COMPETITIVE RATIO AND ITS APPLICATION IN SEQUENTIAL REPRESENTATION LEARNING FOR MULTI-TASK LINEAR BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the competitive ratio between the cumulative loss of Follow-The-Leader (FTL) and that of the best expert in hindsight for online subset and subspace selection. In the *subset selection* problem, the learner chooses a set of s experts from a pool of size K at each step, and we show that FTL is K-competitive. In the *subspace selection* problem, also known as online principal component analysis, the learner chooses an m-dimensional subspace in $\mathbb{R}^d$ at each step, observes a context vector x, and incurs a "compression loss." We show that FTL achieves a competitive ratio of d under some mild assumptions. We apply these results to sequential representation learning in multi-task linear bandits and develop an algorithm `BARON`. We provide regret guarantees in the form of upper and lower bounds, and further demonstrate its computational efficiency empirically on a synthetic dataset.

1 INTRODUCTION

A well-established framework for evaluating online algorithms is through *competitive ratio analysis* (Borodin & El-Yaniv, 2005), which compares its performance to that of an optimal offline algorithm that possesses complete knowledge of the entire input sequence. While competitive ratio has been studied for many online learning tasks, such as the paging problem, the k-server problem, metrical task systems, scheduling, and load balancing (Buchbinder et al., 2012; Borodin & El-Yaniv, 2005; Albers, 2003), its extension to the online expert selection setting (Warmuth & Kuzmin, 2008) or the online subspace selection is still limited.

In particular, the online subspace selection problem, or online principal component analysis (PCA), refers to performing dimensionality reduction in scenarios involving streaming or very large datasets. This setting arises when data arrives sequentially, making it impractical or impossible to store the entire dataset in memory for traditional batch processing. Although much of the literature focuses on reconstruction error at the last step of online PCA algorithms (Mitliagkas et al., 2013; Garber et al., 2015), we are interested in the cumulative loss in the online learning process, which has been studied by Nie et al. (2016) and Warmuth & Kuzmin (2008). To the best of our knowledge, no work has studied the competitive ratio of the greedy Follow-The-Leader (FTL) algorithm in this setting. Thus, we ask the following question:

What is the best achievable competitive ratio of the FTL algorithm for the online subspace selection problem?

In this work, we show that FTL attains the competitive ratio of d, the parameter's dimensionality, which is useful to solve the sequential representation transfer for the multi-task linear bandit problem . Here, a learning agent sequentially encounters N linear bandit tasks (Qin et al., 2022; Duong et al., 2024) in a d-dimensional feature space, each lasting for τ time steps. A key characteristic of these tasks is that their reward parameter vectors reside within an m-dimensional linear subspace ($m \ll d$). This is motivated by applications such as recommender systems, where different customers (tasks) have similar preferences. The learner's objective is to minimize the cumulative regret across all tasks (or meta-regret) by effectively leveraging this shared subspace structure.

A straightforward approach involves treating each task in isolation and applying a standard bandit algorithm, such as LinUCB (Abbasi-Yadkori et al., 2011) or PEGE (Rusmevichientong & Tsitsiklis, 2010). This independent learning strategy, which we will subsequently refer to as the *individual single-task baseline*, would result in a meta-regret upper bound of $\tilde{O}(Nd\sqrt{\tau})$[1]. Conversely, if the m-dimensional subspace common to all tasks were known a priori, the learner would only need to estimate the projection of each task's reward predictor onto this subspace, resulting in a better

[1] $\tilde{O}$ notation hides polylogarithmic factors in τ, N, d

meta-regret of $\tilde{O}(Nm\sqrt{\tau})$. This work focuses on the scenario where N and τ are large, and m is significantly smaller than d ($m \ll d$), a regime where the benefits of representation transfer learning are most pronounced.

While significant progress has been made in multi-task linear bandits in the parallel setting (e.g. Yang et al., 2020; Hu et al., 2021; Yang et al., 2022; Cella et al., 2023), advancements in the sequential setting have been comparatively limited. Assuming that the action sets are well-conditioned ellipsoids, Qin et al. (2022) proposed an efficient algorithm achieving a meta-regret of $\tilde{O}\left(Nm\sqrt{\tau} + dm\sqrt{\tau N}\right)$ with an additional task diversity assumption. Duong et al. (2024) remove such a task diversity assumption by proposing an algorithm that a regret guarantee of $\tilde{O}\big(Nm\sqrt{\tau}+N^{\frac{2}{3}}\tau^{\frac{2}{3}}dm^{\frac{1}{3}}+Nd^2+\tau md\big)$; even then, their algorithm has limited practicality since it requires maintaining a distribution over the set of experts that has size exponential in d. Thus, we ask the following question:

Is it possible to design sequential multi-task linear bandit algorithms that achieve provably low meta-regret while maintaining high computational efficiency?

In this paper, we provide an affirmative answer to this question. Under some the conditions, see Assumption 1, 2, 4, 5, we present an algorithm with a meta-regret of $\tilde{O}\left(Nm\sqrt{\tau} + N^{\frac{2}{3}}\tau^{\frac{2}{3}}d^{\frac{5}{3}} + Nd^3 + \tau d^2\right)$, which improves upon the individual single-task baseline $\tilde{O}(Nd\sqrt{\tau})$ when N is large. We also show that the lower bound regret of the problem is $\Omega\left(Nm\sqrt{\tau} + d\sqrt{m\tau N}\right)$.

Our algorithm, named BARON, reduces the sequential multi-task linear bandit problem to the full-information online subspace selection problem (when restricted to the meta-exploration tasks), for which we leverage the competitive ratio guarantee for the analysis. Specifically, for each new task n, our algorithm selects a subspace, represented by its canonical orthonormal basis $\hat{B}_n$, to approximate the true underlying subspace B and guide the exploration process. One key contribution of this paper is the use of the competitive ratio analysis of FTL to demonstrate the regret guarantee of our approach. We empirically demonstrate the effectiveness of our algorithm in a simulated adversarial environment and demonstrate the efficiency of our algorithm compared to the baselines of Qin et al. (2022) and Duong et al. (2024).

1.1 Related work

1.1.1 Competitive Ratio and Online-PCA

Although there are many works studying the regret analysis of online algorithms, (Sleator & Tarjan, 1985) was the first to suggest comparing an online algorithm to an optimal offline algorithm, which Karlin et al. (1988) later coined as *competitive analysis*. (Buchbinder et al., 2012; Borodin & El-Yaniv, 2005; Albers, 2003) comprehensively study the competitive analysis framework of online algorithms, where the goodness of an algorithm is evaluated relative to a baseline with the performance of the best offline algorithm that has complete knowledge of the future. After much work on the analysis of different settings, (Kakade et al., 2007) studies the competitive ratio for linear optimization problems, both in the full-information setting and the bandit setting. In contrast, our paper focuses on the analysis of FTL in online subspace selection, in which Kakade et al. (2007)'s results cannot be directly applied.

The online subspace selection problem, also known as the online PCA setting, has been studied extensively. Many works focus on reconstruction error at the last step of the online PCA algorithm (Mitliagkas et al., 2013). Different from them, (Nie et al., 2016; Warmuth & Kuzmin, 2008; Garber et al., 2015) study online PCA with the objective of minimizing the cumulative loss. Here, for a stream of N data, the data point $x_n \in \mathbb{R}^d$ ix projected (compressed) to an m-dimensional subspace, represented by the projection matrix P_n (which is a symmetric matrix chosen by the learner such that $P_n \in \mathbb{R}^{d\times d}$ and P_n has m eigenvalues equal 1 and $k = d - m$ eigenvalues equal 0). The goal of *uncentered Online-PCA* is to minimize the cumulative compression loss $\sum_{n=1}^{N} \|x_n - P_n x_n\|^2$. This is different from the *centered Online-PCA* setting, where the algorithm also needs to produce a data center m_{n-1} at step n and incurs the instantaneous compression loss $\ell_n = \|(x_n - m_{n-1}) - P_n(x_n - m_{n-1})\|^2$.

To the best of our knowledge, this paper is the first work that analyzes the competitive ratio, formally defined in Section 2.1, of the FTL algorithm in the online PCA and the subset selection settings.

In Theorem 3, we show that FTL is K-competitive for subset selection, which matches the $\Omega(K)$ competitive ratio lower bound shown by Warmuth & Kuzmin (2008). In Theorem 6, we show that FTL attains the competitive ratio d in the subspace selection online PCA problem under some assumptions (see Assumption 4 and Assumption 5).

Algorithm	Prior info	Additional assumption	Polynomial time	Adv. robustness	Regret guarantee(s)
Lower bound					$\Omega\left(Nm\sqrt{\tau}+d\sqrt{m\tau N}\right)$[3]
Indep. PEGE for each task	**None**	**No**	**Yes**	Bad	$\tilde{O}\left(Nd\sqrt{\tau}\right)$
Qin et al. (2022)	m,τ	Yes	**Yes**	Bad	$\tilde{O}\left(Nm\sqrt{\tau}+dm\sqrt{\tau N}\right)$
Bilaj et al. (2024)	**None**	Yes	**Yes**	N/A	at least $O\left(N\sqrt{\tau}(d-m)\log\left(1+\frac{m}{\lambda^{\mathcal{A}}_{min}(d-m)}\right)\right)$
Duong et al. (2024)	N,m,τ	**No**	No	**Good**	$\tilde{O}\left(Nm\sqrt{\tau}+N^{\frac{2}{3}}\tau^{\frac{2}{3}}dm^{\frac{1}{3}}+Nd^2+\tau md\right)$
BARON (ours)	N,m,τ	Yes	**Yes**	**Good**	$\tilde{O}\left(Nm\sqrt{\tau}+N^{\frac{2}{3}}\tau^{\frac{2}{3}}d^{\frac{5}{3}}+Nd^3+\tau d^2\right)$

Table 1: Comparisons of the settings, assumptions, and regret guarantees in this paper and previous works. The "Additional Assumption" column shows whether a baseline requires any other assumption, besides the low-dimensional structure Assumption 1 and ellipsoid action set Assumption 2 , in their analysis. The "Adv robustness" column indicates whether the algorithm is robust when the environment is designed adversarially.

1.1.2 Sequential Representation Learning for Multi-task Linear Bandit

In this setting, Qin et al. (2022) introduces a task diversity assumption, which posits that any sufficiently large subset of tasks adequately spans the m-dimensional subspace in a well-conditioned manner. Based on this assumption, they derive a meta-regret bound of $\tilde{O}\left(Nm\sqrt{\tau}+dm\sqrt{\tau N}\right)$, which nearly matches the lower bound of $\Omega\left(Nm\sqrt{\tau}+d\sqrt{m\tau N}\right)$ [2]. In this paper, we propose the BAndit Representation transfer by ONline-PCA (BARON) algorithm. Our algorithm has some similarity with Qin et al. (2022) since we both use singular value decomposition (SVD) to estimate the global feature extractor B. The key difference is that they use a deterministic exploration schedule, while BARON explores each task with probability p; thus, Qin et al. (2022) is more susceptible to adversarial choices of task sequences, which we demonstrate in Section 5. Furthermore, Qin et al. (2022)'s guarantee is conditioned on their task diversity assumption, while BARON's guarantee the competitive ratio of FTL under some milder assumptions (see Assumption 7).

Aiming to remove the task diversity assumption, Duong et al. (2024)'s algorithm achieves the upper bound $\tilde{O}\big(Nm\sqrt{\tau}+N^{\frac{2}{3}}\tau^{\frac{2}{3}}dm^{\frac{1}{3}}+Nd^2+\tau md\big)$ by maintaining a distribution over the set of experts that has size exponential in d. At a high level, BARON shares similarities with Duong et al. (2024), where the underlying subspace is explored with a small probability p to facilitate the learning of the global feature extractor B. A key distinction of our approach from Duong et al. (2024) lies in the approach for estimating $\hat{B}_n$: we use a greedy approach where we apply SVD on the data collected so far, whereas their method involves maintaining a probability distribution over an ε-net of potential B matrices.

In a related study, Bilaj et al. (2024) considers a setting where task parameters are independently and identically distributed, exhibiting high variance within an m-dimensional subspace and low variance in the orthogonal directions. From Duong et al. (2024)'s analysis, Bilaj et al. (2024)'s regret guarantee can be as large as $\tilde{O}(Nd\sqrt{\tau})$. For a concise comparison between our work and other relevant studies, please refer to Table 1.

2 Problem Setup

2.1 Competitive analysis of online algorithms

We consider the problem where the learner is faced with an N-round online decision problem. The learner is given a decision space $\mathcal{A}$. In each round n, the learner chooses $w_n \in \mathcal{A}$ and receives a loss $\ell_n : \mathcal{A} \to \mathbb{R}$. The learner suffers a loss of $\ell_n(a_n)$ in round n, and its cumulative loss is $L_N = \sum_{n=1}^{N} \ell_n(a_n)$. We also define the cumulative loss of the

[2] Qin et al. (2022) provides this lower bound without a proof. We provide the proof in Theorem 9.

[3] See footnote 2

benchmark as $L_N^* = \min_{a\in\mathcal{A}} \sum_{n=1}^N \ell_n(a)$. The goal of the learner is to compete with the best fixed action in hindsight (i.e., the benchmark) such that

$$\sum_{n=1}^N \ell_n(a_n) \le a \cdot \min_{a\in\mathcal{A}} \sum_{n=1}^N \ell_n(a) + b,$$

for some constants $a, b > 0$. The follow the leader (FTL) algorithm, in this setting, can be formally stated as

$$a_n = \arg\min_{a\in\mathcal{A}} \sum_{i=1}^{n-1} \ell_i(a).$$

In this paper, we focus on the following settings of decision set $\mathcal{A}$ and loss functions ℓ_n:

Online subset selection. The online subset selection problem, first studied in (Warmuth & Kuzmin, 2008), generalizes the online expert selection problem (Freund & Schapire, 1997). In each round n, the learner chooses A_n, a subset of m experts from $[K]$, and observes a loss vector $(\ell_n^k)_{k=1}^K$; its loss in round n is the total loss over the chosen subset A_n. When $m = 1$, this is the classical online expert selection problem. It can be cast in the above framework by letting $\mathcal{A} = \binom{[K]}{m}$, i.e., all size-m subsets of $[K]$, and $\ell_n(a) = \sum_{k\in a} \ell_n^k$.

Online subspace selection. The online subspace selection, first studied in (Warmuth & Kuzmin, 2008), lifts the above online subset selection problem to online decision making with projection matrices. In each round, the learner chooses a rank-k projection matrix P_n, and sees a loss matrix X_n; its loss at round n is the linear loss $\langle P_n, X_n\rangle$, where the inner product here is over $d \times d$ matrices. The online subspace selection problem is closely related to online PCA, in which case $X_n = x_n x_n^\top$ is the dyad induced by the n-th sample x_n, and $I - P_n$ is the projection matrix the learner uses to compress the data into a smaller subspace. It can be cast in the above framework by letting $\mathcal{A} = \{\text{rank-}k \text{ projection matrices in } \mathbb{R}^d\}$, and $\ell_n(a) = \langle a, X_n\rangle$.

2.2 Sequential Representation Learning for Multi-task Linear Bandit

Consider a sequential learning problem involving N[4] distinct d-dimensional linear bandit tasks, denoted by $\theta_1, \cdots, \theta_N \in \mathbb{R}^d$, each interacted with a horizon of τ steps. A key characteristic of these tasks is that they reside within an m-dimensional linear subspace of $\mathbb{R}^d$ (as detailed in Assumption 1). Specifically, for each task n and at each time t, the algorithm selects an action a_t^n from a predefined action set $\mathcal{A}$ (which adheres to Assumption 2) and observes a reward $r_t^n = \langle a_t^n, \theta_n\rangle + \eta_t^n$, where η_t^n represents independent, zero-mean noise that is 1-sub-Gaussian. Upon completion of task n, the process is repeated for the subsequent task.

The learning agent's objective is to minimize the total (pseudo) regret accumulated across all tasks. The expected pseudo-regret for a single task n over its horizon is defined as $R_\tau^n := \tau \cdot \max_{a\in\mathcal{A}} \langle a, \theta_n\rangle - \mathbb{E}\left[\sum_{t=1}^\tau \langle a_t^n, \theta_n\rangle\right]$. Consequently, the cumulative meta-regret across all N tasks is given by:

$$R_\tau := \sum_{n=1}^N R_\tau^n = \sum_{n=1}^N \left(\tau \cdot \max_{a\in\mathcal{A}} \langle a, \theta_n\rangle - \mathbb{E}\left[\sum_{t=1}^\tau \langle a_t^n, \theta_n\rangle\right]\right). \tag{1}$$

Assumption 1 (Low-Dimensional Structure of Task Parameters)**.** *Suppose $m < d$. For the set of task parameters $\theta_1, \cdots, \theta_N$, there exists: (i) a global feature extractor $B \in \mathbb{R}^{d\times m}$ with orthonormal columns, and (ii) a corresponding set of task vectors $w_1, \ldots, w_N \in \mathbb{R}^m$, such that for every task $i \in [N]$, the parameter θ_i can be expressed as $\theta_i = Bw_i$.*

In line with prior work (Rusmevichientong & Tsitsiklis, 2010; Duong et al., 2024), we also impose structure on the action space $\mathcal{A}$, specifically assuming it to be an ellipsoid, and a bound on the ℓ_2 norms of the task parameters.

Assumption 2 (Linear Bandits with Ellipsoid Action Set)**.** *The feasible action set is given by the ellipsoid $\mathcal{A} := \{x \in \mathbb{R}^d : x^\top M^{-1} x \le 1\}$, where M is a symmetric and positive definite matrix. Additionally, we assume the existence of constants $\theta_{\min}$ and $\theta_{\max} \le 1$ such that for all tasks $n \in [N]$, the ℓ_2 norm of the task parameter is bounded as $\theta_{\min} \le \|\theta_n\|_2 \le \theta_{\max}$.*

[4]We abuse the notation N to mean the number of rounds in the Online subspace selection problem and the number of tasks in the sequential multi-task linear bandit problem. This was done to ensure consistency and make it easy to follow when applying the competitive ratio result in the sequential multi-task linear bandit's analysis.

3 THE COMPETITIVE RATIO OF THE FOLLOW-THE-LEADER ALGORITHM

3.1 ONLINE SUBSET SELECTION

Theorem 3. *In the online subset selection problem, where the learner chooses a set of size s at each step, the cumulative loss of FTL is:*

$$L_{N,Alg} \leq KL_N^* + Km.$$

Recall that $L_N^* = \min_{a\in\mathcal{A}} \sum_{n=1}^N \ell_n(a)$. In prior work, (Freund & Schapire, 1997) propose the Hedge algorithm with the guarantee $L_{N,Alg} \leq \frac{-\log(1-\gamma)}{\gamma} L_N^* + \frac{\ln K}{\gamma}$, for $\gamma \in (0,1)$. When $\gamma = 1$, the Hedge algorithm turns into FTL, and the guarantee becomes vacuous. In contrast, our analysis shows that FTL in fact has a non-vacuous guarantee on its competitive ratio. To the best of our knowledge, this is a novel result. Our proof of Theorem 3 uses induction with the potential function $\Phi_n = \sum_{i=1}^K \phi_{n,i}$, where $\phi_{n,i} = \min\left(L_{n,i}, L_{n,i_n^*} + 1\right)$, to prove the inductive hypothesis $\ell_{n,i_{n-1}^*} \leq \Phi_n - \Phi_{n-1}$. In the interest of space, we defer the proof of Theorem 3 to the Appendix A.

3.2 ONLINE SUBSPACE SELECTION

Here, we will assume that the data $\{X_n\}_{n=1:N}$ has diverse covariance as stated in Assumption 4. We also assume that the instantaneous loss under the concentration event, defined in Assumption 4, is bounded in Assumption 5.

Define $\alpha^i := \frac{d\lambda_{d+1-i}(C_N)+d}{N}$ and let $k := d - m$ be the number of noise direction, where $\lambda_{d+1-i}(C_N)$ is small. Define the covariance matrix $C_n := \sum_{j=1}^n x_j x_j^\top$. Let $\lambda_i(C)$ and $v_i(C)$ be the i-th largest eigenvalues of C and its corresponding eigenvector. In Assumption 4, the Concentration event $D_n^{\alpha^i,i}$ happens when the data x_n lies in the subspace defined by the eigenvectors of C_{n-1} with large eigenvalues. The Diversity event happens when the gap between the largest and smallest eigenvalues of C_{n-1} is less than one. The global bad event E^i happens when the local bad event $G_n^{i,c}$ happens too many times.

Assumption 4 (Covariance Diversity)**.** *For all $i \in [k]$, define*

$$D_n^{\alpha^i,i} := \left\{\lambda_j(C_n) - \lambda_j(C_{n-1}) < \alpha^i \;\forall\, j \in [2, d+1-i] \text{ s.t. } \lambda_j(C_{n-1}) = \lambda_{d+1-i}(C_{n-1})\right\}; \quad \text{(Concentration event)}$$

$$F_n^i := \{\lambda_1(C_{n-1}) < \lambda_{d+1-i}(C_{n-1}) + 1 \text{ AND } \lambda_1(C_n) < \lambda_{d+1-i}(C_n) + 1\}; \quad \text{(Diversity event)}$$

$$G_n^i := D_n^{\alpha^i,i} \cup F_n^i;$$

$$E^i := \left\{\sum_{n=1}^N \mathbb{I}(G_n^{i,c}) \geq d\lambda_{d+1-i}(C_N) + d\right\}.$$

Then, for all $i \in [k]$,

$$\mathbb{I}(E^i) Tr(C_N) \leq d\lambda_{d+1-i}(C_N) + d.$$

Assumption 5 (Bounded instantaneous loss under Concentration event)**.** *Define $\ell_n^i(a) := \left\langle P_n^i a P_n^{i,\top}, X_n \right\rangle$, where P_n^i is the projection matrix mapping to $v_{d+1-i}(C_{n-1})$. Then, there exists a constant c such that, for all $i \in [k]$, under event $D_n^{\alpha^i,i}$, the instantaneous loss is bounded,*

$$\mathbb{I}(D_n^{\alpha^i,i})\ell_n^i \leq c\,\alpha^i.$$

Note: In the special setting where $k = 1$ (i.e., dyad selection), then $i = 1$. Hence, when the context is clear, we omit the superscript i.

Remark: This assumption is mild because

- Conditioned on the sequence of data $\{x_n\}_{n=1:N}$, α can be very large, even larger than one, thus making this assumption redundant.

- The assumption is also redundant when $\mathbb{I}(D_n^\alpha) = 0$.
- In the special case where $\alpha = 0$ or when, for any $n \in [N]$, C_{n-1} and C_n have the same eigensystem, this assumption redundant. This is further elaborated in Section D.
- We also show that removing Assumption 4 is non-trivial by providing an example in Appendix E.

We want to highlight that our proof for the competitive ratio in the Online subset selection is one of our key innovations in this paper. Another key contribution is Assumption 4 and Assumption 5, where we show why it's needed. Removing these assumptions is a non-trivial task that we leave for future work. These assumptions are necessary to show FTL's competitive ratio in Theorem 6.

Theorem 6. *The FTL's performance in the online subspace selection problem, under Assumption 4 and Assumption 5 , satisfies*

$$L_{N,Alg} \leq O\left(dL_N^* + d^2\right).$$

Previously, Warmuth & Kuzmin (2008)'s uncentered online PCA algorithm provided a guarantee of $L_{N,Alg} \leq \frac{-\log(1-\gamma)}{\gamma} L_N^* + \frac{\ln(K/s)}{\gamma}$. With $\gamma = 1$, their algorithm becomes FTL, making their guarantee vacuous. To the best of our knowledge, we are the first to provide a non-vacuous competitive ratio guarantee for the FTL algorithm in online subspace selection.

4 Sequential Representation Learning for Multi-task Linear Bandit

4.1 Algorithm

As stated in Duong et al. (2024), without Qin et al. (2022)'s Task Diversity assumption, it's non-trivial for the learner to efficiently estimate the global feature extractor B. Instead, the learner must infer B based on the knowledge acquired from previously encountered situations while dealing with uncertainty for future tasks.

High-level Overview: Our approach tackles the sequential multi-task bandit problem by employing a bi-level strategy for simultaneously learning and leveraging an estimated B.

- At the upper level, for each task n, the learner faces two key decisions: (1) whether to engage in meta-exploration or meta-exploitation, and (2) if choosing meta-exploitation, which subspace $\hat{B}_n$ to employ to estimate θ_n^* sample-efficiently. To address these decisions, we introduce Algorithm 1, designed to make these choices in a feedback-driven manner to ensure minimal meta-regret.
- At the lower level, for each task n, the learner executes meta-exploration and meta-exploitation depending on the decision made at the upper level. The first option, meta-exploration, performs exploration without relying on prior knowledge of B, using a variant of a full-dimensional linear bandit algorithm (specifically, PEGE as described by Rusmevichientong & Tsitsiklis (2010), detailed in Algorithm 2). The second option, meta-exploitation, involves incorporating a learned subspace $\hat{B}$ as prior information, aiming for a more sample-efficient estimation of θ_n^* that results in reduced regret (Algorithm 3). Since Algorithm 2 and 3 are proposed in previous work (Duong et al., 2024, e.g.), we only give a brief recap of them in Appendix C.1 for completeness.

A more detailed description of the full algorithm is in Appendix C.1. To use the subspace selection's competitive ratio guarantee for regret analysis in Theorem 6, we make the following assumption:

Assumption 7. *Define* $F_n := \left\{Z_n = 0 \vee (Z_n = 1 \wedge \|\hat{\theta}_n - \theta_n\| \leq \tilde{O}\left(\frac{d}{\sqrt{\tau_1}}\right))\right\}$*, and* $F := \cap_{n=1}^N F_n$*. Under event* F*, the sequence of* $\left\{x_n \mid x_n = Z_n\hat{\theta}_n,\ \forall n \in [N]\right\}$*, satisfies Assumption 4 and Assumption 5.*

Assumption 7 assumes that the sequence of estimations $\hat{\theta}_n$ of the meta-exploration tasks satisfies Assumption 4 and Assumption 5, thus enabling us to use the competitive ratio guarantee. Here, since θ_n lies in the same subspace as per Assumption 1, the Concentration event D_n^{i,α^i} would happen for most tasks, making Assumption 4 mild, especially in comparison to Qin et al. (2022)'s Task Diversity assumption. For the same reason, the instantaneous loss under the Concentration even D_n^{i,α^i} can't be too large, thus making Assumption 5 mild.

Algorithm 1 BARON: BAndit Representation transfer by ONline-PCA

Input: Task length τ, number of task N, task dimension d, subspace dimension m, and exploration rate p.
Initialize: A dataset $D_1 = \{\emptyset\}$ over the estimated task parameters
for $n \in [N]$: **do**
    Estimate $\hat{B}_n = \text{SVD}(D_n)$
    With probability p: $Z_n = 1$, otherwise $Z_n = 0$
    if $Z_n = 1$ **then**
        Perform meta-exploration using Algorithm 2 and receive $\hat{\theta}_n$
        Update the dataset $D_{n+1} = D_n \cup \hat{\theta}_n$
    else
        Perform meta-exploitation using Algorithm 3 with $\hat{B}_n$
    end if
end for

Theorem 8. *Let* $p = \min\left[\left(\frac{\tau d}{N^2}\right)^{\frac{1}{3}}, 1\right]$, *the meta-exploration's exploration duration* $\tau_1 = d \cdot \left\lfloor \min\left(d\sqrt{\frac{d\tau}{p}}, \tau\right)/d \right\rfloor$, *and the meta-exploitation's exploration duration* $\tau_2 = m\sqrt{\tau}$. *Under Assumption 1, 2, and 7, the meta-regret of Algorithm 1 satisfies*

$$R_\tau \leq \tilde{O}\left(Nm\sqrt{\tau} + N^{\frac{2}{3}}\tau^{\frac{2}{3}}d^{\frac{5}{3}} + Nd^3 + \tau d^2\right). \tag{2}$$

The proof of Theorem 8 is in Appendix C.2. In particular, we use Theorem 6 in Lemma 16 to show that

$$\sum_{n=1}^{N} Z_n \|\hat{B}_{n,\perp}^\top \hat{\theta}_n\|_2^2 \leq d \cdot \min_B \sum_{n=1}^{N} Z_n \|B_\perp^\top \hat{\theta}_n\|_2^2 + d^2. \tag{3}$$

Using equation 3, we have an upper bound on the subspace estimation error, which can be used to bound the regret of all meta-exploitation tasks, leading to the meta regret upper bound in equation 2. Here, we see that our regret bound is slightly worse than Duong et al. (2024), both in terms of the first two dominant terms and the "burn-in" initial learning costs. Compared with Duong et al. (2024), we think this could be a reasonable exchange for a practical and efficient algorithm.

Comparison Against Individual Single-Task baseline. The meta-regret of learning each task independently is $O(Nd\sqrt{\tau})$. Our derived meta-regret guarantee offers an improvement over this baseline under the conditions that $\tau \gg d^3$ and $N \gg d^2\sqrt{\tau}$. Expanding the parameter ranges where our guarantee outperforms the individual single-task baseline remains an important area for future research.

4.2 Lower bound

In this section, we provide a lower bound for the Sequential Representation Learning for Multi-task Linear Bandit in Theorem 9. This lower bound was given in (Qin et al., 2022), but there was no proof. In this work, for completeness, we provide a proof in Theorem 9.

Theorem 9. *The lower bound for the Sequential Representation Learning for Multi-task Linear Bandit is:*

$$R_\tau \geq \tilde{\Omega}\left(Nm\sqrt{\tau} + d\sqrt{m\tau N}\right).$$

The proof of Theorem 9 is in Appendix C.3. When N is large, our regret upper bound in equation 2 matches this lower bound.

5 EXPERIMENTS

In this section[5], we detail a comparative analysis of our BARON algorithm's performance against the baselines within simulated environments. The algorithms under evaluation are:

- SeqRepL: our implementation of the method presented in (Qin et al., 2022), where the estimate $\hat{B}_n$ is derived via SVD, and tasks for meta-exploration are selected according to the deterministic sequence $n = \frac{i(i+1)}{2}$ for $i = 1, 2, \cdots$.
- BOSS: A computationally efficient approximation of Duong et al. (2024)'s algorithm, utilizing a set of 100,000 experts sampled uniformly at random.
- BOSS-semi-oracle: A computationally efficient approximation of Duong et al. (2024)'s algorithm using oracle information to demonstrate its ideal performance. It has an expert set of 100,000 experts drawn uniformly at random and includes the true subspace B added in.

The experimental setup involves parameters $(N, \tau, d, m) = (4000, 500, 10, 3)$. In this experiment, the action space $\mathcal{A}$ is a unit sphere, i.e., $M = I_d$. For each task n, let $B_n \in \mathbb{R}^{d \times m_n}$ represent the basis of the subspace used by the environment to generate θ_n, with m_n increasing to $1, 2, 3$ at $n = 1, 501, 1001$, respectively. The parameter θ_n is generated as $\theta_n = \lambda_n B_n w_n$, where w_n is sampled uniformly from the unit sphere $\mathbb{S}^{m_n - 1}$ and λ_n is a random scaling factor from the interval $[0.8, 1]$, consistent with Assumption 2 ($\theta_{\min} = 0.8, \theta_{\max} = 1$). The adversary hides all new subspace dimensions according to Qin et al. (2022)'s exploration schedule: when $n = \frac{i(i+1)}{2}$, $\theta_n = \lambda_n B_1 w_n$.

The hyper-parameters are chosen such that $(p, \tau_1, \tau_2) = (0.15, 1000, 300)$. The shaded regions in the figures depict ± 1 standard deviation across 5 independent trials.

Figure 1a illustrates the linear relationship between cumulative regret and the number of tasks N. Notably, BARON outperforms both the BOSS and SeqRepL baselines. The observed performance difference between BOSS-semi-oracle and BOSS arises because the expert set of BOSS employed in this experiment does not fully encompass the true B (the theoretical size of the expert set $\approx 11^{30}$), significantly exceeding the size used in BOSS.

Figure 1b displays $\|\hat{B}_{n,\perp}^\top B_n\|_{\mathrm{F}}$, a metric quantifying the proximity of $\hat{B}_{n,\perp}$ to B_n. Following the environment's introduction of a new subspace dimension at tasks 1, 501, and 1001, the estimated subspace $\hat{B}_n$ for all algorithms undergoes updates and converges after some time. Once again, BARON's superior regret performance can be attributed to a better estimation of both $\hat{\theta}_n$ as demonstrated in Figure 1c and $\hat{B}_n$ as demonstrated in Figure 1b.

In Figure 2, we compare our performance with the baseline algorithms on a benign environment where θ_n is sampled uniform at random from $[-1, 1]^d$, for all $n \in [N]$, such that it satisfies Assumption 1 and Assumption 2. Thus, the subspace B can be estimated efficiently even when only a small number of tasks are seen. While BARON is still competitive with the best algorithms, Duong et al. (2024) showed its limitation when using a computationally efficient approximation, even when dealing with a non-adversarial setting.

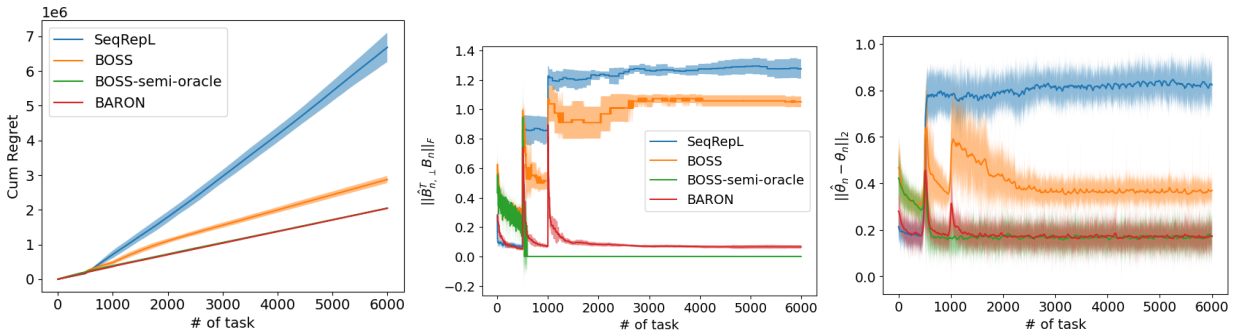
While performing the experiment, we also noticed that Duong et al. (2024) is very sensitive to the hyperparameter choices (p, τ_1, α), while BARON and Qin et al. (2022) have much more robust performance under different hyperparameter choices.

6 DISCUSSION AND FUTURE WORK

In this paper, we show that the Competitive Ratio of FTL in the online subset selection problem is K. For the Online subspace selection (Online PCA) problem, we show that FTL is d competitive under some assumptions.

We demonstrate one application of the Competitive Ratio in the Sequential Representation Transfer in Multi-task Linear Bandit problem. Our BARON algorithm achieves the regret guarantee of $\tilde{O}\left(Nm\sqrt{\tau} + N^{\frac{2}{3}}\tau^{\frac{2}{3}}d^{\frac{5}{3}} + Nd^3 + \tau d^2\right)$ under some assumptions. We also show that the lower bound of this problem is $\Omega\left(Nm\sqrt{\tau} + d\sqrt{m\tau N}\right)$.

[5]The codebase associated with this work is available at https://anonymous.4open.science/r/BOSS-7774/README.md. We build our code on top of Duong et al. (2024).



(a) (b) (c)

Figure 1: Cumulative regret comparison between BARON and other baselines. BARON is both computationally efficient and approaches the performance of Duong et al. (2024)'s BOSS-semi-oracle using oracle information, which is much better over Duong et al. (2024)'s BOSS, a computationally efficient approximation.

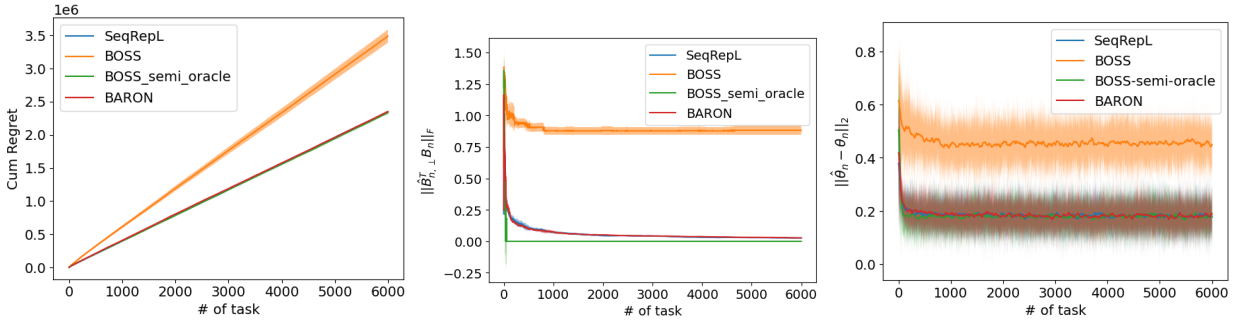


Figure 2: Cumulative regret comparison between BARON and other baselines in a benign setting where the subspace B can be estimated efficiently. BARON performance is still better than BOSS, the computationally efficient approximation of Duong et al. (2024), and competitive with other baselines. The environments are sampled from the subspace uniformly.

In the future, we would like to remove Assumption 4 and Assumption 5 for the competitive ratio analysis, which would lead to removing Assumption 7. We would also want to close the gap with the lower bound and remove the ellipsoid action set Assumption 2 in the Sequential Representation Transfer in Multi-task Linear Bandit problem.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Susanne Albers. Online algorithms: a survey. *Mathematical Programming*, 97:3–26, 2003.

Steven Bilaj, Sofien Dhouib, and Setareh Maghsudi. Meta learning in bandits within shared affine subspaces, 2024.

Allan Borodin and Ran El-Yaniv. *Online computation and competitive analysis*. cambridge university press, 2005.

Niv Buchbinder, Shahar Chen, Joshep Seffi Naor, and Ohad Shamir. Unified algorithms for online learning and competitive analysis. In *Conference on Learning Theory*, pp. 5–1. JMLR Workshop and Conference Proceedings, 2012.

Leonardo Cella, Karim Lounici, Grégoire Pacreau, and Massimiliano Pontil. Multi-task representation learning with stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4822–4847. PMLR, 2023.

Thang Duong, Zhi Wang, and Chicheng Zhang. Beyond task diversity: provable representation transfer for sequential multitask linear bandits. *Advances in Neural Information Processing Systems*, 37:37791–37822, 2024.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Dan Garber, Elad Hazan, and Tengyu Ma. Online learning of eigenvectors. In *International Conference on Machine Learning*, pp. 560–568. PMLR, 2015.

Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, and Liwei Wang. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pp. 4349–4358. PMLR, 2021.

Sham M Kakade, Adam Tauman Kalai, and Katrina Ligett. Playing games with approximation algorithms. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 546–555, 2007.

Anna R Karlin, Mark S Manasse, Larry Rudolph, and Daniel D Sleator. Competitive snoopy caching. *Algorithmica*, 3: 79–119, 1988.

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

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming pca. *Advances in neural information processing systems*, 26, 2013.

Jiazhong Nie, Wojciech Kotlowski, and Manfred K Warmuth. Online pca with optimal regret. *Journal of Machine Learning Research*, 17(173):1–49, 2016.

Yuzhen Qin, Tommaso Menara, Samet Oymak, ShiNung Ching, and Fabio Pasqualetti. Non-stationary representation learning in sequential linear bandits. *IEEE Open Journal of Control Systems*, 1:41–56, 2022.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Daniel D Sleator and Robert E Tarjan. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28(2):202–208, 1985.

Manfred K Warmuth and Dima Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9(Oct):2287–2320, 2008.

Jiaqi Yang, Wei Hu, Jason D Lee, and Simon Shaolei Du. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2020.

Jiaqi Yang, Qi Lei, Jason D Lee, and Simon S Du. Nearly minimax algorithms for linear bandits with shared representation. *arXiv preprint arXiv:2203.15664*, 2022.

A ONLINE SUBSET SELECTION

We first analyze the Online expert selection problem, a special case of the Online subset selection where the subset size is one.

Lemma 10. *In the Online expert selection problem, We define the potential at time step n, Φ_n, as: $\Phi_n = \sum_{i=1}^{K} \phi_{n,i}$, where $\phi_{n,i} := \min(L_{n,i}, L_n^* + 1)$ and $L_{n,i} = \sum_{j=1}^{n} \ell_j(i)$. We also define $L_n^* := \min_{j\in[K]} L_{n,j}$ as the best cumulative loss at time step n.*

Then, for every step n, we have:

$$\ell_{n,i_{n-1}^*} \leq \Phi_n - \Phi_{n-1}.$$

Proof. We look at the right-hand side, which is

$$\Phi_n - \Phi_{n-1} = \sum_{i=1}^{K} (\phi_{n,i} - \phi_{n-1,i}) =: \sum_{i=1}^{K} \Delta\phi_{n,i}$$

For any fixed expert i, its potential $\phi_{n,i}$ is monotonically increasing in n. As a result, every $\Delta\phi_{n,i}$ in the sum above is nonnegative. As a consequence, The above is at least the growth of the potential of the expert i^*_{n-1},

$$\sum_{i=1}^{K} \Delta\phi_{n,i} \geq \Delta\phi_{n,i^*_{n-1}} = \phi_{n,i^*_{n-1}} - \phi_{n-1,i^*_{n-1}}$$

A close examination of the above two terms reveals the following:

- For $\phi_{n-1,i^*_{n-1}}$, it equals $\min\left(L_{n-1,i^*_{n-1}}, L^*_{n-1}+1\right)$, which is exactly $L_{n-1,i^*_{n-1}} = L^*_{n-1}$.
- For $\phi_{n,i^*_{n-1}}$, it equals $\min\left(L_{n,i^*_{n-1}}, L^*_n+1\right)$. The key observation is that the first term in the minimum is always active, making the expression evaluate to $L_{n,i^*_{n-1}}$. The reason is as follows:

$$L_{n,i^*_{n-1}} \leq L_{n-1,i^*_{n-1}} + 1 = \min_{i\in[K]} L_{n-1,i} + 1 \leq \min_{i\in[K]} L_{n,i} + 1 = L^*_n + 1.$$

The last inequality is true because $L_{n-1,i^*_{n-1}} \leq L_{n-1,i^*_n} \leq L_{n,i^*_n}$.

In summary, we just found that

$$\Delta\phi_{n,i^*_{n-1}} = L_{n,i^*_{n-1}} - L_{n-1,i^*_{n-1}} = \ell_{n,i^*_{n-1}}.$$

This concludes the proof of the lemma. □

Lemma 11. *In the Online expert selection problem, the cumulative loss of FTL is:*

$$L_{N,Alg} \leq KL^*_N + K.$$

Proof. From Lemma 10, for every n,

$$\ell_{n,i^*_{n-1}} \leq \Phi_n - \Phi_{n-1}.$$

Summing over all $n \in [N]$, we have

$$L_{N,Alg} = \sum_{n=1}^{N} \ell_{n,i^*_{n-1}} \leq \Phi_N = \sum_{i=1}^{K} \min(L_{N,i}, L^*_N + 1) \leq K(L^*_N + 1).$$

□

We next consider the online subset selection problem with $s > 1$. To this end, we consider the Online m-th best expert selection problem. Denotes $\min_i^m X_i$ as the m-th smallest value of X_i. Then, at each round, FTL chooses i_n to be $i^*_{n-1} := i^*_{n-1}(m) := \arg\min^m_{i\in[K]} L_{n-1,i}$.

Then, for any $m \in [s]$, we aim to use Lemma 11 to show

$$\sum_{n=1}^{N} l_{n,i^*_{n-1}} \leq KL_{N,i^*_N} + K$$

Define:

$$\phi_{n,i} = \phi_{n,i}(m) = \min\left(L_{n,i}(m), L_{n,i^*_n}(m) + 1\right)$$

$$\Phi_n = \Phi^m_n = \sum_{i=1}^{K} \phi_{n,i}(m),$$

$$\Phi_n - \Phi_{n-1} = \Phi_n(m) - \Phi_{n-1}(m) = \sum_{i=1}^{K} \left(\phi_{n,i}(m) - \phi_{n-1,i}(m)\right) =: \sum_{i=1}^{K} \Delta\phi_{n,i}(m) =: \sum_{i=1}^{K} \Delta\phi_{n,i}.$$

We will hide the m in the following proof when the context is clear.

Lemma 12. *For every step $n \in [N]$, we have*

$$\sum_{m=1}^{s} \ell_{n,i^*_{n-1}(m)} \leq \sum_{m=1}^{s} \left(\Phi_n^m - \Phi_{n-1}^m\right).$$

Proof.

$$\Phi_n - \Phi_{n-1} = \sum_{i=1}^{K} \Delta\phi_{n,i} \geq \Delta\phi_{n,i^*_{n-1}} = \phi_{n,i^*_{n-1}} - \phi_{n-1,i^*_{n-1}}$$

We always have: $\phi_{n-1,i^*_{n-1}} = \min\left(L_{n-1,i^*_{n-1}}, L^*_{n-1} + 1\right) = L_{n-1,i^*_{n-1}}$

We also have: $\phi_{n,i^*_{n-1}} = \min\left(L_{n,i^*_{n-1}}, L^*_n + 1\right)$

- If $L_{n-1,i^*_{n-1}} \leq L_{n-1,i^*_n}$, then $\phi_{n,i^*_{n-1}} = \min\left(L_{n,i^*_{n-1}}, L^*_n + 1\right) = L_{n,i^*_{n-1}}$ similar to the previous proof in Lemma 10. Thus, the proof is the same, and we can conclude $\Phi_n - \Phi_{n-1} \geq \ell_{n,i^*_{n-1}}$.
- Otherwise (i^*_n is rank $< m$ at round $n-1$):
  - If $L_{n,i^*_{s-1}} \leq L_{n,i^*_n} + 1$. Then, $\phi_{n,i^*_{n-1}} = \min\left(L_{n,i^*_{n-1}}, L^*_n + 1\right) = L_{n,i^*_{n-1}}$. Thus, the proof is the same as before.
  - $$L_{n,i^*_{n-1}} > L_{n,i^*_n} + 1. \tag{4}$$

    Since i^*_n is rank $< m$ (arm i-th has small rank means $L_{n,i}$ is small) at round $n-1$ and m-th at round n, and i^*_{n-1} is rank m-th at round $n-1$ and $> m$ at round n, there must exists an arm z such that:

    $$\begin{aligned} L_{n-1,z} &> L_{n-1,i^*_{n-1}} \\ L_{n,z} &\leq L_{n,i^*_n}. \end{aligned}$$

    Hence:

    $$\begin{aligned} L_{n,z} &\leq L_{n,i^*_n} \\ &< L_{n,i^*_{n-1}} - 1 && \text{(Eq. (4))} \\ \implies L_{n-1,i^*_{n-1}} + \ell_{n,z} &< L_{n,i^*_{n-1}} - 1 \\ \implies \ell_{n,z} &< \ell_{n,i^*_{n-1}} - 1 \\ \implies 1 &< \ell_{n,i^*_{n-1}} \end{aligned}$$

    Since this is impossible, we conclude that this case never happens.

The proof is finished by summing $\Phi_n^m - \Phi_{n-1}^m \geq \ell_{n,i^*_{n-1}(m)}$ over $m \in [s]$. □

Theorem 3. *In the online subset selection problem, where the learner chooses a set of size s at each step, the cumulative loss of FTL is:*

$$L_{N,Alg} \leq KL^*_N + Km.$$

Proof. From Lemma 12, for every n,

$$\sum_{m=1}^{s} \ell_{n,i^*_{n-1}(m)} \leq \sum_{m=1}^{s} \left(\Phi_n^m - \Phi_{n-1}^m\right).$$

Summing over all $n \in [N]$, we have

$$\begin{aligned}
L_{N,Alg} &= \sum_{n=1}^{N} \sum_{m=1}^{s} \ell_{n,i^*_{n-1}(m)} \\
&\leq \sum_{n=1}^{N} \sum_{m=1}^{s} \left(\Phi_n^m - \Phi_{n-1}^m\right) \\
&\leq \sum_{m=1}^{s} \Phi_N^m \\
&= \sum_{m=1}^{s} \sum_{i=1}^{K} \min(L_{N,i(m)}, L_N^*(m) + 1) \\
&\leq K \sum_{m=1}^{s} L_N^*(m) + Ks \\
&\leq K L_N^* + Ks
\end{aligned}$$

□

B ONLINE SUBSPACE SELECTION

We first analyze the Online dyad selection problem, a special case of the Online subspace selection where the subspace is the outer product of a rank-one unit vector: $B = vv^\top$ ($k = 1$).

Lemma 13. *The FTL's performance in the Online dyad selection problem, under Assumption 4 and Assumption 5, satisfies:*

$$L_{N,Alg} \leq O\left(dL_N^* + d\right)$$

Proof. To prove Lemma 13, we want to show that:

$$\begin{aligned}
O\left(dL_N^* + d\right) &\geq L_{N,Alg} \\
&= \sum_{n=1}^{N} \left\langle v^*_{n-1}(v^*_{n-1})^\top, X_n \right\rangle \qquad (v^*_{n-1} := \arg\min_{v \in \mathbb{S}^{d-1}} L_{n-1,v})
\end{aligned}$$

To do so, we will show that:

- $\sum_{n=1}^{N} \mathbb{I}(G_n) \left\langle v^*_{n-1}(v^*_{n-1})^\top, X_n \right\rangle \leq O\left(dL_N^* + d\right)$
- $\sum_{n=1}^{N} \mathbb{I}(G_n^c) \left\langle v^*_{n-1}(v^*_{n-1})^\top, X_n \right\rangle \leq O\left(dL_N^* + d\right)$

We will start with the second case. Under Assumption 4, we have:

$$\begin{aligned}
\sum_{n=1}^{N} \mathbb{I}(G_n^c) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle &= \mathbb{I}(E) \sum_{n=1}^{N} \mathbb{I}(G_n^c) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle + \mathbb{I}(E^c) \sum_{n=1}^{N} \mathbb{I}(G_n^c) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \\
&\leq \mathbb{I}(E) L_{N,Alg} + \mathbb{I}(E^c) \sum_{n=1}^{N} \mathbb{I}(G_n^c) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \\
&\leq \mathbb{I}(E) Tr(C_N) + \mathbb{I}(E^c) \sum_{n=1}^{N} \mathbb{I}(G_n^c) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \qquad (L_{N,Alg} \leq Tr(C_N)) \\
&\leq d\lambda_d(C_N) + d + \mathbb{I}(E^c) \sum_{n=1}^{N} \mathbb{I}(G_n^c) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \qquad \text{(Assumption 4)} \\
&\leq d\lambda_d(C_N) + d + \mathbb{I}(E^c) \sum_{n=1}^{N} \mathbb{I}(G_n^c) \qquad (\left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \leq 1) \\
&\leq O\left(d\lambda_d(C_N) + d\right). \qquad \text{(By Def } E^c\text{)}
\end{aligned}$$

For the first case, we have:

$$\mathbb{I}(G_n) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \leq \mathbb{I}(D_n^\alpha) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle + \mathbb{I}(F_n) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle$$

By Assumption 5 , we also have: $\mathbb{I}(D_n^\alpha) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \leq c\alpha$

Next, we want to show that $\sum_{n=1}^{N} \mathbb{I}(F_n) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \leq O\left(d\lambda_d(C_N) + d\right)$.

Define $\Phi(C) = \sum_{i=1}^{d} \min\left(\lambda_i(C), \lambda_d(C) + 1\right)$.

We have:

$$\begin{aligned}
\mathbb{I}(F_n)\left[\Phi(C_n) - \Phi(C_{n-1})\right] &= \mathbb{I}(F_n)\left[\sum_{i=1}^{d} \min\left(\lambda_i(C_n), \lambda_d(C_n) + 1\right) - \sum_{i=1}^{d} \min\left(\lambda_i(C_{n-1}), \lambda_d(C_{n-1}) + 1\right)\right] \\
&= \mathbb{I}(F_n)\left[Tr(C_n) - Tr(C_{n-1})\right] \\
&\qquad \text{(By Def } F_n = \{\lambda_1(C_{n-1}) < \lambda_d(C_{n-1}) + 1 \text{ AND } \lambda_1(C_n) < \lambda_d(C_n) + 1\}\text{)} \\
&= \mathbb{I}(F_n)\left[Tr(C_n - C_{n-1})\right] \\
&= \mathbb{I}(F_n)\left[Tr(X_n)\right] \\
&= \mathbb{I}(F_n)\|x_n\|^2 \\
&\geq \mathbb{I}(F_n)\left[(v_{n-1}^*)^\top x_n\right]^2 \qquad \text{(Cauchy-Schwarz)} \\
&= \mathbb{I}(F_n)\left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle
\end{aligned}$$

Hence: $\mathbb{I}(F_n)\left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \leq \mathbb{I}(F_n)\Phi(C_n) - \Phi(C_{n-1})$.

Summing over all n:

$$\begin{aligned}\sum_{n=1}^{N} \mathbb{I}(F_n) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle &\le \sum_{n=1}^{N} \mathbb{I}(F_n) \left[\Phi(C_n) - \Phi(C_{n-1})\right] \\ &\le \sum_{n=1}^{N} \left[\Phi(C_n) - \Phi(C_{n-1})\right] \\ &= \Phi(C_N) \\ &= \sum_{i=1}^{d} \min\left(\lambda_i(C_N), \lambda_d(C_N) + 1\right) \\ &\le d\lambda_d(C_N) + d\end{aligned}$$

Since we have shown $\mathbb{I}(D_n^\alpha) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \le c\frac{d\lambda_d(C_N)+d}{N}$ above, this means:

$$\sum_{n=1}^{N} \mathbb{I}(G_n) \left\langle v_{n-1}^*(v_{n-1}^*)^\top, X_n \right\rangle \le O\left(d\lambda_d(C_N) + d\right)$$

□

Now that we have showed the competitive ratio in the dyad selection setting when $k = 1$, we want to generalize this result in the Online PCA setting in Theorem 6.

Theorem 6. *The FTL's performance in the online subspace selection problem, under Assumption 4 and Assumption 5 , satisfies*

$$L_{N,Alg} \le O\left(dL_N^* + d^2\right).$$

Proof. For $i \in [k]$, we analyze the $FTL(i)$ algorithm which choose chooses v_n^i to be $v_{n-1}^{i,*} := \arg\min_{v \in V^{n-1}}^{d-i+1} L_{n-1,v}^i$ (where V^{n-1} denotes the set of eigenvectors of C_{n-1}, and $\arg\min_{v \in V^{n-1}}^{d-i+1}$ choose the eigenvector with the $(d-i+1)-th$ largest eigenvalue of C_{n-1}) for time step n.

Following the same steps in the proof of the dyad selection setting, for all $i \in [k]$, we have:

$$\sum_{n=1}^{N} \mathbb{I}(G_n^{i,c})\ell_n^i \le O\left(dL_N^{i,*} + d\right) \qquad \text{(Using Assumption 4)}$$

$$\mathbb{I}(D_n^{\alpha,i})\ell_n^i \le c\frac{d\lambda_{d+1-i}(C_N) + d}{N}. \qquad \text{(Using Assumption 5)}$$

Similarly, by modifying the potential function $\Phi^i(C) = \sum_{j=1}^{d+1-i} \min\left(\lambda_j(C), \lambda_{d+1-i}(C) + 1\right)$, following the same steps previously, we have:

$$\sum_{n=1}^{N} \mathbb{I}(F_n^i)\ell_n^i \le d\lambda_{d+1-i}(C_N) + d$$

Hence:

$$
\begin{aligned}
L_{N,Alg} &= \sum_{i=1}^{k}\sum_{n=1}^{N} \ell^i_{n,i^*_{n-1}} \\
&= \sum_{i=1}^{k}\left[\sum_{n=1}^{N} \mathbb{I}(G^i_n)\ell^i_{n,i^*_{n-1}} + \sum_{n=1}^{N} \mathbb{I}(G^{i,c}_n)\ell^i_{n,i^*_{n-1}}\right] \\
&\le \sum_{i=1}^{k}\left[\sum_{n=1}^{N} \mathbb{I}(G^i_n)\ell^i_{n,i^*_{n-1}} + O\left(d\lambda_{d+1-i}(C_N) + d\right)\right] \\
&\le \sum_{i=1}^{k}\left[\sum_{n=1}^{N} \mathbb{I}(D^{\alpha,i}_n)\ell^i_{n,i^*_{n-1}} + \sum_{n=1}^{N} \mathbb{I}(F^i_n)\ell^i_{n,i^*_{n-1}} + O\left(d\lambda_{d+1-i}(C_N) + d\right)\right] \\
&\le \sum_{i=1}^{k}\left[O\left(d\lambda_{d+1-i}(C_N) + d\right)\right] \\
&\le O\left(dL^*_N + d^2\right)
\end{aligned}
$$

□

C Sequential Representation Learning for Multi-task Linear Bandit

C.1 Algorithm details

As described in the paper, Algorithm 1 deploys a bi-level strategy for simultaneously learning and leveraging the estimated B. We now provide a more detailed explanation of each level.

Lower-Level Algorithms. As previously mentioned, Algorithm 2 serves the purpose of meta-exploration. When applied to task n, it simultaneously achieves two objectives: obtaining an unbiased estimate of θ_n and maintaining a reasonable regret bound for the current task. During the initial τ_1 steps (lines 3 to 6), the learner selects actions from the set $\{\lambda_0 e_i\}_{i=1}^d$, which span the action space $\mathcal{A}$. Here, e_i represents the i-th standard basis vector in $\mathbb{R}^d$, and $\lambda_0 = \sqrt{\lambda_{\min}(M)}$ is a constant factor ensuring that $\lambda_0 e_i \in \mathcal{A}$ (as per Assumption 2). Subsequently, the task parameter $\hat{\theta}_n$ is estimated (line 8), and the learner acts greedily for the remainder of the task (line 10). The performance guarantee of this algorithm, originally established by Rusmevichientong & Tsitsiklis (2010), is summarized in Lemma 14.

Conversely, Algorithm 3 is designed for meta-exploitation. It takes a subspace, represented by its orthonormal basis $\hat{B}$, as input. When applied to task n, it can yield a lower regret compared to Algorithm 2 if the provided subspace effectively captures θ_n. Instead of exploring the entire $\mathbb{R}^d$, the learner restricts its exploration to the subspace induced by $\hat{B}_n$ (lines 3 and 6). Following this, the low-dimensional task parameter $\hat{w}_n$ is estimated (line 8), and the learner adopts a greedy strategy for the rest of the task (line 11). The performance guarantee for this algorithm, originally presented by Yang et al. (2020), is detailed in Lemma 15.

Lemma 15 highlights the opportunistic nature of Algorithm 3. If θ_n lies entirely within the subspace spanned by $\hat{B}_n$ (i.e., $\|\hat{B}^\top_{n,\perp}\theta_n\| = 0$), the regret bound becomes $R^n_\tau \le O\left(\tau_2 + \tau \cdot \frac{m^2}{\tau_2}\right)$, potentially as low as $O(m\sqrt{\tau})$. However, in the worst-case scenario, $\|\hat{B}^\top_{n,\perp}\theta_n\|$ can be as large as $\|\theta_n\|$, leading to a trivial linear regret bound. Thus, the effectiveness of this algorithm critically depends on selecting appropriate subspaces $\hat{B}_n$ as input. The proof of Lemma 15 can be found in Appendix C.2.

Upper-Level Strategy. For the upper level, we introduce Algorithm 1, which decides (1) when to engage in meta-exploration and (2) which subspace to utilize during meta-exploitation.

Algorithm 2 Meta-Exploration Routine

1: **Input:** Current task index n, exploration duration τ_1 (a multiple of d)
2: **for** $i \leftarrow 1$ to d **do**
3: Set the action $A_{n,t} = \lambda_0 e_i$ for time steps $t = u(i-1)+1, \cdots, ui$, where $u = \frac{\tau_1}{d}$
4: **end for**
5: **for** time step $t \leftarrow 1$ to τ_1 **do**
6: Choose action $A_{n,t}$ and observe the resulting reward $r_{n,t}$
7: **end for**
8: Compute the estimated task parameter $\hat{\theta}_n := \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{\tau_1} \sum_{t=1}^{\tau_1} (\langle A_{n,t}, \theta \rangle - r_{n,t})^2$
9: **for** time step $t \leftarrow \tau_1 + 1$ to τ **do**
10: Select action $A_{n,t} \leftarrow \arg\max_{a \in \mathcal{A}} \left\langle a, \hat{\theta}_n \right\rangle$
11: **end for**

Algorithm 3 Meta-Exploitation Routine

1: **Input:** Current task index n, exploitation exploration length τ_2 (a multiple of m), and the estimated orthonormal basis of the subspace $\hat{B}_n \in \mathbb{R}^{d \times m}$
2: **for** $i \leftarrow 1$ to m **do**
3: Set the action $A_{n,t} = \lambda_0 \hat{B}_n(:, i)$ for time steps $t = u(i-1)+1, \cdots, ui$, where $u = \frac{\tau_2}{m}$
4: **end for**
5: **for** time step $t \leftarrow 1$ to τ_2 **do**
6: Choose action $A_{n,t}$ and observe the resulting reward $r_{n,t}$
7: **end for**
8: Compute the estimated task vector $\hat{w}_n := \arg\min_{w \in \mathbb{R}^m} \frac{1}{\tau_2} \sum_{t=1}^{\tau_2} (\left\langle A_{n,t}, \hat{B}_n w \right\rangle - r_{n,t})^2$
9: Obtain $\hat{\theta}_n := \hat{B}_n \hat{w}_n$
10: **for** time step $t \leftarrow \tau_2 + 1$ to τ **do**
11: Select action $A_{n,t} \leftarrow \arg\max_{a \in \mathcal{A}} \left\langle a, \hat{\theta}_n \right\rangle$
12: **end for**

Regarding (1), for each task, the learner chooses to perform meta-exploration with a probability p (line 6) or to exploit using the current online estimate of the subspace, $\hat{B}_n$ (line 9).

Concerning (2), we propose selecting the subspace $\hat{B}_n$ using SVD, as shown in line 4.

C.2 Upper bound

Lemma 14. *Let τ_1 be a positive integer that is a multiple of d. Consider running Algorithm 2 on task n with an exploration length of τ_1. Then, there exist positive constants c_1 and c_2, whose values depend on $\lambda_0, \theta_{\max}, \theta_{\min}$, and M, such that the following hold:*

1. *The regret incurred on task n over τ time steps satisfies the bound $R^n_{EXR} \leq c_1 \cdot \left(\tau_1 + \frac{\tau d^2}{\tau_1}\right)$.*

2. *With a probability of at least $1-\delta$, the estimation error of the parameter vector $\hat{\theta}_n$ from the true parameter vector θ_n is bounded by $\|\hat{\theta}_n - \theta_n\| \leq c_2 \cdot d\sqrt{\frac{\ln(d/\delta)}{\tau_1}} =: \alpha$, where $\delta := \frac{d^2}{N\tau_1}$.*

Proof. The proof of this lemma can be derived from Duong et al. (2024)'s Lemma 3. □

Lemma 15. *Let τ_2 be a positive integer that is a multiple of m. Consider running Algorithm 3 on task n, utilizing an input subspace $\hat{B}_n$ and an exploration duration of τ_2. Then, there exists a positive constant c, dependent on*

$\lambda_0, \theta_{\max}, \theta_{\min}$, *and* M, *such that the cumulative regret on task* n *is bounded by:*

$$R_\tau^n \leq c \cdot \left(\tau_2 + \tau \cdot \left(\frac{m^2}{\tau_2} + \|\hat{B}_{n,\perp}^\top \theta_n\|_2^2\right)\right).$$

Proof. The proof of this lemma can be derived from Duong et al. (2024)'s Lemma 4. □

Lemma 16. *Assuming that* $\alpha = \tilde{O}\left(\frac{d}{\sqrt{\tau_1}}\right)$ *as defined in Lemma 14. Then, the expected cumulative regret of all tasks running the Meta-Exploitation routine in Algorithm 3 is:*

$$R_{EXT} \leq \tilde{O}\left(N\tau_2 + N\tau\frac{m^2}{\tau_2} + \tau N\frac{d^3}{\tau_1} + \frac{\tau d^2}{p}\right).$$

Proof. First, we use Lemma 15 to bound the regret of the task n running the Exploitation Routine:

$$R_{\text{EXT}}^n \leq O\left(\tau_2 + \tau \cdot \left(\frac{m^2}{\tau_2} + \|\hat{B}_{n,\perp}^\top \theta_n\|_2^2\right)\right).$$

Hence:

$$\begin{aligned} R_{\text{EXT}} &= \sum_{n=1}^N \mathbb{E}\left[R_{\text{EXT}}^n\right] \mathbb{I}(Z_n = 0) \\ &\leq \sum_{n=1}^N \mathbb{E}\left[R_{\text{EXT}}^n\right] \\ &\leq O\left(N\tau_2 + N\tau\frac{m^2}{\tau_2} + \tau\sum_{n=1}^N \mathbb{E}\left[\|\hat{B}_{n,\perp}^\top \theta_n\|_2^2\right]\right). \end{aligned}$$

To bound the sum on the RHS, we need to upper bound $\min_{\bar{B}} \sum_{n=1}^N \|\bar{B}_{n,\perp}^\top \hat{\theta}_n\|^2$. We will do this below by using the Competitive Ratio for subspace selection in Theorem 6.

First, recall that $F_n = \left\{Z_n = 0 \vee (Z_n = 1 \wedge \|\hat{\theta}_n - \theta_n\| \leq \alpha)\right\}$, and $F = \cap_{n=1}^N F_n$, which was defined in Assumption 7.

Using Lemma 14, we have $P(\|\hat{\theta}_n - \theta_n\| > \alpha \mid Z_n = 1) \leq \delta$. Thus, $P(F_n^c) = P(\|\hat{\theta}_n - \theta_n\| > \alpha \mid Z_n = 1)P(Z_n = 1) \leq p\delta$. Hence $P(F^c) \leq Np\delta \leq N\delta$.

Define $\ell_n(A) = Z_n\|A_\perp^\top \hat{\theta}_n\|^2$. From the Competitive Ratio for subspace selection Theorem 6, we have:

$$
\begin{aligned}
\mathbb{I}(F)\sum_{n=1}^{N}\ell_n(\hat{B}_n) &\le \mathbb{I}(F)O\left(d\min_{\bar{B}}\sum_{n=1}^{N}\ell_n(\bar{B})+d^2\right)\\
&\le \mathbb{I}(F)O\left(d\sum_{n=1}^{N}\ell_n(B)+d^2\right)\\
\implies \sum_{n=1}^{N}\ell_n(\hat{B}_n) &\le \mathbb{I}(F)O\left(d\sum_{n=1}^{N}\ell_n(B)+d^2\right)+\mathbb{I}(F^c)\sum_{n=1}^{N}\ell_n(\hat{B}_n)\\
\implies \sum_{n=1}^{N}\mathbb{E}\left[\ell_n(\hat{B}_n)\right] &\le \mathbb{I}(F)O\left(d\sum_{n=1}^{N}\mathbb{E}\left[\ell_n(B)\right]+d^2\right)+\mathbb{E}\left[\mathbb{I}(F^c)\sum_{n=1}^{N}\ell_n(\hat{B}_n)\right]\\
&\le \mathbb{I}(F)O\left(d\sum_{n=1}^{N}\mathbb{E}\left[\ell_n(B)\right]+d^2\right)+N^2p\delta\\
\implies \sum_{n=1}^{N}\mathbb{E}\left[Z_n\|\hat{B}_{n,\perp}^{\top}\hat{\theta}_n\|^2\right] &\le \mathbb{I}(F)O\left(d\sum_{n=1}^{N}\mathbb{E}\left[Z_n\|B_{\perp}^{\top}\hat{\theta}_n\|^2\right]+d^2\right)+N^2p\delta\\
\implies p\sum_{n=1}^{N}\mathbb{E}\left[\|\hat{B}_{n,\perp}^{\top}\hat{\theta}_n\|^2\right] &\le \mathbb{I}(F)\tilde{O}\left(dp\sum_{n=1}^{N}\mathbb{E}\left[\|B_{\perp}^{\top}\hat{\theta}_n\|^2\right]+d^2\right)+N^2p\delta\\
\implies \sum_{n=1}^{N}\mathbb{E}\left[\|\hat{B}_{n,\perp}^{\top}\hat{\theta}_n\|^2\right] &\le \mathbb{I}(F)\tilde{O}\left(d\sum_{n=1}^{N}\mathbb{E}\left[\|B_{\perp}^{\top}\hat{\theta}_n\|^2\right]+\frac{d^2}{p}\right)+N^2\delta
\end{aligned}
$$

We also have:

$$
\begin{aligned}
\|B_{\perp}^{\top}\hat{\theta}_n\|^2 &\le \|B_{\perp}^{\top}(\theta_n-\hat{\theta}_n)\|^2+\|B_{\perp}^{\top}\theta_n\|^2 && \text{(Triangle inequality)}\\
&\le \|B_{\perp}^{\top}\|^2\|\theta_n-\hat{\theta}_n\|^2+\|B_{\perp}^{\top}\theta_n\|^2\\
&\le \tilde{O}\left(\frac{d^2}{\tau_1}\right)\\
\implies \sum_{n=1}^{N}\|B_{\perp}^{\top}\hat{\theta}_n\|^2 &\le \tilde{O}\left(N\frac{d^2}{\tau_1}\right)\\
\implies \sum_{n=1}^{N}\mathbb{E}\left[\|\hat{B}_{n,\perp}^{\top}\hat{\theta}_n\|^2\right] &\le \tilde{O}\left(N\frac{d^3}{\tau_1}+\frac{d^2}{p}\right)+N^2\delta\\
&\le \tilde{O}\left(N\frac{d^3}{\tau_1}+\frac{d^2}{p}\right) && \left(\delta\le O\left(\tfrac{d^3}{N\tau_1}\right)\right)
\end{aligned}
$$

Thus:

$$
\begin{aligned}
\|\hat{B}_{n,\perp}^{\top}\theta_n\|^2 &\le \|\hat{B}_{n,\perp}^{\top}(\theta_n-\hat{\theta}_n)\|^2+\|\hat{B}_{n,\perp}^{\top}\hat{\theta}_n\|^2 && \text{(Triangle inequality)}\\
&\le \|\hat{B}_{n,\perp}^{\top}\|^2\|\theta_n-\hat{\theta}_n\|^2+\|\hat{B}_{n,\perp}^{\top}\hat{\theta}_n\|^2\\
&\le \tilde{O}\left(\frac{d^2}{\tau_1}\right)+\|\hat{B}_{n,\perp}^{\top}\hat{\theta}_n\|^2\\
\implies \sum_{n=1}^{N}\mathbb{E}\left[\|\hat{B}_{n,\perp}^{\top}\theta_n\|^2\right] &\le \tilde{O}\left(N\frac{d^2}{\tau_1}\right)+\sum_{n=1}^{N}\mathbb{E}\left[\|\hat{B}_{n,\perp}^{\top}\hat{\theta}_n\|^2\right]\\
&\le \tilde{O}\left(N\frac{d^3}{\tau_1}+\frac{d^2}{p}\right)
\end{aligned}
$$

Hence:

$$R_{\text{EXT}} \leq \tilde{O}\left(N\tau_2 + N\tau\frac{m^2}{\tau_2} + \tau\sum_{n=1}^{N}\mathbb{E}\left[\|\hat{B}_{n,\perp}^{\top}\theta_n\|^2\right]\right)$$
$$\leq \tilde{O}\left(N\tau_2 + N\tau\frac{m^2}{\tau_2} + \tau N\frac{d^3}{\tau_1} + \tau\frac{d^2}{p}\right)$$

□

Theorem 8. *Let $p = \min\left[\left(\frac{\tau d}{N^2}\right)^{\frac{1}{3}}, 1\right]$, the meta-exploration's exploration duration $\tau_1 = d\cdot\left\lfloor\min\left(d\sqrt{\frac{d\tau}{p}}, \tau\right)/d\right\rfloor$, and the meta-exploitation's exploration duration $\tau_2 = m\sqrt{\tau}$. Under Assumption 1, 2, and 7, the meta-regret of Algorithm 1 satisfies*

$$R_\tau \leq \tilde{O}\left(Nm\sqrt{\tau} + N^{\frac{2}{3}}\tau^{\frac{2}{3}}d^{\frac{5}{3}} + Nd^3 + \tau d^2\right). \tag{2}$$

Proof. Define R_{EXR} as the cumulative regret of all task running Algorithm 2. Then, we have:

$$\begin{aligned}
R_\tau &\leq \tilde{O}\left(pR_{\text{EXR}} + (1-p)R_{\text{EXT}}\right) && \text{(From Duong et al. (2024)'s Theorem 7)}\\
&\leq \tilde{O}\left(\left(\tau_1 + \tau\cdot\frac{d^2}{\tau_1}\right)Np + R_{\text{EXT}}\right) && \text{(Lemma 14)}\\
&\leq \tilde{O}\left(Np\tau_1 + Np\tau\frac{d^2}{\tau_1} + N\tau_2 + N\tau\frac{m^2}{\tau_2} + \tau N\frac{d^3}{\tau_1} + \tau\frac{d^2}{p}\right) && \text{(Lemma 16)}\\
&= \tilde{O}\left(Nm\sqrt{\tau} + Np\tau_1 + Np\tau\frac{d^2}{\tau_1} + \tau N\frac{d^3}{\tau_1} + \tau\frac{d^2}{p}\right) && (\tau_2 = m\sqrt{\tau})\\
&\leq \tilde{O}\left(Nm\sqrt{\tau} + Np\tau_1 + \tau N\frac{d^3}{\tau_1} + \frac{\tau d^2}{p}\right)\\
&\leq \tilde{O}\left(Nm\sqrt{\tau} + \mathbb{I}(d^3 \geq p\tau)(Np\tau + Nd^3) + \frac{\tau d^2}{p} + Npd\sqrt{\frac{d\tau}{p}} + \tau N\frac{d^3}{d\sqrt{\frac{d\tau}{p}}}\right)\\
& && \text{(Choose } \tau_1 = d\cdot\left\lfloor\min\left(d\sqrt{\tfrac{d\tau}{p}}, \tau\right)/d\right\rfloor\text{)}\\
&\leq \tilde{O}\left(Nm\sqrt{\tau} + \mathbb{I}(d^3 \geq p\tau)(Np\tau + Nd^3) + \frac{\tau d^2}{p} + Nd^{\frac{3}{2}}\sqrt{p\tau}\right)\\
&\leq \tilde{O}\left(Nm\sqrt{\tau} + Nd^3 + \frac{\tau d^2}{p} + Nd^{\frac{3}{2}}\sqrt{p\tau}\right)\\
&\leq \tilde{O}\left(Nm\sqrt{\tau} + Nd^3 + N^{\frac{2}{3}}\tau^{\frac{2}{3}}d^{\frac{5}{3}} + \mathbb{I}(\tau d \geq N^2)(\tau d^2 + Nd^{\frac{3}{2}}\sqrt{\tau})\right) && \text{(Choose } p = \min\left[\left(\tfrac{\tau d}{N^2}\right)^{\frac{1}{3}}, 1\right]\text{)}\\
&\leq \tilde{O}\left(Nm\sqrt{\tau} + N^{\frac{2}{3}}\tau^{\frac{2}{3}}d^{\frac{5}{3}} + Nd^3 + \tau d^2\right)
\end{aligned}$$

□

C.3 Lower bound

Theorem 9. *The lower bound for the Sequential Representation Learning for Multi-task Linear Bandit is:*

$$R_\tau \geq \tilde{\Omega}\left(Nm\sqrt{\tau} + d\sqrt{m\tau N}\right).$$

Proof. Following the construction in (Yang et al., 2020), we prove the two terms separately:

1. By Lemma 17, there exists an N-task linear bandit problem in $\mathbb{R}^m$ with horizon τ, action set $\mathcal{A} = \{x : \|x\| \leq 1\}$, and parameters $(w^1, \ldots, w^N)$ such that $\sum_{n=1}^N R^n(\mathcal{A}, w^n) \geq \Omega\left(Nm\sqrt{\tau}\right)$. Fix any semi-orthogonal matrix $B \in \mathbb{R}^{d\times m}$. We embed this problem into a fixed m-dimensional subspace of $\mathbb{R}^d$. The $\Omega\left(Nm\sqrt{\tau}\right)$ lower bound remains valid.

2. For the second term, it follows from Lemma 17 that there exists an m-task linear bandit problem in $\mathbb{R}^d$ with horizon $\frac{N\tau}{m}$, action set $\mathcal{A} = \{x : \|x\| \leq 1\}$, and parameters $(\theta^1, \ldots, \theta^m)$ such that $\sum_{n=1}^m R^n(\mathcal{A}, \theta^m) \geq \Omega\left(d\sqrt{m\tau N}\right)$. This instance can be regarded as an N-task linear bandit problem where the underlying parameters lie in an m-dimensional subspace and each task has a horizon of τ. To see this, one can group the N tasks into m periods, such that each period comprises N/m tasks and spans $N\tau/m$ rounds. Within each period, the tasks share a common parameter. Since there are only m parameters, they lie in an m-dimensional subspace of $\mathbb{R}^d$.

□

From Section 2.2, we have $r_t^n = \langle a_t^n, \theta_n\rangle + \eta_t^n$, where $\eta_t^n \sim \mathcal{N}(0,1)$.

The meta-regret of a policy is:

$$\sum_{n=1}^N R^n(\mathcal{A}^n, \mu^n) = \tau \sum_{n=1}^N \max_{a\in\mathcal{A}^n} \langle a, \theta_n\rangle - \sum_{n=1}^N \sum_{t=1}^\tau \mathbb{E}\left[r_t^n\right],$$

where the expectation is taken with respect to $\mathbb{P}$, which indicates the measure on outcomes induced by the interaction on the fixed policy and the Gaussian bandit parameterised by θ_n.

Lemma 17. *For any N, τ, l, and action space $\mathcal{A} = \left\{x \in \mathbb{R}^l : \|x\|_2 \leq 1\right\}$, for any algorithm σ, there exists a N-task, horizon-τ multitask bandit instance $\Phi = \left(\phi^1, \cdots, \phi^N \mid \phi^s \in \mathbb{R}^l,\ s \in [N]\right)$ that satisfies Assumption 1, such that $\sum_{n=1}^N R_\tau^n(\mathcal{A}, \Phi, \sigma) \geq \frac{Nl\sqrt{\tau}}{32\sqrt{3}}$*

Proof. Our proof is inspired by Lattimore & Szepesvári (2020), which we extend to the sequential multitask setting.

In the following proof, we denote $R^n(\Phi) = R^n(\phi^n) = R_\tau^n(\mathcal{A}, \Phi, \sigma)$

Fix $n \in [N]$ and $i \in [l]$. Let $\varepsilon = \frac{1}{8\sqrt{3}}\sqrt{\frac{l}{\tau}}$

and $\Phi \in \{\pm\varepsilon\}^{lN}$ and define $\tau^n(i) = \tau \wedge \min\left\{t : \sum_{s=1}^t \left[a_s^n(i)\right]^2 \geq \frac{\tau}{l}\right\}$, where $a_t^n(i)$ is the value at dimension i of a_t^n.

By Lemma 18, we have:

$$R^n(\Phi) \geq \frac{\varepsilon\sqrt{l}}{2} \sum_{i=1}^l \mathbb{E}_\Phi\left[\sum_{t=1}^{\tau^n(i)} \left(\frac{1}{\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right)^2\right]$$

For $x \in \{\pm 1\}$, define $U^n(i,x) = \sum_{t=1}^{\tau^n(i)} \left(\frac{1}{\sqrt{l}} - a_t^n(i)x\right)^2$ and let $\Phi = \left(\phi^1, \cdots, \phi^N\right)$ and $\Phi^{-,n,i} = \left(\phi^{1'}, \cdots, \phi^{N'}\right)$ such that $\phi^{s'} = \phi^s$ for all $s \in [n-1] \cup [n+1, N]$, $\phi^{n'}(s) = \phi^n(s)$ for all $s \in [l] \setminus \{i\}$, and $\phi^{n'}(i) = -\phi^n(i)$. Assume without loss of generality that $\phi^n(i) > 0$.

Let $\mathbb{P}$ and $\mathbb{P}'$ be the laws of $U^n(i,1)$ with respect to the bandit/learner interaction measure induced by Φ and $\Phi^{-,n,i}$, respectively. Let $\mathcal{F}_t^j = \sigma(r_1^j, a_1^j, \ldots, r_t^j, a_t^j, \{r_1^s, a_1^s, \ldots, r_T^s, a_T^s\}_{s=1}^{j-1})$ Then, $\tau^n(i)$ is an $(\mathcal{F}_t^n)_{t=1}^\tau$-measurable stopping time.

We have:

$$\begin{aligned}
U^n(i,1) &= \sum_{t=1}^{\tau^n(i)} \left(\frac{1}{\sqrt{l}} - a_t^n(i)\right)^2 \\
&\leq 2\sum_{t=1}^{\tau^n(i)} \frac{1}{l} + 2\sum_{t=1}^{\tau^n(i)} a_t^n(i)^2 \\
&\leq \frac{4\tau}{l} + 2 && \text{(By definition of of } \tau^n(i))
\end{aligned}$$

Denote $\mathcal{X}_t^j = \left\{r_1^j, a_1^j, \ldots, r_t^j, a_t^j, \{r_1^s, a_1^s, \ldots, r_T^s, a_T^s\}_{s=1}^{j-1}\right\}$. Then,

$$\begin{aligned}
\mathbb{E}_{\Phi}\left[U^n(i,1)\right] - \mathbb{E}_{\Phi^{-,n,i}}\left[U^n(i,1)\right] &= \int_x U^n(i,1)d\mathbb{P}(x) - \int_x U^n(i,1)d\mathbb{P}'(x) \\
&\geq -\left(\frac{4\tau}{l} + 2\right)\left\|\mathbb{P} - \mathbb{P}'\right\|_1 && (\ell_1/\ell_\infty \text{ Holder's inequality}) \\
&\geq -\left(\frac{4\tau}{l} + 2\right)\sqrt{\frac{1}{2}D(\mathbb{P}', \mathbb{P})} \\
& && \text{(Pinsker's inequality)} \\
&\geq -\varepsilon\left(\frac{4\tau}{l} + 2\right)\sqrt{\mathbb{E}_{\Phi^{-,n,i}}\left[\sum_{t=1}^{\tau^n(i)}(a_t^n(i))^2\right]} && \text{(Lemma 19 below)} \\
&\geq -\varepsilon\left(\frac{4\tau}{l} + 2\right)\sqrt{\frac{\tau}{l} + 1} && \text{(By def of } \tau^n(i)) \\
&\geq -\frac{8\sqrt{3}\varepsilon\tau}{l}\sqrt{\frac{\tau}{l}} && \text{(Assume that } d \leq 2\tau)
\end{aligned}$$

We have:

$$\begin{aligned}
&\mathbb{E}_{\Phi^{-,n,i}}\left[U^n(i,-1)\right] = \mathbb{E}_{\Phi^{-,n,i}}\left[U^n(i,1) + U^n(i,-1)\right] - \mathbb{E}_{\Phi^{-,n,i}}\left[U^n(i,1)\right] \\
\implies &\mathbb{E}_{\Phi}\left[U^n(i,1)\right] + \mathbb{E}_{\Phi^{-,n,i}}\left[U^n(i,-1)\right] \geq \mathbb{E}_{\Phi^{-,n,i}}\left[U^n(i,1) + U^n(i,-1)\right] + \mathbb{E}_{\Phi}\left[U^n(i,1)\right] - \mathbb{E}_{\Phi^{-,n,i}}\left[U^n(i,1)\right] \\
&\geq \mathbb{E}_{\Phi^{-,n,i}}\left[U^n(i,1) + U^n(i,-1)\right] - \frac{8\sqrt{3}\varepsilon\tau}{l}\sqrt{\frac{\tau}{l}} \\
&= 2\mathbb{E}_{\Phi^{-,n,i}}\left[\frac{\tau^n(i)}{l} + \sum_{t=1}^{\tau^n(i)}(a_t^n(i))^2\right] - \frac{8\sqrt{3}\varepsilon\tau}{l}\sqrt{\frac{\tau}{l}} \\
& && (U^n(i,x) = \textstyle\sum_{t=1}^{\tau^n(i)}\left(\frac{1}{\sqrt{l}} - a_t^n(i)x\right)^2) \\
&\geq \frac{2\tau}{l} - \frac{8\sqrt{3}\varepsilon\tau}{l}\sqrt{\frac{\tau}{l}} \\
& && \text{(By definition of } \tau^n(i) = \tau \wedge \min\left\{t : \textstyle\sum_{s=1}^t \left[a_s^n(i)\right]^2 \geq \frac{\tau}{l}\right\}) \\
&= \frac{\tau}{l} && (\varepsilon = \tfrac{1}{8\sqrt{3}}\sqrt{\tfrac{l}{\tau}})
\end{aligned}$$

Following Lattimore & Szepesvári (2020), proof of Theorem 24.2, using the randomization hammer, for the $R^n(\Phi)$ inequality above from Lemma 18, we have:

$$
\begin{aligned}
\sum_{\Phi\in\{\pm\varepsilon\}^{lN}}\sum_{n=1}^{N} R^n(\Phi) &\geq \frac{\varepsilon\sqrt{l}}{2}\sum_{n=1}^{N}\sum_{i=1}^{l}\sum_{\Phi\in\{\pm\varepsilon\}^{lN}} \mathbb{E}_\Phi\left[\sum_{t=1}^{\tau^n(i)}\left(\frac{1}{\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right)^2\right] \\
&= \frac{\varepsilon\sqrt{l}}{2}\sum_{n=1}^{N}\sum_{i=1}^{l}\sum_{\Phi\in\{\pm\varepsilon\}^{lN}} \mathbb{E}_\Phi\left[U^n(i,\mathrm{sign}(\phi^n(i)))\right] && \text{(By definition of } U^n(\cdot)\text{)} \\
&= \frac{\varepsilon\sqrt{l}}{2}\sum_{n=1}^{N}\sum_{i=1}^{l}\sum_{\{\Phi\setminus\{\phi^n\}\}\in\{\pm\varepsilon\}^{l(N-1)}}\ \sum_{\{\phi^n(j)\}_{j\in[l]\setminus\{i\}}\in\{\pm\varepsilon\}^{l-1}}\ \sum_{\phi^n(i)\in\pm\varepsilon} \mathbb{E}_\Phi\left[U^n(i,\mathrm{sign}(\phi^n(i)))\right] \\
&\geq \frac{\varepsilon\sqrt{l}}{2}\sum_{n=1}^{N}\sum_{i=1}^{l}\sum_{\{\Phi\setminus\{\phi^n\}\}\in\{\pm\varepsilon\}^{l(N-1)}}\ \sum_{\{\phi^n(j)\}_{j\in[l]\setminus\{i\}}\in\{\pm\varepsilon\}^{l-1}} \frac{\tau}{l} && \text{(From above)} \\
&= 2^{Nl-2}N\tau\varepsilon\sqrt{l} && \text{(Since } |\left\{\phi^n(j)\right\}_{j\in[l]\setminus\{i\}}| = 2^{l-1} \text{ and } |\left\{\Phi\setminus\{\phi^n\}\right\}| = 2^{l(N-1)}\text{)}
\end{aligned}
$$

Hence, there exists a $\Phi\in\{\pm\varepsilon\}^{lN}$ such that $\sum_{n=1}^{N} R^n(\Phi) \geq \frac{N\tau\varepsilon\sqrt{l}}{4} = \frac{Nl\sqrt{\tau}}{32\sqrt{3}}$ □

Lemma 18. *For any N,τ,l, and action space $\mathcal{A}=\left\{x\in\mathbb{R}^l : \|x\|_2\leq 1\right\}$, for any algorithm σ, there exists a N-task, horizon-τ multitask bandit instance $\Phi=\left(\phi^1,\cdots,\phi^N \mid \phi^s\in\mathbb{R}^l,\ s\in[N]\right)$ that satisfies Assumption 1. Assume that $l\leq 2\tau$, let $\phi^n\in\{\pm\varepsilon\}^l$, and define $\tau^n(i)=\tau\wedge\min\left\{t:\sum_{s=1}^{t}\left[a_s^n(i)\right]^2\geq\frac{\tau}{l}\right\}$, where $a_t^n(i)$ is the value at dimension i of a_t^n. Then, the regret at task n is:*

$$
R_\tau^n(\mathcal{A},\Phi,\sigma)\geq\frac{\varepsilon\sqrt{l}}{2}\sum_{i=1}^{l}\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau^n(i)}\left(\frac{1}{\sqrt{l}}-a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right)^2\right]
$$

Proof. In the following proof, we denote $R^n(\Phi)=R^n(\phi^n)=R_\tau^n(\mathcal{A},\Phi,\sigma)$

$$
\begin{aligned}
R^n(\phi^n) &= \mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}(a_*^n - a_t^n)^\top \phi^n\right] \\
&= \mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}\|\phi^n\| - (a_t^n)^\top \phi^n\right] && (\tfrac{\phi^n}{\|\phi^n\|} = \arg\max_{a\in\mathcal{A}} a^\top \phi^n) \\
&= \varepsilon\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}\sum_{i=1}^{l}\left(\frac{1}{\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right)\right] && (\|\phi^n\| = \varepsilon\sqrt{l} \text{ and } \phi^n(i) = \varepsilon\ \mathrm{sign}(\phi^n(i))) \\
&= \varepsilon\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}\sum_{i=1}^{l}\left(\frac{1}{2\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right) + \sum_{t=1}^{\tau}\frac{\sqrt{l}}{2}\right] \\
&\geq \varepsilon\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}\sum_{i=1}^{l}\left(\frac{1}{2\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right) + \frac{\sqrt{l}}{2}\sum_{t=1}^{\tau}\|a_t^n\|_2^2\right] && (\|a_t^n\|_2^2 \leq 1 \text{ by definition of } \mathcal{A}) \\
&= \varepsilon\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}\sum_{i=1}^{l}\left(\frac{1}{2\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right) + \frac{\sqrt{l}}{2}\sum_{t=1}^{\tau}\sum_{i=1}^{l}(a_t^n(i))^2\right] \\
&= \varepsilon\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}\sum_{i=1}^{l}\left(\frac{1}{2\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i)) + \frac{\sqrt{l}}{2}(a_t^n(i))^2\right)\right] \\
&= \frac{\varepsilon\sqrt{l}}{2}\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}\sum_{i=1}^{l}\left(\frac{1}{l} - \frac{2}{\sqrt{l}}a_t^n(i)\ \mathrm{sign}(\phi^n(i)) + (a_t^n(i)\,\mathrm{sign}(\phi^n(i)))^2\right)\right] \\
&= \frac{\varepsilon\sqrt{l}}{2}\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau}\sum_{i=1}^{l}\left(\frac{1}{\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right)^2\right] \\
&\geq \frac{\varepsilon\sqrt{l}}{2}\sum_{i=1}^{l}\mathbb{E}_\Phi\left[\sum_{t=1}^{\tau^n(i)}\left(\frac{1}{\sqrt{l}} - a_t^n(i)\ \mathrm{sign}(\phi^n(i))\right)^2\right]
\end{aligned}
$$

□

Lemma 19 (Sequential multitask Divergence decomposition for Linear bandit with Stopping time)**.** *Let $\tau^n(i)$ be a stopping time for $i \in [l]$ and $n \in [N]$. For a_t^n, the action taken at task n and step t, let $a_t^n(i)$ be the value at dimension i. Define $\Phi = \left\{\phi^s \mid \phi^s \in \mathbb{R}^l\right\}_{s=1:N} \in \{\pm\varepsilon\}^{lN}$ and $\Phi' = \left\{\phi^{1'}, \cdots, \phi^{N'}\right\}$ such that $\phi^{s'} = \phi^s$ for all $s \in [n-1] \cup [n+1, N]$, $\phi^{n'}(s) = \phi^n(s)$ for all $s \in [l] \setminus i$, and $\phi^{n'}(i) = -\phi^n(i)$. For $x \in \{\pm 1\}$, define $U^n(i,x) = \sum_{t=1}^{\tau^n(i)}\left(\frac{1}{\sqrt{l}} - a_t^n(i)x\right)^2$ and let $\mathbb{P}$ and $\mathbb{P}'$ be the laws of $U^n(i,1)$ with respect to the bandit/learner interaction measure induced by $\Phi = \{\phi^s\}_{s=1:N}$ and $\Phi' = \left\{\phi^{s'}\right\}_{s=1:N}$, respectively.*

Then:

$$
D(\mathbb{P}', \mathbb{P}) = 2\varepsilon^2\mathbb{E}_{\Phi^{0,n}}\left[\sum_{t=1}^{\tau^n(i)}(a_t^n(i))^2\right]
$$

Proof. When $N = 1$, the proof is trivial since it's the classical Stopping time version of Divergence Decomposition in the traditional Bandit setting (Lattimore & Szepesvári (2020)'s Exercise 15.7).

When $N > 1$, by denoting $\mathcal{X}_0^n = \mathcal{X}_T^{n-1}$ and $\mathcal{X}_0^0 = \{\emptyset\}$, we have:

$$
\begin{aligned}
D(\mathbb{P}', \mathbb{P}) &= \sum_{\mathcal{X}^n_{\tau^n(i)}} \mathbb{P}'(\mathcal{X}^n_{\tau^n(i)}) \log\left(\frac{\mathbb{P}'(\mathcal{X}^n_{\tau^n(i)})}{\mathbb{P}(\mathcal{X}^n_{\tau^n(i)})}\right) \\
&= \sum_{\mathcal{X}^n_{\tau^n(i)}} \mathbb{P}'(\mathcal{X}^n_{\tau^n(i)}) \log\left(\Pi_{s=1}^{n-1}\Pi_{t=1}^{T}\frac{\mathbb{P}'(x_t^s \mid \mathcal{X}_{t-1}^s)}{\mathbb{P}(x_t^s \mid \mathcal{X}_{t-1}^s)}\right) + \sum_{\mathcal{X}^n_{\tau^n(i)}} \mathbb{P}'(\mathcal{X}^n_{\tau^n(i)}) \log\left(\Pi_{t=1}^{\tau^n(i)}\frac{\mathbb{P}'(x_t^n \mid \mathcal{X}_{t-1}^n)}{\mathbb{P}(x_t^n \mid \mathcal{X}_{t-1}^n)}\right) \\
&\quad \text{(Stopping time version of Divergence Decomposition. Lattimore \& Szepesvári (2020)'s Exercise 15.7)} \\
&= \sum_{\mathcal{X}^n_{\tau^n(i)}} \mathbb{P}'(\mathcal{X}^n_{\tau^n(i)}) \sum_{s=1}^{n-1}\sum_{t=1}^{T} \log\left(\frac{\mathbb{P}'(x_t^s \mid \mathcal{X}_{t-1}^s)}{\mathbb{P}(x_t^s \mid \mathcal{X}_{t-1}^s)}\right) + \sum_{t=1}^{\tau^n(i)} \sum_{\mathcal{X}^n_{\tau^n(i)}} \log\left(\frac{\mathbb{P}'(x_t^n \mid \mathcal{X}_{t-1}^n)}{\mathbb{P}(x_t^n \mid \mathcal{X}_{t-1}^n)}\right) \\
&= \sum_{s=1}^{n-1}\sum_{t=1}^{T}\sum_{\mathcal{X}^n_{\tau^n(i)}} \mathbb{P}'(\mathcal{X}^n_{\tau^n(i)}) \log\left(\frac{\mathbb{P}'(x_t^s \mid \mathcal{X}_{t-1}^s)}{\mathbb{P}(x_t^s \mid \mathcal{X}_{t-1}^s)}\right) + \sum_{\mathcal{X}^n_{\tau^n(i)}} \mathbb{P}'(\mathcal{X}^n_{\tau^n(i)}) \sum_{t=1}^{\tau^n(i)} \log\left(\frac{\mathbb{P}'(x_t^n \mid \mathcal{X}_{t-1}^n)}{\mathbb{P}(x_t^n \mid \mathcal{X}_{t-1}^n)}\right) \\
&= \sum_{\mathcal{X}^n_{\tau^n(i)}} \mathbb{P}'(\mathcal{X}^n_{\tau^n(i)}) \sum_{t=1}^{\tau^n(i)} \log\left(\frac{\mathbb{P}'(x_t^n \mid \mathcal{X}_{t-1}^n)}{\mathbb{P}(x_t^n \mid \mathcal{X}_{t-1}^n)}\right)
\end{aligned}
$$

The last equality holds because $\mathbb{P}'(x_t^s \mid \mathcal{X}_{t-1}^s) = \mathbb{P}(x_t^s \mid \mathcal{X}_{t-1}^s)\ \forall s \in [n-1]$ since the two environments are similar to each other.

Hence:

$$
\begin{aligned}
D(\mathbb{P}', \mathbb{P}) &= \sum_{\mathcal{X}^n_{\tau^n(i)}} \mathbb{P}'(\mathcal{X}^n_{\tau^n(i)}) \sum_{t=1}^{\tau^n(i)} \log\left(\frac{\mathbb{P}'(x_t^n \mid \mathcal{X}_{t-1}^n)}{\mathbb{P}(x_t^n \mid \mathcal{X}_{t-1}^n)}\right) \\
&= \sum_{t=1}^{\tau^n(i)} \sum_{\mathcal{X}^n_{t-1}} \mathbb{P}'(\mathcal{X}^n_{t-1}) D\left(\mathbb{P}'(x_t^n \mid \mathcal{X}_{t-1}^n), \mathbb{P}(x_t^n \mid \mathcal{X}_{t-1}^n)\right) \\
&= \sum_{t=1}^{\tau^n(i)} \sum_{\mathcal{X}^n_{t-1}} \mathbb{P}'(\mathcal{X}^n_{t-1}) D\left[\mathcal{N}\left(\langle a_t^n, \phi^n\rangle, 1\right), \mathcal{N}\left(\left\langle a_t^n, \phi^{n'}\right\rangle, 1\right)\right] \\
&= \frac{1}{2}\sum_{t=1}^{\tau^n(i)} \sum_{\mathcal{X}^n_{t-1}} \mathbb{P}'(\mathcal{X}^n_{t-1}) \mathbb{E}_{\Phi}\left[\left\langle a_t^n, \phi^n - \phi^{n'}\right\rangle^2\right] \qquad \text{(Lattimore \& Szepesvári (2020) Eq. (24.1))} \\
&= \frac{1}{2}\sum_{t=1}^{\tau^n(i)} \sum_{\mathcal{X}^n_{t-1}} \mathbb{P}'(\mathcal{X}^n_{t-1}) \mathbb{E}_{\Phi'}\left[(a_t^n(i))^2 \cdot \left(\phi^n(i) - \phi^{n'}(i)\right)^2\right] \\
&\qquad\qquad (\phi^n \text{ is the same as } \phi^{n'} \text{ except at dimension } i) \\
&= 2\varepsilon^2 \sum_{t=1}^{\tau^n(i)} \sum_{\mathcal{X}^n_{t-1}} \mathbb{P}'(\mathcal{X}^n_{t-1}) \mathbb{E}_{\Phi'}\left[(a_t^n(i))^2\right] \\
&= 2\varepsilon^2 \mathbb{E}_{\Phi'}\left[\sum_{t=1}^{\tau^n(i)} (a_t^n(i))^2\right]
\end{aligned}
$$

□

D ASSUMPTION 5'S REMARK

In this section, we want to elaborate why Assumption 5 is redundant when $\alpha = 0$ or when the eigensystem of C_{n-1} and C_n are the same, for any $n \in [N]$

D.1 When $\alpha = 0$

By the definition of D_n^α in Assumption 4, when $\alpha = 0$, there is no change in the eigenvalues of the subspace $\mathcal{V}_{n-1}$ with the smallest eigenvalue between step $n-1$ and step n. Thus, the instantaneous loss of the FTL algorithm is $l_n = P_{\mathcal{V}_{n-1}} x_n = 0$, hence, Assumption 5 is redundant.

D.2 When the eigensystem of C_{n-1} and C_n are the same

The eigensystem of C_{n-1} and C_n are the same when $x_n = c\, v_j(C_{n-1})$ for some constant c and the eigenvector v_j of C_{n-1}. Recall that, from Assumption 5, $\ell_n^i(a) := \left\langle P_n^i a P_n^{i,\top}, X_n \right\rangle$, where P_n^i is the projection matrix mapping to $v_{d+1-i}(C_{n-1})$.

If $j \neq d+1-i$, then $\ell_n^i(a) = c^2 \left\langle P_n^i a P_n^{i,\top}, v_j(C_{n-1}) v_j^\top(C_{n-1}) \right\rangle = 0 \leq \alpha^i$.

Otherwise:

$$\begin{aligned} & \ell_n^i = \lambda_{d+1-i}(C_n) - \lambda_{d+1-i}(C_{n-1}) \\ \implies & \mathbb{I}(D_n^{\alpha^i,i}) \ell_n^i = \mathbb{I}(D_n^{\alpha^i,i}) \left[\lambda_{d+1-i}(C_n) - \lambda_{d+1-i}(C_{n-1})\right] \\ & \leq \alpha^i \qquad \text{(By definition of } \alpha^i\text{)} \end{aligned}$$

E Justification for Assumption 4

We have:

$$\begin{aligned} D_n^{\alpha^i,i,c} &= \left\{\exists i \in [2,d] \text{ s.t. } \lambda_i(C_{n-1}) = \lambda_d(C_{n-1}) \text{ AND } \lambda_i(C_n) - \lambda_i(C_{n-1}) \geq \alpha^i\right\} && \text{(Surprise events)} \\ F_n^c &= \left\{\lambda_1(C_{n-1}) \geq \lambda_d(C_{n-1}) + 1 \text{ OR } \lambda_1(C_n) \geq \lambda_d(C_n) + 1\right\} && \text{(Big-gap events)} \\ G_n^c &= D_n^{\alpha^i,i,c} \cap F_n^c \end{aligned}$$

Intuitively, whenever Surprise event happens, the gap between $\lambda_i(\cdot)$ and $\lambda_d(\cdot)+1$ got smaller, thus, the number of times both Surprise and Big-gap happen simultaneously should not be too large, making Assumption 4 mild. Even when both events happens $O(N)$ times, this would inflate $\lambda_d(C_N) = O(N)$, thus, still making Assumption 4 mild.

Even then, we provide an example to show that Theorem 6 is not true, in general, without Assumption 4.

Let $d = 2$ and $N = 16$. Then, an adversarial environment is shown in Table 2:

Sanity check:

$$\begin{aligned} 1 &\geq 4\alpha && (\tfrac{1}{4} \geq \alpha \text{ to satisfy } D_n^{\alpha,c}) \\ &= \frac{4}{16}\left(2\lambda_2(C_{16}) + 2\right) \\ &= \frac{1}{4}\left(2+2\right) \\ &= 1 \end{aligned}$$

Recall that $E = \left\{\sum_{n=1}^N \mathbb{I}(G_n^c) \geq d\lambda_d(C_N) + d\right\}$. Since $\sum_{n=1}^N \mathbb{I}(G_n^c) = 4 \geq d\lambda_d(C_N) + d = 4$, then $\mathbb{I}(E) = 1$. Thus, the condition in Assumption 4 is violated: $\mathbb{I}(E) Tr(C_N) = 10 > d\lambda_d(C_N) + d = 4$.

n	x_n	ℓ_n	λ_1	λ_2	$\sum_{i=1}^{n} \mathbb{I}(G_i^c)$
1	e_1	1	1	0	0
2	$e_1/2$	0	1+ 1/4	0	0
3	$e_2/2$	1/4	1+ 1/4	1/4	1
4	$e_1/2$	0	1+ 1/2	1/4	1
5	$e_2/2$	1/4	1+ 1/2	1/2	2
6	$e_1/2$	0	1+ 3/4	1/2	2
7	$e_2/2$	1/4	1+ 3/4	3/4	3
8	$e_1/2$	0	2	3/4	3
9	$e_2/2$	1/4	2	1	4
10	e_1	0	3	1	4
11	e_1	0	4	1	4
12	e_1	0	5	1	4
13	e_1	0	6	1	4
14	e_1	0	7	1	4
15	e_1	0	8	1	4
16	e_1	0	9	1	4

Table 2: An example to justify the Relaxed Task Diversity Assumption 4. Here, $\mathbb{I}(E)Tr(C_N) = 10 > d\lambda_d(C_N) + d = 4$.