# Meta Learning in Bandits within Shared Affine Subspaces

**Steven Bilaj**
Ruhr University Bochum

**Sofien Dhouib**
University of Tübingen

**Setareh Maghsudi**
Ruhr University Bochum

## Abstract

We study the problem of meta-learning several contextual stochastic bandits tasks by leveraging their concentration around a low dimensional affine subspace, which we learn via online principal component analysis to reduce the expected regret over the encountered bandits. We propose and theoretically analyze two strategies that solve the problem: One based on the principle of optimism in the face of uncertainty and the other via Thompson sampling. Our framework is generic and includes previously proposed approaches as special cases. Besides, the empirical results show that our methods significantly reduce the regret on several bandit tasks.

## 1 Introduction

In several real-world applications, such as website design and healthcare, the system recommends an item to a user upon observing some side information depending on the user and the corresponding item. Upon receiving the recommendation, the user sends feedback to the system that captures his interest in the recommendation (Glowacka et al., 2019; Bouneffouf et al., 2020; Atan et al., 2023). One can interpret the feedback as a reward that characterizes the suitability of the selected recommendation or action with the final objective of maximizing the cumulative payoff over time. At the same time, such a selection might be suboptimal due to the incomplete knowledge of the environment. This *exploration-exploitation* tradeoff, along with the side information, is formalized by the *contextual multi-armed bandit (CMAB)* problem (Langford and Zhang, 2007; Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Nourani-Koliji et al., 2022), a notable extension of the *multi-armed bandit (MAB)* problem (Thompson, 1933; Robbins, 1952).

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

In the applications mentioned above, the tasks often relate to each other despite being different. For instance, subgroups of patients have comparable features. As another example, holidays or discount periods promote similar interests in the products of an e-commerce website. That observation motivates us to look beyond a single task to uncover a relation between different ones to accelerate learning on newly encountered tasks. That problem, referred to as *meta-learning* or *learning-to-learn (LTL)*, has mainly appeared in the offline learning literature so far (Hutter et al., 2019). Nevertheless, an emergent body of literature combines LTL and MAB to accelerate learning and reduce the average regret per task (Cella et al., 2020; Cella and Pontil, 2021; Bilaj et al., 2023). In the linear contextual setting, an assumption about the preference vectors captures the relation between the tasks.

In this work, we assume that the feature vectors stem from a distribution that concentrates around a low dimensional subspace, *i.e.,* its variance is explained by a limited number of principal components. We propose learning this structure using online *principal component analysis (PCA)*. We then exploit that knowledge to develop two decision-making policies. The first policy relies on the principle of optimism in the face of uncertainty for linear bandits (OFUL) (Chu et al., 2011; Abbasi-Yadkori et al., 2011), and the second is a Thompson sampling policy (Russo et al., 2018; Agrawal and Goyal, 2013). Analytically, we establish per-task regret upper bounds for both strategies that theoretically prove the benefit of learning such a structure. Moreover, our empirical evaluations of our methods using simulated and real-world data sets confirm their benefits.

Our paper is organized as follows. We review meta-learning and related themes for bandit problems in Section 2. Then we formulate our problem in Section 3. We describe the subspace learning procedure in Section 4 to use in our proposed algorithms in Sections 5 and 6. Finally, we empirically assess our algorithms in Section 8.

## 2 Related Work

Learning to learn was first developed for offline learning (Thrun, 1998; Baxter, 2000; Hutter et al., 2019) as a sub-

field of transfer learning. In this paradigm, one seeks to learn a structure shared by many tasks to generalize to new ones. That structure can be encoded in several ways such as a prior over the task distribution (Amit and Meir, 2018; Rothfuss et al., 2021), a kernel (Aiolli, 2012), a common mean around which tasks concentrate (Denevi et al., 2018), or an approximate low dimensional manifold (Jiang et al., 2022), to name a few.

Recently, meta-learning received attention in the online setting (Finn et al., 2019), more precisely, in the case of bandit feedback. The main idea is that the learner interacts sequentially with bandit problems, so the meta-learned shared structure accelerates exploration for upcoming tasks. In this setting, the objective is to improve the regret guarantees compared to those achievable by considering each task separately. The notion of regret can capture such guarantees; nevertheless, it has several definitions depending on the line of work. We distinguish mainly two regret types in a multi-task scenario: *transfer regret* and *meta regret*. The former depends on the number of learned tasks, whereas the latter takes an expectation on a possibly infinite number of tasks.

Concerning transfer regret, the goal is to prove sub-linear regret in the number of tasks. If the learner considers each task independently, the total regret over tasks is linear in the number of tasks. Within a task, the expected transfer regret is linear in the number of rounds. References Cella and Pontil (2021) and Cella et al. (2022a) prove that if preference vectors have a low-rank structure, then learning it improves performance.

In the setting of Bayesian bandits, instead of assuming that the agent knows the true prior over tasks, a recent line of work proposes to learn that distribution. For example, Bastani et al. (2019) studies the dynamic pricing problem and proposes a Thompson sampling approach. Reference Kveton et al. (2021) generalizes the scope of the stochastic MAB problem by developing a meta-Thompson Sampling (meta-TS) algorithm. Basu et al. (2021) improves the guarantees of Kveton et al. (2021) via a modification of meta-TS. It also generalizes the core idea to other bandit settings, such as linear and combinatorial bandits. While Basu et al. (2021) and Kveton et al. (2021) study learning the mean of the tasks with a known covariance matrix, Peleg et al. (2022) relaxes that assumption. It proposes a general multivariate Gaussian prior learning framework that applies to several prior-update-based bandit algorithms. In the non-linear contextual bandit case, Kassraie et al. (2022); Schur et al. (2022) investigate learning a shared kernel. Concerning the second type of guarantees, Cella et al. (2020) proves that the regret expectation over a potentially infinite number of tasks shrinks to 0 provided that the ridge regularization parameter is inversely proportional to the tasks' variance, and that said variance approaches 0.

Another line of work (Boutilier et al., 2020; Kveton et al., 2020; Yang and Toni, 2020) takes inspiration from the policy gradient methods (Williams, 1992) and aims to learn hyperparameters of policies to maximize the expected cumulative reward. Besides, meta-learning is also applicable to solve problems in other settings concerning the reward generating mechanism, such as the non-stationary case (Azizi et al., 2022), and more generally the adversarial case (Balcan et al., 2022).

Multi-task learning is a field closely related to meta-learning. The main difference between the two is the following: The former is about simultaneously learning over a finite family of bandit tasks without being concerned with generalization over future ones. That method is applied to solve the unstructured stochastic bandit case (Azar et al., 2013), where although the interaction with tasks is sequential, they are finite. Therefore, the agent might encounter the same bandit problem more than once and can leverage the previous experience. Besides, In the case of contextual bandits, a low dimensional structure (Cella and Pontil, 2021; Cella et al., 2022b; Yang et al., 2020a) or prior knowledge of the relations between tasks (Yang et al., 2020b) provably reduces the regret

In this work, we borrow the concept of low dimensional structure from multi-task learning and leverage it with the concentration of tasks around some space region to improve the regret bound over a family of contextual linear bandits tasks. Indeed, assumptions such as high task concentration around a mean or strictly belonging to a low-dimensional subspace are restrictive. Thus, we aim at relaxing them. Finally, our approach is interpretable as learning an approximation of the covariance matrix of the tasks where the total variance is dominated by the contributions of a few principal components that span the subspace so it tightly relates to Peleg et al. (2022); Nevertheless, one of our proposed algorithms does not rely on the prior update.

## 3 Problem Formulation

We consider an agent (learner, interchangeably) that sequentially interacts with several contextual bandit tasks. While learning one task over $n$ rounds, at each round $k$, the learner selects an arm $a_k$ from a dynamic set of arms $\mathcal{A}_k$ with associated context vector $\mathbf{x}_{a_k} \in \mathbb{R}^d$ satisfying $\|\mathbf{x}_{a_k}\| \leq 1$. Then it receives a reward $r_k = \mathbf{x}_{a_k}^\top \boldsymbol{\theta}^* + \epsilon_k$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is the true task parameter to estimate. For different tasks, $\boldsymbol{\theta}^*$ is independently drawn from a probability distribution $\rho$ over $\mathbb{R}^d$ (i.i.d.) with mean $\boldsymbol{\mu}$. Besides, they are bounded, formally, $\|\boldsymbol{\theta}^*\| \leq V$ for some $V > 0$.[1] Moreover, $\epsilon_k$ is the zero-mean 1-subgaussian noise such that $\{\epsilon_k\}_{k=1}^n$ are independent and identically distributed (i.i.d).

[1] Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm.

Our main assumption is that the distribution $\rho$ has low variance along certain directions in space which ought to be learnt. Assumption 1 states this requirement formally. Besides, an illustration of a sampling from such a task distribution in 3 dimensions appears in Figure 1. Finally, we denote the covariance of $\rho$ as $\mathbf{\Sigma}$ with ordered eigenvalues $\sigma_1 \geq \cdots \geq \sigma_d$.

**Assumption 1.** *There exists an orthogonal projection matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ with rank $p$ such that:*

$$\begin{aligned}\mathrm{Var}_\rho &\coloneqq \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|(\mathbf{I}-\mathbf{P})\left(\boldsymbol{\theta}^*-\boldsymbol{\mu}\right)\right\|^2\right] \\ &\ll \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\mathbf{P}\left(\boldsymbol{\theta}^*-\boldsymbol{\mu}\right)\right\|^2\right] \\ &\leq \mathrm{Var}_{\max} \coloneqq \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\boldsymbol{\theta}^*-\boldsymbol{\mu}\right\|^2\right], \\ \mathrm{Var}_\rho &\ll \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\mathbf{P}\boldsymbol{\theta}^*\right\|^2\right].\end{aligned}$$

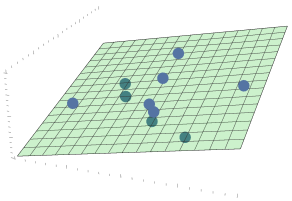


Figure 1: Sample of task parameters (blue points) from a distribution with low variance along one dimension.

Our goal is to learn $\mathbf{P}$ and $\boldsymbol{\mu}$ as well as bound the expected transfer regret $\mathcal{R}(n)$ adapted from Cella et al. (2020) defined as:

$$\mathcal{R}(n) = \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\mathbb{E}\left[\sum_{k=1}^{n}\left(\mathbf{x}_{a_k^*}-\mathbf{x}_{a_k}\right)^\top\boldsymbol{\theta}^*\right]\right], \quad (1)$$

where $a_k^* \coloneqq \arg\max_{a\in\mathcal{A}_k}\mathbf{x}_a^\top\boldsymbol{\theta}^*$ is the optimal arm at round $k$. We propose two different approaches to exploit the knowledge of $\mathbf{P}$. First we present a variation of the standard LinUCB algorithm (Abbasi-Yadkori et al., 2011) by adjusting the regularization term in the regularized least squares optimization problem. Our second approach is a variation of the linear Thompson Sampling algorithm (Agrawal and Goyal, 2013), where we adjust the covariance term of the normal distribution from which a task parameter is sampled from after every task according to the learned projection.

## 4 Subspace Learning

We use an online PCA version, namely, Candid Covariance-Free Incremental Principal Component Analysis (CCIPCA) Cardot and Degras (2015), to learn the underlying subspace structure from estimated task parameters. The core idea is to find an approximation of a set of orthonormal vectors that represent the principal components of a vertically concatenated data set $\mathbf{\Theta} = [(\boldsymbol{\theta}(i)) - \bar{\boldsymbol{\theta}})^\top]_{i\in\{1,\dots,t\}}$, with $\bar{\boldsymbol{\theta}} \coloneqq \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\theta}(i)$. Vector $\boldsymbol{\theta}(i)$ denotes the $i$th task, which was estimated after a total of at least $n$ rounds. Upon finishing a task after $n$ rounds, the agent updates the learned projection matrix. Nevertheless, applying PCA is costly in the long run, whereas an online estimation mitigates the costs while offering sufficient estimations on the learned projection. Starting with a set of orthonormal eigenvectors $\{\mathbf{u}_1,\dots,\mathbf{u}_d\}$ and their corresponding eigenvalues $\{\sigma_1,\dots,\sigma_d\}$ based on the covariance matrix $\frac{1}{t}\mathbf{\Theta}^\top\mathbf{\Theta}$, we define $\mathbf{v}_j \coloneqq \sigma_j\mathbf{u}_j$ for $j\in\{1,\dots,d\}$ as the set of scaled principal components. Here we assume $\bar{\boldsymbol{\theta}} = \mathbf{0}$, for the general case, the task parameters have to be centralized. Each additional task parameter $\boldsymbol{\theta}(i)$ adjusts the estimation of $\mathbf{v}_j$ *i.e.* after every round, every principal component $\mathbf{v}_j$ will be updated as

$$\mathbf{v}_{j,i+1} = \frac{i}{i+1}\mathbf{v}_{j,i} + \frac{1}{i+1}\mathbf{z}_{i+1}\mathbf{z}_{i+1}^\top\frac{\mathbf{v}_{j,i}}{\|\mathbf{v}_{j,i}\|}, \quad (2)$$

with $\mathbf{z}_i$ determined to ensure orthogonality of the eigenvector estimations. Formally, to compute $\mathbf{v}_{j,i+1}$ we have:

$$\mathbf{z}_{i+1} = \boldsymbol{\theta}(i+1) - \sum_{j'=1}^{j-1}\left(\boldsymbol{\theta}^\top(i+1)\mathbf{u}_{j',i}\right)\mathbf{u}_{j',i}.$$

CCIPCA is especially beneficial as it is hyperparameter-free. Besides, it estimates the eigenvalues and the corresponding eigenvectors of all principal components. The eigenvalue estimations are essential when choosing the rank $p$ of the projection, which is generally unknown. The vectors $\mathbf{u}_i$ with the $p$ highest values $\sigma_i$ are selected as principal components. We define their horizontal concatenation as $\mathbf{U} = [\mathbf{u}_j]_{j\in\{1,\dots,p\}} \in \mathbb{R}^{d\times p}$.

**Remark 1.** *The choice of $p$ depends on the respective eigenvalues, a common choice would be to maximize the eigengap, thus $p = \arg\max_{p'}\sigma_{p'} - \sigma_{p'+1}$.*

The projection matrix $\mathbf{P}$ with rank $p$ as well as the orthogonal projection $\mathbf{P}^\perp$ with rank $q = d - p$ can then be constructed using of the principal components as

$$\mathbf{P} = \mathbf{U}\mathbf{U}^\top, \quad \mathbf{P}^\perp = \mathbf{I} - \mathbf{P} \quad (3)$$

We will use the learned projections to exploit the low dimensional subspace structure in both LinUCB and Thompson sampling setting.

## 5 Projection Meta learning with LinUCB

In this section, we present our contextual bandit algorithms based on LinUCB.

### 5.1 Basics of LinUCB

In classic LinUCB, at each round $k$, the agent uses the collection of previously selected actions $\mathbf{D}_k = [\mathbf{x}_{a_i}^\top]_{i\in\{0,\dots,k-1\}}$ and the corresponding rewards $\mathbf{y}_k = [r_i]_{i\in\{0,\dots,k-1\}}$ to estimate the task parameter $\boldsymbol{\theta}_k$ by solving the following regularized least squares optimization problem:

$$\boldsymbol{\theta}_k = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \|\mathbf{D}_k\boldsymbol{\theta} - \mathbf{y}_k\|^2 + \lambda\|\boldsymbol{\theta}\|^2, \tag{4}$$

with $\lambda > 0$ being the regularization parameter. The solution of (4) is the *ridge estimator*. Given that, the learner selects an action that maximizes the UCB index

$$\mathrm{UCB}(a) = \mathbf{x}_a^\top\boldsymbol{\theta}_k + \gamma_k\|\mathbf{x}_a\|_{\mathbf{A}_k^{-1}},^2 \tag{5}$$

with $\mathbf{A}_k \coloneqq \lambda\mathbf{I} + \mathbf{D}_k^\top\mathbf{D}_k$, $\boldsymbol{\theta}_k = \mathbf{A}_k^{-1}\mathbf{D}_k^\top\mathbf{y}_k$ and $\gamma_k > 0$ as an upper bound on the confidence set radius proposed in Abbasi-Yadkori et al. (2011). The additional term scaling is essential for exploration as $\|\mathbf{x}\|_{\mathbf{A}_k^{-1}}$ is maximized for context vectors that have the least correlation with already explored arms.

### 5.2 LinUCB with Projection Bias

In our first proposal, we enhance the LinUCB by including the knowledge of the projection matrix $\hat{\mathbf{P}}$. The agent learns $\hat{\mathbf{P}}$ by an online PCA algorithm using the parameters of $t$ already learned tasks. To enforce the knowledge of the affine subspace during learning, we formulate the following optimization problem for a given task, where we define $\hat{\boldsymbol{\theta}}_k$ as the minimizer over $\boldsymbol{\theta}\in\mathbb{R}^d$ in the following objective:

$$\|\mathbf{D}_k\boldsymbol{\theta} - \mathbf{y}_k\|^2 + \lambda_1\left\|\hat{\mathbf{P}}^\perp(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\right\|^2 + \lambda_2\left\|\hat{\mathbf{P}}\boldsymbol{\theta}\right\|^2, \tag{6}$$

with $\hat{\mathbf{P}}^\perp \coloneqq \mathbf{I} - \hat{\mathbf{P}}$, $\lambda_1 > 0$ and $\lambda_2 > 0$. Besides, $\bar{\boldsymbol{\theta}} \coloneqq \frac{1}{t}\sum_{i=1}^t \boldsymbol{\theta}(i)$ is the mean of the ridge regression estimators of the $t$ previous tasks. We justify the explicit choice of the regularization parameters in the analysis. Problem (6) has a closed form solution given by

$$\hat{\boldsymbol{\theta}}_k = (\mathbf{D}_k^\top\mathbf{D}_k + \lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}})^{-1}\left(\mathbf{D}_k^\top\mathbf{y}_k + \lambda_1\mathbf{w}\right), \tag{7}$$

with $\mathbf{w} \coloneqq \hat{\mathbf{P}}^\perp\bar{\boldsymbol{\theta}}$. The second regularization term in eq. (6) scaling with $\lambda_2$ is necessary so that our closed form solution (7) is well defined i.e., it enables us to determine the inverse of

$$\mathbf{B}_k \coloneqq \mathbf{D}_k^\top\mathbf{D}_k + \lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}}^\perp. \tag{8}$$

The case $\hat{\mathbf{P}} = \mathbf{I}$, which implies that all tasks are highly concentrated around the vector $\mathbf{w} = \bar{\boldsymbol{\theta}}$, would correspond to the setting of Cella et al. (2020).

[2]Throughout the paper, $\|\cdot\|_{\mathbf{A}}$ denotes the weighted norm: $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top\mathbf{A}\mathbf{x}}$.

Action selection is based on the principle of optimism in the face of uncertainty (OFUL), we propose an alternative UCB index by estimating the difference between mean reward $r$ and estimated reward $\hat{r}$:

$$\begin{aligned}|\hat{r} - \mathbb{E}(r|\mathbf{x})| &= |\mathbf{x}^\top(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)| \\ &\le \left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right\|_{\mathbf{B}_k}\|\mathbf{x}\|_{\mathbf{B}_k^{-1}} \le \gamma_k\|\mathbf{x}\|_{\mathbf{B}_k^{-1}},\end{aligned}$$

with $\gamma_k \ge \left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right\|_{\mathbf{B}_k}$. We provide an upper bound on the confidence set of the current estimation $\hat{\boldsymbol{\theta}}_k$ in Section 5.3. The UCB function is then given by

$$\mathrm{UCB}(a) = \mathbf{x}_a^\top\hat{\boldsymbol{\theta}}_k + \gamma_k\|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}}. \tag{9}$$

### 5.3 Analysis

We start by providing a confidence set bound on the current estimation of our task parameter. We make use of an adapted concentration inequality provided by Abbasi-Yadkori et al. (2011) in the following lemma.

**Lemma 1** (Self-normalized bound for vector-valued martingales)**.** *Let $\tau$ be a stopping time with respect to a filtration $\{\mathcal{F}_k\}_{k=1}^\infty$ and define $\boldsymbol{\eta}_k = \mathbf{D}_k^\top\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon}\in\mathbb{R}^k$ as subgaussian noise vector. Then, for every $\delta\in(0,1)$, with probability at least $1-\delta$ we have*

$$\|\boldsymbol{\eta}_k\|^2_{\mathbf{B}_k^{-1}} \le \log\left(\frac{\det(\mathbf{B}_k)}{\delta^2\lambda_1^q\lambda_2^p}\right).$$

To emphasize the dimension of the subspace and the residual, in the lemma below, we bound the ratio of determinants in Lemma 1.

**Lemma 2.** *Let $\lambda_1, \lambda_2 > 0$ and $\mathbf{B}$ be defined as in eq.* (8)*. Then*

$$\log\left(\frac{\det(\mathbf{B}_k)}{\det\left(\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}}\right)}\right) \le S_k^{\lambda_1,\lambda_2} \coloneqq p\log\left(1 + \frac{k}{p\lambda_2}\right) + q\log\left(1 + \frac{k}{q\lambda_1}\right).$$

In the following lemma, we formulate the confidence set bound in our setting.

**Lemma 3.** *At round $k$, and with probability of at least $1-\delta$, the confidence set bound for $\hat{\boldsymbol{\theta}}_k$ is given by*

$$\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right\|_{\mathbf{B}_k} \le \sqrt{S_k^{\lambda_1,\lambda_2} + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda_2}V + \frac{\lambda_1}{\sqrt{\lambda_2}}W,$$

*where $W \coloneqq \left\|\hat{\mathbf{P}}^\perp(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})\right\|$.*

Ideally, we want to show that the confidence set bound is tightened with the knowledge of the shared subspace and the corresponding projection matrix. That can be observed in the regularization terms scaling with $\left\|\hat{\mathbf{P}}^{\perp}\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*}\right)\right\|$, which is small with high probability due to Assumption 1. The second regularization based term scales with $\left\|\hat{\mathbf{P}}\boldsymbol{\theta}^{*}\right\|$ and $\lambda_2$ and guarantees our problem to be well posed. In addition, the choice of $\lambda_2 \ll \lambda_1$ enforces leveraging our assumption on $\rho$.

Before establishing an upper bound for the expected transfer regret, we deliver an error estimation on our projection matrices. For this purpose, we use the eigengap $\Delta_\sigma \coloneqq \sigma_p - \sigma_{p+1}$, which is assumed to be positive, where $p$ is the dimension of the low dimensional subspace. A projection $\mathbf{P}$ depends on the number $p$ of selected eigenvectors, thus it can be assigned a specific eigengap. The following results shows the benefit of large eigengaps.

**Lemma 4.** *Let $\bar{\boldsymbol{\theta}} = \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\theta}(i)$ be the empirical mean of $L_2$ regularized task parameter estimations $\boldsymbol{\theta}(i)$ of true parameters $\boldsymbol{\theta}^{*}(i) \sim \rho$. Assume that each $\boldsymbol{\theta}(i)$ was estimated after the selection of at least $n$ arms. Let $\hat{\mathbf{P}}^{\perp}$ and $\Delta_\sigma > 0$ be the estimation of $\mathbf{P}^{\perp}$ and the eigengap of $\boldsymbol{\Sigma}$, respectively. We have*

$$\mathbb{E}_{\boldsymbol{\theta}^{*}\sim\rho}\left[\mathbb{E}\left[\left\|\hat{\mathbf{P}}^{\perp}\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*}\right)\right\|\right]\right] = \mathcal{O}\left(\sqrt{\mathrm{Var}_\rho + b^2\beta_d^2 + \epsilon_{\boldsymbol{\mu}}^2 + c\epsilon_{\boldsymbol{\Sigma}}^2}\right),$$

*with $\epsilon_{\boldsymbol{\mu}} = \frac{2\log(2t)}{t} + \sqrt{\frac{2\log(2t)\mathrm{Var}_\rho}{t}}$, $\epsilon_{\boldsymbol{\Sigma}}^2 = \frac{C\log(2t)}{t}$, $C$ is an absolute constant, $b = 1 + 64\sqrt{2p}\frac{V^2}{\Delta_\sigma}$, $c = \frac{128pV^2}{\Delta_\sigma^2}$, $\beta_d = \frac{1}{\sqrt{\lambda_{\min}}}\left[\sqrt{d\log\left(1+\frac{n^2V^2}{d}\right)+2} + \sqrt{\frac{1}{n}}\right]$ and $\lambda_{\min}$ is the smallest of the minimal eigenvalues of matrices $\mathbf{A}_n$.*

In what follows, we define

$$Y \coloneqq \mathrm{Var}_\rho + \beta_d^2\left(1 + 64\sqrt{2p}\frac{V^2}{\Delta_\sigma}\right)^2 + \epsilon_{\boldsymbol{\mu}}^2 + \frac{128p\epsilon_{\boldsymbol{\Sigma}}^2V^2}{\Delta_\sigma^2}.$$

The mean concentration error is $\epsilon_{\boldsymbol{\mu}}$ that converges to zero for a sufficiently large number of tasks. Besides, $\epsilon_{\boldsymbol{\Sigma}}$ gives us the concentration error bound of the covariance estimated by the true task parameters $\boldsymbol{\theta}^{*}$ and converges to zero. The bound depends heavily on the eigengap of the true covariance. Larger eigengaps reduce the expected error term and increase the reliability of the projection estimation. For the analysis, we assume $\Delta_\sigma > 0$ for the chosen value of $p$. By assumption, $\mathrm{Var}_\rho$ is relatively small. Thus, the complete term is mostly dominated by $\beta_d^2$, which is an upper bound on the mean squared error (MSE) of the ridge estimator from the standard linUCB case. By selecting $\lambda \sim \frac{1}{n}$, the MSE of the ridge estimator converges to the estimator's variance. This variance, in turn, scales with the subgaussian noise term added on the rewards and also depends on the singular values of the respective data covariance matrix $\mathbf{D}^{\top}\mathbf{D}$. Ideally, we would prefer non-zero singular values. That implies that the set of context vectors yield information along any dimension, which would minimize the variance of the ridge estimator; Nevertheless, our setting does not guarantee this.

We establish an upper bound on the transfer regret in the following theorem.

**Theorem 1.** *Assuming that $\mathbf{P}$ and $\boldsymbol{\mu}$ are known, the expected transfer regret of the projected LinUCB algorithm is upper bounded by*

$$\mathcal{R}(n) = \mathcal{O}\left(\sqrt{n}\left(p\log\left(1+\frac{nV^2}{p}\right) + q\log\left(1+\frac{n\sqrt{\mathrm{Var}_\rho}}{q}\right)\right)\right).$$

*If the assumptions of Lemma 4 hold, the expected transfer regret is upper bounded by*

$$\mathcal{R}(n) = \mathcal{O}\left(\sqrt{n}\left(p\log\left(1+\frac{nV^2}{p}\right) + q\log\left(1+\frac{n\sqrt{Y}}{q}\right)\right)\right).$$

**Remark 2.** *The case $p = d, q = 0$ yields the expected transfer regret $\mathcal{O}\left(\sqrt{n}d\log\left(1+\frac{nV^2}{d}\right)\right)$, when no actual meta learning takes place and each task is learnt independently by the LinUCB algorithm.*

The results show that our approach is at most beneficial when $p$ is as low as possible such that Assumption 1 still holds. In that case, increasing $\lambda_1$ in the algorithm reduces the overall regret bound, further supporting the argument that we made while discussing the confidence set bound in Lemma 3. By setting $\lambda_1 = \frac{1}{\sqrt{Y}}$, the term $S_n^{\lambda_1,\lambda_2}$ defined in Lemma 2 changes such that only the $p$ dependent term becomes relevant as $Y$ significantly decreases and in turn $\log\left(1+\frac{n\sqrt{Y}}{q}\right)$ as well, essentially reducing the effective dimension of the problem to $p$ and indicating that less exploration is required within the $q$-dimensional subspace.

# 6 Projection Meta Learning with Linear Thompson Sampling

## 6.1 Basics of Linear Thompson Sampling

LinUCB and linear Thompson sampling have the same requirements and assumptions concerning the linear relation between expected rewards and context vectors. Their difference lies in the decision-making process: In the former, the learner maximizes a UCB function by selecting the action at every round, whereas in the latter, it utilizes a Gaussian posterior calculated as $\mathcal{N}(\boldsymbol{\theta}_k, v^2\mathbf{A}_k^{-1})$, with $\boldsymbol{\theta}_k$ estimated through solving the regularized least squares as done

in LinUCB. From which, the learner then samples a parameter vector $\tilde{\boldsymbol{\theta}}_k$. It then selects the actions as

$$a = \arg\max \mathbf{x}_a^\top \tilde{\boldsymbol{\theta}}_k.$$

The posterior is built from the prior of the previous instance given by $\mathcal{N}(\boldsymbol{\theta}_{k-1}, v^2 \mathbf{A}_{k-1}^{-1})$. This means that at $k = 1$, during the initialization, we have $\mathbf{A}_0 = \mathbf{I}$. The sampling process reflects the uncertainty of the current estimation $\boldsymbol{\theta}_k$ and directly indicates the exploration behaviour of the learner. A low variance across a specified dimension indicates a high confidence of the current estimation and vice versa. Thus during initialization with $\mathbf{A}_0 = \mathbf{I}$, there is equal exploration potential along any direction.

### 6.2 Thompson Sampling with Linear Payoffs within an Affine Subspace

Our second proposal is a variation of the linear Thompson sampling: We change the posterior from which $\tilde{\boldsymbol{\theta}}$ is sampled. The mean of the new distribution is the biased regularization solution $\hat{\boldsymbol{\theta}}$ of eq. (6), and its covariance matrix is $\mathbf{B}^{-1}$. Thus,

$$\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, v^2 \mathbf{B}^{-1}). \tag{10}$$

In eq. (10), $v$ is a hyper-parameter that we determine in the analysis. During initialization, we have $\mathbf{B}_0 = \lambda_1 \hat{\mathbf{P}}^\perp + \lambda_2 \hat{\mathbf{P}}$ and its inverse as covariance for the prior distribution. By choosing $\lambda_1 \gg \lambda_2$ similar to the projected LinUCB setting, we embed our knowledge of the affine subspace into the prior. That way, the sampling process of $\tilde{\boldsymbol{\theta}}$ incorporates the low variance along the orthogonal subspace.

### 6.3 Analysis

The analysis in inspired by Agrawal and Goyal (2013). First, we define the following two events:

**Definition 1.** *The event $E_r$ occurs if*

$$\forall a \in \mathcal{A}_k : \left| \mathbf{x}_a^\top \left( \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^* \right) \right| \leq l_n \|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}},$$

*with* $l_n \coloneqq \sqrt{2 \log\left(\frac{1}{\delta}\right)(d+2)\log(n) + 2K^2}$ *and* $K \coloneqq \frac{\lambda_2}{\sqrt{\lambda_1}} W + \sqrt{\lambda_1} V$.
*The event $E_{\boldsymbol{\theta}}$ occurs if*

$$\forall a \in \mathcal{A}_k : \left| \mathbf{x}_a^\top \left( \hat{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k \right) \right| \leq \sqrt{2d + 6\log(n)} v \|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}},$$

*with* $v \coloneqq 4\sqrt{\log\left(\frac{1}{\delta}\right)\frac{d+2}{\alpha}}$ *and* $\alpha \in (0,1)$.

The event $E_r$ essentially reflects the confidence set bound discussed for the projected LinUCB case. It gives the probability that our current estimation $\hat{\boldsymbol{\theta}} K$ lies within the bound. The event $E_{\boldsymbol{\theta}}$ is directly linked to the sampling procedure of $\tilde{\boldsymbol{\theta}}_k$. It gives the probability that the reward estimation of the sampled $\tilde{\boldsymbol{\theta}}_k$ is within some limited range of the estimated reward of $\hat{\boldsymbol{\theta}}_k$. Below, we define a filtration, containing all necessary information for the algorithm.

**Definition 2.** *We define the filtration $\{\mathcal{F}_k\}_{k \in \{1,\dots,n\}}$ with sub-$\sigma$-algebras $\mathcal{F}_{k-1}$ at round $k$ generated by the current action set and the history up to round $k-1$: $\mathcal{F}_{k-1} = \{\mathcal{A}_k, \mathcal{H}_{k-1}\}$, with the history being recursively defined as:*

$$\mathcal{H}_k = \{\mathcal{A}_k, \hat{\boldsymbol{\theta}}_k, \mathbf{B}_k, \|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}}, \mathcal{N}(\hat{\boldsymbol{\theta}}_k, \mathbf{B}_k^{-1})\} \cup \mathcal{H}_{k-1}.$$

The next lemma states the probability of the events $E_r$ and $E_{\boldsymbol{\theta}}$.

**Lemma 5.** *For all $\delta \in (0,1)$, the probability of event $E_r$ is bounded from below as follows: $\Pr(E_r) \geq 1 - \frac{\delta}{n^2}$. Moreover, for all possible filtrations $\mathcal{F}_{k-1}$, the probability of event $E_{\boldsymbol{\theta}}$ is bounded from below as follows: $\Pr(E_{\boldsymbol{\theta}} | \mathcal{F}_{k-1}) \geq 1 - \frac{1}{n^2}$.*

In the following theorem, we establish an upper-bound for the transfer regret of the projected Thompson sampling algorithm

**Theorem 2.** *The expected transfer regret of the projected Thompson sampling algorithm verifies*

$$\mathcal{R}(n) = \mathcal{O}\left( \left( d^{\frac{3}{2}} \log(n) + \sqrt{d} \log(n)^2 \right) \sqrt{n S_n^{\frac{1}{\sqrt{Y}}, \frac{1}{V^2}}} \right).$$

**Remark 3.** *With $p = d, q = 0$, the meta learning does not take place, i.e., the agent learns each task independently by the linear TS algorithm. As such, the expected transfer regret yields $\mathcal{O}\left( \left( d^2 \log(n) + d \log(n)^2 \right) \sqrt{n \log\left(1 + \frac{nV^2}{d}\right)} \right)$.*

The results shows a the dependency on the dimensions $p$ and $q$, and the variance related term $Y$. For a sufficiently small $Y$, the terms scaling with $p$ would dominate the regret, so we expect greater improvements with decreasing $p$. The term scaling with $q$ would benefit from the low variance within the respective subspace. As suggested in Agrawal and Goyal (2013), we chose $\alpha = \frac{1}{\log(n)}$ in the proofs.

## 7 Algorithms

The projected LinUCB and projected TS algorithms share many steps. Thus, we unify them and use sub-procedures. We introduce an initialization phase for learning the subspace, as it may only be well-defined after including sufficient task parameters. Enforcing the subspace learning already from the first task might lead to zero-dimensional subspace with $\hat{\mathbf{P}} = \mathbf{0}$ that would degrade the overall performance. In the projected LinUCB algorithm, we require the estimation of $\gamma_k$ taken from Lemma 2, which in turn requires the value of $W = \left\| \hat{\mathbf{P}}^\perp \left( \boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}} \right) \right\|$, which is intractable but we work around this issue by using the simple bound $W \leq 2V$. Since $\gamma_k$ acts more as exploration scaling

**Algorithm 1:** Projected LinUCB/Thompson Sampling

Initialize: $v > 0$, $\delta \in (0,1)$, $\lambda_1 > \lambda_2 > 0$, $\lambda > 0$, $\delta \in (0,1)$;
**for** $t \in \{1, ..., T\}$ **do**
  Initialize: $\mathbf{A}_0 = \lambda \mathbf{I}$, $\mathbf{b}'_0 = \mathbf{0}$;
  Sample new task $\boldsymbol{\theta}^* \sim \rho$;
  **if** $t < d$ **then**
    $\hat{\mathbf{P}} = \mathbf{I}$, $\hat{\mathbf{P}}^{\perp} = \mathbf{0}$, $\mathbf{w} = \mathbf{0}$;
  **else**
    Determine principal components and calculate $\hat{\mathbf{P}}$ and $\hat{\mathbf{P}}^{\perp}$ with $[\boldsymbol{\theta}(i)]_{i \in \{1,...,t\}}$ according to eqs. (2) and (3) and $\mathbf{w} = \frac{1}{t-1}\hat{\mathbf{P}}^{\perp} \sum_{i=1}^{t-1} \boldsymbol{\theta}(i)$;
  Initialize $\mathbf{B}_0 = \lambda_1 \hat{\mathbf{P}}^{\perp} + \lambda_2 \hat{\mathbf{P}}$, $\mathbf{b}_0 = \lambda_1 \hat{\mathbf{P}}^{\perp} \mathbf{w}$, $\hat{\boldsymbol{\theta}}_0 = \mathbf{B}_0^{-1} \mathbf{b}_0$;
  **for** $k \in \{0, ..., n-1\}$ **do**
    Select arm $a_k$ according to respective arm selection strategy (Algorithm 2 or 3);
    Collect immediate reward $r_k$;
    $\mathbf{B}_{k+1} = \mathbf{B}_k + \mathbf{x}_{a_k} \mathbf{x}_{a_k}^{\top}$;
    $\mathbf{A}_{k+1} = \mathbf{A}_k + \mathbf{x}_{a_k} \mathbf{x}_{a_k}^{\top}$;
    $\mathbf{b}_{k+1} = \mathbf{b}_k + r_k \mathbf{x}_{a_k}$;
    $\mathbf{b}'_{k+1} = \mathbf{b}'_k + r_k \mathbf{x}_{a_k}$;
    $\hat{\boldsymbol{\theta}}_{k+1} = \mathbf{B}_{k+1}^{-1} \mathbf{b}_{k+1}$;
  $\boldsymbol{\theta}(t) = \mathbf{A}_n^{-1} \mathbf{b}'_n$;

**Algorithm 2:** Projected LinUCB Arm Selection Routine

Input: $\hat{\boldsymbol{\theta}}_k$, $\mathbf{B}_k$;
$\gamma_k = \log\left(\frac{\det(\mathbf{B}_k)}{\delta^2 \lambda_1^q \lambda_2^p}\right) + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W$;
Select arm $a_k = \arg\max_a \mathrm{UCB}(a)$ from (9);

**Algorithm 3:** Projected TS arm selection routine

Input: $\hat{\boldsymbol{\theta}}_k$, $\mathbf{B}_k$, $\alpha \in (0,1)$;
$v = 4\sqrt{\log\left(\frac{1}{\delta}\right) \frac{d+2}{\alpha}}$;
sample $\tilde{\boldsymbol{\theta}}_k \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_k, v^2 \mathbf{B}_k^{-1})$;
Select arm $a_k = \arg\max_a \mathbf{x}_a^{\top} \tilde{\boldsymbol{\theta}}_k$;

factor, we do not lose any benefit from the meta learning as the actual knowledge transfer becomes relevant in the calculations of $\mathbf{B}_k$ and $\hat{\boldsymbol{\theta}}_k$

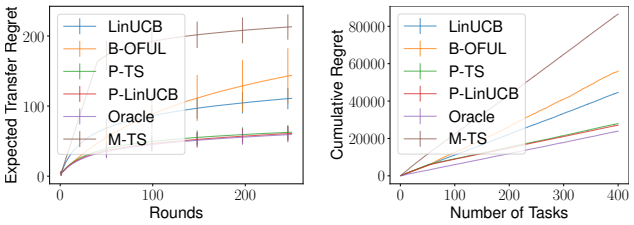
## 8 Numerical Experiments

We test our algorithms experimentally on synthetic data and on real world data taken from the MovieLens data set.

### 8.1 Synthetic Data Experiments

We sampled the context vectors from a zero mean normal distribution with a diagonal covariance matrix whose elements followed a uniform distribution. Following Mezzadri (2006), we used a randomly generated orthogonal matrix to define a subspace. We project the randomly generated task parameters onto the subspace and add a multivariate Gaussian noise term in the orthogonal direction to the given subspace to simulate the variance of the task distribution. One drawback of this approach is that it misses the benefits of subspace learning during the first tasks. That is because a subspace with dimension $p$ that ought to be learned requires at least $q = d - p$ data points or task parameters to use the PCA algorithms successfully. Thus, we also implement an initialization phase to prevent subspace learning until learning at least $d$ task parameters. Note that we require at least $d$ tasks as we do not use our knowledge of $p$. We consider a task as finished after at least $n = 250$ rounds.

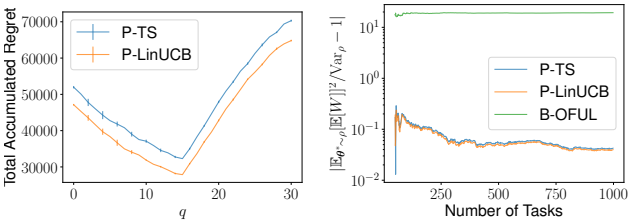
Figure 2a shows the expected transfer regret for $d = 30$ and $p = 15$, with the oracle and the algorithms of Cella et al. (2020) (B-OFUL), Peleg et al. (2022) (M-TS) as benchmarks. The projected Thompson sampling (P-TS) approach performs as well as projected LinUCB (P-LinUCB), while the oracle using the true projection and mean is the most efficient one. Our algorithms are significant improvements the other baselines. The superiority of our approach mainly stems from its generality compared to Cella et al. (2020). The algorithm provided by Peleg et al. (2022) has the worst performance mainly due to the regret contributed by the required forced exploration within a task. Additionally, Figure 2b shows the total cumulative regret over tasks, which does not suffer from overlapping error bars as they become negligible. To further emphasize the benefit of exploiting the knowledge on any dimensional low variance, Figure 3a shows the total accumulated regret of the projected LinUCB algorithm after $T$ tasks with $n$ rounds of learning each as a function of $q = \mathrm{rank}(\hat{\mathbf{P}}^{\perp})$. Note that




(a) Synthetic data plots of the expected transfer regret as a function of number of rounds.

(b) Synthetic data plots of the cumulative regret over the number of tasks.

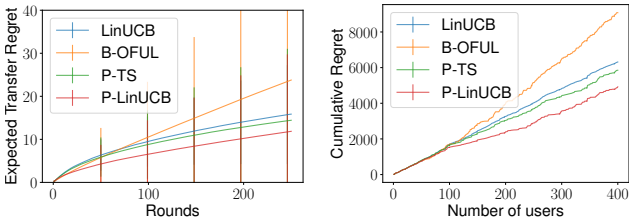
Figure 2: Expected transfer and total cumulative regret plots of the LinUCB and Thompson sampling methods compared to their projection counterparts and additional baselines.




(a) Total accumulated regret after 400 tasks as a function of $\text{rank}(\hat{\mathbf{P}}^{\perp})$.

(b) Relative error of $\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho}[\mathbb{E}[W]]^2$ as a function of number of tasks in a logarithmic scale.

Figure 3




(a) Real data plots of the expected transfer regret as a function of number of rounds.

(b) Real data plots of the cumulative regret over tasks/users.

Figure 4: Expected transfer regret and total regret plots of our algorithms and baselines applied to the MovieLens data set. We have included 400 users in the simulations.

at $q = 0$, the plot shows the total regret using classic LinUCB. As expected, the regret reaches its minimum when $q$ is equal to the rank of the true projection, which is $q = 15$ in this case. Nevertheless even for different values of $q$, there is a clear benefit over the classic approach. In Figure 3b we plot $\left|\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho}[\mathbb{E}[W]]^2 / \text{Var}_\rho - 1\right|$ as a function of number of tasks. As expected the curves for P-LinUCB and P-TS imply that $\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho}[\mathbb{E}[W]]^2$ is close to $\text{Var}_\rho$. The curve for the B-OFUL algorithm assumes $\mathbf{P}^{\perp} = \mathbf{I}$, disregarding the covariance and thus resulting into higher values. Note that lower values imply greater transfer in between sequential tasks as the projection matrix would be well estimated.

### 8.2 Real Data Experiments

We use MovieLens data to test our algorithms in a real-world environment. MovieLens data contains information about over 6000 users that represent the tasks in our setting. Besides, it includes over 3000 movies, which are the arms with their corresponding context vectors. The context vectors are 18-dimensional, each denoting a possible genre. If a movie has a label for a specific genre, the corresponding entry for that genre in the context vector is 1. With at most six different genres assigned to a single movie, we normalize the context vectors such that we have $\|\mathbf{x}_a\| \leq 1$. Each movie has some available ratings between 1 and 5, given by a user who has watched that movie. Each rating represents a reward for our algorithm. We normalize all such ratings so that $r \in [0, 1]$. We further process the data by grouping the users by their profession or gender and run the algorithm within that set of users. That method stems from the assumption that groups of similar users might share an affine subspace. For every user (task), we run the algorithms for at least $n = 250$ rounds. We do not include the algorithm developed in Peleg et al. (2022) as baseline for experimentation using the real data set as it requires contexts from a distribution with an invertible covariance. The reason is that the authors do not use a regularizer on the minimum least squares solution for $\boldsymbol{\theta}$, and thus find the inverse of $\mathbf{D}^\top \mathbf{D}$ after every finished task. However, in the MovieLens data set, that condition does not hold for many users, making the estimation of $\boldsymbol{\theta}$ ill-posed. In the real data experiments, we observe significant improvements of our models over the baselines in Figure 4a showcasing the expected transfer regret, and in Figure 4b, showcasing the total cumulative regret over users, which does not suffer from overlapping error bars. A significant point is that we did not perform data preprocessing besides normalizing the rewards and dividing the users into male and female. That would also explain the performance gap to the algorithm of Cella et al. (2020), as our assumption is more general and widely applicable.

## 9 Discussion and Outlook

Our work shows that obtaining knowledge about the underlying subspace structure in a meta-learning setting improves sequential task learning. More precisely, assuming a low variance along certain dimensions in the task distribution, we proposed two decision-making policies that exploit the knowledge of the subspace structure for sequential arm selection and significantly improve the performance of widely used algorithms, namely LinUCB and linear Thompson sampling. We provided an improved regret bound that manifests the dependency on the lower dimension, the low variance term, and the eigengap at the considered low dimension. We evaluated our methods numerically through experimentations on synthetic and real-world datasets, confirming their better performance than traditional benchmarks. The results are significant in the real data environments as the rewards do not necessarily follow a linear relation.

Possible extensions of this work include further generalization of our model by learning the variance of the task distribution along all dimensions. Another direction is to generalize our methods to non-linear settings, *i.e.*, when tasks concentrate around a low dimensional manifold.

## Acknowledgements

This work was supported by Grant 01IS20051 from the German Federal Ministry of Education and Research (BMBF). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Steven Bilaj.

## References

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

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*.

Aiolli, F. (2012). Transfer learning by kernel meta-learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*.

Amit, R. and Meir, R. (2018). Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *Proceedings of the 35th International Conference on Machine Learning*.

Atan, O., Ghoorchian, S., Maghsudi, S., and van der Schaar, M. (2023). Data-driven online recommender systems with costly information acquisition. *IEEE Trans. Serv. Comput.*

Azar, M., Lazaric, A., and Brunskill, E. (2013). Sequential transfer in multi-armed bandit with finite set of models. *Advances in Neural Information Processing Systems*.

Azizi, M., Duong, T., Abbasi-Yadkori, Y., György, A., Vernade, C., and Ghavamzadeh, M. (2022). Non-stationary bandits and meta-learning with a small set of optimal arms. *arXiv preprint arXiv:2202.13001*.

Balcan, M.-F., Harris, K., Khodak, M., and Wu, Z. S. (2022). Meta-learning adversarial bandits. *arXiv preprint arXiv:2205.14128*.

Bastani, H., Simchi-Levi, D., and Zhu, R. (2019). Meta Dynamic Pricing: Transfer Learning Across Experiments.

Basu, S., Kveton, B., Zaheer, M., and Szepesvari, C. (2021). No Regrets for Learning the Prior in Bandits. In *Advances in Neural Information Processing Systems*.

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*.

Bilaj, S., Dhouib, S., and Maghsudi, S. (2023). Hypothesis transfer in bandits by weighted models. In *Machine Learning and Knowledge Discovery in Databases*.

Bouneffouf, D., Rish, I., and Aggarwal, C. (2020). Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*.

Boutilier, C., Hsu, C.-W., Kveton, B., Mladenov, M., Szepesvári, C., and Zaheer, M. (2020). Differentiable meta-learning of bandit policies. *Advances in Neural Information Processing Systems*.

Cardot, H. and Degras, D. (2015). Online principal component analysis in high dimension: Which algorithm to choose?

Cella, L., Lazaric, A., and Pontil, M. (2020). Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*. PMLR.

Cella, L., Lounici, K., and Pontil, M. (2022a). Meta representation learning with contextual linear bandits. *arXiv preprint arXiv:2205.15100*.

Cella, L., Lounici, K., and Pontil, M. (2022b). Multi-task representation learning with stochastic linear bandits.

Cella, L. and Pontil, M. (2021). Multi-task and meta-learning with sparse linear bandits. In *Uncertainty in Artificial Intelligence*. PMLR.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *AISTATS*.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. (2018). Learning to learn around a common mean. *Advances in Neural Information Processing Systems*.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019). Online meta-learning. In *International Conference on Machine Learning*. PMLR.

Glowacka, D. et al. (2019). Bandit algorithms in information retrieval. *Foundations and Trends® in Information Retrieval*.

Hanson, D. L. and Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*.

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Springer Nature.

Jiang, W., Kwok, J., and Zhang, Y. (2022). Subspace learning for effective meta-learning. In *International Conference on Machine Learning*. PMLR.

Kassraie, P., Rothfuss, J., and Krause, A. (2022). Meta-Learning Hypothesis Spaces for Sequential Decision-making. *ArXiv*.

Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-w., Mladenov, M., Boutilier, C., and Szepesvari, C. (2021). Meta-thompson sampling. In *International Conference on Machine Learning*. PMLR.

Kveton, B., Mladenov, M., Hsu, C.-W., Zaheer, M., Szepesvari, C., and Boutilier, C. (2020). Meta-learning bandit policies by gradient ascent. *arXiv e-prints*.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*.

Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*.

Mezzadri, F. (2006). How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*.

Nourani-Koliji, B., Ghoorchian, S., and Maghsudi, S. (2022). Linear combinatorial semi-bandit with causally related rewards. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*.

Peleg, A., Pearl, N., and Meir, R. (2022). Metalearning Linear Bandits by Prior Update. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*.

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. (2021). Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*. PMLR.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*.

Schur, F., Kassraie, P., Rothfuss, J., and Krause, A. (2022). Lifelong Bandit Optimization: No Prior and No Regret.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*.

Thrun, S. (1998). Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. (2020a). Impact of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*.

Yang, K. and Toni, L. (2020). Differentiable linear bandit algorithm. *arXiv preprint arXiv:2006.03000*.

Yang, K., Toni, L., and Dong, X. (2020b). Laplacian-regularized graph bandits: Algorithms and theoretical analysis. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis—kahan theorem for statisticians. *Biometrika*.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# A Notations

| Notation | Meaning |
|---|---|
| $a$, $a^*$ | Arm and optimal arm yielding highest mean reward respectively |
| $\mathcal{A}_k$ | Set of available arms |
| $\mathbf{x}_a$ | Context vector associated with arm $a$ |
| $d$ | Dimension of the context vectors |
| $n$ | Horizon |
| $r_k$ | Immediate reward at round $k$ |
| $\epsilon_k$ | Subgaussian noise added to the reward at round $k$ |
| $\mathbf{D}_k$ | Vertical concatenation of up to round $k$ collected context vectors $\mathbf{x}_a^\top$ |
| $\mathbf{y}_k$ | Concatenation of up to round $k$ collected rewards |
| $\gamma_k$ | Upper confidence set bound of LinUCB or projected LinUCB algorithm in round $k$ |
| $v$ | Scaling factor for the covariance of the Thompson Sampling posterior |
| $\lambda$, $\lambda_1$, $\lambda_2$ | Regularization parameters for ridge and projection based estimators |
| $\rho$ | Task distribution |
| $\mathbf{\Sigma}$ | True covariance of $\rho$ |
| $\{\sigma_j\}_{j\in\{1,\dots,d\}}$ | Eigenvalues of $\mathbf{\Sigma}$ |
| $\Delta\sigma$ | Eigengap of $\mathbf{\Sigma}$ |
| $\mathbf{P}$, $\hat{\mathbf{P}}$ | True subspace projection and its estimation respectively |
| $\mathbf{P}^\perp$, $\hat{\mathbf{P}}^\perp$ | $\mathbf{I}-\mathbf{P}$ and $\mathbf{I}-\hat{\mathbf{P}}$ respectively |
| $S_k^{\lambda_1,\lambda_2}$ | $p\log\left(1+\frac{k}{p\lambda_2}\right)+q\log\left(1+\frac{k}{q\lambda_1}\right)$ |
| $\boldsymbol{\theta}^*$, $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ | True task parameter, its ridge estimator and projections based estimator respectively |
| $\bar{\boldsymbol{\theta}}$ | Mean of $t$ collected ridge estimations of true task parameters: $\frac{1}{t}\sum_i^t \boldsymbol{\theta}(i)$ |
| $p$ | Rank of $\hat{\mathbf{P}}$ |
| $q$ | Rank of $\hat{\mathbf{P}}^\perp$ |
| $\mathbf{w}$ | $\hat{\mathbf{P}}^\perp\bar{\boldsymbol{\theta}}$ |
| $\mathbf{A}_k$ | $\lambda\mathbf{I}+\mathbf{D}_k^\top\mathbf{D}_k$ |
| $\mathbf{B}_k$ | $\lambda_1\hat{\mathbf{P}}^\perp+\lambda_2\hat{\mathbf{P}}+\mathbf{D}_k^\top\mathbf{D}_k$ |
| $\mathbf{b}'_k$ | $\mathbf{D}_k^\top\mathbf{y}_k$ |
| $\mathbf{b}_k$ | $\mathbf{D}_k^\top\mathbf{y}_k+\lambda_1\hat{\mathbf{P}}^\perp\bar{\boldsymbol{\theta}}$ |
| $V$ | Upper Bound on the norm of any true task parameter |
| $W$ | $\left\|\hat{\mathbf{P}}^\perp(\boldsymbol{\theta}^*-\bar{\boldsymbol{\theta}})\right\|$ |
| $Y$ | Upper bound of $\mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\mathbb{E}\left[\left\|\hat{\mathbf{P}}^\perp\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|\right]\right]^2$ |
| $\alpha$ | Hyper parameter of Thompson Sampling algorithm |
| $l_n$ | $\sqrt{2\log\left(\frac{1}{\delta}\right)(d+2)\log(n)+2K^2}$ |
| $g_n$ | $\sqrt{2d+6\log(n)}v+l_n$ |
| $\mathcal{R}$ | Expected transfer regret |
| $\|\cdot\|$ | Euclidean norm |
| $\|\cdot\|_{\mathbf{A}}$ | Weighted norm: $\|\mathbf{x}\|_{\mathbf{A}}=\sqrt{\mathbf{x}^\top\mathbf{A}\mathbf{x}}$ |

Table 1: Table of Notations

## B Proof of Theorem 1

In order to prove Theorem 1 in the main paper we provide proofs of additional Lemmas here or refer to the original works:

*Proof of Lemma 1.* Given Lemma 9 of Abbasi-Yadkori et al. (2011), we have:

$$\|\boldsymbol{\eta}_k\|^2_{\mathbf{B}_k^{-1}} \leq \log\left(\frac{\det(\mathbf{B}_k)}{\delta^2 \det\left(\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}}\right)}\right), \tag{11}$$

where the term $\det\left(\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}}\right)$ can be further evaluated knowing the eigenvalues of the matrix $\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}}$. With orthogonal projections $\hat{\mathbf{P}}$ and $\hat{\mathbf{P}}^\perp$ and $\hat{\mathbf{P}}^\perp = \mathbf{I} - \hat{\mathbf{P}}$ it holds that for any eigenvector $\mathbf{e_P}$ of $\hat{\mathbf{P}}$ we have: $\hat{\mathbf{P}}^\perp\mathbf{e_P} = (\mathbf{I} - \hat{\mathbf{P}})\mathbf{e_P} = \mathbf{0}$ and vice versa for any eigenvector $\mathbf{e}_{\mathbf{P}^\perp}$ of $\hat{\mathbf{P}}^\perp$: $\hat{\mathbf{P}}\mathbf{e}_{\mathbf{P}^\perp} = \mathbf{0}$. Thus any eigenvector of $\hat{\mathbf{P}}$ or $\hat{\mathbf{P}}^\perp$ is also an eigenvector of $\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}}$:

$$\begin{aligned}(\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}})\mathbf{e_P} &= (0 + \lambda_2)\mathbf{e_P},\\ (\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}})\mathbf{e}_{\mathbf{P}^\perp} &= (\lambda_1 + 0)\mathbf{e}_{\mathbf{P}^\perp},\end{aligned}$$

with eigenvalues $\lambda_1$ and $\lambda_2$. Lastly we require the multiplicities of both eigenvalues given by the dimension of nullspaces of the matrices $\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}} - \lambda_1\mathbf{I} = (\lambda_2 - \lambda_1)\hat{\mathbf{P}}$ for $\lambda_1$ and $\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}} - \lambda_2\mathbf{I} = (\lambda_1 - \lambda_2)\hat{\mathbf{P}}^\perp$ for $\lambda_2$, which are $q = \operatorname{rank}(\hat{\mathbf{P}}^\perp)$ and $p = \operatorname{rank}(\hat{\mathbf{P}})$ respectively. Thus we get:

$$\det\left(\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}}\right) = \lambda_1^q\lambda_2^p, \tag{12}$$

finalizing our proof. □

*Proof of Lemma 2.* Let $\lambda_i'$ be the singular values of $\mathbf{D}^\top\mathbf{D}$ and $\|\mathbf{x}_a\| \leq 1$ then we have:

$$\begin{aligned}\log\left(\frac{\det(\mathbf{B})}{\det\left(\lambda_1\hat{\mathbf{P}}^\perp + \lambda_2\hat{\mathbf{P}}\right)}\right) &\leq \sum_{i=1}^{p}\log\left(1 + \frac{\lambda_i'}{\lambda_2}\right) + \sum_{i=p+1}^{d}\log\left(1 + \frac{\lambda_i'}{\lambda_1}\right)\\ &\leq p\log\left(1 + \frac{1}{p\lambda_2}\sum_{i=1}^{p}\lambda_i'\right) + q\log\left(1 + \frac{1}{q\lambda_1}\sum_{i=p+1}^{d}\lambda_i'\right)\\ &\leq p\log\left(1 + \frac{k}{p\lambda_2}\right) + q\log\left(1 + \frac{k}{q\lambda_1}\right)\end{aligned}$$

where we applied the Jensen inequality in the second inequality and bounded the trace by $k\|\mathbf{x}_{a_k}\|^2 \leq k$ in the last inequality. □

*Proof of Lemma 3.* We leave out the subscript $k$ during the proof for readability purposes. Our estimation of $\boldsymbol{\theta}^*$ for the projected LinUCB algorithm yields:

$$\hat{\boldsymbol{\theta}} = \mathbf{B}^{-1}\left(\mathbf{D}^\top\mathbf{y} + \lambda_1\mathbf{w}\right), \tag{13}$$

thus we can write:

$$\begin{aligned}
\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_{\mathbf{B}} &= \left\|\mathbf{B}^{-1}\left(\mathbf{D}^\top\mathbf{y}+\lambda_1\mathbf{w}\right)-\boldsymbol{\theta}^*\right\|_{\mathbf{B}} \\
&= \left\|\mathbf{B}^{-1}\left(\mathbf{D}^\top(\mathbf{D}\boldsymbol{\theta}^*+\boldsymbol{\epsilon})+\lambda_1\mathbf{w}\right)-\boldsymbol{\theta}^*\right\|_{\mathbf{B}} \\
&= \left\|\mathbf{B}^{-1}\left(\mathbf{D}^\top\boldsymbol{\epsilon}+\lambda_1\mathbf{w}\right)-\mathbf{B}^{-1}\left(\lambda_1\hat{\mathbf{P}}^\perp+\lambda_2\hat{\mathbf{P}}\right)\boldsymbol{\theta}^*\right\|_{\mathbf{B}} \\
&= \left\|\mathbf{D}^\top\boldsymbol{\epsilon}+\lambda_1\mathbf{w}-\left(\lambda_1\hat{\mathbf{P}}^\perp+\lambda_2\hat{\mathbf{P}}\right)\boldsymbol{\theta}^*\right\|_{\mathbf{B}^{-1}} \\
&\leq \left\|\mathbf{D}^\top\boldsymbol{\epsilon}\right\|_{\mathbf{B}^{-1}}+\lambda_1\left\|\hat{\mathbf{P}}^\perp\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|_{\mathbf{B}^{-1}}+\lambda_2\left\|\hat{\mathbf{P}}\boldsymbol{\theta}^*\right\|_{\mathbf{B}^{-1}} \\
&\leq \sqrt{2\log\left(\frac{\sqrt{\det(\mathbf{B})}}{\delta\sqrt{\det\left(\lambda_1\hat{\mathbf{P}}^\perp+\lambda_2\hat{\mathbf{P}}\right)}}\right)}+\frac{\lambda_1}{\lambda_{\min}(\mathbf{B})}\left\|\hat{\mathbf{P}}^\perp\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|+\frac{\lambda_2}{\lambda_{\min}(\mathbf{B})}\left\|\hat{\mathbf{P}}\boldsymbol{\theta}^*\right\| \\
&\leq \sqrt{2\log\left(\frac{\sqrt{\det(\mathbf{B})}}{\delta\sqrt{\det\left(\lambda_1\hat{\mathbf{P}}^\perp+\lambda_2\hat{\mathbf{P}}\right)}}\right)}+\sqrt{\lambda_2}V+\frac{\lambda_1}{\sqrt{\lambda_2}}W \\
&\leq \sqrt{p\log\left(1+\frac{k}{p\lambda_2}\right)+q\log\left(1+\frac{k}{q\lambda_1}\right)+\log\left(\frac{1}{\delta^2}\right)}+\sqrt{\lambda_2}V+\frac{\lambda_1}{\sqrt{\lambda_2}}W,
\end{aligned}$$

where we used Lemma 1 in the second inequality. Here, $\lambda_{\min}()$ is a function returning the minimal eigenvalue of a given matrix. □

For the upper bound on the projection based error term in Lemma 4, need to make some definitions: We denote $\boldsymbol{\mu}$ as the true mean of the distribution of tasks $\rho$, $\bar{\boldsymbol{\theta}}^* = \frac{1}{t}\sum_{i=1}^t \boldsymbol{\theta}^*(i)$ as the mean estimated by the true task parameters and $\bar{\boldsymbol{\theta}} = \frac{1}{t}\sum_{i=1}^t \boldsymbol{\theta}(i)$ as the mean estimated by the $L2-$regularized ridge estimators. We define $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_d$ as ordered eigenvalues of the true covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{\Delta} = \boldsymbol{\theta}^* - \boldsymbol{\mu}$ as random variable with mean zero and $\boldsymbol{\xi} = \boldsymbol{\mu} - \bar{\boldsymbol{\theta}}$ as difference between the estimated and true mean, furthermore we define the covariance matrices $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^* = \frac{1}{t}\sum_{i=1}^t(\boldsymbol{\theta}^*(i) - \bar{\boldsymbol{\theta}}^*)(\boldsymbol{\theta}^*(i) - \bar{\boldsymbol{\theta}}^*)^\top, \hat{\boldsymbol{\Sigma}} = \frac{1}{t}\sum_{i=1}^t(\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}})^\top$ as the true covariance matrix, the covariance estimated by $\boldsymbol{\theta}^*(i)$ and the covariance estimated by $\hat{\boldsymbol{\theta}}(i)$. We also define vertical concatenations $\mathbf{U} = [\mathbf{u}_j^\top]_{j\in\{1,...,p\}}^\top$, $\mathbf{U}^* = [\mathbf{u}_j^{*\top}]_{j\in\{1,...,p\}}^\top$ and $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_j^\top]_{j\in\{1,...,p\}}^\top$, with $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$, $\{\mathbf{u}_1^*, ..., \mathbf{u}_p^*\}$, $\{\hat{\mathbf{u}}_1, ..., \hat{\mathbf{u}}_p\}$ being the eigenvectors corresponding to the $p$ largest eigenvalues of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^*$ and $\hat{\boldsymbol{\Sigma}}$ respectively. Similarly we define $\mathbf{P} = \mathbf{U}\mathbf{U}^\top, \mathbf{P}^* = \mathbf{U}^*\mathbf{U}^{*\top}, \hat{\mathbf{P}} = \hat{\mathbf{U}}\hat{\mathbf{U}}^\top$ as the true projection, the projection estimated by the true task parameters $\boldsymbol{\theta}^*(i)$ and the projection estimated by $\boldsymbol{\theta}(i)$ respectively. For the following parts we need to define the matrix norms: We denote the matrix norm $\|\cdot\|$ as the spectral norm and $\|\cdot\|_F$ as the Frobenius norm. We also require some auxiliary Lemmas:

**Lemma 6** (Smale and Zhou (2007)). *Let $\boldsymbol{\theta}^*(1), ..., \boldsymbol{\theta}(t)^* \in \mathbb{R}^d$ be vector valued random variables sampled from a distribution $\rho$ with true mean $\boldsymbol{\mu}$ and $\|\boldsymbol{\theta}^*(i)\| \leq V$, $\forall i \in \{1, ..., t\}$. Then the following holds with probability $1-\delta$:*

$$\left\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\mu}\right\| \leq \frac{2\log\left(\frac{2}{\delta}\right)V}{t} + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)\mathrm{Var}_{\max}}{t}},$$

*with $\mathrm{Var}_{\max} = \mathbb{E}\left[\|\boldsymbol{\Delta}\|^2\right] = \mathrm{tr}(\boldsymbol{\Sigma})$ as the total variance of distribution $\rho$.*

**Lemma 7** (Corollary 5.50 of Vershynin (2012)). *Consider a subgaussian distribution in $\mathbb{R}^d$ with true covariance $\boldsymbol{\Sigma}$ and the covariance $\boldsymbol{\Sigma}^*$ estimated from $t$ samples as it was defined above. Let $\delta \in (0,1)$, then we have with probability $1-\delta$:*

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^*\| \leq \sqrt{C\frac{\log(2/\delta)}{t}},$$

*with $C$ as an absolute constant.*

**Lemma 8** (Theorem 2 in Yu et al. (2015)). *Let $\mathbf{\Sigma}, \hat{\mathbf{\Sigma}} \in \mathbb{R}^{d\times d}$ be two symmetric matrices with eigenvalues $\sigma_1 \geq ... \geq \sigma_d$ and $\hat{\sigma}_1 \geq ... \geq \hat{\sigma}_d$ respectively. Fix $1 \leq p \leq d$ and let $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_p]$ and $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1, ..., \hat{\mathbf{u}}_p]$ with eigenvectors $\mathbf{u}_i$ and $\hat{\mathbf{u}}_i$ of matrices $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ respectively. Assume that the eigengap satisfies $\Delta_\sigma = \sigma_p - \sigma_{p+1} > 0$, then there exists an orthogonal matrix $\mathbf{O}$ such that the the following holds:*

$$\left\|\mathbf{U} - \hat{\mathbf{U}}\mathbf{O}\right\|_F \leq \frac{\sqrt{8}\min\left(\sqrt{p}\left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\right\|, \left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\right\|_F\right)}{\Delta_\sigma},$$

*Proof of Lemma 4.* For the proof we will use the triangular inequality to express the bound in terms of the true variance along the orthogonal subspace, the projected mean estimation error and the projection estimation error. For the mean estimation error we apply an additional triangular inequality in order to estimate it with respect to the true mean estimation error $\left\|\mathbf{P}^\perp(\boldsymbol{\mu} - \bar{\boldsymbol{\theta}}^*)\right\|$ and the error $\left\|\bar{\boldsymbol{\theta}}^* - \bar{\boldsymbol{\theta}}\right\|$, with the former being a simple concentration bound and the latter being estimated from the oracle inequality for $\boldsymbol{\theta}$. We intend to express the projection error with respect to the estimation error on the covariance matrix. Bounding the term $\left\|\mathbf{P} - \hat{\mathbf{P}}\right\|$ requires the Davis-Kahan Theorem. Thus we begin the proof:

$$\begin{aligned}
\left\|\hat{\mathbf{P}}^\perp(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})\right\| &\leq \left\|\hat{\mathbf{P}}^\perp - \mathbf{P}^\perp\right\|\left\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}\right\| + \left\|\mathbf{P}^\perp(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})\right\| && (14)\\
&\leq \left\|\hat{\mathbf{P}}^\perp - \mathbf{P}^\perp\right\|\left\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}\right\| + \left\|\mathbf{P}^\perp\boldsymbol{\Delta}\right\| + \left\|\mathbf{P}^\perp\boldsymbol{\xi}\right\| && (15)\\
&\leq 2V\left\|\hat{\mathbf{P}} - \mathbf{P}\right\| + \left\|\mathbf{P}^\perp\boldsymbol{\Delta}\right\| + \left\|\mathbf{P}^\perp\boldsymbol{\xi}\right\|, && (16)
\end{aligned}$$

where we used $\mathbf{P}_i^\perp - \mathbf{P}_j^\perp = \mathbf{P}_i - \mathbf{P}_j$ for all projection matrices $\mathbf{P}_i, \mathbf{P}_j$ in the last inequality. We deliver an upper bound on all of these terms separately: The second term being straight forward with $\mathbf{P}^\perp$ as the true orthogonal projection:

$$\mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\mathbf{P}^\perp\boldsymbol{\Delta}\right\|^2\right] = \mathrm{Var}_\rho, \tag{17}$$

with $\mathrm{Var}_\rho$ denoting the low variance of distribution $\rho$ along the orthogonal subspace. This holds simply due to our problem setting.

The last term yields the mean estimation error of tasks along the orthogonal subspace which was similarly discussed in Cella et al. (2020):

$$\left\|\mathbf{P}^\perp(\boldsymbol{\mu} - \bar{\boldsymbol{\theta}})\right\| \leq \left\|\mathbf{P}^\perp(\boldsymbol{\mu} - \bar{\boldsymbol{\theta}}^*)\right\| + \left\|(\bar{\boldsymbol{\theta}}^* - \bar{\boldsymbol{\theta}})\right\| \tag{18}$$

The first term can simply be bounded by a concentration inequality which was also discussed in Lemma 3 of Cella et al. (2020) by using Lemma 6 we state with probability of $1 - \delta$ that the following holds:

$$\left\|\mathbf{P}^\perp(\boldsymbol{\mu} - \bar{\boldsymbol{\theta}}^*)\right\| \leq \frac{2\log\left(\frac{2}{\delta}\right)}{t} + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)\mathrm{Var}_\rho}{t}}.$$

By choosing $\delta = 1/t$ and taking the expectation value with respect to the task distribution, we have:

$$\mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\mathbf{P}^\perp\left(\boldsymbol{\mu} - \bar{\boldsymbol{\theta}}^*\right)\right\|\right] = \mathcal{O}\left(\frac{2\log(2t)}{t} + \sqrt{\frac{2\log(2t)\mathrm{Var}_\rho}{t}}\right).$$

We will denote $\epsilon_{\boldsymbol{\mu}} \coloneqq \frac{2\log(2t)}{t} + \sqrt{\frac{2\log(2t)\mathrm{Var}_\rho}{t}}$ for the rest of the proof As for the second term in eq. (18), we assume that all previously learnt tasks were running for at least $n$ rounds and use the subscript $i \in \{1, ..., t\}$ to refer to a given task:

$$\left\|\bar{\boldsymbol{\theta}}^* - \bar{\boldsymbol{\theta}}\right\| \leq \max_i \|\boldsymbol{\theta}(i)^* - \boldsymbol{\theta}(i)\| \tag{19}$$

$$\leq \max_i \frac{\|\boldsymbol{\theta}^*(i) - \boldsymbol{\theta}(i)\|_{\mathbf{A}_n(i)}}{\sqrt{\lambda_{\min}(\mathbf{A}_n(i))}} \tag{20}$$

$$\leq \frac{1}{\sqrt{\log(n)}} \left( \sqrt{d \log\left(1 + \frac{n}{d\lambda}\right) + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda} V \right), \tag{21}$$

where we used a linear regression result $\lambda_{\min}(\mathbf{A}_n) \geq \log(n)$ from Lai and Wei (1982). For the most general case, we will keep $\lambda_{\min} = \min_i \lambda_{\min}(\mathbf{A}_n(i))$ Choosing $\delta = 1/n$, $\lambda = \frac{1}{nV^2}$ and taking the expectation with respect to the arm selection process yields:

$$\mathbb{E}\left[\left\|\bar{\boldsymbol{\theta}}^* - \bar{\boldsymbol{\theta}}\right\|\right] \leq \mathcal{O}\left( \frac{1}{\lambda_{\min}} \sqrt{d \log\left(1 + \frac{n^2 V^2}{d}\right) + 2} + \sqrt{\frac{1}{n}} \right). \tag{22}$$

We denote $\beta_d \coloneqq \frac{1}{\lambda_{\min}} \sqrt{d \log\left(1 + \frac{n^2 V^2}{d}\right) + 2} + \sqrt{\frac{1}{n}}$ We note that this upper bound is independent from the task distribution.

What is left is to upper bound the term $\left\|\mathbf{P} - \hat{\mathbf{P}}\right\|$:

$$\begin{aligned}
\left\|\mathbf{P} - \hat{\mathbf{P}}\right\| &= \left\|\mathbf{U}\mathbf{U}^\top - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\right\| \\
&= \left\|\mathbf{U}\mathbf{U}^\top - \hat{\mathbf{U}}\mathbf{O}\mathbf{O}^\top\hat{\mathbf{U}}^\top\right\| \\
&= \left\|\mathbf{U}\mathbf{U}^\top + \hat{\mathbf{U}}\mathbf{O}\mathbf{U}^\top - \hat{\mathbf{U}}\mathbf{O}\mathbf{U}^\top - \hat{\mathbf{U}}\mathbf{O}\mathbf{O}^\top\hat{\mathbf{U}}^\top\right\| \\
&= \left\|\hat{\mathbf{U}}\mathbf{O}(\mathbf{U}^\top - \mathbf{O}^\top\hat{\mathbf{U}}^\top) + (\mathbf{U} - \hat{\mathbf{U}}\mathbf{O})\mathbf{U}^\top\right\| \\
&\leq \left\|\hat{\mathbf{U}}\mathbf{O}(\mathbf{U}^\top - \mathbf{O}^\top\hat{\mathbf{U}}^\top)\right\| + \left\|(\mathbf{U} - \hat{\mathbf{U}}\mathbf{O})\mathbf{U}^\top\right\| \\
&\leq 2\left\|\mathbf{U} - \hat{\mathbf{U}}\mathbf{O}\right\|_F,
\end{aligned}$$

where we used Cauchy-Schwarz in the last inequality and the fact that $\mathbf{O}$ is a orthogonal matrix and $\mathbf{U}^\top\mathbf{U} = \hat{\mathbf{U}}^\top\hat{\mathbf{U}} = \mathbf{I}$. Now we are able to apply Lemma 8:

$$\left\|\mathbf{P} - \hat{\mathbf{P}}\right\| \leq \frac{\sqrt{32}\min\left(\sqrt{p}\left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\right\|, \left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\right\|_F\right)}{\Delta_\sigma} \tag{23}$$

Using the triangular inequality we bound the term $\left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\right\|$:

$$\left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\right\| \leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^*\| + \left\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\right\|. \tag{24}$$

The first term in eq. (24) is a simple concentration inequality for covariance matrices. Using Lemma 7 we have with probability $1 - \delta$:

$$\|\boldsymbol{\Sigma}^* - \boldsymbol{\Sigma}\| \leq \sqrt{\frac{C \log\left(\frac{2}{\delta}\right)}{t}},$$

with an absolute constant $C$. Setting $\delta = 1/t$ and taking the expectation value yields:

$$\mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\|\boldsymbol{\Sigma}^* - \boldsymbol{\Sigma}\|\right] = \mathcal{O}\left(\sqrt{\frac{C\log(2t)}{t}}\right).$$

We denote $\epsilon_{\boldsymbol{\Sigma}} \coloneqq \sqrt{\frac{C\log(2t)}{t}}$. Finally we need to bound $\left\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\right\|$. We denote $\boldsymbol{\Theta}^* = \left[\left(\boldsymbol{\theta}^*(i) - \bar{\boldsymbol{\theta}}^*\right)^\top\right]_{\{i\in 1,\dots,t\}}$ and $\hat{\boldsymbol{\Theta}} = \left[\left(\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}}\right)^\top\right]_{i\in\{1,\dots,t\}}$, with vertically concatenated vectors, such that we have:

$$\begin{aligned}
\left\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\right\| &\leq \left\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\right\|_F \\
&= \frac{1}{t}\left\|\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}}^\top\hat{\boldsymbol{\Theta}}\right\|_F \\
&= \frac{1}{t}\left\|\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^* - \boldsymbol{\Theta}^{*\top}\hat{\boldsymbol{\Theta}} + \boldsymbol{\Theta}^{*\top}\hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}^\top\hat{\boldsymbol{\Theta}}\right\|_F \\
&= \frac{1}{t}\left\|\boldsymbol{\Theta}^{*\top}\left(\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}}\right) + \left(\boldsymbol{\Theta}^{*\top} - \hat{\boldsymbol{\Theta}}^\top\right)\hat{\boldsymbol{\Theta}}\right\|_F \\
&\leq \frac{1}{t}\left(\|\boldsymbol{\Theta}^*\|_F + \left\|\hat{\boldsymbol{\Theta}}\right\|_F\right)\left\|\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}}\right\|_F
\end{aligned}$$

We can further bound this while also taking the expectation, using:

$$\mathbb{E}\left[\left\|\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}}\right\|_F\right] \leq \mathbb{E}\left[\sqrt{t\beta_d^2 + \sqrt{\sum_{i=1}^{t}\|\boldsymbol{\theta}^*(i) - \boldsymbol{\theta}(i)\|^2}}\right] \leq 2\sqrt{t}\beta_d.$$

where we used the result of eq. (22). The same estimation can be done for the term $\|\boldsymbol{\Theta}^*\|_F + \left\|\hat{\boldsymbol{\Theta}}\right\|_F$:

$$\|\boldsymbol{\Theta}^*\|_F + \left\|\hat{\boldsymbol{\Theta}}\right\|_F \leq \sqrt{t}\left(\max_i\|\boldsymbol{\theta}^*(i) - \bar{\boldsymbol{\theta}}^*\| + \max_i\|\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}}\|\right] \leq 4\sqrt{t}V$$

Thus we conclude:

$$\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\right\|\right] \leq 8V\beta_d$$

Inserting the results into eq. (23) gives:

$$\mathbb{E}\left[\left\|\mathbf{P} - \hat{\mathbf{P}}\right\|\right] \leq \sqrt{32p}\frac{8V\beta_d + \epsilon_{\boldsymbol{\Sigma}}}{\Delta_\sigma} \tag{25}$$

After estimation of every term of our original expression we can summarize it by taking the expectation and applying Jensen's inequality:

$$\mathbb{E}_{\theta^*\sim\rho}\left[\mathbb{E}\left[\left\|\hat{\mathbf{P}}^\perp(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})\right\|\right]\right] = \mathcal{O}\left(\sqrt{\mathrm{Var}_\rho + \beta_d^2\left(1 + 64\sqrt{2p}\frac{V^2}{\Delta_\sigma}\right)^2 + \epsilon_{\boldsymbol{\mu}}^2 + \frac{128p\epsilon_{\boldsymbol{\Sigma}}^2V^2}{\Delta_\sigma^2}}\right) \tag{26}$$

□

**Lemma 9.** *(Abbasi-Yadkori et al., 2011, Lemma 11) Let $\mathbf{x}_{a_k}$ be a sequence in $\mathbb{R}^d$ with $\|\mathbf{x}_{a_k}\| \leq 1$ and $\mathbf{B}$ defined as usual. Then we have:*

$$\sum_{k=1}^{n}\left\|\mathbf{x}_{a_{k-1}}\right\|^2_{\mathbf{B}_{k-1}^{-1}} \leq 2S_{k-1}^{\lambda_1,\lambda_2}$$

*Proof of Theorem 1.* First we denote $\xi \coloneqq \|\mathbf{x}\|_{\mathbf{B}^{-1}} \left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_{\mathbf{B}}$ as exploration term. Then we continue by estimating the peudo regret $R(n)$:

$$
\begin{aligned}
R(n) &= \sum_{k=1}^{n} \left( \mathbf{x}_{a^*_{k-1}} - \mathbf{x}_{a_{k-1}} \right)^\top \boldsymbol{\theta}^* \\
&\leq \sum_{k=1}^{n} \mathbf{x}_{a_{k-1}}^\top \hat{\boldsymbol{\theta}}_{k-1} + \xi_{k-1} - \mathbf{x}_{a_{k-1}}^\top \boldsymbol{\theta}^* \\
&\leq \sum_{k=1}^{n} \mathbf{x}_{a_{k-1}}^\top \hat{\boldsymbol{\theta}}_{k-1} + \xi_{k-1} - \mathbf{x}_{a_{k-1}}^\top \hat{\boldsymbol{\theta}}_{k-1} + \xi_{k-1} \\
&= \sum_{k=1}^{n} 2\xi_{k-1} \\
&\leq \sum_{k=1}^{n} \left( \sqrt{p \log\left( \frac{1}{\delta} + \frac{k-1}{p\lambda_2 \delta} \right) + q \log\left( \frac{1}{\delta} + \frac{k-1}{q\lambda_1 \delta} \right)} + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W \right) \left\| \mathbf{x}_{a_{k-1}} \right\|_{\mathbf{B}_{k-1}^{-1}} \\
&\leq \left( \sqrt{p \log\left( \frac{1}{\delta} + \frac{n}{p\lambda_2 \delta} \right) + q \log\left( \frac{1}{\delta} + \frac{n}{q\lambda_1 \delta} \right)} + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W \right) \\
&\quad \sqrt{n \sum_{k=1}^{n} \left\| \mathbf{x}_{a_{k-1}} \right\|^2_{\mathbf{B}_{k-1}^{-1}}} \\
&\leq \left( \sqrt{p \log\left( 1 + \frac{n}{p\lambda_2} \right) + q \log\left( 1 + \frac{n}{q\lambda_1} \right) + \log\left( \frac{1}{\delta^2} \right)} + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W \right) \\
&\quad \sqrt{2n \left( p \log\left( 1 + \frac{n}{p\lambda_2} \right) + q \log\left( 1 + \frac{n}{q\lambda_1} \right) \right)}
\end{aligned}
$$

The first and second inequality make use of the OFUL principle and the definition of the UCB function. We used Lemma 3 in the third inequality and Lemma 9 in the last inequality. This regret holds with probability $1 - \delta$.

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho} \left[ \mathbb{E} \left[ R(n) \right] \right] \leq & \left( \sqrt{p \log\left( 1 + \frac{nV^2}{p} \right) + q \log\left( 1 + \frac{n\sqrt{Y}}{q} \right) + \log\left( \frac{1}{\delta^2} \right)} + 1 + V \right) \\
& \sqrt{2n \left( p \log\left( 1 + \frac{nV^2}{p} \right) + q \log\left( 1 + \frac{n\sqrt{Y}}{q} \right) \right)}
\end{aligned}
$$

We obtain the final results by setting $\delta = 1/n$, take the expectation value, followed by an additional expectation value with respect to the task distribution: $\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho} \left[ \mathbb{E} \left[ R(n) \right] \right]$, setting $\lambda_1 = \frac{1}{\sqrt{Y}}$, $\lambda_2 = \frac{1}{V^2}$ and application of Jensen's inequality. □

## C Proof of Theorem 2

The following proofs are adapted from Agrawal and Goyal (2013) and are required to finish the proof on the regret bound. Before proceeding, we define the concept of a saturated arm, which is basically a measurement of the required exploration for any arm.

**Definition 3.** *We call an arm $a$ saturated if $g_n \|\mathbf{x}_a\|_{\mathbf{B}^{-1}} < l_n \|\mathbf{x}_{a^*}\|_{\mathbf{B}^{-1}}$ and unsaturated otherwise, with $g_n = \sqrt{2d + 6\log(n)} v + l_n$. The set of saturated arms at round $k$ is denoted as $\mathcal{C}_k$.*

We will also utilize the following Lemma from Hsu et al. (2012), which is a special case of the inequality in Hanson and Wright (1971):

**Lemma 10** (Proposition 1.1 in Hsu et al. (2012)). *Let $\mathbf{x} \in \mathbb{R}^d$ be a $d$-dimensional standard normal variable and $\mathbf{C} \in \mathbb{R}^{d\times d}$ a matrix. Then we have for all $t > 0$:*

$$\Pr\left(\|\mathbf{C}\mathbf{x}\|^2 > \operatorname{tr}\big(\mathbf{C}^\top\mathbf{C}\big) + 2\sqrt{\operatorname{tr}\Big(\big(\mathbf{C}^\top\mathbf{C}\big)^2\Big)t} + 2\big\|\mathbf{C}^\top\mathbf{C}\big\|t\right) \leq e^{-t}$$

*Proof of Lemma 5.* The probability of event $E_r$ is determined using Lemma 3: we have with probability $1-\delta$:

$$\begin{aligned}
|\mathbf{x}_a^\top\left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right)| &\leq \|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}}\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right\|_{\mathbf{B}_k} \\
&\leq \|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}}\left(\sqrt{S_k^{\lambda_1,\lambda_2} + \log\left(\frac{1}{\delta^2}\right)} + \frac{\lambda_2}{\sqrt{\lambda_1}}W + \sqrt{\lambda_1}V\right),
\end{aligned}$$

by substituting $\delta \to \frac{\delta}{n^2}$ and further upper bounding $S_k^{\lambda_1,\lambda_2}$ we get:

$$\begin{aligned}
\sqrt{S_k^{\lambda_1,\lambda_2} + \log\left(\frac{n^2}{\delta^2}\right)} &= \sqrt{p\log\left(1 + \frac{k}{p\lambda_2}\right) + q\log\left(1 + \frac{k}{q\lambda_1}\right) + \log\left(\frac{n^2}{\delta^2}\right)} \\
&\leq \sqrt{p\log\left(n\left(\frac{n}{\delta}\right)^{2/d}\right) + q\log\left(n\left(\frac{n}{\delta}\right)^{2/d}\right)} \\
&\leq \sqrt{\log\left(\frac{1}{\delta}\right)(d+2)\log(n)},
\end{aligned}$$

and therefore we have:

$$\begin{aligned}
|\mathbf{x}_a^\top\left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right)| &\leq \left(\sqrt{\log\left(\frac{1}{\delta}\right)(d+2)\log(n)} + \frac{\lambda_2}{\sqrt{\lambda_1}}W + \sqrt{\lambda_1}V\right)\|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}} \\
&\leq \sqrt{2\log\left(\frac{1}{\delta}\right)(d+2)\log(n) + 2K^2}\|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}},
\end{aligned}$$

with $K = \frac{\lambda_2}{\sqrt{\lambda_1}}W + \sqrt{\lambda_1}V$. Since we substituted $\delta \to \frac{\delta}{n^2}$, this event has a probability of at least $1 - \frac{\delta}{n^2}$.
For proof of the bound on the probability of event $E_\theta$ we have for all $a \in \mathcal{A}_k$:

$$\begin{aligned}
\left|\mathbf{x}_a^\top\left(\hat{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k\right)\right| &= \left|\mathbf{x}_a^\top\mathbf{B}_k^{-\frac{1}{2}}\mathbf{B}_k^{\frac{1}{2}}\left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)\right| \\
&\leq v\sqrt{\mathbf{x}_a^\top\mathbf{B}_k^{-1}\mathbf{x}_a}\left\|\frac{1}{v}\mathbf{B}_k^{\frac{1}{2}}\left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)\right\|
\end{aligned}$$

.

By definition, the term $\frac{1}{v}\mathbf{B}_k^{\frac{1}{2}}\left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)$ is $d$-dimensional standard normal variable, such that we can apply Lemma 10, where we set $\mathbf{C} = \mathbf{I}$ and $t = 2\log(n)$:

$$\Pr\left(\left\|\frac{1}{v}\mathbf{B}_k^{\frac{1}{2}}\left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)\right\| > \sqrt{d + \sqrt{8d\log(n)} + 4\log(n)}\right) \leq \frac{1}{n^2}.$$

Thus the following inequality holds with probability of at least $1 - \frac{1}{n^2}$ for all $a \in \mathcal{A}_k$:

$$\begin{aligned}\left|\mathbf{x}_a^\top \left(\hat{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k\right)\right| &\leq v\|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}} \sqrt{d + \sqrt{8d\log(n)} + 4\log(n)} \\ &\leq v\|\mathbf{x}_a\|_{\mathbf{B}_k^{-1}} \sqrt{2d + 6\log(n)},\end{aligned}$$

where we used the inequality of arithmetic and geometric means in the last step. □

**Lemma 11.** *For any filtration $\mathcal{F}_{k-1}$ such that $E_r$ is true, we have:*

$$\Pr(\mathbf{x}_{a_k^*}^\top \tilde{\boldsymbol{\theta}}_k > \mathbf{x}_{a_k^*}^\top \boldsymbol{\theta}^* + l_n \left\|\mathbf{x}_{a_k^*}\right\|_{\mathbf{B}_k^{-1}}) \geq c_n$$

*and:*

$$\Pr(a_k \in \mathcal{C}_k | \mathcal{F}_{k-1}) \leq \frac{1}{c_n} \Pr(a_k \notin \mathcal{C}_k | \mathcal{F}_{k-1}) + \frac{1}{c_n n^2},$$

*with $c_n = \frac{1}{4e\sqrt{\pi n^\alpha}}$.*

*Proof.* Assuming the event $E_r$ holds and $\mathbf{x}_{a^*}^\top \tilde{\boldsymbol{\theta}}$ is a Gaussian random variable with mean $\mathbf{x}_{a^*}^\top \hat{\boldsymbol{\theta}}$ and variance $v\|x_{a^*}\|_{\mathbf{B}^{-1}}$, we can apply the anti-concentration inequality such that:

$$\begin{aligned}\Pr(\mathbf{x}_{a_k^*}^\top \tilde{\boldsymbol{\theta}}_k \geq \mathbf{x}_{a_k^*}^\top \hat{\boldsymbol{\theta}}_k + l_n \left\|\mathbf{x}_{a_k^*}\right\|_{\mathbf{B}_k^{-1}} | \mathcal{F}_{k-1}) &= \Pr\left( \frac{\mathbf{x}_{a_k^*}^\top \left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)}{v\left\|\mathbf{x}_{a_k^*}\right\|_{\mathbf{B}_k^{-1}}} \geq \frac{\mathbf{x}_{a_k^*}^\top \left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right) + l_n \left\|\mathbf{x}_{a_k^*}\right\|_{\mathbf{B}_k^{-1}}}{v\left\|\mathbf{x}_{a_k^*}\right\|_{\mathbf{B}_k^{-1}}} \middle| \mathcal{F}_{k-1} \right) \\ &\geq \frac{1}{4\sqrt{\pi}} \exp(-Z^2),\end{aligned}$$

with

$$\begin{aligned}Z &= \frac{\mathbf{x}_{a_k^*}^\top \left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right) + l_n \left\|x_{a_k^*}\right\|_{\mathbf{B}_k^{-1}}}{v\left\|x_{a_k^*}\right\|_{\mathbf{B}_k^{-1}}} \\ &\leq \frac{2l_n \left\|x_{a_k^*}\right\|_{\mathbf{B}_k^{-1}}}{v\left\|x_{a_k^*}\right\|_{\mathbf{B}_k^{-1}}} \\ &\leq \frac{2\sqrt{2\log\left(\frac{1}{\delta}\right)(d+2)\log(n) + 2K^2}}{4\sqrt{\log\left(\frac{1}{\delta}\right)\frac{d+2}{\alpha}}} \\ &\leq \sqrt{\frac{\alpha}{2}\log(n) + \frac{\alpha K^2}{8\log\left(\frac{1}{\delta}\right)(d+2)}} \\ &\leq \sqrt{\frac{\alpha}{2}\log(n) + 1}.\end{aligned}$$

Thus we have

$$\Pr(\mathbf{x}_{a_k^*}^\top \tilde{\boldsymbol{\theta}}_k \geq \mathbf{x}_{a_k^*}^\top \hat{\boldsymbol{\theta}}_k + l_n \left\|\mathbf{x}_{a_k^*}\right\|_{\mathbf{B}_k^{-1}} | \mathcal{F}_{k-1}) \geq \frac{1}{4e\sqrt{\pi n^\alpha}}$$

The proof of the second inequality is provided in Lemma 3 of Agrawal and Goyal (2013).

□

**Lemma 12.** *Let* $\text{regret}_k = (\mathbf{x}_{a^*}^\top - \mathbf{x}_{a_k}^\top)\boldsymbol{\theta}^*$ *be defined as the instantaneous regret at round* $k$ *and* $\text{regret}'_k = \text{regret}_k I(E_r)$. *Define*

$$X_k = \text{regret}'_{\text{k}} - \frac{g_n}{c_n} I\left(a_k \notin \mathcal{C}\right) \|x_{a^*}\|_{\mathbf{B}_k^{-1}} - \frac{2g_n}{c_n n^2} - \frac{2g_n^2}{l_n}\|x_{a_k}\|_{\mathbf{B}_k^{-1}}$$

*and*

$$Y_k = \sum_{t=1}^{k} X_k,$$

*then* $(Y_k; k = 1, ..., n)$ *is a super-martingale process with respect to filtration* $\mathcal{F}_{k-1}$.

*Proof.* The proof is provided in Lemma 4 of Agrawal and Goyal (2013). □

*Proof of Theorem 2.* Each value in $X_k$ is bounded by $\frac{2g_n^2}{c_n l_n}$ which implies a bounded difference on the super-martingale $Y_k$ with $|Y_k - Y_{k-1}| \leq \frac{8g_n^2}{c_n l_n}$, allowing us to apply Azuma-Hoeffding's inequality during the proof. Thus we have with probability $1 - \frac{\delta}{2}$:

$$\begin{aligned}
\sum_{k=1}^{n} \text{regret}'_k &\leq \sum_{k=1}^{n} \left( \frac{g_n}{c_n} I(a_k \notin \mathcal{C}_k \|x_{a^*}\|_{\mathbf{B}_k^{-1}}) \right) + \frac{2g_n}{c_n n^2} \\
&+ \frac{2g_n^2}{c_n l_n} \sum_{k=1}^{n} \|x_{a_k}\|_{\mathbf{B}_k^{-1}} + \frac{8g_n^2}{c_n l_n} \sqrt{2n \log\left(\frac{2}{\delta}\right)} \\
&\leq \sum_{k=1}^{n} \left( \frac{g_n^2}{c_n l_n} I(a_k \notin \mathcal{C}_k \|x_{a_k}\|_{\mathbf{B}_k^{-1}}) \right) + \frac{2g_n}{c_n n^2} \\
&+ \frac{2g_n^2}{c_n l_n} \sum_{k=1}^{n} \|x_{a_k}\|_{\mathbf{B}_k^{-1}} + \frac{8g_n^2}{c_n l_n} \sqrt{2n \log\left(\frac{2}{\delta}\right)} \\
&\leq \frac{3g_n^2}{c_n l_n} \sum_{k=1}^{n} \|x_{a_k}\|_{\mathbf{B}_k^{-1}} + \frac{2g_n}{c_n n^2} + \frac{8g_n^2}{c_n l_n} \sqrt{2n \log\left(\frac{2}{\delta}\right)} \\
&\leq \frac{3g_n^2}{c_n l_n} \sqrt{2n S_n^{\lambda_1, \lambda_2}} + \frac{2g_n}{c_n n^2} + \frac{8g_n^2}{c_n l_n} \sqrt{2n \log\left(\frac{2}{\delta}\right)} \\
&= \frac{g_n^2}{c_n l_n} \left( \sqrt{18n S_n^{\lambda_1, \lambda_2}} + \sqrt{128n \log\left(\frac{2}{\delta}\right)} \right) + \frac{2g_n}{c_n n^2} \\
&= \left( \frac{l_n + 2\sqrt{2d + 6\log(n)}v + (2d + 6\log(n))v^2/l_n}{c_n} \right) \left( \sqrt{18n S_n^{\lambda_1, \lambda_2}} + \sqrt{128n \log\left(\frac{2}{\delta}\right)} \right) + \frac{2g_n}{c_n n^2} \\
&= \mathcal{O}\left( \left( 2\sqrt{\frac{2d^2 \log\left(\frac{1}{\delta}\right) + 6d \log\left(\frac{1}{\delta}\right) \log(n)}{\alpha}} + \frac{2d^{\frac{3}{2}} \sqrt{\log\left(\frac{1}{\delta}\right)}}{\alpha \sqrt{\log(n)}} + \frac{6\sqrt{d \log\left(\frac{1}{\delta}\right) \log(n)}}{\alpha} \right) \sqrt{n^{1+\alpha} S_n^{\lambda_1, \lambda_2}} \right)
\end{aligned}$$

We used our definition for saturated arms in the second inequality and Lemma 9 in the fourth inequality. Now similar as done in Theorem 1 we set $\delta = \frac{1}{n}$ and take the expectation value. Additionally we set $\alpha = \frac{1}{\log(n)}$:

$$\mathbb{E}\left[ \sum_{k=1}^{n} \text{regret}'_k \right] = \mathcal{O}\left( \left( d^{\frac{3}{2}} \log(n) + \sqrt{d} \log(n)^2 \right) \sqrt{n S_n^{\lambda_1, \lambda_2}} \right)$$

Inserting $\lambda_2 = 1/V^2$ and $\lambda_1 = \frac{1}{\sqrt{Y}}$, while taking the second expectation value with respect to the task distribution and applying Jensen's inequality, gives the final result:

$$\mathcal{R}(n) = \mathcal{O}\left(\left(d^{\frac{3}{2}}\log(n) + \sqrt{d}\log(n)^2\right)\sqrt{n\left(p\log\left(1+\frac{nV^2}{p}\right) + q\log\left(1+\frac{n\sqrt{Y}}{q}\right)\right)}\right)$$

□