Recycling History: Efficient Recommendations from Contextual Dueling Bandits

Suryanarayana Sankagiri SURYANARAYANA.SANKAGIRI@EPFL.CH
School of Communication and Computer Science
EPFL, Switzerland

Jalal Etesami J.ETESAMI@TUM.DE
Department of Computer Science
Munich Institute of Robotics and Machine Intelligence
TU Munich, Germany

Pouria Fatemi POURIA.FATEMI@TUM.DE
Department of Mathematics
TU Munich, Germany

Matthias Grossglauser MATTHIAS.GROSSGLAUSER@EPFL.CH
School of Communication and Computer Science
EPFL, Switzerland

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

The contextual dueling bandit problem models adaptive recommender systems, where at each step the algorithm presents a set of items to the user, and the user's choice reveals their preference. This setup is well suited for implicit choices users make when navigating a content platform, but does not capture other possible comparison queries. Motivated by the fact that users provide more reliable feedback after consuming items, we propose a new bandit model that can be described as follows. The algorithm recommends one item per time step; after consuming that item, the user is asked to compare it with another item chosen from the user's consumption history. Importantly, in our model, this comparison item can be chosen without incurring any additional regret, potentially leading to better performance. However, the regret analysis is challenging because of the temporal dependency in the user's history. To overcome this challenge, we first show that the algorithm can construct informative queries provided the history is rich, *i.e.*, satisfies a certain diversity condition. We then show that a short initial random exploration phase is sufficient for the algorithm to accumulate a rich history with high probability. This result, proven via matrix concentration bounds, yields $O(\sqrt{T})$ regret guarantees. Additionally, our simulations show that reusing past items for comparisons can lead to significantly lower regret than only comparing between simultaneously recommended items.

Keywords: contextual bandits, dueling bandits, recommender systems

1. Introduction

Recommender systems are central to digital platforms, helping users navigate the vast set of options by filtering items based on their preferences (Bobadilla et al., 2013). A user's taste profile is typically learned from the feedback they provide. One fundamental question in the design of recommender systems is how to elicit feedback effectively. A basic distinction can be drawn between *implicit* and *explicit* feedback. The former is obtained by observing user actions (deleting vs saving in a wishlist) and the latter through explicit prompts (rating an item from one to five stars). In

contrast to implicit modes, explicit feedback is typically obtained after the user has consumed the items. Therefore, they provide more accurate reflection of preferences. Another dimension of elicitation design is whether to seek *ordinal* (ranking- or comparison-based) or *cardinal* (rating-based) feedback (Wang and Shah, 2019). While ratings are the most common form of feedback, seeking comparisons instead could offer several advantages. For one, it naturally eliminates the effect of dynamically varying user biases, arising due to mood, experiences, etc. Equally importantly, it gets around the discretization problem: it is difficult to distinguish among many items with five stars, but a comparison between two such items still extracts useful information.

Model Contribution. Motivated by the above considerations, we aim to formulate a recommender system that learns user tastes through explicit comparisons among consumed items. We adopt the classical contextual bandit (CB) framework for this task. Specifically, we assume there is a single user who is repeatedly served recommendations by the bandit algorithm. The user and the items are endowed with feature vectors, representing their tastes and characteristics respectively. While the user's feature vector $\boldsymbol{\theta}^*$ is unknown, the items' features $\mathbf{x}$ are assumed to be known and fixed. At every time step t, the algorithm is provided a set of items $\mathcal{X}_t$ (potentially the result of a search query). The algorithm picks a single item, $\mathbf{x}_t$, from this set and recommends it to the user. After the user has consumed the item, the system asks the user to compare it with another item $\mathbf{y}_t$ *consumed by the user in the past*. That is, $\mathbf{y}_t$ must belong to $\mathcal{H}_t$, the user's consumption *history*. Based on the user's response to this comparison query, the algorithm updates its estimate of $\boldsymbol{\theta}^*$ and refines its future recommendations. The algorithm's performance is measured in terms of the regret, *i.e.*, the suboptimality in utility, of the recommended items $\mathbf{x}_t$.

The contextual dueling bandits (CDB) problem, introduced by Saha (2021) and later studied by Bengs et al. (2022), shares several features with our setup. However, there are also important differences, which we highlight below. The model adopted by Saha (2021) and Bengs et al. (2022) assumes that at each time t, two fresh items $(\mathbf{x}_t, \mathbf{y}_t)$ are drawn from $\mathcal{X}_t$ and presented to the user, from which the user picks one. The quality of the recommendation is measured in terms of the average regret over both $\mathbf{x}_t$ and $\mathbf{y}_t$. In contrast, our model assumes that $\mathbf{x}_t$ is chosen from $\mathcal{X}_t$ and $\mathbf{y}_t$ from $\mathcal{H}_t$ (the user's consumption history). Regret is measured only in terms of $\mathbf{x}_t$, because the regret for $\mathbf{y}_t$ has already been accounted for when it was consumed. We refer to our model as the *history-constrained CDB model* and to the model in the literature as the *concurrent CDB model*.

Practically speaking, the concurrent CDB models a scenario where a user is offered two options and asked to pick a movie to watch, while the history-constrained model asks the user whether they prefer the movie they have just watched over one that they watched in the past. The concurrent model is well suited for learning from the *implicit* comparisons that users make at the theatre or on a streaming platform (Saha (2021) motivates their model through a similar example). Although, technically, the model does not preclude the possibility of seeking explicit comparisons, it is often impractical, or wasteful, to force the user to consume two items just to elicit a comparison. The history-constrained model overcomes this limitation of the concurrent model, providing a framework to seek *explicit comparisons among consumed items*.

This seemingly minor difference in formulation has interesting consequences in terms of the algorithm design as well as regret bounds. On the one hand, simultaneously choosing $\mathbf{x}_t$ and $\mathbf{y}_t$ from the same set $\mathcal{X}_t$ makes it easier to formulate informative queries in the concurrent model. The asymmetry introduced in our model complicates this key step, which makes it challenging to bound the regret. In particular, the history set $\mathcal{H}_t$ cannot be designed knowing $\mathbf{x}_t$. On the other hand,

the history-constrained model allows the user additional flexibility in navigating the exploration-exploitation trade-off. To elaborate, because the choice of $\mathbf{y}_t$ doesn't incur any regret, it can be chosen purely with an exploration objective. In contrast, optimal algorithms for the concurrent model, such as `MaxInP` (Saha, 2021) and `CoLSTIM` (Bengs et al., 2022) must delicately balance exploration and exploitation when selecting both $\mathbf{x}_t$ and $\mathbf{y}_t$, as the regret depends on both items. Thus, a well-designed algorithm for the history-constrained model may obtain better regret than what is possible in the concurrent model.

Algorithmic Contribution. In this work, we present an algorithm called `ROAM` (Regret Once, Ask Many), that provably achieves $O(\sqrt{T})$ cumulative regret under the history-constrained model. To elaborate, while choosing what query to ask the user, the algorithm optimizes a metric quantifying the uncertainty in the estimate $\hat{\boldsymbol{\theta}}$. Intuitively, the algorithm needs to probe the user along different dimensions (axes), in order to narrow down on the user's tastes. Thus, there needs to be sufficient diversity in the choice of comparison queries, so that the algorithm can frame a query probing any arbitrary direction. We make this notion precise by defining a property called *rich history*. Once the history is rich, the algorithm can pick $\mathbf{y}_t$ among these items at every step t in order to learn about the user's tastes. The algorithm's name reflects this key feature.

Accumulating a rich history requires exploration (essentially, recommending items at random) and therefore incurs regret. For optimal regret bounds, it is important to ensure that this period of exploration remains small. We use matrix concentration bounds to show that a short exploration phase lasting $\tilde{O}(d)$ steps is sufficient for the algorithm to accumulate a rich history with high probability. Notably, this degree of initial exploration is typically used by concurrent CDB algorithms to get a reasonable estimate $\hat{\boldsymbol{\theta}}$ to start with (Saha, 2021; Bengs et al., 2022). Thus, accumulating a rich history does not lead to additional regret. Using this result, we obtain $O(\sqrt{T})$ regret guarantees. Although our theoretical regret bounds may have suboptimal coefficients on the $O(\sqrt{T})$, we find through simulations on synthetic data that the regret suffered by `ROAM` scales gracefully with the model parameters, such as dimension. In particular, with appropriate parameters for a fair comparison, we demonstrate that `ROAM` (applied to the history-constrained setting) has significantly lower regret than `CoLSTIM` (Bengs et al., 2022), the state-of-the art algorithm for the concurrent setting. This demonstrates that the flexibility of recycling items from the user's consumption history to elicit comparisons has tangible benefits.

Related Work. As mentioned above, our model shares many similarities with the concurrent CDB model proposed by Saha (2021) and followed upon by Bengs et al. (2022). This model, in turn, is a natural merger of two classical models: the stochastic linear bandit model (Abbasi-yadkori et al., 2011) and the K-armed dueling bandit model (Yue et al., 2012). (See (Lattimore and Szepesvári, 2020) for an introduction to linear bandits and (Bengs et al., 2021) for a review of dueling bandits). In particular, the proofs in (Saha, 2021), (Bengs et al., 2022), as well as our work, rely on some key technical lemmas for generalised linear bandits, proven by Li et al. (2017). We note that (Dudík et al., 2015) and (Saha and Krishnamurthy, 2022) also deal with contextual dueling bandits, but their model has a completely different interpretation. Instead of referring to item features, context here refers to a global variable set arbitrarily, which influences the choice probabilities.

Although linear and generalized linear bandits have been widely used in practical recommender systems (Pereira et al., 2019; Bendada et al., 2020; He et al., 2020), the same cannot be said for contextual dueling bandits (CDBs). In fact, to the best of our knowledge, there are no papers that test CDB algorithms on real data. (One instance of testing a CB algorithm on preference data is

presented in Agnihotri et al. (2024), but their method does not fit the CDB model). Moreover, the prior work on concurrent CDBs by Saha (2021) and Bengs et al. (2022) have a limited discussion on the potential of this model in recommender systems. This is perhaps in part because learning from comparisons (instead of ratings) is, to this date, not widely accepted in the recommender systems community. We hope that our theoretical modeling and analysis, combined with the empirically demonstrated efficacy of comparison-based learning in the context of LLM alignment (Ouyang et al., 2022; Rafailov et al., 2023), leads to the design of comparison-based recommender systems in the near future.

2. Model and Algorithm

Basic Setup. We begin by stating the modelling assumptions of our history-constrained contextual duelling bandit (CDB) framework. The model describes the sequential interaction of a single user with a recommender system (also referred to as the platform or the algorithm). The user and all the items are endowed with d-dimensional time-invariant features. The user's feature vector, $\boldsymbol{\theta}^*$, is unknown to the bandit algorithm, whereas item features are assumed to be known. We refer to items by their feature vector $\mathbf{x}$. The user's utility for any item $\mathbf{x}$ is taken to be $\langle \mathbf{x}, \boldsymbol{\theta}^* \rangle$, the inner product of the user and item feature vectors. Without loss of generality, we may assume $\|\boldsymbol{\theta}^*\|_2 = 1$, as utilities can be scaled by scaling the item features.

At each round t, the platform is presented with a set of candidate items $\mathcal{X}_t$ (called the context set). The context set should be interpreted as the result of a filtering out of the vast universe of items, based on a search query or other factors such as location, time, etc. While, in principle, the context set size could vary over time, we take it to be constant over the execution horizon. We assume $\mathcal{X}_t$ is stochastic; in particular, each item in $\mathcal{X}_t$ is drawn i.i.d. from a d-dimensional distribution $\mathcal{P}$. We make two restrictions on $\mathcal{P}$; First, we assume any $\mathbf{x} \sim \mathcal{P}$ satisfies $\|\mathbf{x}\|_2 \leq r$ almost surely. Second, we assume that the matrix $\Sigma = \mathbb{E}_{\mathbf{x},\mathbf{y} \sim \mathcal{P} \text{ i.i.d.}}[(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^\top]$ is invertible, that is, it is positive definite.

Given $\mathcal{X}_t$, the platform picks an item $\mathbf{x}_t$ from $\mathcal{X}_t$ and recommends it to the user. This recommendation is based on the platform's estimate of the user's feature vectors. After the user has consumed $\mathbf{x}_t$, the platform seeks feedback by posing a comparison query to the user. Let $\mathcal{H}_t$ denote the set of all items consumed up to time t: $\mathcal{H}_t = \{\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}\}$; we refer to $\mathcal{H}_t$ as the *history*. The platform then picks $\mathbf{y}_t$ from $\mathcal{H}_t$ and asks the user to compare $\mathbf{x}_t$ and $\mathbf{y}_t$. We do not impose any other restriction on the choice of $\mathbf{y}_t$; in particular, an item can be chosen for comparison arbitrarily often. Note the difference from the concurrent CDB model, where $\mathbf{y}_t$ is chosen from $\mathcal{X}_t$. Consequently, our notion of regret, the key performance measure of any bandit algorithm, also differs from the concurrent case. Let $\mathbf{x}_t^* = \arg\max_{\mathbf{x} \in \mathcal{X}_t} \langle \mathbf{x}, \boldsymbol{\theta}^* \rangle$ denote the best item at time t. Then $r_t = \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle - \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle$, the instantaneous regret at time t, measures the suboptimality of the recommended item $\mathbf{x}_t$. The concurrent CDB model, on the other hand, measures the regret in terms of both $\mathbf{x}_t$ and $\mathbf{y}_t$. The cumulative regret, $R_T = \sum_{t \in [T]} r_t$, measures the performance of the algorithm over a time horizon T. Our goal is to design an algorithm where R_T scales as $O(\sqrt{T})$ with high probability (ignoring logarithmic terms).

We assume that the user makes comparison decisions according to a simple probabilistic model called the linear stochastic transitivity model (Bengs et al., 2022), which we describe below. Denote the outcome of the user's comparison by the binary variable o_t; we say $o_t = 1$ if the user picks $\mathbf{x}_t$ (denoted by the event $\mathbf{x}_t \succ \mathbf{y}_t$) and $o_t = 0$ otherwise. When asked to compare $\mathbf{x}_t$ and $\mathbf{y}_t$, the

user picks $\mathbf{x}_t$ with probability $F(\langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle - \langle \mathbf{y}_t, \boldsymbol{\theta}^* \rangle)$; more succinctly, $\mathbb{P}(o_t = 1) = F(\langle \mathbf{z}_t, \boldsymbol{\theta}^* \rangle)$, where $\mathbf{z}_t$ denotes $\mathbf{x}_t - \mathbf{y}_t$. The function $F(\cdot)$ is called the *link function*, and determines the noise in the comparisons. This choice model conveys the intuition that users are more likely to pick an item with larger utility. A prototypical example of $F(\cdot)$ is the sigmoid function: $\sigma(x) = 1/(1 + \exp(-x))$. This special case corresponds to the popular Bradley-Terry choice model (Saha, 2021). More generally, $F(\cdot)$ is a smooth, strictly increasing function from $\mathbb{R} \to (0, 1)$, satisfying the symmetry condition $F(u) + F(-u) = 1$.

We assume that the link function is known to the algorithm. Knowing the link function allows the algorithm to calculate the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_t$ of $\boldsymbol{\theta}^*$, given the collected data up to time t. Let $\mathcal{D}_t$ denote the dataset $\{\mathbf{z}_i, o_i\}_{i \in [t-1]}$ (recall $\mathbf{z}_i = \mathbf{x}_i - \mathbf{y}_i$). Then the maximum likelihood estimator over $\mathcal{D}_t$ is obtained by solving $\sum_{i \in [t-1]} (F(\langle \mathbf{z}_i, \theta \rangle) - o_i)\mathbf{z}_i = 0$. In practice, this can be solved using Newton's method or any variant of it. For the special case of F being the sigmoid function, this estimation problem is identical to the logistic regression problem over the dataset $\mathcal{D}_t$. This estimation step is common in many generalized linear bandit and CDB algorithms (Bengs et al., 2022; Li et al., 2017).

Algorithm. We now present a simple and efficient algorithm for the history-constrained contextual dueling bandits model, called `ROAM` (see pseudocode of Algorithm 1 below). The name reflects the fact that once an item is consumed and its regret paid for, it can be reused many times to pose a comparison query to the user without incurring further regret.

Algorithm 1: Regret Once, Ask Many (`ROAM`)

Input: time horizon T, pure exploration horizon τ
Initialization: $\mathcal{H}_1 = \emptyset$, $\mathcal{D}_1 = \emptyset$
for $t = 1, \ldots, \tau$ **do**
  Pick $\mathbf{x}_t \in \mathcal{X}_t$ uniformly at random and recommend it to the user
  Set $\mathbf{y}_t \leftarrow \mathbf{x}_{t-1}$ and ask the user to compare $\mathbf{x}_t$ to $\mathbf{y}_t$
  Update $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{\mathbf{x}_t - \mathbf{y}_t, o_t\}$ based on the comparison outcome o_t
  Update $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \cup \{\mathbf{x}_t\}$
end
Set $V_{\tau+1} \leftarrow \sum_{i=1}^{\tau} (\mathbf{x}_i - \mathbf{y}_i)(\mathbf{x}_i - \mathbf{y}_i)^\top$
for $t = \tau + 1, \tau + 2, \ldots, T$ **do**
  Calculate the MLE $\hat{\boldsymbol{\theta}}_t$ over $\mathcal{D}_t$
  Select $\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x} \in \mathcal{X}_t} \langle \mathbf{x}, \hat{\boldsymbol{\theta}}_t \rangle$ and recommend it to the user
  Set $\mathbf{y}_t \leftarrow \arg\max_{\mathbf{y} \in \mathcal{H}_t} \|\mathbf{x}_t - \mathbf{y}\|_{V_t^{-1}}$ and ask the user to compare $\mathbf{x}_t$ to $\mathbf{y}_t$
  Update $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{\mathbf{x}_t - \mathbf{y}_t, o_t\}$ based on the comparison outcome o_t
  Update $V_{t+1} \leftarrow V_t + (\mathbf{x}_t - \mathbf{y}_t)(\mathbf{x}_t - \mathbf{y}_t)^\top$
  Update $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \cup \{\mathbf{x}_t\}$
end

The algorithm begins with a pure exploration phase that lasts τ rounds. In this phase, the platform picks an item uniformly at random from the context set and recommends it to the user. Then, it asks the user to compare this item with the one recommended just before. This exploration phase serves two purposes. First, it provides enough data such that $\hat{\boldsymbol{\theta}}_\tau$ is a good-enough estimate of $\boldsymbol{\theta}^*$; this helps the algorithm to perform well in subsequent rounds. This aspect is fairly standard

in the literature (Saha, 2021; Bengs et al., 2022; Li et al., 2017). The second aspect is novel; the random exploration phase populates the history $\mathcal{H}_\tau$ with a diverse set of items. This diverse history, in turn, is crucial for formulating informative queries in subsequent rounds. While some minimum degree of exploration is essential to satisfy these two criteria, too large of an exploration can lead to poor regret. Thus, τ must be carefully chosen based on the system parameters. In the derivation of our regret bounds, we specify a suitable value of τ.

The main phase of the algorithm is described in lines 12 and 13 of Algorithm 1. The algorithm picks $\mathbf{x}_t$, the new item to recommend to the user, by optimizing the *estimated utility* $\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_t \rangle$ over the arms. This strategy can be interpreted as being greedy or purely exploitative, as it picks the item that is most likely to yield the lowest regret. It is useful to contrast this step with the corresponding step in CoLSTIM, the algorithm proposed in Bengs et al. (2022), where the objective being optimized is $\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_t \rangle + \epsilon \|\mathbf{x}\|_{V_t^{-1}}$ (ϵ being a bounded random variable). This additional term reflects the uncertainty in the utility of item $\mathbf{x}$. The weighted sum is thus a careful balance of exploration and exploitation. Indeed, in the UCB algorithm for generalized linear bandits, a near-identical objective is optimized (Li et al., 2017).

To interpret the choice of the comparison arm $\mathbf{y}_t$, it is instructive to draw an analogy with linear regression in order to better understand the process of estimating $\boldsymbol{\theta}^*$. Upon asking a comparison query $(\mathbf{x}_t, \mathbf{y}_t)$ and recording the corresponding outcome o_t, the algorithm essentially 'probes' the unknown variable $\boldsymbol{\theta}^*$ in the direction of $\mathbf{z}_t = \mathbf{x}_t - \mathbf{y}_t$. For this reason, we refer to $\mathbf{z}_t$ as the probing vector. Repeated probes in the same direction leads to a better estimate of $\boldsymbol{\theta}^*$ in that direction. The matrix V_t, being the sum of the outer products of the probing vectors so far (see lines 9 and 15 of Algorithm 1), summarizes the combined effect of all probes on $\boldsymbol{\theta}^*$. Put differently, V_t^{-1} reflects the uncertainty in the estimate $\hat{\boldsymbol{\theta}}_t$. The weighted norm $\|\cdot\|_{V_t^{-1}}$ gives a higher emphasis to vectors aligned along directions of larger uncertainty.

With the above interpretation in mind, it is easy to interpret the choice of $\mathbf{y}_t$ as a pure exploration step. By optimising the term $\|\mathbf{x}_t - \mathbf{y}\|_{V_t^{-1}}$, we choose $\mathbf{y}_t$ such that the comparison query probes along the direction of maximum uncertainty. Once again, it is instructive to compare this action against the corresponding action in CoLSTIM, in which $\mathbf{y}_t$ is chosen by maximizing a weighted sum of $\|\mathbf{x}_t - \mathbf{y}\|_{V_t^{-1}}$ and $\langle \mathbf{y} - \mathbf{x}_t, \hat{\boldsymbol{\theta}}_t \rangle$, reflecting a balance of exploration and exploitation. In our model, we have the freedom to optimize purely for exploration while choosing $\mathbf{y}_t$ because we do not (directly) suffer any regret in the choice of $\mathbf{y}_t$. Thus, it is natural to choose $\mathbf{y}_t$ in a manner that would lead to the best improvement in the estimate $\hat{\boldsymbol{\theta}}_t$. This, in turn, contributes to reducing the cumulative regret.

An important consequence of our algorithm's design is that, with the exception of the length of the initial exploration period, there is no hyperparameter to be optimised. Unlike our method, most bandit algorithms (not just CoLSTIM) optimize a weighted sum of the exploitation and exploration terms. Theoretically, the level of exploration is dictated by the concentration bounds on $\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t$. Often, one observes that a lower degree of exploration works better empirically than what is suggested in theory; this is because the concentration bounds can be quite loose. Thus, the relative weight of exploration to exploitation (*e.g.*, C_{thresh}, c_1 in CoLSTIM) is often a hyper-parameter that needs to be optimized. No such hyper-parameter tuning is required in our algorithm.

3. Main Result

In this section, we state and prove our main result (Theorem 1): a high probability bound on the regret of our algorithm ROAM under the history-constrained CDB model. Our proof also prescribes the length of the initial exploration time τ. Before stating the theorem, we recall the parameters d, r, and Σ from Section 2. Also recall that we have assumed Σ to be positive definite, which implies $\lambda_{\min}(\Sigma)$ is strictly positive.

Theorem 1 (Regret Bound for ROAM) *Let $\delta \in (0,1)$ be given. Suppose ROAM is run with $\tau = O(r^2/\lambda_{\min}(\Sigma)) \log(d/\delta)$. Then, with probability at least $1 - 3\delta$, ROAM suffers a cumulative regret of*

$$R_T = O\left(\frac{r}{\kappa\sqrt{\lambda_{\min}(\Sigma)}} d\sqrt{T} \log\left(\frac{T}{d\delta}\right)\right)$$

where

$$\kappa = \inf_{\mathbf{z},\boldsymbol{\theta}:\|\mathbf{z}\|_2 \le 2r, \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \le 1} F'(\langle \mathbf{z}, \boldsymbol{\theta}\rangle) \tag{1}$$

and the $O(\cdot)$ notation hides absolute constants.

It is useful to compare this result with the $O(\kappa^{-1} d\sqrt{T} \log(T/d\delta))$ regret bounds for similar algorithms in the generalised linear bandit model (Li et al., 2017) and the concurrent contextual dueling bandit setting (Saha, 2021; Bengs et al., 2022). The results in the literature are proven under the assumption $r = 1$; thus, our bounds include an extra factor of $1/\sqrt{\lambda_{\min}(\Sigma)}$. As we had discussed in Section 1, the fact that $\mathbf{x}_t$ and $\mathbf{y}_t$ are not chosen from the same set poses some difficulties in the regret analysis; this extra factor is a manifestation of this difficulty (see Section 3.1 below). In many practical scenarios, this term may be reasonably small. For example, if we assume that the distribution $\mathcal{P}$ is the uniform distribution over the unit ball in d dimensions, then $1/\sqrt{\lambda_{\min}(\Sigma)}$ scales as $\sqrt{d}$. Thus, the regret guarantee in Theorem 1 scales as $O(d^{3/2}\sqrt{T})$, which is slightly looser than the regret guarantees of $O(d\sqrt{T})$ given in Saha (2021) and Bengs et al. (2022). It remains an open problem whether Theorem 1 can be tightened or whether the extra $\sqrt{d}$ factor is necessary in the history-constrained setting.

The proof of Theorem 1 rests on several key lemmas. In particular, Lemmas 3 and 4 highlight the two roles played by the initial exploration period: constructing a rich history and obtaining a good estimate $\hat{\boldsymbol{\theta}}_t$. Among these, the former is a novel result in the bandit literature, which is what we state and discuss next.

3.1. Critical Ratio and Rich History

The following bound is crucial in the proof of Theorem 1: $\|\mathbf{x}_t - \mathbf{x}_t^*\|_{V_t^{-1}} \le \beta \|\mathbf{x}_t - \mathbf{y}_t\|_{V_t^{-1}}$. A smaller value of β leads to a tighter regret bound. For this reason, we refer to the ratio $\|\mathbf{x}_t - \mathbf{x}_t^*\|_{V_t^{-1}} / \|\mathbf{x}_t - \mathbf{y}_t\|_{V_t^{-1}}$ as the *critical ratio*. Observe that if we were to run ROAM in the concurrent CDB model, *i.e.,* $\mathbf{y}_t$ were to be chosen from $\mathcal{X}_t$, then $\beta = 1$ would suffice (a similar observation is used by Saha (2021) in their proof of the regret bound). However, since $\mathbf{y}_t$ is chosen from $\mathcal{H}_t$, controlling the critical ratio is more challenging as it depends on the entire trajectory up to time t via V_t^{-1} and $\mathcal{H}_t$. Our technique for proving a suitable bound is to consider the worst-case scenario over $\mathbf{x}_t, \mathbf{x}_t^*$, and V_t^{-1}. The following definition captures this notion.

Definition 2 (Rich history) *$\mathcal{H}_t$, the history at time t, is said to be β-rich if, for any $\mathbf{x}, \mathbf{x}'$ such that $\|\mathbf{x}\|_2 \leq r$ and $\|\mathbf{x}'\|_2 \leq r$, and for any positive definite matrix A, the following bound holds:*

$$\|\mathbf{x} - \mathbf{x}'\|_A \leq \beta \max_{\mathbf{y} \in \mathcal{H}_t} \|\mathbf{x} - \mathbf{y}\|_A,$$

The following lemma shows that after a suitable number of steps of pure exploration, the history remains rich forever after (with high probability).

Lemma 3 *Let $\delta \in (0,1)$ be given. There exists a universal constant $C > 0$ such that if we run* ROAM *with*

$$\tau = \frac{Cr^2}{\lambda_{\min}(\Sigma)} \log(d/\delta),$$

then with probability at least $1 - \delta$, the history $\mathcal{H}_t$ at any time $t > \tau$ is $\frac{8r}{\sqrt{\lambda_{\min}(\Sigma)}}$-rich.

The main intuition behind this result is that all the 'richness' of the history comes from the initial pure exploration rounds. Indeed, with sufficient exploration, we are likely to have items with features spanning 'all directions'. Given this diversity of items in the history, for any $\mathbf{x}$ and A, it is possible to choose a $\mathbf{y}$ from the history that is (reasonably) well-aligned with the direction of the principal eigenvector of A (corresponding to the largest eigenvalue). This gives a lower bound on the denominator of the term $\max_{\mathbf{y} \in \mathcal{H}_t} \|\mathbf{x} - \mathbf{y}\|_A$. The remaining steps to get an upper bound on the critical ratio are not hard. The full derivation of this result is given in the appendix.

3.2. Other Key Lemmas

The three lemmas that we state in this section are standard results in the contextual bandit literature. Near-identical lemmas are stated in Saha (2021) (Lemmas 13, 1 and 2 respectively), and these results in turn are straightforward extensions of lemmas proven in Li et al. (2017).

Lemma 4 *Let $\delta \in (0,1)$ be given. Suppose the initial exploration phase of* ROAM *is run for τ rounds, where $\tau \geq c(r^2/\lambda_{\min}(\Sigma)) \log(d/\delta)$; c being a universal constant. Then, with probability at least $1 - \delta$, $\lambda_{\min}(V_{\tau+1}) \geq 1$.*

Lemma 4 states that given a sufficient initial exploration duration, with high probability, $\lambda_{\min}(V_{\tau+1})$ is at least one. Effectively, it translates to the fact that all directions are explored sufficiently well. Thus, it is similar in spirit to Lemma 3. Note that the exploration duration τ required for both Lemmas 3 and 4. Ultimately, both lemmas stem from a key matrix concentration inequality borrowed from Vershynin (2012). The proof of this Lemma is given in the Appendix. Although Li et al. (2017) also prove a very similar result using the same concentration inequality, our proof is slightly different; in particular, our proof of Lemma 4 is mirrors the proof of Lemma 3.

The following two lemmas use the condition $\lambda_{\min}(V_{\tau+1}) \geq 1$ as a starting point. Lemma 5 is a concentration inequality, giving a guarantee that the estimate $\hat{\boldsymbol{\theta}}_t$ remains close to the ground truth $\boldsymbol{\theta}^*$ at all times. Lemma 6 is an algebraic bound on the sum of the norms of the probing vectors $\mathbf{x}_t - \mathbf{y}_t$. Lemmas 5 and 6 are near-identical to Lemmas 1 and 2 in Saha (2021); the only difference being that we assume the covariates $\mathbf{x}_t$ have norm bounded by r rather than 1. For this reason, we do not prove these results here.

Lemma 5 *Suppose for some $\tau \in [T]$, $\lambda_{\min}(V_{\tau+1}) \geq 1$. Fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$,*

$$\forall\, t > \tau, \quad \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\|_{V_t} \leq \alpha, \;\; \textit{where} \;\; \alpha = \frac{1}{\kappa}\sqrt{\frac{d}{2}\log\left(1+\frac{2t}{d}\right) + \log\left(\frac{1}{\delta}\right)}.$$

Lemma 6 *Suppose for some $\tau \in [T]$, $\lambda_{\min}(V_{\tau+1}) \geq 1$. Then, for all $t > 0$,*

$$\sum_{i=\tau+1}^{t+\tau} \|\mathbf{x}_t - \mathbf{y}_t\|_{V_t^{-1}} \leq \sqrt{2td\log\left(\frac{4r^2\tau + t}{d}\right)}.$$

3.3. Proof of Main Theorem

The proof of Theorem 1 proceeds in a manner that is quite similar to the derivation of regret bounds in the literature. We begin by obtaining a bound on the instantaneous regret r_t, following which we bound the cumulative regret.

Proof of Theorem 1 Throughout this proof, we assume that the concentration inequalities of Lemmas 3, 4, and 5 hold. Each of these results hold individually with probability $1-\delta$; thus, by the union bound, all three of them hold with probability $1-3\delta$.

Recall that, for any t, $r_t = \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle - \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle$. For $t \leq \tau$, the instantaneous regret can be bounded by using the Cauchy-Schwarz inequality, the triangle inequality, and the fact that $\|\mathbf{x}\|_2 \leq r$ and $\|\boldsymbol{\theta}^*\|_2 = 1$.

$$r_t = \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle - \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle = \langle \mathbf{x}_t^* - \mathbf{x}_t, \boldsymbol{\theta}^* \rangle \leq \|\mathbf{x}_t^* - \mathbf{x}_t\|_2 \|\boldsymbol{\theta}^*\|_2 \leq 2r$$

Further, for any $t > \tau$, we obtain:

$$\begin{aligned} r_t = \langle \mathbf{x}_t^*, \boldsymbol{\theta}^* \rangle - \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle &= \langle \mathbf{x}_t^* - \mathbf{x}_t, \hat{\boldsymbol{\theta}}_t \rangle + \langle \mathbf{x}_t^* - \mathbf{x}_t, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t \rangle \overset{(i)}{\leq} \langle \mathbf{x}_t^* - \mathbf{x}_t, \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t \rangle \\ &\overset{(ii)}{\leq} \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\|_{V_t} \|\mathbf{x}_t^* - \mathbf{x}_t\|_{V_t^{-1}} \overset{(iii)}{\leq} \alpha \|\mathbf{x}_t^* - \mathbf{x}_t\|_{V_t^{-1}} \overset{(iv)}{\leq} \alpha\beta \|\mathbf{x}_t - \mathbf{y}_t\|_{V_t^{-1}} \end{aligned}$$

where inequality (i) holds because of the choice of $\mathbf{x}_t$ in `ROAM` implies $\langle \mathbf{x}_t^* - \mathbf{x}_t, \hat{\boldsymbol{\theta}}_t \rangle \leq 0$, (ii) holds because of Cauchy-Schwartz inequality, (iii) uses Lemma 5, and (iv) uses Lemma 3.

Using these bounds on the instantaneous regret, we get

$$R_T = \sum_{t=1}^{\tau} r_t + \sum_{t=\tau+1}^{T} r_t \overset{(i)}{\leq} 2r\tau + \alpha\beta \sum_{t=\tau+1}^{T} \|\mathbf{x}_t - \mathbf{y}_t\|_{V_t^{-1}} \overset{(ii)}{\leq} 2r\tau + \alpha\beta\sqrt{2dT\log\left(\frac{4r^2\tau + T}{d}\right)},$$

where (i) uses the bounds on r_t derived above and (ii) follows from Lemma 6.

Finally, plugging in the values of α from Lemma 5, β from Lemma 3, and τ from Lemmas 3 and 4 respectively, we get:

$$R_T \leq \frac{cr^2}{\lambda_{\min}(\Sigma)}\log(d/\delta) + \frac{8r}{\kappa\sqrt{\lambda_{\min}(\Sigma)}}\sqrt{\frac{d}{2}\log\left(1+\frac{2t}{d}\right) + \log\left(\frac{1}{\delta}\right)}\sqrt{2dT\log\left(\frac{4r^2\tau + T}{d}\right)}.$$

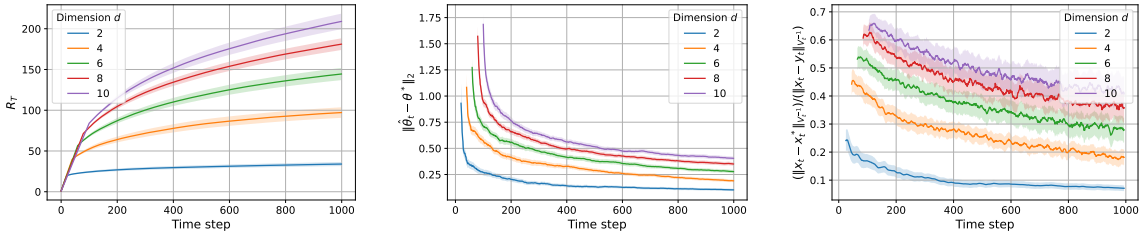
Simplifying this expression yields $R_T = O\left((r/\kappa\sqrt{\lambda_{\min}(\Sigma)})d\sqrt{T}\log(T/d\delta)\right)$ as claimed in Theorem 1. ∎

4. Experimental Results

In this section, we present experimental results illustrating the behavior of our algorithm ROAM on synthetic data. The default parameters used in the default important parameters of the experiments are given below. The ambient dimension d is set to be five. We generate both $\boldsymbol{\theta}^*$ and the context vectors uniformly on the unit ball of radius one. Thus, $r = 1$ and $1/\sqrt{\lambda_{\min}(\Sigma)}$ is $\sqrt{d}$. The size of the context set, $|\mathcal{X}_t|$ is chosen to be 1000, and the context vectors are sampled i.i.d. uniformly at random at each step. We simulate the user via a Plackett-Luce choice model; in other words, the link function $F(\cdot)$ is taken to be the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$. With this choice, we are able to use SciKitLearn's inbuilt method for Logistic Regression to calculate $\hat{\theta}_t$. The initial exploration period is set to be $10d$, and the total time horizon T is set to 1000. Note that these are default values; in the experiments that follow, we explore the performance of the algorithm by varying one parameter at a time.

For each experiment, we plots three quantities: the cumulative regret R_T, the error of the MLE estimate, $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\|_2$, and the critical ratio $\|\mathbf{x}_t - \mathbf{x}_t^*\|_{V_t^{-1}} / \|\mathbf{x}_t - \mathbf{y}_t\|_{V_t^{-1}}$. While the regret is the main performance metric of bandit algorithms, the estimation error and critical ratio are key quantities that influence the regret of our algorithm. Each plot is averaged over $n = 100$ independent runs. The solid curves shown reflect the mean over n runs and the shaded region shows a 95% confidence interval on the estimate of the mean ($\pm 2\sigma/\sqrt{n}$; σ being the empirical standard deviation over the n runs). In addition, the critical ratio plots are smoothened by a moving average window of length ten. All experiments were run on locally on a MacBook with the M1 Pro chip and 16GB RAM. For all experiments, a single run (over $T = 1000$ timesteps) took approximately one second.

Our first experiment, Figure 1, examines the variation of these three quantities with the underlying dimension d. We choose five different values of d: $\{2, 4, 6, 8, 10\}$. The rest of the parameters are chosen as per the default values; in particular, τ grows linearly with d. We plot the regret, error, and the critical ratio as a function of time in Figures 1(a), 1(b), and 1(c) respectively. As expected, the regret increases linearly in the initial pure exploration phase, followed by a sublinear increase in the main explore-exploit phase. We observe that the regret increases with the dimension, roughly linearly. The error in the estimate, $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$, decreases gracefully over time, but increases with d. The critical ratio is significantly smaller in practice than the theoretical bound, but increases with the dimension d, consistent with the analysis. It also decreases slightly over time.

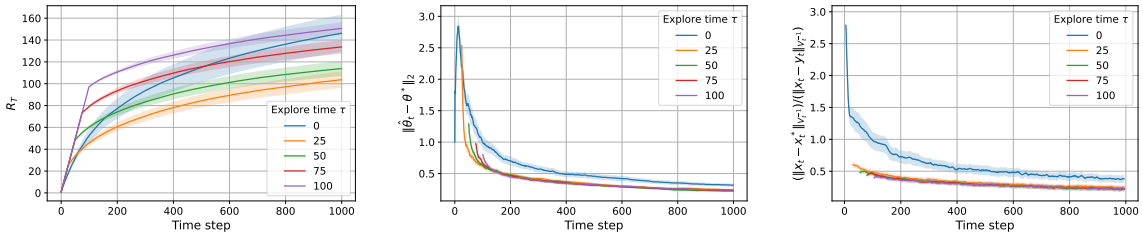


(a) Cumulative Regret (b) Error in MLE (c) Critical Ratio

Figure 1: Variation of regret, error, and the critical ratio as a function of dimension d.

In Figure 2, we examine the role of the initial pure exploration phase. We vary τ among the values $\{0, 25, 50, 75, 100\}$, keeping all other parameters to their default values listed above. In the case where $\tau = 0$, we add a small regularizer parameter of $\lambda = 0.1$ while calculating the maximum likelihood estimator. (We experimented with different values of λ and found this to be the best). We

observe in Figure 2(*a*) that as τ increases from 25 to 100, the regret progressively worsens. This is because the extra initial exploration affects neither the error in $\hat{\boldsymbol{\theta}}$, nor the critical ratio. Thus, the extra exploration penalty is not compensated for (see Figures 2(*b*), 2(*c*)). However, when $\tau = 0$, the regret is large. This is because both the error and the critical ratio are significantly larger for this case. This experiment demonstrates that a small amount of initial exploration is optimal for ROAM.



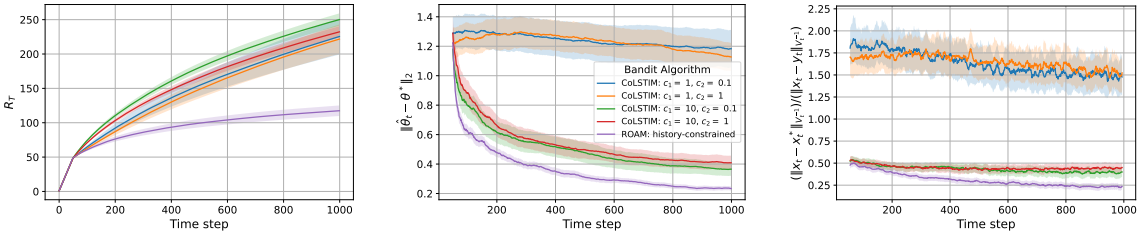
(*a*) Cumulative Regret (*b*) Error in MLE (*c*) Critical Ratio

Figure 2: Variation of regret, error, and the critical ratio as a function of exploration time τ.

Our final experiment (Figure 3) compares ROAM, an algorithm designed for the history-constrained CDB model, with CoLSTIM, the state-of-the-art algorithm for concurrent CDBs (Bengs et al., 2022). For CoLSTIM, we experiment with a few values of the hyper-parameters. c_1 controls the exploration-exploitation tradeoff in the choice of $\mathbf{y}$, and c_2 (called C_{thresh} in Bengs et al. (2022)) controls a similar tradeoff in the choice of $\mathbf{x}$. We vary these hyperparameters, choosing $c_1 \in \{1, 10\}$ and $c_2 \in \{0.1, 1\}$. Recall that for ROAM, there is no such hyper-parameter. We plot the regret in terms of $\mathbf{x}$ alone for ROAM and the average regret in terms of $\mathbf{x}$ and $\mathbf{y}$ for CoLSTIM. This is a fair comparison, for one counts a unit of regret for every item recommended and for every comparison made. Figure 3(*a*) shows that ROAM outperforms CoLSTIM over all values of the hyperparameters. The performance can be better understood through Figures 3(*b*) and 3(*c*). On the one hand, for $c_1 = 1$ (low exploration), we observe that CoLSTIM does not learn $\boldsymbol{\theta}^*$ quickly, leading to large regret. On the other hand, for $c_1 = 10$ (large exploration), the exploratory nature of $\mathbf{y}$ leads to a large regret, despite $\hat{\boldsymbol{\theta}}_t$ converging quickly to $\boldsymbol{\theta}^*$. A deeper understanding can be obtained by casting the concurrent model in the history-constrained setting. Assume the context set remains unchanged for two successive time steps (say t and $t + 1/2$). CoLSTIM first recommends $\mathbf{x}_t$ and then $\mathbf{y}_t$ (paying regret for both), and then compares the latest item ($\mathbf{y}_t$) to *the one consumed just before* ($\mathbf{x}_t$). ROAM, on the other hand, recommends $\mathbf{x}_t$ at both successive time steps (paying twice the regret), but compares the latest item to *any suitable item in the past*. This experiment clearly demonstrates the benefit of recycling items from the user's consumption history.

5. Discussion

Summary In this work, we propose a new contextual bandit model that learns a user's tastes by seeking explicit comparisons between consumed items. The key idea is to reuse previously recommended items–incurring no new regret–for future comparisons. Our algorithm, ROAM, exploits this flexibility to accumulate a rich history within a short exploration time. This rich history enables informative queries subsequently, yielding $O(\sqrt{T})$ regret. Our simulations show that ROAM outperforms the state-of-the-art in concurrent CDBs. Our work opens up several avenues of future research, which we discuss below.



(*a*) Cumulative Regret (*b*) Error in MLE (*c*) Critical Ratio

Figure 3: A comparison of ROAM with CoLSTIM, keeping all parameters identical.

Model Interpretation We assume that $\boldsymbol{\theta}^*$ encodes a single user's fixed preferences, and item features are known and time-invariant. These assumptions ensure a consistent notion of a user's history, useful for selecting past items for comparison. In alternative contextual bandit interpretations, item features may vary over time (*e.g.*, by incorporating user metadata), and $\boldsymbol{\theta}^*$ may consequently reflect global feature weights. While our analysis depends on the particular viewpoint we adopt, the broader insight, that comparing with items consumed in the past is valuable, is likely to hold more generally. Formalizing this intuition in other contextual bandit models remains an open direction.

Handling Nonlinearity Like prior work on CDBs (Saha, 2021; Bengs et al., 2022), our regret bounds scale with $1/\kappa$, which can grow exponentially with r. This is a result of the analysis technique, which uses a uniform lower bound on $F'(\cdot)$ (see Lemma 5). Recent work on generalized linear bandits avoids this dependence on κ by more refined handling of the nonlinearity (Dong et al., 2019; Faury et al., 2020; Abeille et al., 2021). Extending such techniques to the CDB setting could improve regret bounds, but remains an open challenge.

Comparison Horizon Our algorithm selects comparison items from the entire history, often reusing those from the exploration phase. In practice, it may be preferable to compare with recently consumed items. A caveat is that once the user's preferences are reasonably estimated, such items are likely to be similar. In terms of the richness of the history, this could be severely detrimental (see Section 3.1). However, comparisons among similar queries could actually prove useful to resolve fine-grained rankings among items. Alternate algorithms, potentially guided by criteria like the information ratio (Dong et al., 2019) may prove to be effective in this scenario.

Choice Model Like prior work, we assume that choice probabilities depend only on utility differences. This idealization may fail in practice: some item pairs may not be comparable, noise may be temporally correlated, and preferences may be shaped by context or prior experiences (e.g., anchoring) (Tversky, 1972; Tversky and Kahneman, 1974). Capturing such effects requires richer choice models. Developing practical bandit algorithms under more realistic user behavior remains an important direction for future work.

Holistic Cost In our model, we treat item consumption as costly (incurring regret) and comparisons as essentially free. However, taken to an extreme, this would suggest querying repeatedly with every new item recommended, which is impractical. A more realistic model would assign costs to queries and optimize the total cost: regret plus query burden.

Acknowledgments

This research has been supported in part by the Swiss National Science Foundation (SNSF) under grant IZBRZ2_186313.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, 2011.

Marc Abeille, Louis Faury, and Clement Calauzenes. Instance-wise minimax-optimal algorithms for logistic bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Proceedings of Machine Learning Research. PMLR, 2021.

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. Online bandit learning with offline preference data. *arXiv preprint arXiv:2406.09574*, 2024.

Walid Bendada, Guillaume Salha, and Théo Bontempelli. Carousel personalization in music streaming apps with contextual bandits. In *Proceedings of the 14th ACM Conference on Recommender Systems*. Association for Computing Machinery, 2020.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research (JMLR)*, 22(7):1–108, 2021.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the performance of thompson sampling on logistic bandits. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT)*, Proceedings of Machine Learning Research. PMLR, 2019.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, Proceedings of Machine Learning Research. PMLR, 2015.

Louis Faury, Marc Abeille, Clement Calauzenes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research. PMLR, 2020.

Xu He, Bo An, Yanghua Li, Haikai Chen, Qingyu Guo, Xin Li, and Zhirong Wang. Contextual user browsing bandits for large-scale online mobile recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. Association for Computing Machinery, 2020.

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

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research. PMLR, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Bruno L. Pereira, Alberto Ueda, Gustavo Penha, Rodrygo L. T. Santos, and Nivio Ziviani. Online learning to rank for sequential music recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. Association for Computing Machinery, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory (ALT)*, Proceedings of Machine Learning Research. PMLR, 2022.

Amos Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281, 1972.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1974.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing: theory and applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

Jingyan Wang and Nihar B. Shah. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Appendix A. Proofs of Lemmas 3 and 4

A.1. Basics of Symmetric Matrices

Let $\mathcal{S}_+$ denote the set of positive definite matrices of dimension d. Any $A \in \mathcal{S}_+$ can be expressed in terms of its eigenvalues and eigenvectors as follows:

$$A = \sum_{i=1}^{d} \lambda_i(A) v_i(A) v_i(A)^\top, \tag{2}$$

where $\lambda_1(A) \geq \lambda_2(A) \ldots \geq \lambda_d(A) \geq 0$ are the eigenvalues of A. $v_1(A), \ldots, v_d(A)$ are the corresponding eigenvectors, forming an orthonormal basis of $\mathbb{R}^d$. In particular, $\|v_i(A)\|_2 = 1 \; \forall \; i, \; \forall \; A$. When the matrix A is clear from context, we use the simpler notation λ_i and v_i.

For any $A \in \mathcal{S}_+$, the matrix-induced norm $\|\cdot\|_A$ is defined as:

$$\|z\|_A = \sqrt{z^\top A z} \; \forall \; z, \; \forall \; A \tag{3}$$

By the eigen-decomposition of the matrix, it follows that

$$\|z\|_A^2 \geq \lambda_1(A) \langle z, v_1(A) \rangle^2 \; \forall \; z, \; \forall \; A \tag{4}$$

We can also derive the following upper bound:

$$\|z\|_A^2 \leq \lambda_1(A) \|z\|_2^2 \; \forall \; z, \; \forall \; A \tag{5}$$

For any matrix A, let $\|A\|_2$ denote the induced ℓ_2 norm of the matrix. If $A \in \mathcal{S}_+$, we have the identity

$$\|A\|_2 = \sup_{v : \|v\|_2 = 1} v^\top A v = \lambda_1(A) \tag{6}$$

Let Σ be an arbitrary positive definite matrix and z an arbitrary vector (both in d dimensions). Let $y = \Sigma^{-1/2} z$ and $B = \Sigma^{1/2} A \Sigma^{1/2}$. Then $\|z\|_A = \|y\|_B$. Further,

$$\lambda_1(B) = \lambda_1(\Sigma^{1/2} A \Sigma^{1/2}) = \sup_{v : \|v\|_2 = 1} v^\top \Sigma^{1/2} A \Sigma^{1/2} v \geq \left(\lambda_{\min}(\Sigma^{1/2}) \right)^2 \sup_{v : \|v\|_2 = 1} v^\top A v = \lambda_1(A) \lambda_{\min}(\Sigma) \tag{7}$$

A.2. Basic Concentration Results

Definition 7 (Isotropic Distribution) *A random vector $z \in \mathbb{R}^d$ is said to satisfy an isotropic distribution if $\mathbb{E}[z z^\top] = I_d$; here, I_d refers to the identity matrix in d dimensions.*

Lemma 8 (Key Concentration Inequality) *Let $z_1, z_2, \ldots$ be an i.i.d. sequence of random vectors from an isotropic distribution, satisfying the bound $\|z\|_2 \leq r$ almost surely. Let $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$ be given. Suppose $\tau \geq c(r^2/\epsilon^2) \log(d/\delta)$, where $c > 0$ is a universal constant. Then, with probability at least $1 - \delta$,*

$$\left\| \frac{1}{\tau} \sum_{t=1}^{\tau} z_t z_t^\top - I_d \right\|_2 \leq \epsilon$$

Proof This lemma follows in a straightforward fashion from a matrix concentration result proven in Vershynin (2012), namely Theorem 5.41. We reproduce this result here, in our notation. This result states that there exists an absolute constant $c' > 0$ such that the following statement holds. For any $s > 0$, with probability at least $1 - 2d\exp(-c's^2)$,

$$\left\| \frac{1}{\tau} \sum_{t=1}^{\tau} z_t z_t^\top - I_d \right\|_2 \leq \max(\varepsilon, \varepsilon^2) =: \epsilon \text{ where } \varepsilon = s\frac{r}{\sqrt{\tau}}.$$

Since we have chosen ϵ to be less than one, $\max(\varepsilon, \varepsilon^2)$ equals ε. It remains to deduce a value of τ as a function of ϵ and δ.

Setting $2d\exp(-c's^2)$ to δ gives us $s = O(\sqrt{\log(d/\delta)})$. Setting $sr/\sqrt{\tau}$ to ϵ gives us $\tau = O(r^2/\epsilon^2)\log(d/\delta)$. The constant hidden in the $O(\cdot)$ notation is a universal constant. ■

Lemma 9 (Existence of a Good Vector) *Let $z_1, z_2, \ldots$ be an i.i.d. sequence of random vectors from an isotropic distribution, satisfying the bound $\|z\|_2 \leq r$ almost surely. Let $\epsilon \in (0,1)$ and $\delta \in (0,1)$ be given. There exists a universal constant $c > 0$ such that if $\tau \geq c(r^2/\epsilon^2)\log(d/\delta)$, then with probability at least $1 - \delta$,*

$$\inf_{v:\|v\|_2=1} \max_{t\in[\tau]} \langle z_t, v\rangle^2 \geq 1 - \epsilon$$

Proof This result follows easily from Lemma 8. We outline the steps below. Suppose the concentration result holds (which happens with probability $1 - \delta$). Then

$$\left\| \frac{1}{\tau} \sum_{t=1}^{\tau} z_t z_t^\top - I_d \right\|_2 \leq \epsilon$$

$$\Rightarrow \forall\, v : \|v\|_2 = 1,\ \left| v^\top \left(\frac{1}{\tau} \sum_{t=1}^{\tau} z_t z_t^\top - I_d \right) v \right| \leq \epsilon \quad \text{(by (6))}$$

$$\Rightarrow \inf_{v:\|v\|_2=1} \left| \frac{1}{\tau} \sum_{t=1}^{\tau} \langle v, z_t\rangle^2 - \|v\|_2^2 \right| \leq \epsilon$$

$$\Rightarrow \inf_{v:\|v\|_2=1} \frac{1}{\tau} \sum_{t=1}^{\tau} \langle v, z_t\rangle^2 \geq 1 - \epsilon \quad (\because \|v\|_2 = 1)$$

$$\Rightarrow \inf_{v:\|v\|_2=1} \max_{t\in[\tau]} \langle v, z_t\rangle^2 \geq 1 - \epsilon \quad (\because \text{maximum is larger than average})$$

■

Lemma 10 (Lower Bound on Matrix Norm) *Let $z_1, z_2, \ldots$ be an i.i.d. sequence of random vectors from an isotropic distribution, satisfying the bound $\|z\|_2 \leq r$ almost surely. Let $\epsilon \in (0,1)$ and $\delta \in (0,1)$ be given. Suppose $\tau \geq c(r^2/\epsilon^2)\log(d/\delta)$, where c is some universal constant. Then, with probability at least $1 - \delta$,*

$$\forall\, A \in \mathcal{S}_+,\ \max_{t\in[\tau]} \|z_t\|_A^2 \geq (1 - \epsilon)\lambda_1(A)$$

Proof This result follows easily from Lemma 9 and the basic inequalities concerning positive semidefinite matrices presented in Section A.1. Starting from (4), we get:

$$\begin{aligned}&\forall\, A \in \mathcal{S}_+,\ \forall\, z \in \mathbb{R}^d, \|z\|_A^2 \geq \lambda_1(A)\langle z, v_1(A)\rangle^2 \\ \Rightarrow\ &\forall\, A \in \mathcal{S}_+, \forall\, t \in [\tau], \|z_t\|_A^2 \geq \lambda_1(A)\langle z_t, v_1(A)\rangle^2\end{aligned}$$

By definition, $\|v_1(A)\|_2 = 1$ for all $A \in \mathcal{S}_+$. Therefore, By Lemma 9, with probability at least $1-\delta$,

$$\forall\, A \in \mathcal{S}_+, \max_{t\in[\tau]} \langle v_1(A), z_t\rangle^2 \geq 1-\epsilon$$

Putting these two inequalities together, we get that with probability at least $1-\delta$,

$$\begin{aligned}\forall\, A \in \mathcal{S}_+,\ \max_{t\in[\tau]} \|z_t\|_A^2 &\geq \max_{t\in[\tau]} \lambda_1(A)\langle z_t, v_1(A)\rangle^2 \\ &\geq (1-\epsilon)\lambda_1(A).\end{aligned}$$

■

The next lemma extends this result to the case of nonisotropic random vectors.

Lemma 11 (Lower Bound for Nonisotropic Vectors) *Let $z_1, z_2, \ldots$ be an i.i.d. sequence of random vectors, with a distribution such that $\mathbb{E}[zz^\top] = \Sigma$ is invertible and the bound $\|z\|_2 \leq r$ holds almost surely. Let $\epsilon \in (0,1)$ and $\delta \in (0,1)$ be given. Suppose $\tau \geq c(r^2/\lambda_{\min}(\Sigma)\epsilon^2)\log(d/\delta)$, where c is some universal constant. Then, with probability at least $1-\delta$,*

$$\forall\, A \in \mathcal{S}_+,\ \max_{t\in[\tau]} \|z_t\|_A^2 \geq (1-\epsilon)\lambda_1(A)\lambda_{\min}(\Sigma)$$

Proof Define $w_i = \Sigma^{-1/2} z_i$. Then $w_1, w_2, \ldots$ is an i.i.d. sequence of isotropic random vectors satisfying the bound $\|w\|_2 \leq \|\Sigma^{-1/2}\|_2 r$, which equals $r/\sqrt{\lambda_{\min}(\Sigma)}$. Invoking Lemma 10, with $\tau = c(r^2/\lambda_{\min}(\Sigma)\epsilon^2)\log(d/\delta)$, we see that with probability at least $1-\delta$,

$$\forall\, A \in \mathcal{S}_+,\ \max_{t\in[\tau]} \|w_t\|_A^2 \geq (1-\epsilon)\lambda_1(A)$$

The desired result follows by noting that $\|z_t\|_A^2 = \|w_t\|_B^2$, where $B = \Sigma^{1/2} A \Sigma^{1/2}$, and (7). ■

A.3. Proof of Lemma 3

The proof of Lemma 3 follows from the concentration result in Lemma 11 and the following algebraic result.

Lemma 12 (Triangle Inequality) *For any τ, for any $\mathbf{x} \in \mathbb{R}^d$, and for any $A \in \mathcal{S}_+$,*

$$\max_{\mathbf{y}\in\mathcal{H}_{2\tau+1}} \|\mathbf{x}-\mathbf{y}\|_A \geq (1/2) \max_{t\in[\tau]} \|\mathbf{x}_{2t-1} - \mathbf{x}_{2t}\|_A$$

Proof The result follows from a straightforward inequality of triangle inequality. Recall, by definition, $\mathcal{H}_{2\tau+1} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{2\tau}\}$. It follows that

$$\begin{aligned}\max_{\mathbf{y}\in\mathcal{H}_{2\tau+1}} \|\mathbf{x}-\mathbf{y}\|_A &= \max_{t\in[2\tau]} \|\mathbf{x}-\mathbf{x}_t\|_A \\ &= \max_{t\in[\tau]} \left(\max\{\|\mathbf{x}-\mathbf{x}_{2t-1}\|_A, \|\mathbf{x}-\mathbf{x}_{2t}\|_A\}\right) \\ &\geq \max_{t\in[\tau]} (\|\mathbf{x}-\mathbf{x}_{2t-1}\|_A + \|\mathbf{x}-\mathbf{x}_{2t}\|_A)/2 \\ &\geq \max_{t\in[\tau]} (\|\mathbf{x}_{2t-1}-\mathbf{x}_{2t}\|_A)/2\end{aligned}$$

The last step uses the triangle inequality with respect to the norm $\|\cdot\|_A$. ■

We now have all the ingredients to prove the main result of this section, which states that the history at time $2\tau+1$ is β-rich with high probability, for $\beta = 8r/\sqrt{\lambda_{\min}(\Sigma)}$.

Lemma 13 (Main Inequality) *Let $\delta \in (0,1)$ be given. Suppose the initial exploration phase of* ROAM *is run for τ rounds, where $\tau \geq c(r^2/\lambda_{\min}(\Sigma))\log(d/\delta)$; c being a universal constant. Then, with probability at least $1-\delta$, for any $\mathbf{x}, \mathbf{x}'$ satisfying $\|\mathbf{x}\|_2 \leq r, \|\mathbf{x}'\|_2 \leq r$ and for any $A \in \mathcal{S}_+$*

$$\|\mathbf{x}-\mathbf{x}'\|_A \leq \frac{8r}{\sqrt{\lambda_{\min}(\Sigma)}} \sup_{\mathbf{y}\in\mathcal{H}_{2\tau+1}} \|\mathbf{x}-\mathbf{y}\|_A.$$

Proof The proof consists of deriving an upper bound for the term on the left hand side ($\|\mathbf{x}-\mathbf{x}'\|_A$) and a lower bound for the right hand side ($\sup_{\mathbf{y}\in\mathcal{H}_{2\tau+1}} \|\mathbf{x}-\mathbf{y}\|_A$).

Combining (5) with the fact that both $\mathbf{x}$ and $\mathbf{x}'$ have norm at most r, we get the following upper bound:

$$\|\mathbf{x}-\mathbf{x}'\|_A \leq \sqrt{\lambda_1(A)}\|\mathbf{x}-\mathbf{x}'\|_2 \leq 2r\sqrt{\lambda_1(A)} \tag{8}$$

By Lemma 12, we get that

$$\max_{\mathbf{y}\in\mathcal{H}_{2\tau+1}} \|\mathbf{x}-\mathbf{y}\|_A \geq (1/2) \max_{t\in[\tau]} \|\mathbf{x}_{2t-1}-\mathbf{x}_{2t}\|_A \tag{9}$$

Denote $\mathbf{x}_{2t-1} - \mathbf{x}_{2t}$ by $\mathbf{z}_t$. Then $\mathbf{z}_1, \ldots, \mathbf{z}_\tau$ are i.i.d. random vectors satisfying $\|\mathbf{z}\|_2 \leq r$ almost surely and $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \Sigma$. Invoking Lemma 11 with $\epsilon = 3/4$, we get that if $\tau \geq O(r^2/\lambda_{\min}(\Sigma))\log(d/\delta)$, then with probability at least $1-\delta$,

$$\max_{t\in[\tau]} \|\mathbf{x}_{2t-1}-\mathbf{x}_{2t}\|_A \geq \frac{1}{2}\sqrt{\lambda_1(A)\lambda_{\min}(\Sigma)} \tag{10}$$

Combining (8), (9), and (10), the stated result follows. ■

Lemma 3 follows from Lemma 13 by an appropriate change of notation from τ to 2τ and noting that once the history at time $\tau+1$ is β rich, the history at all subsequent times is also β rich. This is because for all $t \geq \tau+1$, $\mathcal{H}_{\tau+1} \subseteq \mathcal{H}_t$, which implies $\sup_{\mathbf{y}\in\mathcal{H}_{\tau+1}} \|\mathbf{x}-\mathbf{y}\|_A \leq \sup_{\mathbf{y}\in\mathcal{H}_t} \|\mathbf{x}-\mathbf{y}\|_A$.

A.4. Proof of Lemma 4

Lemma 4 states that with high probability, $\lambda_{\min}(V_{\tau+1}) \geq 1$. Before we get to this result, we prove a simple result regarding the relation between eigenvalues of positive definite matrices A and B that are congruent (see Lemma 14). Let $\mathbb{B}$ denote the unit ball in d-dimensions, *i.e.*, $\mathbb{B} = \{\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$. Recall that for any $A \in \mathcal{S}_+$,

$$\lambda_{\max}(A) = \max_{\mathbf{x}\in\mathbb{B}} \mathbf{x}^\top A\mathbf{x}\,; \quad \lambda_{\min}(A) = \min_{\mathbf{x}\in\mathbb{B}} \mathbf{x}^\top A\mathbf{x}$$

Using these definitions of the eigenvalues, we prove the following lemma.

Lemma 14 *Let $A, \Sigma \in \mathcal{S}_+$ be given. Let $B \triangleq \Sigma^{1/2} A \Sigma^{1/2}$. Then $B \in \mathcal{S}_+$. Further, the following inequalities hold:*

$$\lambda_{\max}(B) \geq \lambda_{\min}(\Sigma)\lambda_{\max}(A) \tag{11}$$
$$\lambda_{\min}(B) \geq \lambda_{\min}(\Sigma)\lambda_{\min}(A) \tag{12}$$

Proof The claim that B is positive definite is easily verified from the definition. We proceed to prove (12) first. Let $\underline{\mathbf{x}}$ denote the eigenvector of B corresponding to its smallest eigenvalue (of norm one). Let $\underline{\mathbf{y}}$ be the vector of norm one aligned along $\Sigma^{1/2}\underline{\mathbf{x}}$, that is,

$$\underline{\mathbf{y}} = \Sigma^{1/2}\underline{\mathbf{x}} / \left\|\Sigma^{1/2}\underline{\mathbf{x}}\right\|_2 .$$

Thus, $\underline{\mathbf{y}} \in \mathbb{B}$. Furthermore, we have the inequality

$$\left\|\Sigma^{1/2}\underline{\mathbf{x}}\right\|_2 = \sqrt{\underline{\mathbf{x}}\Sigma\underline{\mathbf{x}}} \geq \sqrt{\lambda_{\min}(\Sigma)} \quad (\because\ \Sigma \in \mathcal{S}_+ \text{ and } \underline{\mathbf{x}} \in \mathbb{B})$$

We now have all the ingredients to prove (12).

$$\begin{aligned}\lambda_{\min}(B) = \underline{\mathbf{x}}^\top B\underline{\mathbf{x}} = \underline{\mathbf{x}}^\top \Sigma^{1/2} A \Sigma^{1/2}\underline{\mathbf{x}} = \left\|\Sigma^{1/2}\underline{\mathbf{x}}\right\|_2^2 \underline{\mathbf{y}}^\top A\underline{\mathbf{y}} \\ \geq \lambda_{\min}(\Sigma)\underline{\mathbf{y}}^\top A\underline{\mathbf{y}} \geq \lambda_{\min}(\Sigma) \inf_{\mathbf{y}\in\mathbb{B}} \mathbf{y}^\top A\mathbf{y} = \lambda_{\min}(\Sigma)\lambda_{\min}(A)\end{aligned}$$

The proof of (11) follows similar steps, but with a careful reordering of the arguments. Let $\overline{\mathbf{y}}$ denote the eigenvector of A corresponding to its largest eigenvalue (of norm one). Let $\overline{\mathbf{x}}$ be the vector of norm one aligned along $\Sigma^{-1/2}\overline{\mathbf{y}}$, that is,

$$\overline{\mathbf{x}} = \Sigma^{-1/2}\overline{\mathbf{y}} / \left\|\Sigma^{-1/2}\overline{\mathbf{y}}\right\|_2 .$$

Thus, $\overline{\mathbf{x}} \in \mathbb{B}$. Also note that

$$\left\|\Sigma^{-1/2}\overline{\mathbf{y}}\right\|_2 = \sqrt{\overline{\mathbf{y}}^\top \Sigma^{-1}\overline{\mathbf{y}}} \leq \sqrt{\lambda_{\max}(\Sigma^{-1})} = 1/\sqrt{\lambda_{\min}(\Sigma)}$$

Finally, note that

$$B = \Sigma^{1/2} A \Sigma^{1/2} \ \Leftrightarrow\ A = \Sigma^{-1/2} B \Sigma^{-1/2}$$

Putting these equations together, we get:

$$\lambda_{\max}(A) = \overline{\mathbf{y}}^\top A \overline{\mathbf{y}} = \overline{\mathbf{y}}^\top \Sigma^{-1/2} B \Sigma^{-1/2} \overline{\mathbf{y}} = \left\| \Sigma^{-1/2} \overline{\mathbf{y}} \right\|_2^2 \overline{\mathbf{x}}^\top B \overline{\mathbf{x}}$$
$$\leq \lambda_{\min}^{-1}(\Sigma) \overline{\mathbf{x}}^\top B \overline{\mathbf{x}} \leq \lambda_{\min}^{-1}(\Sigma) \sup_{\mathbf{x} \in \mathbb{B}} \mathbf{x}^\top B \mathbf{x} = \lambda_{\min}^{-1}(\Sigma) \lambda_{\max}(B)$$

This proves (11), and also provides a more detailed justification for (7) (which is used in the proof of Lemma 11). ■

Proof of Lemma 4 Recall that, in the pure exploration phase, $\mathbf{y}_t = \mathbf{x}_{t-1}$. Therefore,

$$V_{\tau+1} = \sum_{t=1}^{\tau} (\mathbf{x}_t - \mathbf{y}_t)(\mathbf{x}_t - \mathbf{y}_t)^\top = \sum_{t=1}^{\tau} (\mathbf{x}_t - \mathbf{x}_{t-1})(\mathbf{x}_t - \mathbf{x}_{t-1})^\top$$

We can express $V_{\tau+1}$ as the sum of two matrices, $V_{\tau+1}^{\text{even}} + V_{\tau+1}^{\text{odd}}$, where:

$$V_{\tau+1}^{\text{even}} = \sum_{t=1}^{\tau/2} (\mathbf{x}_{2t} - \mathbf{x}_{2t-1})(\mathbf{x}_{2t} - \mathbf{x}_{2t-1})^\top; \quad V_{\tau+1}^{\text{odd}} = \sum_{t=1}^{\tau/2} (\mathbf{x}_{2t-1} - \mathbf{x}_{2t-2})(\mathbf{x}_{2t-1} - \mathbf{x}_{2t-2})^\top$$

Recall that $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are i.i.d. random vectors with distribution $\mathcal{P}$. Denoting $\mathbf{x}_{2t} - \mathbf{x}_{2t-1}$ by $\mathbf{z}_{2t}$, we observe that $\mathbf{z}_2, \mathbf{z}_4, \ldots$ are i.i.d. random vectors satisfying $\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \Sigma$ and $\|\mathbf{z}\|_2 \leq 2r$ almost surely.

Let $\mathbf{w}_i \triangleq \Sigma^{-1/2} \mathbf{z}_{2i}$. Then $\mathbf{w}_1, \mathbf{w}_2, \ldots$ are i.i.d. random vectors satisfying $\mathbb{E}[\mathbf{w}\mathbf{w}^T] = I$ and $\|\mathbf{z}\|_2 \leq 2r\sqrt{\|\Sigma^{-1/2}\|_2}$ almost surely. We also know that $\sqrt{\|\Sigma^{-1/2}\|_2} = 1/\sqrt{\lambda_{\min}(\Sigma)}$, which gives us $\|\mathbf{z}\|_2 \leq 2r/\sqrt{\lambda_{\min}(\Sigma)}$.

Define the matrix

$$U_\tau = \left(\frac{2}{\tau}\right) \sum_{i \in [\tau/2]} \mathbf{w}_i \mathbf{w}_i^\top$$

Observe that $\mathbf{w}_i = \Sigma^{-1/2} \mathbf{z}_{2i}$ implies $\mathbf{z}_{2i} = \Sigma^{1/2} \mathbf{w}_i$. Thus,

$$V_{\tau+1}^{\text{even}} = \sum_{t=1}^{\tau/2} \mathbf{z}_{2t} \mathbf{z}_{2t}^\top = \left(\frac{\tau}{2}\right) \Sigma^{1/2} U_\tau \Sigma^{1/2}$$

Using Lemma 14 (in particular, (12)), we get

$$\lambda_{\min}(V_{\tau+1}^{\text{even}}) = \left(\frac{\tau}{2}\right) \lambda_{\min}(\Sigma^{1/2} U_\tau \Sigma^{1/2}) \geq \left(\frac{\tau}{2} \lambda_{\min}(\Sigma)\right) \lambda_{\min}(U_\tau)$$

Applying Lemma 8 to U_τ with $\epsilon = 1/2$, we conclude that with probability at least $1 - \delta$,

$$\lambda_{\min}(U_\tau) \geq 1/2 \Rightarrow \lambda_{\min}(V_{\tau+1}^{\text{even}}) \geq \frac{\tau}{4} \lambda_{\min}(\Sigma) \Rightarrow \lambda_{\min}(V_{\tau+1}) \geq \frac{\tau}{4} \lambda_{\min}(\Sigma)$$

The last step follows from the fact that adding the positive semidefinite matrix $V_{\tau+1}^{\text{odd}}$ to $V_{\tau+1}^{\text{even}}$ can only raise its minimum eigenvalue. Finally, we know that $\tau \geq c(r^2/\lambda_{\min}(\Sigma)) \log(d/\delta)$. Choosing c large enough, we get that $\tau/4 \geq 1/\lambda_{\min}(\Sigma)$, which implies $\lambda_{\min}(V_{\tau+1}) \geq 1$. ■