\text{off}} - \frac{2\sqrt{\rho_t}}{m}\sum_{a\in\mathcal{A}} m_a \max_{t'\leq t}\|a\|_{V_{t'-1}^{-1}} && \text{(since } \rho_t > \rho_{t'} \text{ for all } t'\leq t\text{)} \\
&= \mu_{\text{off}} - \frac{2\sqrt{\rho_t}}{m}\sum_{a\in\mathcal{A}} m_a \|a\|_{V_0^{-1}} && \text{(since } V_{t'-1}^{-1} \preceq V_0^{-1} \text{ for all } t'\text{)} \\
&\geq \mu_{\text{off}} - 2\sqrt{\frac{d\rho_t}{m}}\,. && \text{(using Lemma 5)}
\end{aligned}
$$

Rearranging, the $R_n^{\text{off}}(\text{LCB})$ is bounded as

$$
R_n^{\text{off}}(\text{LCB}, \nu) = \sum_{t=1}^{n}\left(\frac{1}{m}\sum_{a\in\mathcal{A}} m_a\mu_a - \mu_{L_t}\right) \leq 2\sqrt{\frac{d}{m}}\sum_{t=1}^{n}\sqrt{\rho_t} \leq 2n\sqrt{\frac{d\rho_n}{m}}.
$$

This establishes Proposition 3. □

E Additional Experimental Results

Figure 2 shows the regret of LinUCB, LinLCB, LinOtO, and a random policy by varying the number of dimensions $d \in \{10, 20, 50\}$. The other problem parameters are set to the default values as described in Section 6. In Figure 3, we plot the sensitivity of LinOtO with respect to the tuning parameter α (see eq. (4)) for different number of offline samples m.

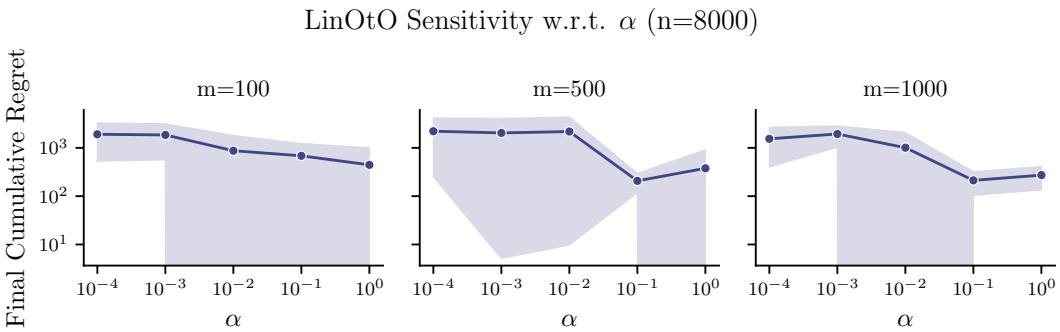


Figure 3: Sensitivity of LinOtO with respect to α for varying number of offline samples m. The vertical axis is on log scale.