Multi-task Representation Learning with Stochastic Linear Bandits

Leonardo Cella
CSML, Italian Institute of Technology

Karim Lounici
CMAP, Ecole Polytechnique

Grégoire Pacreau
CMAP, Ecole Polytechnique

Massimiliano Pontil
CSML, Italian Institute of Technology
& Dept of Computer Science, UCL

Abstract

We study the problem of transfer-learning in the setting of stochastic linear contextual bandit tasks. We consider that a low dimensional linear representation is shared across the tasks, and study the benefit of learning the tasks jointly. Following recent results to design Lasso stochastic bandit policies, we propose an efficient greedy policy based on trace norm regularization. It implicitly learns a low dimensional representation by encouraging the matrix formed by the task regression vectors to be of low rank. Unlike previous work in the literature, our policy does not need to know the rank of the underlying matrix, nor does it requires the covariance of the arms distribution to be invertible. We derive an upper bound on the multi-task regret of our policy, which is, up to logarithmic factors, of order $T\sqrt{rN} + \sqrt{rNTd}$, where T is the number of tasks, r the rank, d the number of variables and N the number of rounds per task. We show the benefit of our strategy over an independent task learning baseline, which has a worse regret of order $T\sqrt{dN}$. We also argue that our policy is minimax optimal and, when $T \geq d$, has a multi-task regret which is comparable to the regret of an oracle policy which knows the true underlying representation.

1 INTRODUCTION

Contextual bandits (Abbasi-Yadkori et al., 2011; Li et al., 2010; Auer, 2002) are a prominent learning framework to study sequential decision problems with partial feedback. They find applications in numerous fields, ranging from recommender systems (Li et al., 2010), to finance (Shen et al., 2015) and to adaptive routing (Awerbuch and Kleinberg, 2008), among others. This methodology was originally motivated by applications in clinical trials (Woodroofe, 1979), whereby a doctor has to decide which among available treatments is best suited for a patient, through a sequence of trials. A fundamental aspect in bandit problems is to control the trade-off between *exploration* and *exploitation*, namely, the balance between the need of acquiring more information and the temptation to act optimally according to the already available knowledge.

In this paper we study multi-task learning with stochastic linear contextual bandit tasks (Lu et al., 2021; Li et al., 2017; Filippi et al., 2010). Within this setting, each task is associated with a regression vector and proceeds sequentially. At each trial an agent observes a set of different alternatives (*arms*) linked to a feature (*context*) vector. The agent then selects one context vector and subsequently observes a stochastic reward generated by a noisy linear regression associated to the chosen vector. The goal is to design an algorithm (policy) that learns, interacting with the tasks, how to select contexts that are most aligned with the underlying task (regression vector), hence maximizing the cumulative rewards across all the tasks.

A central idea of multi-task learning is to leverage similarities between tasks in order to facilitate learning. In this paper, we consider that the tasks share a low dimensional representation, that is, the task regression vectors span a low dimensional subspace. The benefit of learning such representation has been widely investigated in both the standard supervised learning and in the reinforcement learning settings (see Lounici et al., 2011; Koltchinskii et al., 2011; Negahban and Wainwright, 2011; Calandriello et al., 2014, and references therein). In the bandit setting, this problem presents additional difficulties, since contexts vectors are no longer independently distributed. Indeed, they are collected sequentially depending on past observations and on the adopted bandit policy. This entails two main challenges.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia. PMLR: Volume 206.

On the one hand, we would like the collected contexts to span the whole feature space, as this would facilitate the estimation of the unknown regression vectors. On the other hand, collecting contexts which are misaligned with respect to the regression vector results in poor performance for the bandit strategy.

Contributions. In this work we present an efficient policy based on the trace norm regularization estimator (see Koltchinskii et al., 2011; Negahban and Wainwright, 2011; Maurer and Pontil, 2013, and references therein) which leverages the tasks common representation to improve learning. We provide oracle inequalities for this policy under the restricted strong convexity condition with correct and explicit dependencies (Lemma 1 and Proposition 1) that are based based on a novel martingale concentration argument. Next, we provide an upper bound for the proposed policy (Theorem 1) which is valid without boundedness assumption on the arms and is minimax optimal (up to a logarithmic factor). We noticed that when the number of tasks is larger than the ambient dimension our policy is comparable to the oracle policy which knows the true underlying representation a-priori. A key novelty of the proposed policy over the state-of-the-art (Yang et al., 2020) is that it does not need to know the rank of the representation matrix, it does not require the arms covariance distribution to be invertible and its regret bound is non-trivial already when the time horizon is $O(d)$ in the worse case and potentially $O(\log d)$ in favorable scenarios.

Notation. For a real vector $\mathbf{x} \in \mathbb{R}^d$, we use $\|\mathbf{x}\|$ to denote its Euclidean norm. Given a pair of symmetric matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$, the expression $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semi-definite. We respectively use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to refer to the minimum and maximum eigenvalues of a square symmetric matrix $\mathbf{A}$. Similarly, we respectively use $\sigma_{\min}(\mathbf{A})$ and $\sigma_{\max}(\mathbf{A})$ for the smallest and the largest singular values of a generic matrix $\mathbf{A}$. Given a positive definite matrix $\mathbf{A} \succ \mathbf{0}$, we indicate with $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$ the corresponding weighted Euclidean norm. For any matrix $\mathbf{A} \in \mathbb{R}^{d\times T}$, we let $\|\mathbf{A}\|_* = \sum_{i=1}^d \sigma_i(\mathbf{A})$ be the trace norm (sum of its singular values), $\|\mathbf{A}\|_{\mathrm{F}}$ the Frobenius norm and $\|\mathbf{A}\|_{\mathrm{op}} = \sigma_{\max}(\mathbf{A})$ the operator norm (largest singular value). We denote with $[\mathbf{A}]^i$ its i-th row and with $[\mathbf{A}]_j$ its j-th column. Finally, we use diag$(\lambda_1, \ldots, \lambda_d)$ for the $d \times d$ diagonal matrix with values $\lambda_1, \ldots, \lambda_d$ on the diagonal. Given a random event Υ, we denote with Υ^c its complement. Finally, given a scalar $\epsilon \in \mathbb{R}$ we denote with ϵ_d the d-dimensional vector having value ϵ in each component. Additional notation is introduced on the way; a table summarizing the notation used throughout the paper is reported in the appendix.

2 RELATED WORKS

In the last two decades many efforts have been devoted to designing contextual bandit policies (Abbasi-Yadkori et al., 2011, 2012; Ariu et al., 2020; Auer, 2002; Bastani and Bayati, 2020; Chu et al., 2011; Kim and Paik, 2019; Li et al., 2017; Oh et al., 2021; Wang et al., 2018; Kuzborskij et al., 2019; Foster and Rakhlin, 2020). When restricting to the high-dimensional setting an appealing approach is based on sparse linear models, i.e. the number s of non-zero components of the regression vector is assumed to be much smaller than the input dimension d. Many different strategies have been investigated (Kim and Paik, 2019; Ariu et al., 2020; Bastani and Bayati, 2020). Among the proposed approaches, one of the most recent and valuable works is (Oh et al., 2021), where they design a greedy policy based on the Lasso estimator. Interestingly, their approach does not require knowledge of s and it does not perform random pulls in order to have i.i.d. data. Inspired by this work, we propose a greedy algorithm which does not need to know the rank index r (dimensionality of the underlying representation). A technical challenge that we are facing is how to obtain an accurate estimator relying on non i.i.d. samples, while considering a more complex matrix estimator. We observe that their regret bound argument is not accurate as there might be a hidden dependency in the number of features (the same inaccuracy can be found in (Cella and Pontil, 2021; Kim and Paik, 2019; Calandriello et al., 2014).

Multi-task and meta-learning frameworks have been studied primarily in the supervised-learning setting (Tripuraneni et al., 2021; Denevi et al., 2018, 2019; Lounici et al., 2011; Argyriou et al., 2008; Baxter, 2000). Specifically, the impact of representation learning with trace-norm regularization has been widely investigated when considering i.i.d. data (Koltchinskii et al., 2011; Negahban and Wainwright, 2011; Maurer and Pontil, 2013). More recently different authors have investigated the combination of multi-task and meta-learning with interactive learning settings (e.g. bandits and reinforcement learning) (Hu et al., 2021; Basu et al., 2021; Cella and Pontil, 2021; Kveton et al., 2021; Simchowitz et al., 2021; Calandriello et al., 2014; D'Eramo et al., 2019; Cella et al., 2020; Yang et al., 2020). Among the latter category, the most relevant works are (Yang et al., 2020; Hu et al., 2021; Cella and Pontil, 2021). Cella and Pontil (2021) considered both the multi-task and the meta-learning frameworks but assuming the task vector parameters to be *jointly sparse*, which is more restrictive than the low rank assumption considered here. Differently, (Yang et al., 2020) considers a similar multi-task problem both in the finite and infinite action settings, whereas (Hu et al., 2021) considers only the latter setting. Note that these two references impose stronger conditions and assumptions. Indeed, the policies considered in those papers require knowledge of the low-rank parameter r and are not sample efficient. In this work, we develop a different and more general analysis for the finite action setting. Notably, our algorithm is computationally efficient and does not requires knowledge of low-rank parameter r. Our idea is similar to the last approach investigated in (Calandriello et al., 2014), but their work considers the simpler case where samples are independently distributed. At last we mention that (Kveton et al., 2017) studied low-rank stochastic bandits. However they consider a completely different protocol, regret and arms definitions. Thus it is not possible to draw a quantitative comparison to this work.

3 PROBLEM SETTING

In this section we recall the stochastic linear contextual bandit problem and then introduce the multi-task learning framework.

3.1 Stochastic Linear Contextual Bandits

The contextual linear bandit problem consists of a sequence of N interactions between a learning policy π and the environment. At each round $n \in [N]$, the agent is given a set $\mathcal{D}_n \subseteq \mathbb{R}^d$ formed of K arms (context vectors) from which it has to select one arm $\mathbf{x}_n \in \mathcal{D}_n$ among the available ones. Then, the agent observes the corresponding reward Y_n via a noisy linear regression,

$$Y_n = \mathbf{x}_n^\top \mathbf{w} + \eta_n, \tag{1}$$

where η_n is a random variable whose statistical distribution will be specified later. By knowing the true regression vector $\mathbf{w}$, at each round $n \in [N]$ the optimal policy π^* selects $\mathbf{x}_n^* = \arg\max_{\mathbf{x} \in \mathcal{D}_n} \mathbf{x}^\top \mathbf{w}$, maximizing the instantaneous expected reward. The learning objective is then to maximize the cumulative reward, or equivalently, to minimize the *pseudo-regret*

$$R(N, \mathbf{w}) = \sum_{n=1}^{N} r_n = \sum_{n=1}^{N} (\mathbf{x}_n^* - \mathbf{x}_n)^\top \mathbf{w}.$$

We make the following additional assumptions on the arms, the regression vector and the noise variables.

Assumption 1 (Contexts Distribution). *At each round $n \in [N]$, the decision set $\mathcal{D}_n \subset \mathbb{R}^d$ consists of K d-dimensional vectors $\mathbf{x}_k$ admitting representation $\mathbf{x}_k = \Sigma_k^{1/2} \mathbf{z}_k$ where Σ_k is the covariance operator of $\mathbf{x}_k$ and $\mathbf{z}_k$ is a d-dimensional vector with independent sub-Gaussian entries with zero mean and variance 1. We assume the tuples $\mathcal{D}_1, \ldots, \mathcal{D}_N$ to be drawn i.i.d. from a fixed unknown zero mean sub-Gaussian joint distribution p on $\mathbb{R}^{Kd}$.*

Let $C_{\mathbf{z}}$ be a positive constant satisfying

$$\max_{1 \le k \le K} \|\mathbf{z}_k\|_{\psi_2} < C_{\mathbf{z}}.$$

We denote by $C_{\mathbf{x}}$ the sub-Gaussian Orlicz norm of the K marginal distributions of p corresponding to the K arms, that is $C_{\mathbf{x}} := \max_{1 \le k \le K} \|\mathbf{x}_k\|_{\psi_2}$. An obvious computation gives

$$C_{\mathbf{x}} \le C_{\mathbf{z}} \max_{1 \le k \le K} \left\{ \|\Sigma_k\|_{\text{op}}^{1/2} \right\} < \infty.$$

The above assumption is quite standard when considering high-dimensional linear bandits (Bastani and Bayati, 2020; Hao et al., 2020; Kim and Paik, 2019; Rusmevichientong and Tsitsiklis, 2010). Note that vectors associated to different arms are allowed to be correlated between each other. Moreover, Assumption 1 implies that each arm k admits zero mean, square-integrable marginal distribution with $d \times d$ covariance Σ_k.

Finally, similarly to (Bastani et al., 2021; Oh et al., 2021), we introduce the following assumption on the arms distribution. This mild condition is necessary for our analysis in order to control the fulfillment of a specific regularity property (the RSC condition, see Definition 1 below) by the empirical covariance matrix. Specifically, it will allow the designed policy to avoid interleaving its arm selection strategy with random plays.

Assumption 2 (Arms distribution). *There exists a constant $\nu < \infty$ such that $p(-\overline{\mathbf{x}})/p(\overline{\mathbf{x}}) \le \nu$ $\forall \overline{\mathbf{x}} \in \mathbb{R}^{Kd}$. Moreover, there exists a constant $\omega_{\mathcal{X}} < \infty$ such that, for any permutation $(a_1, \ldots, a_K)$ of $[K]$, any integer $i \in \{2, \ldots, K-1\}$ and any fixed vector $\mathbf{w} \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}\left[\mathbf{x}_{a_i}\mathbf{x}_{a_i}^\top \mathbb{I}\left\{\mathbf{x}_{a_1}^\top \mathbf{w} < \cdots < \mathbf{x}_{a_K}^\top \mathbf{w}\right\}\right] \preceq$$
$$\preceq \omega_{\mathcal{X}} \mathbb{E}\left[\left(\mathbf{x}_{a_1}\mathbf{x}_{a_1}^\top + \mathbf{x}_{a_K}\mathbf{x}_{a_K}^\top\right) \mathbb{I}\left\{\mathbf{x}_{a_1}^\top \mathbf{w} < \cdots < \mathbf{x}_{a_K}^\top \mathbf{w}\right\}\right].$$

Parameter ν characterizes the skewness of the arms distribution p; for symmetrical distributions $\nu = 1$. Notice that Assumption 2 is satisfied for a large class of distributions both discrete and continuous (e.g. Gaussian, multi-dimensional uniform and Rademacher). The value $\omega_{\mathcal{X}}$ captures dependencies between arms: the more positively correlated they are, the smaller $\omega_{\mathcal{X}}$ will be. The extreme scenario is given by perfectly correlated arms, in which case we have $\omega_{\mathcal{X}}$ independent of any problem parameters. Finally, when arms are generated i.i.d. from a multivariate Gaussian or a multivariate uniform distribution over the sphere we have $\omega_{\mathcal{X}} = O(1)$.

3.2 Low-Rank Linear Contextual Bandits

In this paper, we address the problem of simultaneously solving T linear contextual bandit tasks. Each task $t \in [T]$ lasts for N rounds and is associated with a regression vector $\mathbf{w}_t \in \mathbb{R}^d$. We denote with

$$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_T]$$

the $d \times T$ matrix, whose columns are formed by the T regression vectors, which are unknown to the learner. The bandit tasks are explored in parallel. At each round $n \in [N]$

each task simultaneously receives a decision set $\mathcal{D}_{t,n}$ from the environment, upon which the learner (policy π) picks an arm $\mathbf{x}_{t,n}$, and the associated feedback $y_{t,n}$ is observed via the linear regression

$$y_{t,n} = \mathbf{x}_{t,n}^\top \mathbf{w}_t + \eta_{t,n} \tag{2}$$

where $\eta_{t,n}$ is a noise random variable which we specify below.

We make the following assumption on the task regression vectors.

Assumption 3 (Low-Rank Assumption). *The matrix $\mathbf{W} \in \mathbb{R}^{d\times T}$ has rank $\rho(\mathbf{W}) = r$, with $r \ll \min(d, T)$.*

The above assumption implies that there exists a low dimensional representation $\mathbf{B} \in \mathbb{R}^{d\times r}$ with orthonormal columns and a matrix $\mathbf{C} \in \mathbb{R}^{r\times T}$ such that $\mathbf{W} = \mathbf{BC}$. This is inline with the standard high-dimensional setting where many features are redundant.

Our principal objective is to minimize the multi-task pseudo-regret,

$$\begin{aligned} \overline{R}(T,N) &= \sum_{t=1}^{T} R(N,\mathbf{w}_t) \\ &= \sum_{t=1}^{T}\sum_{n=1}^{N} (\mathbf{x}_{t,n}^* - \mathbf{x}_{t,n})^\top \mathbf{w}_t, \end{aligned} \tag{3}$$

where $\mathbf{x}_{t,n}^* = \arg\max_{\mathbf{x}\in\mathcal{D}_{t,n}} \mathbf{x}^\top \mathbf{w}_t$.

4 LOW-RANK MATRIX ESTIMATION

In the standard supervised learning setting a natural estimator suited to Assumption 3 is given by the trace (nuclear) norm regularized estimator (see Argyriou et al., 2008; Bühlmann and Van De Geer, 2011; Lounici et al., 2011; Negahban and Wainwright, 2011, and references therein). In particular a large body of works have shown that bounds on the estimation error is controlled by the rank of the underlying regression matrix. Here we adapt this methodology to the bandit setting. At each round $n \in [N]$, we estimate matrix $\mathbf{W}$ via $\widehat{\mathbf{W}}_{n+1}$ as the solution of the following trace norm regularization problem

$$\arg\min_{\mathbf{A}\in\mathbb{R}^{d\times T}} \frac{1}{n}\sum_{t=1}^{T} \|\mathbf{y}_{t,n} - \mathbf{X}_{t,n}[\mathbf{A}]_t\|_2^2 + \lambda_n \|\mathbf{A}\|_* \tag{4}$$

where the design matrix $\mathbf{X}_{t,n} \in \mathbb{R}^{n\times d}$ contains the context vectors $\mathbf{x}_{t,i} \in \mathbb{R}^d$, $i \in [n]$ as its rows, the vector $\mathbf{y}_{t,n} \in \mathbb{R}^n$ is formed by the rewards for task t after n interactions, sampled from (2), and $\|\mathbf{A}\|_*$ is the trace norm of the matrix $\mathbf{A}$, that is the sum of its singular values. If compared to the Lasso estimator, the objective function (4) encourages low-rank matrices instead of sparse vectors. Before presenting the technical results we need to introduce the following additional notation.

Covariance matrices. We indicate the theoretical averaged $d \times d$ covariance matrix as

$$\mathbf{\Sigma} = \frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K} \mathbf{x}_k \mathbf{x}_k^\top\right] = \frac{1}{K}\sum_{k=1}^{K} \Sigma_k, \tag{5}$$

where the expectation is over the decision set sampling distribution $p_{\mathcal{X}}$ which is assumed to be shared between the tasks.

For every $t \in [T]$, we denote the empirical covariance matrix for task t as

$$\widehat{\mathbf{\Sigma}}_{t,n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{t,i}\mathbf{x}_{t,i}^\top. \tag{6}$$

Moreover, we use the notation $\overline{\mathbf{\Sigma}}$ and $\widehat{\overline{\mathbf{\Sigma}}}_n \in \mathbb{R}^{dT\times dT}$ for the theoretical and the empirical multi-task matrices respectively. They are both block diagonal and composed by the corresponding T single task $d \times d$ matrices on the diagonal, that is

$$\overline{\mathbf{\Sigma}} = \mathrm{diag}(\mathbf{\Sigma}, \dots, \mathbf{\Sigma})$$

and

$$\widehat{\overline{\mathbf{\Sigma}}}_n = \mathrm{diag}(\widehat{\mathbf{\Sigma}}_{1,n}, \dots, \widehat{\mathbf{\Sigma}}_{T,n}).$$

We introduce now the restricted strong convexity (RSC) condition on covariance $\overline{\mathbf{\Sigma}}$. To this end, we denote by $\mathrm{Vec}(\mathbf{\Delta})$, the vector in $\mathbb{R}^{dT}$ obtained by stacking together the columns of $\mathbf{\Delta} \in \mathbb{R}^{d\times T}$. Let

$$\mathbf{W} = \mathbf{UDV}^\top$$

be the singular value decomposition of matrix $\mathbf{W}$ of rank r where $\mathbf{U} \in \mathbb{R}^{d\times r}$, $\mathbf{V} \in \mathbb{R}^{T\times r}$ are the matrices formed by the left and right singular vectors, respectively and $\mathbf{D}$ is the $r \times r$ diagonal matrix of singular values.

Definition 1 (RSC Condition). *We say that the restricted strong convexity (RSC) condition is met for the theoretical multi-task matrix $\overline{\mathbf{\Sigma}} \in \mathbb{R}^{dT\times dT}$, with positive constant $\kappa(\overline{\mathbf{\Sigma}})$ if*

$$\min\left\{\frac{\|\mathrm{Vec}(\mathbf{\Delta})\|_{\overline{\mathbf{\Sigma}}}^2}{2\,\|\mathrm{Vec}(\mathbf{\Delta})\|_2^2} : \mathbf{\Delta} \in \mathcal{C}(r)\right\} \geq \kappa(\overline{\mathbf{\Sigma}}), \tag{7}$$

where

$$\mathcal{C}(r) = \left\{\mathbf{\Delta} \in \mathbb{R}^{d\times T} : \|\Pi(\mathbf{\Delta})\|_* \leq 3\,\|\mathbf{\Delta} - \Pi(\mathbf{\Delta})\|_*\right\} \tag{8}$$

and $\Pi(\mathbf{\Delta})$ is the projection onto set

$$\left\{\mathbf{\Delta} \in \mathbb{R}^{d\times T} : \mathrm{Col}(\mathbf{\Delta}) \perp \mathbf{U}, \mathrm{Row}(\mathbf{\Delta}) \perp \mathbf{V}\right\}.$$

The RSC condition allows us to control the error $\mathbf{\Delta}_n = \widehat{\mathbf{W}}_n - \mathbf{W}$ as it guarantees that the considered empirical loss is strictly convex on a restricted subset of approximately low rank matrices defined by the cone $\mathcal{C}(r)$.

Remark 1 (Value of $\kappa(\Sigma)$)**.** *As discussed in (Calandriello et al., 2014) Definition 1 can be compared with the restricted eigenvalue and the compatibility conditions that have been investigated for the group Lasso (Bühlmann and Van De Geer, 2011; Lounici et al., 2011) and Lasso (Oh et al., 2021; Kim and Paik, 2019; Calandriello et al., 2014) estimators. It is standard in high-dimensional statistics or in compressed sensing to assume the existence of an absolute constant $\kappa > 0$ such that $\kappa(\overline{\mathbf{\Sigma}}) > \kappa > 0$. Such conditions are satisfied w.h.p. for instance by context vectors with i.i.d. zero mean, variance 1, sub-Gaussian entries.*

4.1 Oracle Inequality with non i.i.d. Data

We can now state our first result which controls the Forbeniou-norm estimation error

$$\left\|\widehat{\mathbf{W}}_{n+1} - \mathbf{W}\right\|_{\mathrm{F}}$$

assuming the RSC condition to hold for the empirical multi-task matrix $\widehat{\overline{\mathbf{\Sigma}}}_n$. In Proposition 2 we will show that such condition is satisfied with high probability under our Assumptions 1, 2 and the RSC condition on $\overline{\mathbf{\Sigma}}$.

We denote by $\{\mathbf{e}_1, \ldots, \mathbf{e}_T\}$ the standard canonical basis of $\mathbb{R}^T$. We have then to upper bound the operator norm of the following matrix

$$\mathbf{D}_n = \sum_{t=1}^{T}\sum_{i=1}^{n} \eta_{t,i}\, \mathbf{x}_{t,i}\mathbf{e}_t^\top$$

in order to set the regularization parameter at round n so that the estimation bound in the following lemma holds with high probability.

Lemma 1. *Let $\{\mathbf{x}_{t,i} : t \in [T], i \in [N]\}$ be the sequence generated by Algorithm (1) Suppose the RSC condition holds for the empirical multi-task matrix $\widehat{\overline{\mathbf{\Sigma}}}_n$ with constant $\kappa\big(\widehat{\overline{\mathbf{\Sigma}}}_n\big)$. For $\delta \in (0, 1)$, define the regularization parameter $\lambda_n = \lambda_n(\delta)$ such that with probability at least $1 - \delta$*

$$\frac{1}{n}\left\|\mathbf{D}_n\right\|_{\mathrm{op}} \le \lambda_n. \tag{9}$$

Then with probability at least $1 - \delta$ the trace-norm regularized estimate $\widehat{\mathbf{W}}_n$ defined in (4) satisfies

$$\left\|\widehat{\mathbf{W}}_{n+1} - \mathbf{W}\right\|_{\mathrm{F}} \le \frac{32\lambda_n\sqrt{r}}{\kappa\big(\widehat{\overline{\mathbf{\Sigma}}}_n\big)}. \tag{10}$$

The proof, which is presented in the appendix, follows along the lines of (Negahban and Wainwright, 2011, Theorem 1).

Remark 2. *We note that a similar result can be also be found in Koltchinskii et al. (2011). In addition, as pointed out in these references, it is possible to extend this result to the case of approximately low rank matrices $\mathbf{W}$ by introducing an additional mispecification error in (10).*

4.2 Controlling the Noise Term

The result in Lemma 1 requires us to set the regularization parameter at round n as in (9). To this end we exploit deviation bounds for martingales. Let us define the filtration $(\overline{\mathcal{F}}_n)_{n\ge 0}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ as follows: $\overline{\mathcal{F}}_0$ is the trivial σ-field $\{\emptyset, \Omega\}$, and for any $n \ge 1$,

$$\overline{\mathcal{F}}_n = \sigma\big(\mathbf{X}_{1,i}, \eta_{1,i}, \ldots, \mathbf{X}_{T,i}, \eta_{T,i},\ i \in [n]\big).$$

We consider the standard noise assumption (Abbasi-Yadkori et al., 2011; Bastani and Bayati, 2020; Cella et al., 2020).

Assumption 4 (subGaussian noise)**.** *The noise variables $(\eta_{t,n})_{t,n}$ are a sequence of sub-Gaussian random variables adapted to the filtration $\{\overline{\mathcal{F}}_n\}_{n\ge 0}$ and such that for any $1 \le t \le T$ and $n \ge 1$,*

$$\mathbb{E}[\eta_{t,n}|\overline{\mathcal{F}}_{n-1}] = 0, \quad \textit{and} \quad \mathbb{E}[\eta_{t,n}^2|\overline{\mathcal{F}}_{n-1}] \le \sigma^2,$$

and $\eta_{1,n}, \ldots, \eta_{T,n}$ are mutually independent conditionally on $\overline{\mathcal{F}}_{n-1}$. We denote by c_η the sub-Gaussian norm of the $\eta_{t,n}$'s, that is, $\max_{t,n}\{\|\eta_{t,n}\|_{\psi_2}\} \le c_\eta$.

Proposition 1 (SubGaussian noise)**.** *Let Assumptions 1 and 4 be satisfied. Then, with probability at least $1 - \delta$*

$$\frac{1}{n}\left\|\mathbf{D}_n\right\|_{\mathrm{op}} \le \lambda_n$$

where

$$\lambda_n = \max_{k\in[K]} \|\mathbf{\Sigma}_k\|_{\mathrm{op}}^{1/2}\left[\sigma\sqrt{\frac{(d+T)\,\log\frac{2N(d+T)}{\delta}}{n}} \bigvee \frac{C\sqrt{\left(T + d + \log\frac{4N}{\delta}\right)\log^3\frac{8N(d+T)}{\delta}}}{n}\right] \tag{11}$$

where $C > 0$ can depend only on $c_\eta, C_{\mathbf{z}}$.

Proof Sketch. The complete proof is given in Appendix B. The proof relies on Freedman's inequality for martingale (Tropp, 2011, Corollary 1.3) combined with a stopping time argument to handle unbounded arms.

□

5 MULTI-TASK BANDITS VIA TRACE NORM REGULARIZATION

In this section we present our proposed algorithm for multi-task learning with linear stochastic bandits. The algorithm

Algorithm 1 Trace-Norm Bandit

Require: Confidence parameter δ, noise variance σ^2

At round $n = 1$ arms are picked randomly
Observe $Y_{1,1}, \ldots, Y_{T,1}$
for $n \in 2, \ldots, N$ **do**
  update $\widehat{\mathbf{W}}_n$ and λ_n according to (4) and (11)
  for $t \in 1, \ldots, T$ **do**
    observe $\mathcal{D}_{t,n}$
    pick $\mathbf{x}_{t,n} \in \arg\max_{\mathbf{x} \in \mathcal{D}_{t,n}} \mathbf{x}^\top [\widehat{\mathbf{W}}_n]_t$
    observe reward $Y_{t,n}$
  end for
end for

relies on the trace-norm regularized estimator in (4) and is displayed in Algorithm 1. It does not require to know any input parameter besides an upper bound on the noise variance σ^2 and the confidence value δ, which are necessary to specify the regularization parameter[1] λ_n. Indeed this is the only parameter required by our policy and it should be chosen large enough (in theory) as prescribed by Proposition 1; see also equation (11).

Algorithm 1 updates the task parameters via a trace-norm regularized estimation using all the already observed data. Similarly to (Oh et al., 2021), our policy is completely greedy and does not need any forced-sampling strategy. This phenomenon was recently highlighted by (Bastani et al., 2021) which proved that greedy policies do not need an explicit exploration and can even be rate-optimal if there is sufficient randomness in the observed contexts (relying on i.i.d. arms assumption). This is unlike the UCB-like policies (like OFUL) or forced-exploration policies Bastani and Bayati (2020). Our trace-norm bandit policy (Alg. 1) belongs to the former class of "exploration-free" greedy policies. More precisely, our policy performs a natural initial exploration thanks to the arms randomness condition (Assumptions 1 and 2) until the empirical covariance matrix satisfies the (RSC) condition. In the regret analysis below this can be seen in the definition of the first N_0 rounds in (12).

5.1 Multi-Task Learning

In this subsection we present our main result which is a high-probability upper bound on the multi-task regret incurred by the trace-norm bandit policy.

For any $\delta \in (0, 1)$, let $N_0(\delta)$ be the smallest integer such that

$$N \geq C \max_{k \in [K]} \|\Sigma_k\|_{\mathrm{op}}^2 \left((r \log d) \log \frac{4TN}{\delta} \right)^2, \qquad (12)$$

[1] Using a regularization is quite common in the bandit literature (e.g. Abbasi-Yadkori et al. (2011); Bastani and Bayati (2020).

where $C > 0$ is some large enough numerical constant that may depend only on the distribution of the arms, in particular $\nu, \omega_{\mathcal{X}}$ and $C_{\mathbf{z}}$.

Note that (12) is a sufficient condition on the minimum number of rounds for our regret bound to be valid. Interestingly it depends on the ambient dimension d only logarithmically. This means that for our greedy policy, we can guarantee that the duration of the implicit exploration phase is at most r^2 (up to logarithmic factors). In situation where the $\mathbf{W}^*$ matrix is low-rank with $r \ll \sqrt{d}$, this represents another benefit of our trace norm policy procedure over (Yang et al., 2020) which requires a much longer exploration phase of at least d^2 rounds for their regret bound to be valid.

We now state the main result.

Theorem 1. *Let Assumptions 1, 2, 3, 4 be satisfied. Assume that $\max_{t \in [T]} \|\mathbf{w}_t\|_2 \leq L$, for some constant $L > 0$ and that $N \geq N_0(\delta)$ for some $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, the multi-task regret of Algorithm 1 is upper bounded (up to logarithmic factors) by*

$$\mathcal{O}\left((T\sqrt{rN} + \sqrt{rdTN}) \right).$$

Our proof is inspired by the one proposed in (Oh et al., 2021) for the single task lasso bandit approach. We present here a sketch summarizing the key steps. Full technical details are provided in Appendix D.

Proof Sketch of Theorem 1. We consider the instantaneous multi-task regret at round n

$$\overline{R}_n = \sum_{t=1}^{T} R_{t,n} = \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}^*, \mathbf{w}_t \rangle - \langle \mathbf{x}_{t,n}, \mathbf{w}_t \rangle.$$

During the first $N_0(\delta)$ rounds, the RSC condition may not be satisfied. During this phase, using a simple conditioning argument, we obtain the following bound. We have with probability at least $1 - \delta$, for any $n \in [N_0]$

$$\sum_{n=1}^{N_0} \overline{R}_n \lesssim T\, L\, N_0 \sqrt{\log(eTN_0K\delta^{-1})}.$$

Starting from the $N_0 + 1$ round, we bound the instantaneous regret as follows:

$$\overline{R}_n \leq \sum_{n=N_0+1}^{N} \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}^* - \mathbf{x}_{t,n}, \mathbf{w}_t - \widehat{\mathbf{w}}_{t,n} \rangle.$$

Using another conditioning, we get with probability at least $1 - \delta$, for any $t \in [T]$, $n \in [N]$,

$$\overline{R}_n \lesssim \sqrt{T} \left\| \mathbf{W} - \widehat{\mathbf{W}}_n \right\|_F \sqrt{\log(eTNK\delta^{-1})}.$$

This bound is of interest if we can guarantee that $\widehat{\mathbf{W}}_n$ is an accurate estimate of $\mathbf{W}$, meaning that $\left\|\mathbf{W} - \widehat{\mathbf{W}}_n\right\|_F$ is small.

To this end, we first prove that $\widehat{\overline{\mathbf{\Sigma}}}_n$ satisfies the RSC condition whp starting from round N_0. The argument is based on concentration bounds for martingales combined with a novel geometric analysis of the cone of matrix $\mathcal{C}_r$; see Lemma 2.

Proposition 2. *Let Assumptions 1 and 2 be satisfied. Assume in addition that $\overline{\mathbf{\Sigma}}$ satisfies the RSC condition. Then, for any $\delta \in (0,1)$ and any $n \geq N_0(\delta)$, with probability at least $1-\delta$, the multi-task empirical matrix $\widehat{\overline{\mathbf{\Sigma}}}_n$ satisfies the RSC condition with constant*

$$\kappa(\widehat{\overline{\mathbf{\Sigma}}}_n) \geq \frac{\kappa(\overline{\mathbf{\Sigma}})}{4\nu\omega_{\mathcal{X}}} > 0.$$

Consequently, this means that we can now use Lemma 1 and Proposition 1 which guarantee with probability at least $1-\delta$, simultaneously for any $n \in [N_0, N]$

$$\left\|\widehat{\mathbf{W}}_n - \mathbf{W}\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{r}{n}} \left(\sqrt{T + d + \log\left(\frac{4N(d+T)}{\delta}\right)} \cdot \log^{3/2}\left(\frac{8N(d+T)}{\delta}\right)\right).$$

Summing over n, we obtain with probability at least $1-\delta$ up to logarithmic factors

$$\sum_{n=N_0+1}^{N} \overline{R}_s \leq C\sqrt{rT(T+d)N},$$

where $C = C\left(L, \eta, \sigma, C_{\mathbf{z}}, \kappa(\overline{\mathbf{\Sigma}}), \max_{1\leq k\leq K}\left\{\|\Sigma_k\|_{\mathrm{op}}^{1/2}\right\}\right)$ is a finite constant under our assumptions.

An union bound summing the regrets for the first phase $n \leq N_0$ and the second phase $n > N_0$ gives the result (up to a rescaling of the constants).

□

5.2 Result Discussion

We now discuss the implication of Theorem 1 and compare it to previous approaches to multi-task representation learning in the bandit setting.

Advantage over ITL. Notice that running **any** T independent policies (ITL) with the linear contextual bandit setting defined in Section 3.1 would yield at best a regret bound of order $T\sqrt{dN}$, up to logs; see e.g. (Lattimore and Szepesvári, 2020, Chapter 19.4, comment 5). Since this regret bound is always larger than the upper bound in Theorem 1 for the proposed MTL strategy, there is a gain in using our method. In particular, if $T > d$, discarding logarithmic factors, the bound for our method is smaller by a factor of order $O(\sqrt{d/r})$, while for $d > T$ the gain is of order $O(\sqrt{T/r})$.

Minimax Optimality. Theorem 2 in (Yang et al., 2020) provides a matching minimax lower bound to our Theorem 1 (up to logarithmic terms). In particular this implies that in the regime $T \geq d$, our policy achieves the minimax regret $T\sqrt{rN}$ (up to logs). This corresponds to the performance of the oracle policy which knows the true underlying representation a-priori.

Comparison to SOTA Approaches. Our upper bound compares favorably to Theorem 1 in (Yang et al., 2020) for the finitely many arms setting. Their result assumes Gaussian arms with non singular covariance matrices. They also assume that $\mathrm{rank}(\mathbf{W}) = r$ with **known** r and $K, T \leq \mathrm{poly}(d)$, $N \geq d^2$. Then Theorem 1 in (Yang et al., 2020) guarantees for their MLinGreedy policy that

$$\mathbb{E}[\overline{R}(T,N)] = O\Big(\left(T\sqrt{rN} + \sqrt{rdTN}\right) \cdot \sqrt{\log(NKT)\log(NTdr)}\log\log N\Big).$$

We stress out that their policy requires the knowledge of the rank r whereas our policy does not. Notably, their regret bound requires $N \geq d^2$ rounds to be valid whereas our regret bound is valid as soon as $N \gtrsim r^2$ (up to logs). Moreover, their analysis requires the invertibility of the arms covariance, whereas we only need the less restrictive RSC condition. Finally we also extend the result to sub-Gaussian arm distributions. With our notation, Hu et al. (2021) obtained a regret bound of order $O(T\sqrt{drN} + d\sqrt{rTN})$ up to logs for their MTLR-OFUL policy in the infinitely many arms setting, provided the rank r is known to their policy.

6 EXPERIMENTS

In this section we validate the policy proposed in Section 5. The experiments displayed in Figures 1 and 2 compare 3 different policies: the Trace-Norm Bandit approach of Algorithm 1 for different choices of the regularization parameter, the Oracle Policy which knows the low dimensional representation $\mathbf{B} \in \mathbb{R}^{d\times r}$ to select the arm to play at each round and the ITL policy which solves each d-dimensional task separately. In order to compute the trace-norm estimator (Eq. 4) we adopt the accelerated gradient method proposed in (Ji and Ye, 2009). In all the experiments, we report results averaged over 5 repetitions.

Data Processing. We conduct numerical experiments on different configurations. Specifically, we analyze the impact of the number of tasks T, the dimension d, noise variance σ^2 and the choice of the regularization parameter. As metric we consider the cumulative reward averaged on all the tasks $\overline{R}(T,N)/T$ as a function of the number of rounds N. The arm set is generated randomly from a standard Gaussian distribution . The task matrix $\mathbf{W}$ has been chosen to be of rank $r \ll d$ with randomly generated gaussian entries.

Result Discussion. Figures 1 and 2 indicate that the pro-

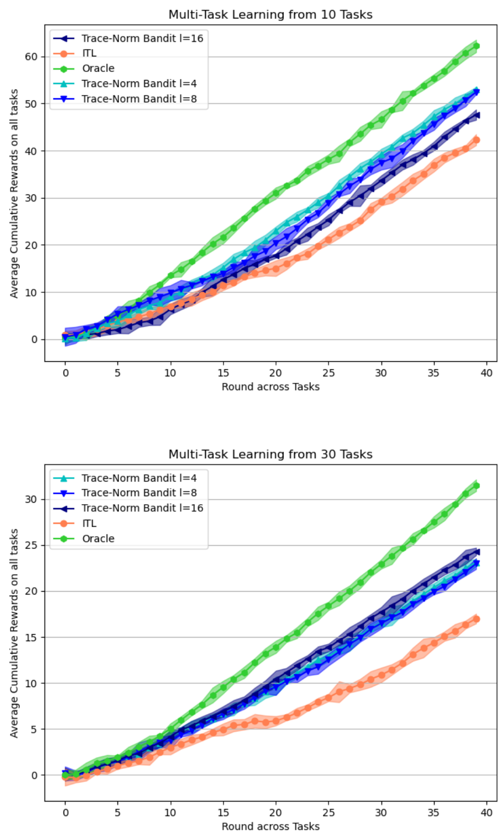

Figure 1: Averaged cumulative reward over all tasks for $T = 10$ (top) and $T = 30$ (bottom). Each task lasts for $N = 40$ rounds, has $K = 10$ arms with $d = 20$ features, noise variance $\sigma^2 = 1$.

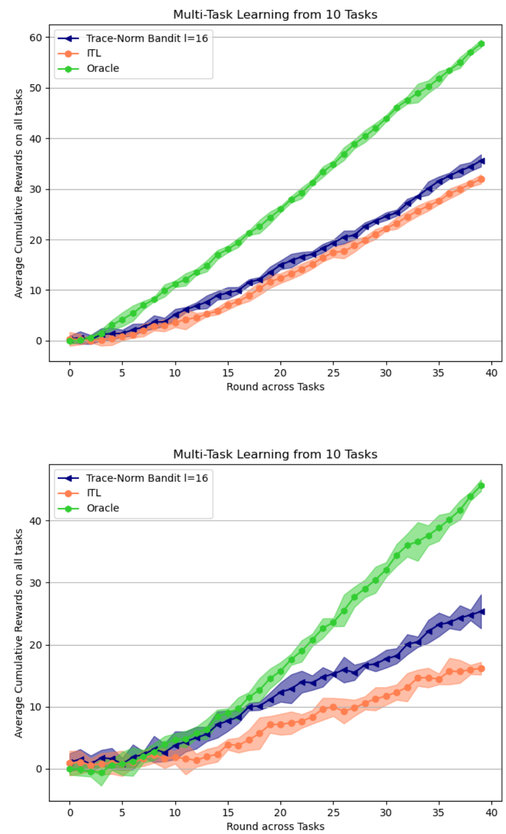

Figure 2: Averaged cumulative reward over all tasks. Each task lasts for $N = 40$ rounds, has $K = 10$ arms with $d = 50$ features and noise variance $\sigma^2 = 1$ (top), $\sigma^2 = 9$ (bottom).

posed multi-task approach performs favorably over independent task learning. Figure 1 highlights that our trace-norm policy performs well and significantly better than ITL policy as long as the regularization parameter used in Eq. (4) is taken large enough according to theory:

$$\lambda_n = l\left[\left(\frac{T+d}{n} + \frac{\log\frac{2}{\delta}}{n}\right) \bigvee \left(\sqrt{\frac{T+d}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{n}}\right)\right].$$

In addition, we also observe that the performance of the trace norm bandit improves and tends to that of the oracle policy as the number of tasks increases. In Figure 2, we investigate the impact of noise variance on the performance of the different policies. We used a higher value of dimension $d = 50$ as it is well-known that the impact of the noise becomes more problematic in high-dimension. We observe that the trace-norm policy is significantly less impacted by increased noise variance than the ITL policy. Indeed the performance of the ITL policy degrades by 48% as the noise variance increases from 1 to 9. Comparatively, the oracle policy and the trace-norm policy only incur a degradation of about 25% as the noise variance increases. This may be due to the fact that the oracle works in a subspace of dimension r, and that the trace-norm policy is able to perform dimensionality reduction through the nuclear norm regularization whereas the ITL policy works in the whole d-dimensional space.

7 LIMITATIONS AND FUTURE WORK

We have studied the benefit of multi-task representation learning in the setting of linear contextual bandit tasks. We proposed a novel bandit policy based on trace-norm regularization which is computationally efficient and does not need knowledge of the rank of the underlying task matrix. We derived an upper bound for the multi-task regret of the proposed policy, showing that it is effective in comparison to learning the tasks independently. Additionally, this regret bound is minimax optimal and, in the regime $T \geq d$, our policy's regret matches the one of the oracle policy which knows the low-rank common representation.

In this work we have restricted our analysis to the case of linear feature learning. Additionally, we still require the designed policy to know one problem parameter which is the variance associated to the noisy term. Relying on (Maurer et al., 2016), in the future an interesting extension would be to go beyond the linearity of the shared representation. Secondly, inspired by (Belloni et al., 2011) a challenge one may try to solve is to design a fully parameter free policy. Namely, a policy which does not require to know the noise variance σ^2. Finally an important future direction is to consider the meta-learning setting, in which the tasks are observed sequentially and the goal is to minimize the regret on future yet-unseen tasks.

Acknowledgements

This work was supported in part by the Chaire Business Analytic for Future Banking, PNRR MUR project PE0000013-FAIR, and the European Union (Project 101070617).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9, 2012.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

Kaito Ariu, Kenshi Abe, and Alexandre Proutière. Thresholded lasso bandit. *arXiv preprint arXiv:2010.11994*, 2020.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Soumya Basu, Branislav Kveton, Manzil Zaheer, and Csaba Szepesvári. No regrets for learning the prior in bandits. *arXiv preprint arXiv:2107.06196*, 2021.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Daniele Calandriello, Alessandro Lazaric, and Marcello Restelli. Sparse multi-task reinforcement learning. In *NIPS*, 2014.

Leonardo Cella and Massimiliano Pontil. Multi-task and meta-learning with sparse linear bandits. In *The Conference on Uncertainty in Artificial Intelligence*, 2021.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1360–1370. PMLR, 13–18 Jul 2020.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Incremental learning-to-learn with statistical guarantees. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 457–466, 2018.

Giulia Denevi, Dimitris Stamos, Carlo Ciliberto, and Massimiliano Pontil. Online-within-online meta-learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, and Jan Peters. Sharing knowledge in multi-task deep reinforcement learning. In *International Conference on Learning Representations*, 2019.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *NIPS*, volume 23, pages 586–594, 2010.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10753–10763. Curran Associates, Inc., 2020.

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

Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *International Conference on Machine Learning*, pages 457–464, 2009.

Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32:5877–5887, 2019.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*. Lecture Notes in Mathematics. Springer, 2011. ISBN 9783642221460. URL https://books.google.fr/books?id=D5Jxen3_xkAC.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Ilja Kuzborskij, Leonardo Cella, and Nicolò Cesa-Bianchi. Efficient linear bandits through matrix sketching. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 177–185. PMLR, 2019.

Branislav Kveton, Csaba Szepesvári, Anup Rao, Zheng Wen, Yasin Abbasi-Yadkori, and S Muthukrishnan. Stochastic low-rank bandits. *arXiv preprint arXiv:1712.04644*, 2017.

Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-thompson sampling. *arXiv preprint arXiv:2102.06129*, 2021.

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

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080, 2017.

Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204, 2011.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 460–468. PMLR, 13–15 Apr 2021.

Andreas Maurer and Massimiliano Pontil. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pages 55–76. PMLR, 2013.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17 (81):1–32, 2016.

Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. In *Proceedings of the 38-th International Conference on Machine Learning*, volume 139, pages 8271–8280, 2021.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, page 395–411, 2010.

Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

Max Simchowitz, Christopher Tosh, Akshay Krishnamurthy, Daniel Hsu, Thodoris Lykouris, Miroslav Dudík, and Robert E Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. *arXiv preprint arXiv:2107.01509*, 2021.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16(none):262 – 270, 2011. doi: 10.1214/ECP.v16-1624. URL https://doi.org/10.1214/ECP.v16-1624.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Xue Wang, Mingcheng Wei, and Tao Yao. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pages 5200–5208, 2018.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

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

APPENDIX

This appendix provides full proofs of the results stated in the main body of the paper. It is organized as follows:

- Appendix A contains the proof of Lemma 1, which is the oracle inequality associated to the error $\Delta_{n+1} = \widehat{\mathbf{W}}_{n+1} - \mathbf{W}$ considering non i.i.d. data.
- Appendix B presents the proof of Proposition 1 on the control to the operator norm of matrix $\mathbf{D}_n = \sum_{t=1}^{T}\sum_{i=1}^{n} \eta_{t,i}\, \mathbf{x}_{t,i} \otimes \mathbf{e}_t$.
- Appendix C contains the proof of Proposition 2. It relies on a matrix perturbation argument and a novel analysis of the uniform deviation of the empirical arms covariance on the cone $\mathcal{C}_r$ of approximately low-rank matrices (Lemmas 2 and 3). These results are then combined to relate the RSC constant $\kappa(\overline{\mathbf{\Sigma}})$ of $\overline{\mathbf{\Sigma}}$ to that of its empirical counterpart $\kappa(\widehat{\overline{\mathbf{\Sigma}}}_n)$.
- Appendix D contains the proof for the regret upper bound associated to the policy of Alg. 1.
- In Appendix E, we provide additional numerical experiments.

A PROOF OF LEMMA 1

By definition of $\widehat{\mathbf{W}}_n$, we have for any $\mathbf{W} \in \mathbb{R}^{d\times T}$,

$$\frac{1}{n}\sum_{t=1}^{T}\left\|\mathbf{y}_{t,n} - \mathbf{X}_{t,n}[\widehat{\mathbf{W}}_{n+1}]_t\right\|_2^2 + \lambda_n\left\|\widehat{\mathbf{W}}_{n+1}\right\|_* \leq \frac{1}{n}\sum_{t=1}^{T}\left\|\mathbf{y}_{t,n} - \mathbf{X}_{t,n}[\mathbf{W}]_t\right\|_2^2 + \lambda_n\left\|\mathbf{W}\right\|_* .$$

We define the error matrix $\Delta_{n+1} = \widehat{\mathbf{W}}_{n+1} - \mathbf{W}$ and introduce the following operator $\mathcal{A} : \mathbb{R}^{d\times T} \to \mathbb{R}^{n\times T}$ and its adjoint $\mathcal{A}^* : \mathbb{R}^{n\times T} \to \mathbb{R}^{d\times T}$ as

$$[\mathcal{A}(\mathbf{W})]_n^t = \langle \mathbf{x}_{t,n}\mathbf{e}_t^\top, \mathbf{W}\rangle = \mathrm{Tr}(\mathbf{W}\mathbf{e}_t\mathbf{x}_{t,n}) = \mathbf{x}_{t,n}^\top\mathbf{w}_t$$
$$\mathcal{A}^*(\mathbf{H}_n) = \sum_{t=1}^{T}\sum_{i=1}^{n}\mathbf{x}_{t,i}\mathbf{e}_t^\top\eta_{t,i} = \mathbf{D}_n \quad \in \mathbb{R}^{d\times T},$$

where $\boldsymbol{\eta}_{t,n} = (\eta_{t,1},\ldots,\eta_{t,n})^\top \in \mathbb{R}^n$ and $\mathbf{H}_n := (\boldsymbol{\eta}_{1,n},\ldots,\boldsymbol{\eta}_{T,n}) \in \mathbb{R}^{n\times T}$.

Using this notation the following hold

$$\begin{aligned}\frac{1}{n}\left\|\mathcal{A}(\Delta_{n+1})\right\|_{\mathrm{F}}^2 = \frac{1}{n}\sum_{t=1}^{T}\left\|[\mathcal{A}(\Delta_{n+1})]^t\right\|_2^2 &\leq \frac{1}{n}\langle\mathbf{H}_n, \mathcal{A}(\Delta_{n+1})\rangle + \lambda_n\left(\|\mathbf{W}\|_* - \left\|\widehat{\mathbf{W}}_{n+1}\right\|_*\right)\\ &\leq \frac{1}{n}\langle\mathbf{H}_n, \mathcal{A}(\Delta_{n+1})\rangle + \lambda_n\left(\left\|\widehat{\mathbf{W}}_{n+1} + \Delta_{n+1}\right\|_* - \left\|\widehat{\mathbf{W}}_{n+1}\right\|_*\right)\\ &\leq \frac{1}{n}\langle\mathbf{H}_n, \mathcal{A}(\Delta_{n+1})\rangle + \lambda_n\left(\|\Delta_{n+1}\|_*\right).\end{aligned}$$

Considering now the first term on the RHS and applying Holder's inequality we have

$$\frac{1}{n}|\langle\mathbf{H}_n, \mathcal{A}(\Delta_{n+1})\rangle| = \frac{1}{n}|\langle\mathcal{A}^*(\mathbf{H}_n), \Delta_{n+1}\rangle| \leq \frac{1}{n}\left\|\mathcal{A}^*(\mathbf{H}_n)\right\|_{\mathrm{op}}\left\|\Delta_{n+1}\right\|_* .$$

Now, considering $\lambda_n \geq \frac{1}{n}\left\|\mathcal{A}^*(\mathbf{H}_n)\right\|_{\mathrm{op}}$ the following holds:

$$\frac{1}{n}\left\|\mathcal{A}(\Delta_{n+1})\right\|_{\mathrm{F}}^2 \leq 2\lambda_n\left\|\Delta_{n+1}\right\|_* . \tag{13}$$

Relying on (Negahban and Wainwright, 2011, Lemma 1) we can decompose the error matrix Δ_{n+1} as $\Delta'_{n+1} + \Delta''_{n+1}$ such that Δ'_{n+1} is of rank at most $2r$ and

$$\left\|\Delta_{n+1}\right\|_* \leq 4\left\|\Delta'_{n+1}\right\|_* . \tag{14}$$

Assuming the RSC condition to be met with constant $\kappa\left(\widehat{\overline{\boldsymbol{\Sigma}}}_n\right)$, starting from equation (13) we get

$$\|\Delta_{n+1}\|_{\mathrm{F}}^2 \leq= \frac{\|\mathcal{A}(\Delta_{n+1})\|_{\mathrm{F}}^2}{2n\kappa\left(\widehat{\overline{\boldsymbol{\Sigma}}}_n\right)} \leq \frac{\lambda_n \ \|\Delta_{n+1}\|_*}{\kappa\left(\widehat{\overline{\boldsymbol{\Sigma}}}_n\right)}.$$

These last two results combined with (Negahban and Wainwright, 2011, Lemma 1) give that $\left\|\Delta'_{n+1}\right\|_* \leq \sqrt{2r}\left\|\Delta'_{n+1}\right\|_{\mathrm{F}}$, from which we conclude that

$$\|\Delta_{n+1}\|_{\mathrm{F}} \leq \frac{32\lambda_n\sqrt{r}}{\kappa\left(\widehat{\overline{\boldsymbol{\Sigma}}}_n\right)}.$$

Symbol	Description
$[n]$	The set $\{1,\ldots,n\}$, given a positive integer n
T	Number of tasks
N	Time horizon associated to each single task
d	Dimension of context vectors
$a \vee b$	The maximum between a and b ($\max(a,b)$)
$\|\cdot\|_{\psi_2}$	The sub-Gaussian norm with $\psi_2(s) = e^{s^2} - 1$ (See e.g. page 215 in Koltchinskii (2011))
$\mathbf{e}_1,\ldots,e_T \in \mathbb{R}^T$	The standard basis indicator vectors, i.e. $\mathbf{e}_{t,j} = 1$, if $j = t$ and 0 otherwise, for all $t, j \in [T]$
$\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_T] \in \mathbb{R}^{d\times T}$	Matrix of T regression tasks (we also use the notation $[\mathbf{W}]_t \equiv w_t,\ t \in [T]$)
r	Rank of the task matrix $\mathbf{W}$
K	Number of arms
p	Joint distribution on $\mathbb{R}^{dK}$ (from which K arm vectors are sample)
$\mathcal{D}_{t,n},\ t\in[T], n\in[N]$	Decision sets (each containing K arm vectors) sampled i.i.d. from p
$\mathbf{x}_{t,n} \in \mathcal{D}_{t,n}$	Arm vector chosen in task $t \in [T]$ at round $n \in [N]$
$\mathbf{x}^*_{t,n} \in \mathcal{D}_{t,n}$	Optimal arm vector in task $t \in [T]$ during round $n \in [N]$
$\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$	Theoretical covariance matrix - see eq. (5)
$\boldsymbol{\Sigma}_{t,n} \in \mathbb{R}^{d\times d}$	Adapted covariance matrix for task t at round n; see eq. (32)
$\widehat{\boldsymbol{\Sigma}}_{t,n} \in \mathbb{R}^{d\times d}$	Empirical covariance matrix for task t at round n; see eq. (6)
$\overline{\boldsymbol{\Sigma}} \in \mathbb{R}^{dT\times dT}$	T-block diagonal matrix $\mathrm{diag}(\boldsymbol{\Sigma},\ldots,\boldsymbol{\Sigma})$
$\overline{\boldsymbol{\Sigma}}_n \in \mathbb{R}^{dT\times dT}$	T-block diagonal matrix $\mathrm{diag}(\boldsymbol{\Sigma}_{1,n},\ldots,\boldsymbol{\Sigma}_{T,n})$
$\widehat{\overline{\boldsymbol{\Sigma}}}_n \in \mathbb{R}^{dT\times dT}$	T-block diagonal matrix $\mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_{1,n},\ldots,\widehat{\boldsymbol{\Sigma}}_{T,n})$
$\|\mathbf{x}\|,\ \|\mathbf{x}\|_1,\ \|\mathbf{x}\|_\infty$	Euclidean, ℓ_1 and maximum norm associated to a vector $\mathbf{x}$
$[\mathbf{A}]_t$	The t-th column of matrix $\mathbf{A} \in \mathbb{R}^{d\times T}$
$\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})$	Minimum and maximum eigenvalues of a square symmetric matrix $\mathbf{A}$
$\sigma_{\min}(\mathbf{A}), \sigma_{\max}(\mathbf{A})$	Minimum and maximum singular values of matrix $\mathbf{A}$
$\|\mathbf{A}\|_*$	Trace norm of matrix $\mathbf{A}$ (sum of its singular values)
$\|\mathbf{A}\|_{\mathrm{F}}$	Frobenius norm of matrix $\mathbf{A}$ (ℓ_2 norm o matrix elements / singular values)
$\|\mathbf{A}\|_{\mathrm{op}}$	Operator norm of matrix $\mathbf{A}$ (maximum singular value)

B PROOF OF PROPOSITION 1

B.1 Preliminary results

We consider the stochastic process $\{\mathbf{M}_n\}_{n\geq 0}$ defined as $\mathbf{M}_0 = 0$ a.s. and for any $n \geq 1$

$$\mathbf{M}_n = \sum_{t=1}^{T} \mathbf{x}_{t,n} \otimes \mathbf{x}_{t,n} - \mathbb{E}\left[\mathbf{x}_{t,n} \otimes \mathbf{x}_{t,n} | \overline{\mathcal{F}}_{n-1}\right].$$

By definition of the Trace-Norm bandit (Algorithm 1), $\{M_n\}_{n\geq 0}$ is a $\overline{\mathcal{F}}_{n-1}$-martingale. Furthermore, given the past history $\overline{\mathcal{F}}_{n-1}$, we select at round n for each task t the arm $\mathbf{x}_{t,n} \in \mathcal{D}_{t,n}$ where the sets $\mathcal{D}_{t,n}, t \in [T]$, are mutually independent. This means that the arms $\mathbf{x}_{t,n}$, $t \in [T]$, are mutually independent given the past history $\overline{\mathcal{F}}_{n-1}$.

Next, we use a standard argument to control the operator norm of $\mathbb{M}_n$. Fix $\epsilon \in (0, 1/2)$. An ϵ-net $\mathcal{N}_\epsilon$ of $\mathbb{S}^d$ is a subset of $\mathbb{S}^d$ such that for any $\mathbf{u} \in \mathbb{S}^d$, there exists $\mathbf{v} \in \mathcal{N}_\epsilon$ such that $\|\mathbf{u} - \mathbf{v}\| \leq \epsilon$. Corollary 4.2.13 in Vershynin (2018) guarantees the existence of an ϵ-net $\mathcal{N}_\epsilon$ of $\mathbb{S}^d$ such that

$$|\mathcal{N}_\epsilon| \leq \left(1 + \frac{2}{\epsilon}\right)^d. \tag{15}$$

Similarly to the first step of the proof of Theorem 4.4.5 in Vershynin (2018), we get

$$\|\mathbf{M}_n\|_{\mathrm{op}} \leq \frac{1}{1-2\epsilon} \max_{\mathbf{u}\in\mathcal{N}_\epsilon} \{\langle \mathbf{M}_n \mathbf{u}, \mathbf{u}\rangle\} \leq \frac{1}{1-2\epsilon} \max_{\mathbf{u}\in\mathcal{N}_\epsilon} \left\{ \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}, u\rangle^2 - \mathbb{E}[\langle \mathbf{x}_{t,n}, u\rangle^2 | \overline{\mathcal{F}}_{n-1}] \right\}. \tag{16}$$

In view of Assumption 1, we have $\langle \mathbf{x}_{t,n}, u\rangle = \langle \mathbf{z}_{t,n}, \mathbf{\Sigma}_k^{1/2}\mathbf{u}\rangle$ and $\mathbb{E}[\langle \mathbf{x}_{t,n}, \mathbf{u}\rangle^2 | \overline{\mathcal{F}}_{n-1}] = \langle \mathbf{\Sigma}_{k(t)}\mathbf{u}, \mathbf{u}\rangle$ for some $k(t) \in [K]$. We apply now the Hanson-Wright's inequality conditionally on $\overline{\mathcal{F}}_{n-1}$ to get for any $x > 0$

$$\mathbb{P}\left(\sum_{t=1}^{T} \langle \mathbf{z}_{t,n}, \mathbf{\Sigma}_{k(t)}^{1/2}\mathbf{u}\rangle^2 - \langle \mathbf{\Sigma}_{k(t)}\mathbf{u}, \mathbf{u}\rangle \geq C\left(\sqrt{T \max_{k\in[K]}\{\left\|\mathbf{\Sigma}_k^{1/2}\mathbf{u}\otimes\mathbf{u}\mathbf{\Sigma}_k^{1/2}\right\|_F^2\}\, x} + \max_{k\in[K]}\{\|\mathbf{\Sigma}_k\mathbf{u}\otimes\mathbf{u}\mathbf{\Sigma}_k\|_{\mathrm{op}}\}\, x\right) |\overline{\mathcal{F}}_{n-1}\right) \leq e^{-x},$$

where $C > 0$ is a numerical constant which can depend only on $C_{\mathbf{z}}$. Note that $\left\|\mathbf{\Sigma}_k^{1/2}\mathbf{u}\otimes\mathbf{u}\mathbf{\Sigma}_k^{1/2}\right\|_F = \left\|\mathbf{\Sigma}_k^{1/2}\mathbf{u}\otimes\mathbf{u}\mathbf{\Sigma}_k^{1/2}\right\|_{\mathrm{op}} \|\mathbf{u}\|^2 \leq \|\mathbf{\Sigma}_k\|_{\mathrm{op}}$. Hence we get for any $x > 0$

$$\mathbb{P}\left(\sum_{t=1}^{T} \langle \mathbf{z}_{t,n}, \mathbf{\Sigma}_{k(t)}\mathbf{u}\rangle^2 - \langle \mathbf{\Sigma}_{k(t)}\mathbf{u}, \mathbf{u}\rangle \geq C \max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\} \left(\sqrt{T\, x} + x\right) |\overline{\mathcal{F}}_{n-1}\right) \leq e^{-x}.$$

We define now the event

$$\Omega_n = \bigcap_{\mathbf{u}\in\mathcal{N}_\epsilon} \left\{ |\langle \mathbf{M}_n \mathbf{v}, \mathbf{u}\rangle| \leq C \max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\} \left(\sqrt{T\, x} + x\right)\right\}.$$

Set $\epsilon = 1/4$. A simple union bound combining (15) with the last two displays gives

$$\mathbb{P}\left(\Omega_n^c\right) \leq |\mathcal{N}_\epsilon|\, e^{-t'} \leq 2e^{d\log(9) - x}.$$

Now we set $x = \log(4\delta^{-1}) + d\log(9)$ for some $\delta \in (0,1)$. Consequently, we obtain that

$$\mathbb{P}\left(\Omega_n\right) \geq 1 - \delta/4.$$

It follows, in view of (37), with probability at least $1 - \delta$

$$\|\mathbf{M}_n\|_{\mathrm{op}} \leq C' \max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\} \left(\sqrt{T\,(d + \log(4\delta^{-1}))} + d + \log(4\delta^{-1})\right), \tag{17}$$

for some numerical constant $C' > 0$ that can depend only on $C_{\mathbf{z}}$.

We note that
$$\left\|\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{x}_{t,n}\otimes\mathbf{x}_{t,n}|\overline{\mathcal{F}}_{n-1}\right]\right\|_{\mathrm{op}}\leq T\max_{k}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\},$$
since $\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{x}_{t,n}\otimes\mathbf{x}_{t,n}|\overline{\mathcal{F}}_{n-1}\right]=\sum_{t=1}^{T}\mathbf{\Sigma}_{k(t)}$, where $k(t)\in[K]$. Combining this observation with (17), we prove that the following event

$$\Omega_n=\left\{\left\|\sum_{t=1}^{T}\mathbf{x}_{t,n}\otimes\mathbf{x}_{t,n}\right\|_{\mathrm{op}}\leq\max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\}\left(T+C'\left(\sqrt{T\left(d+\log(4N\delta^{-1})\right)}+d+\log(4N\delta^{-1})\right)\right)\right\}\tag{18}$$

satisfies $\mathbb{P}(\Omega_n)\geq 1-\delta/(4N)$. Next we also introduce the event
$$\overline{\Omega}_n=\bigcap_{l=0}^{n}\Omega_l\in\overline{\mathcal{F}}_n,\tag{19}$$
with Ω_0 being the whole sample space. The Bayes rule gives
$$\mathbb{P}\left(\overline{\Omega}_N\right)=\prod_{n=1}^{N}\mathbb{P}\left(\Omega_n|\bigcap_{k=0}^{n-1}\Omega_k\right)\geq\left(1-\frac{\delta}{4N}\right)^N.$$
Bernoulli's inequality ($(1+x)^n\geq 1+nx$ for any $x>-1$ and integer $n\geq 1$) gives
$$\mathbb{P}\left(\overline{\Omega}_N\right)=\prod_{n=1}^{N}\mathbb{P}\left(\Omega_n|\bigcap_{k=0}^{n-1}\Omega_k\right)\geq 1-\frac{\delta}{4}.\tag{20}$$

Define for any $n\in[N]$ the events
$$\Omega'_n=\bigcap_{t=1}^{T}\left\{|\eta_{t,n}|\leq c_\eta\sqrt{\log(8TN\delta^{-1})}\right\},\tag{21}$$
and
$$\overline{\Omega}'_n=\bigcap_{l=0}^{n}\Omega'_l,\tag{22}$$
with Ω'_0 being the whole sample space.

Assumption 4 guarantees that for any $\delta\in(0,1)$, with probability at least $1-\delta$
$$\mathbb{P}\left(\Omega'_n|\overline{\mathcal{F}}_{n-1}\right)\geq 1-\frac{\delta}{4N}.\tag{23}$$
and
$$\mathbb{P}\left(\overline{\Omega}'_N|\overline{\mathcal{F}}_{N-1}\right)\geq 1-\frac{\delta}{4}.\tag{24}$$

B.2 Main proof

Checking the martingale structure. Let us define the stochastic process $(\mathbf{D}_n)_{n\geq 0}$ as $\mathbf{D}_0=0$ a.s. and
$$\mathbf{D}_n:=\sum_{t=1}^{T}\sum_{i=1}^{n}\eta_{t,i}\,\mathbf{x}_{t,i}\mathbf{e}_t^\top\in\mathbb{R}^{d\times T}.\tag{25}$$

By definition of the Trace-Norm bandit (Algorithm 1), given the past history $\overline{\mathcal{F}}_{n-1}$, we select at round n the arms for the T tasks: $\mathbf{x}_{t,n}\in\mathcal{D}_{t,n}$ with $\mathcal{D}_{t,n}$ independent of $(\eta_{t,n})_{1\leq t\leq T}$. This means that
$$(\mathbf{x}_{t,n})_{1\leq t\leq T}\perp\!\!\!\perp(\eta_{t,n})_{1\leq t\leq T}\Big|\overline{\mathcal{F}}_{n-1}.$$

Consequently, under Assumption 1 and 4, $(\mathbf{D}_n)_{n\geq 0}$ is a square-root integrable martingale adapted to the filtration $\{\overline{\mathcal{F}}_n\}_{n\geq 0}$.

We would like to apply the Freedman inequality for matrix martingales (Tropp, 2011, Corollary 1.3). However this result is for bounded martingales. Therefore it is not directly applicable to $(\mathbf{D}_n)_{n\geq 0}$. To remedy this difficulty. We introduce the following stopping times:

$$\tau_1 = \inf\left\{ n \geq 0 \,:\, \left\|\sum_{t=1}^{T} \mathbf{x}_{t,n} \otimes \mathbf{x}_{t,n}\right\|_{\mathrm{op}} \geq \max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\}\left(T + C'\left(\sqrt{T\,(d+\log(4N\delta^{-1}))} + d + \log(4N\delta^{-1})\right)\right)\right\}, \tag{26}$$

and

$$\tau_2 = \inf\left\{ n \geq 0 \,:\, \min_{t\in[T]}\{|\eta_{t,n}|\} \geq c_\eta \sigma \sqrt{\log(8TN\delta^{-1})}\right\}. \tag{27}$$

Again by definition of the Trace-Norm bandit and assumptions on the noise, τ_1 and τ_2 are both stopping times relative to the filtration $\{\overline{\mathcal{F}}_n\}_{n\geq 0}$, so is $\tau = \tau_1 \wedge \tau_2$. Hence the stopped process $\{\mathbf{D}^{\tau}_n\}_{n\geq 0}$ defined as $\mathbf{D}^{\tau}_n = \mathbf{D}_{n\wedge\tau}$ is also a martingale adapted to the filtration $\{\overline{\mathcal{F}}_n\}_{n\geq 0}$. Furthermore $\{\mathbf{D}^{\tau}_n\}_{n\geq 0}$ is a bounded martingale. Hence we can apply the Freedman inequality to it.

By definition of τ_1 and (19), we have $\{\tau_1 > N\} = \overline{\Omega}_N$. Similarly for τ_2 in view of (22), we have $\{\tau_2 > N\} = \overline{\Omega}'_N$. Hence

$$\begin{aligned}
\mathbb{P}\left(\|\mathbf{D}_n\|_{\mathrm{op}} \geq t\right) &= \mathbb{P}\left(\{\mathbf{D}_n \geq t\} \cap \{\tau > N\}\right) + \mathbb{P}\left(\{\|\mathbf{D}_n\|_{\mathrm{op}} \geq t\} \cap \{\tau \leq N\}\right) \\
&\leq \mathbb{P}\left(\|\mathbf{D}^{\tau}_n\|_{\mathrm{op}} \geq t\right) + \mathbb{P}\left(\{\tau_1 \leq N\} \cup \{\tau_2 \leq N\}\right) \\
&= \mathbb{P}\left(\|\mathbf{D}^{\tau}_n\|_{\mathrm{op}} \geq t\right) + \mathbb{P}\left(\tau_1 \leq N\right) + \mathbb{P}\left(\tau_2 \leq N\right) \\
&\leq \mathbb{P}\left(\|\mathbf{D}^{\tau}_n\|_{\mathrm{op}} \geq t\right) + \delta/2,
\end{aligned} \tag{28}$$

where we have used (20) and (24) in the last line.

Application of Freedman's inequality. We now use (Tropp, 2011, Corollary 1.3) to control $\{\mathbf{D}^{\tau}_n\}_{n\geq 0}$.

Theorem 2 (Corollary 1.3 in Tropp (2011))**.** *Consider a matrix martingale $\{\boldsymbol{Y}_n \,:\, n = 0,1,2,\dots\}$ whose values are matrices with dimension $d_1 \times d_2$, and let $\{\boldsymbol{X}_k : n = 1,2,3,\dots\}$ be the difference sequence. Assume that the difference sequence is uniformly bounded:*

$$\|\boldsymbol{X}_n\|_{\mathrm{op}} \leq R \quad \textit{almost surely} \quad \textit{for } n = 1,2,3,\dots.$$

Define two predictable quadratic variation processes for this martingale:

$$\begin{aligned}
\boldsymbol{W}_{\mathrm{col},\,k} &:= \sum\nolimits_{j=1}^{n} \mathbb{E}_{j-1}\left(\boldsymbol{X}_j \boldsymbol{X}_j^*\right) \quad \textit{and} \\
\boldsymbol{W}_{\mathrm{row},\,k} &:= \sum\nolimits_{j=1}^{n} \mathbb{E}_{j-1}\left(\boldsymbol{X}_j^* \boldsymbol{X}_j\right) \quad \textit{for } n = 1,2,3,\dots.
\end{aligned}$$

Then, for all $t \geq 0$ and $\overline{\sigma}^2 > 0$,

$$\mathbb{P}\left(\exists n \geq 0 : \|\boldsymbol{Y}_n\| \geq t \textit{ and } \max\{\|\boldsymbol{W}_{\mathrm{col},\,n}\|_{\mathrm{op}}, \|\boldsymbol{W}_{\mathrm{row},\,n}\|_{\mathrm{op}}\} \leq \overline{\sigma}^2\right) \leq (d_1 + d_2)\cdot \exp\left\{-\frac{-t^2/2}{\overline{\sigma}^2 + Rt/3}\right\}.$$

We now check the conditions of Theorem 2. In view of (18)-(19) and (21)-(22), we have on the event $\{\tau > N\} = \{\tau_1 >$

$N\} \cap \{\tau_2 > N\}$, for any $n \in [N]$,

$$\begin{aligned}
\left\|\sum_{t=1}^{T} \eta_{t,n} \mathbf{x}_{t,n} \otimes \mathbf{e}_t\right\|_{\mathrm{op}} &= \sqrt{\left\|\left(\sum_{t=1}^{T} \eta_{t,n} \mathbf{x}_{t,n} \otimes \mathbf{e}_t\right)\left(\sum_{t=1}^{T} \eta_{t,n} \mathbf{e}_t \otimes \mathbf{x}_{t,n}\right)\right\|_{\mathrm{op}}} \\
&= \sqrt{\left\|\sum_{t=1}^{T} \eta_{t,n}^2 \mathbf{x}_{t,n} \otimes \mathbf{x}_{t,n}\right\|_{\mathrm{op}}} \leq \max_{t\in[T]}\{|\eta_{t,n}|\} \sqrt{\left\|\sum_{t=1}^{T} \mathbf{x}_{t,n} \otimes \mathbf{x}_{t,n}\right\|_{\mathrm{op}}} \\
&\leq c_\eta \sqrt{\log(8TN\delta^{-1})} \sqrt{\max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\} \left(T + C'\left(\sqrt{T\left(d+\log(4N\delta^{-1})\right)} + d + \log(4N\delta^{-1})\right)\right)} \\
&\leq c_\eta \sqrt{1 + C'/2} \sqrt{\log(8TN\delta^{-1})} \sqrt{\max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\}\left(T + d + \log(4N\delta^{-1})\right)}.
\end{aligned}$$

Next we have

$$\boldsymbol{W}_{\mathrm{col}} = \sum_{s=1}^{n} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{n} \eta_{t,s}^2 \mathbf{x}_{t,s} \otimes \mathbf{x}_{t,s} | \overline{\mathcal{F}}_{s-1}\right] = \sigma^2 \sum_{t=1}^{T}\sum_{s=1}^{n} \mathbf{\Sigma}_{k(t,s)},$$

where $k(t,s) \in [K]$ for any t, s. Hence

$$\|\boldsymbol{W}_{\mathrm{col}}\|_{\mathrm{op}} \leq Tn\,\sigma^2 \max_{k\in[K]} \left\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\right\}.$$

We proceed similarly for $\boldsymbol{W}_{\mathrm{row}}$

$$\boldsymbol{W}_{\mathrm{row}} = \sum_{s=1}^{n} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{n} \eta_{t,s}^2 \left\|\mathbf{x}_{t,s}\right\|^2 \mathbf{e}_t \otimes \mathbf{e}_t | \overline{\mathcal{F}}_{s-1}\right] \leq n\sigma^2 \max_{k\in[K]} \left\{\mathrm{tr}(\mathbf{\Sigma}_k)\right\} I_T.$$

Hence, we get

$$\|\boldsymbol{W}_{\mathrm{row}}\|_{\mathrm{op}} \leq n\sigma^2 \max_{k\in[K]} \left\{\mathrm{tr}(\Sigma_k)\right\} \leq nd\,\sigma^2 \max_{k\in[K]} \left\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\right\}.$$

Define the event

$$\begin{aligned}
A_n = \Bigg\{ \frac{\|\mathbf{D}_n\|_{\mathrm{op}}}{n} \leq \max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}^{1/2}\} &\left(\sigma \sqrt{\frac{(d+T)\left(\log(2\delta^{-1}N(d+T))\right)}{n}}\right. \\
&\left. \bigvee \frac{c_\eta\sqrt{1+C'/2}\sqrt{\log(8TN\delta^{-1})}\sqrt{(T+d+\log(4N\delta^{-1}))}\left(\log(2\delta^{-1}N(d+T))\right)}{n}\right\}.
\end{aligned}$$

Applying Theorem 2, we get for any $t > 0$

$$\mathbb{P}\left(A_n^c | \bar{\Omega}_n\right) \leq \frac{\delta}{2N}. \tag{29}$$

An elementary argument combining the previous display with (55) gives

$$\mathbb{P}\left(A_n^c\right) = \mathbb{P}\left(A_n^c \cap \bar{\Omega}_n\right) + \mathbb{P}\left(A_n^c \cap \bar{\Omega}_n^c\right) \leq \mathbb{P}\left(A_n^c | \bar{\Omega}_n\right) + \mathbb{P}\left(\bar{\Omega}_n^c\right) \leq \frac{\delta}{N}.$$

From the previous display and an union bound, we immediately deduce, with probability at least $1-\delta$,

$$\begin{aligned}
\max_{1\leq n\leq N} \frac{\|\mathbf{D}_n\|_{\mathrm{op}}}{n} \leq \max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}^{1/2}\} &\left(\sigma \sqrt{\frac{(d+T)\left(\log(2\delta^{-1}N(d+T))\right)}{n}}\right. \\
&\left. \bigvee \frac{c_\eta\sqrt{1+C'/2}\sqrt{(T+d+\log(4N\delta^{-1}))}\left(\log^{3/2}(8\delta^{-1}N(d+T))\right)}{n}\right).
\end{aligned} \tag{30}$$

C PROOF OF PROPOSITION 2

C.1 Concentration bounds on the arms covariance.

For every $t \in [T]$, we recall that the empirical covariance matrix for task t as

$$\widehat{\mathbf{\Sigma}}_{t,n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{t,i}\mathbf{x}_{t,i}^{\top} \tag{31}$$

and the corresponding adapted covariance matrix (w.r.t. the filtration $(\overline{\mathcal{F}}_n)_{n\geq 0}$) as

$$\mathbf{\Sigma}_{t,n}=\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{x}_{t,i}\mathbf{x}_{t,i}^{\top}\middle|\overline{\mathcal{F}}_{i-1}\right]. \tag{32}$$

Here *adapted* means that the covariance matrix $\mathbf{\Sigma}_{t,n}$ is fully known at time n since it depends only on the past history $\overline{\mathcal{F}}_{n-1}$.

Moreover, we use the notation $\overline{\mathbf{\Sigma}}, \widehat{\overline{\mathbf{\Sigma}}}_n, \overline{\mathbf{\Sigma}}_n \in \mathbb{R}^{dT\times dT}$ for the theoretical, the empirical and the adapted multi-task matrices, respectively. They are all block diagonal and composed by the corresponding T single task $d \times d$ matrices on the diagonal (e.g. $\overline{\mathbf{\Sigma}}_n = \mathrm{diag}(\overline{\mathbf{\Sigma}}_{1,n}, \ldots, \overline{\mathbf{\Sigma}}_{T,n})$).

The following lemma gives a high probability bound on the deviation of $\widehat{\overline{\mathbf{\Sigma}}}_n$ from $\overline{\mathbf{\Sigma}}_n$ according to the operator norm restricted to the cone $\mathcal{C}_r$ defined in (7). We define for any symmetric $dT \times dT$ matrix A

$$\|A\|_{\mathrm{op},\mathcal{C}_{\mathrm{r}}} = \max_{\mathbf{\Delta}\in\mathcal{C}_r\,:\,\|\mathbf{\Delta}\|_F=1} \{\langle A\mathrm{Vec}(\mathbf{\Delta}), \mathrm{Vec}(\mathbf{\Delta})\rangle\}.$$

Lemma 2 (Concentration)**.** *Let Assumption 1 be satisfied. For any $n \geq 1$ and $\delta \in (0,1)$, with probability at least $1-\delta$, for any $n \in [N]$*

$$\left\|\widehat{\overline{\mathbf{\Sigma}}}_n - \overline{\mathbf{\Sigma}}_n\right\|_{\mathrm{op},\mathcal{C}_{\mathrm{r}}} \leq C \max_{1\leq k\leq K}\left\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\right\} r\left(\frac{(1+\log(4NT\delta^{-1}))\log^2(d)}{\sqrt{n}} + \frac{(1+\log(4NT\delta^{-1}))^2\log^2(d)}{n}\right). \tag{33}$$

for some constant $C = C(C_{\mathbf{z}}) > 0$.

Proof. For any $n \in [N]$, we define the $dT \times dT$ block-diagonal matrix

$$\mathbf{M}_n := \mathrm{diag}(\mathbf{M}_{1,n}, \ldots, \mathbf{M}_{T,n}),$$

where the diagonal $d \times d$ diagonal block matrices are defined as follows:

$$\mathbf{M}_{t,n} := n\left(\widehat{\mathbf{\Sigma}}_{t,n} - \mathbf{\Sigma}_{t,n}\right) = \sum_{s=1}^{n} \mathbf{x}_{t,s}\mathbf{x}_{t,s}^{\top} - \mathbb{E}\left[\mathbf{x}_{t,s}\mathbf{x}_{t,s}^{\top}\middle|\overline{\mathcal{F}}_{s-1}\right]. \tag{34}$$

We also set $\mathbf{M}_0 = \mathbf{0}_{dT\times dT}$ a.s. By construction $(\mathbf{M}_n)_{n\geq 0}$ is a $\{\overline{\mathcal{F}}_n\}_n$-martingale satisfying

$$\|\mathbf{M}_n\|_{\mathrm{op},\mathcal{C}_{\mathrm{r}}} = \max_{\mathbf{\Delta}\in\mathcal{C}_r\,:\,\|\mathbf{\Delta}\|_F=1}\{\langle \mathbf{M}_n\mathrm{Vec}(\mathbf{\Delta}), \mathrm{Vec}(\mathbf{\Delta})\rangle\} = \max_{\mathbf{\Delta}\in\mathcal{C}_r\,:\,\|\mathbf{\Delta}\|_F=1}\left\{\sum_{t=1}^{T}\langle \mathbf{M}_{t,n}\mathbf{\Delta}_t, \mathbf{\Delta}_t\rangle\right\}. \tag{35}$$

We derive several geometric properties of matrices in the cone $\mathcal{C}_r$. We then exploit these properties via an improved union bound argument to derive a sharper uniform deviation bound of the martingale $(\mathbf{M}_n)_{n\geq 0}$ on the cone $\mathcal{C}_r$ as compared to the standard operator norm bound on the whole space of $d \times T$ matrices.

Uniform deviation bound on $\langle \mathbf{M}_{t,n}\mathbf{u}, \mathbf{v}\rangle$ for vectors $\mathbf{u}$ $\mathbf{v}$ in lower-dimensional balls. We consider two linear subspaces $\mathbf{U}$, $\mathbf{V}$ of $\mathbb{R}^d$ of respective dimension $d_{\mathbf{U}}$ and $d_{\mathbf{V}}$. We then introduce the ball $B_{\mathbf{U}}(0, r_{\mathbf{U}}) \in \mathbf{U}$ centered at 0 of radius $r_{\mathbf{U}}$. We define similarly the ball $B_{\mathbf{V}}(0, r_{\mathbf{V}}) \in \mathbf{V}$. Fix $\mathbf{u} \in B_{\mathbf{U}}(0, r_{\mathbf{U}})$ and $\mathbf{v} \in B_{\mathbf{V}}(0, r_{\mathbf{V}})$. We consider now the $\{\overline{\mathcal{F}}_n\}_n$-martingale $(\langle \mathbf{M}_{t,n}\mathbf{u}, \mathbf{v}\rangle)_{n\geq 0}$. Note that

$$\max_{\mathbf{u}\in B_{\mathbf{U}}(0,r_{\mathbf{U}}),\mathbf{v}\in B_{\mathbf{V}}(0,r_{\mathbf{V}})} \langle \mathbf{M}_{t,n}\mathbf{u}, \mathbf{v}\rangle \leq r_{\mathbf{U}}\, r_{\mathbf{V}} \max_{\mathbf{u}\in B_{\mathbf{U}}(0,1),\mathbf{v}\in B_{\mathbf{V}}(0,1)} \langle \mathbf{M}_{t,n}\mathbf{u}, \mathbf{v}\rangle \leq r_{\mathbf{U}}\, r_{\mathbf{V}}\, \|P_{\mathbf{U}}\mathbf{M}_{t,n}P_{\mathbf{V}}\|_{\mathrm{op}}. \tag{36}$$

Using Lemma 4 with δ replaced by $\delta/(2T)$. gives with probability at least $1-\delta$, for any $n \in [N]$ and any $t \in [T]$

$$
\begin{aligned}
&\max_{\mathbf{u}\in B_{\mathbf{U}}(0,r_{\mathbf{U}}),\mathbf{v}\in B_{\mathbf{V}}(0,r_{\mathbf{V}})} \left\langle \frac{\mathbf{M}_{t,n}}{n}\mathbf{u},\mathbf{v}\right\rangle \\
&\quad\le C r_{\mathbf{U}}\, r_{\mathbf{V}} \max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\}\Bigg(\sqrt{\frac{\big(d_{\mathbf{U}}\vee d_{\mathbf{V}}+\log(4NT\delta^{-1})\big)\log(8NT\delta^{-1}d)}{n}} \\
&\qquad\qquad + \frac{\sqrt{(d_{\mathbf{U}}+\log(4NT\delta^{-1}))\,(d_{\mathbf{V}}+\log(4NT\delta^{-1}d))}\log(4NT\delta^{-1})}{n}\Bigg). \qquad (37)
\end{aligned}
$$

Geometric properties of $\mathcal{C}_r$. Set $\bar r = \mathrm{rank}(\mathbf{\Delta})$. Note that $\bar r \le d\wedge T$. Taking the SVD of $\mathbf{\Delta}$, we have

$$
\mathbf{\Delta} = \sum_{j=1}^{\bar r}\sigma_j(\mathbf{\Delta})\mathbf{u}_j(\mathbf{\Delta})\otimes \mathbf{v}_j(\mathbf{\Delta}),
$$

with singular values $\sigma_1(\mathbf{\Delta})\ge\sigma_2(\mathbf{\Delta})\ge\cdots\ge\sigma_{\bar r}(\mathbf{\Delta})>0$ and orthonormal families $\{\mathbf{u}_j(\mathbf{\Delta})\}_{j=1}^{\bar r}\in\mathbb{R}^d$, $\{\mathbf{v}_j(\mathbf{\Delta})\}_{j=1}^{\bar r}\in\mathbb{R}^T$. For the sake of brevity, we set $\sigma_j(\mathbf{\Delta})=\sigma_j$, $\mathbf{u}_j(\mathbf{\Delta})=\mathbf{u}_j$ and $\mathbf{v}_j(\mathbf{\Delta})=\mathbf{v}_j$ for any $j\in[\bar r]$. From the previous displays, we immediately get the following representation for the columns of $\mathbf{\Delta}$:

$$
\mathbf{\Delta}_t = \sum_{j=1}^{\bar r}\sigma_j\mathbf{v}_{j,t}\mathbf{u}_j,\quad \forall t\in T, \qquad (38)
$$

where $\mathbf{v}_j$ admits components $\mathbf{v}_j=(\mathbf{v}_{j,1},\cdots,\mathbf{v}_{j,T})^\top$.

By definition of the cone $\mathcal{C}_r$ in (7), for any $\mathbf{\Delta}\in\mathcal{C}_r$, we have $\|\Pi(\mathbf{\Delta})\|_*\le 3\,\|\mathbf{\Delta}-\Pi(\mathbf{\Delta})\|_*$. Note that $\mathrm{rank}(\mathbf{\Delta}-\Pi(\mathbf{\Delta}))\le 2r$ by definition of Π and the cone $\mathcal{C}_r$. Hence we have, for any $\mathbf{\Delta}\in\mathcal{C}_r$ with $\|\mathbf{\Delta}\|_F=1$,

$$
\|\mathbf{\Delta}\|_*\le\|\Pi(\mathbf{\Delta})\|_*+\|\mathbf{\Delta}-\Pi(\mathbf{\Delta})\|_*\le 4\,\|\mathbf{\Delta}-\Pi(\mathbf{\Delta})\|_*\le 4\sqrt{2r}\,\|\mathbf{\Delta}-\Pi(\mathbf{\Delta})\|_F\le 4\sqrt{2r}\,\|\mathbf{\Delta}\|_F=4\sqrt{2r}. \qquad (39)
$$

We deduce from (39) that $\sum_{j=1}^{\bar r}\sigma_j\le 4\sqrt{2r}$ and consequently for any $\mathbf{\Delta}\in\mathcal{C}_r$

$$
\sigma_j(\mathbf{\Delta})\le\frac{4\sqrt{2r}}{j},\quad\forall j\in[\bar r]. \qquad (40)
$$

We conclude this paragraph with some elementary facts on orthonormal basis. We complete the orthonormal family $\{\mathbf{v}_j\}_{j=1}^{\bar r}$ into an orthonormal basis $\{\mathbf{v}_j\}_{j=1}^{T}$ of $\mathbb{R}^T$. By properties of orthonormal basis, we have $\langle\mathbf{v}_j,\mathbf{v}_{j'}\rangle=\sum_{t=1}^T\mathbf{v}_{j,t}\mathbf{v}_{j',t}=\delta_{j,j'}$ where $\delta_{j,j'}=1$ if $j=j'$ and 0 otherwise.

Uniform bound over $\mathbf{\Delta}\in\mathcal{C}_r$. Our goal is to control $\|\mathbf{M}_n\|_{\mathrm{op},\mathcal{C}_r}$. To this end, we set $m_*=\lceil\log_2(d\wedge T)\rceil$ and, for any $m\in[m_*]$, define $J_m=\{j\in[m_*]\,:\,2^{m-1}\le j<2^m\}$. Next we propose the following decomposition of $\mathbf{\Delta}$:

$$
\mathbf{\Delta}_t=\sum_{m=1}^{m_*}\mathbf{\Delta}_{t,m},\quad\text{where}\quad\mathbf{\Delta}_{t,m}=\sum_{j\in J_m}\sigma_j\mathbf{v}_{j,t}\mathbf{u}_j,\quad\forall t\in[T].
$$

By construction and (40), for any $m\in[m_*]$, all the vectors $\mathbf{\Delta}_{t,m}$, $t\in[T]$ live in the same subspace of dimension at most 2^{m-1} and

$$
\|\mathbf{\Delta}_{t,m}\|^2\le\sigma^2_{2^{m-1}}\sum_{j\in J_m}\mathbf{v}^2_{j,t}\le\sigma^2_{2^{m-1}}\le\frac{32r}{2^{2(m-1)}},
$$

and

$$
\sum_{t=1}^T\|\mathbf{\Delta}_{t,m}\|^2\le\sigma^2_{2^{m-1}}\sum_{j\in J_m}\sum_{t=1}^T\mathbf{v}^2_{j,t}\le\sigma^2_{2^{m-1}}|J_m|\le\frac{32r}{2^{(m-1)}}. \qquad (41)
$$

Combining the last two displays with (38), we get that

$$\sum_{t=1}^{T}\langle \mathbf{M}_{t,n}\boldsymbol{\Delta}_t, \boldsymbol{\Delta}_t\rangle = \sum_{t=1}^{T}\sum_{m,m'=1}^{m_*}\langle \mathbf{M}_{t,n}\boldsymbol{\Delta}_{t,m}, \boldsymbol{\Delta}_{t,m'}\rangle. \tag{42}$$

An union bound combining the last two displays with (35) and (37) gives for any $\delta \in (0,1)$ with probability at least $1-\delta$, for any $n \in [N]$

$$\frac{1}{n}\left\|\mathbf{M}_n\right\|_{\mathrm{op},\mathcal{C}_{\mathrm{r}}} \leq C \max_{1\leq k\leq K}\left\{\left\|\boldsymbol{\Sigma}_k\right\|_{\mathrm{op}}\right\}\sum_{m,m'=1}^{m_*}\sum_{t=1}^{T}\left\|\boldsymbol{\Delta}_{t,m}\right\|\left\|\boldsymbol{\Delta}_{t,m'}\right\|\epsilon_{m,m'},$$

where

$$\epsilon_{m,m'} = \sqrt{\frac{\left(2^{m\vee m'-1}+\log(4NT\delta^{-1})\right)\log(4NT\delta^{-1}d)}{n}} \\ + \frac{\sqrt{\left(2^{m-1}+\log(4NT\delta^{-1})\right)\left(2^{m'-1}+\log(4NT\delta^{-1}d)\right)}\log(4NT\delta^{-1})}{n}.$$

The Cauchy-Schwartz inequality and (41) give

$$\begin{aligned}\left\|\mathbf{M}_n\right\|_{\mathrm{op},\mathcal{C}_{\mathrm{r}}} &\leq C \max_{1\leq k\leq K}\left\{\left\|\boldsymbol{\Sigma}_k\right\|_{\mathrm{op}}\right\}\sum_{m,m'=1}^{m_*}\left(\sum_{t=1}^{T}\left\|\boldsymbol{\Delta}_{t,m}\right\|^2\right)^{1/2}\left(\sum_{t=1}^{T}\left\|\boldsymbol{\Delta}_{t,m'}\right\|^2\right)^{1/2}\epsilon_{m,m'} \\ &\leq 32C \max_{1\leq k\leq K}\left\{\left\|\boldsymbol{\Sigma}_k\right\|_{\mathrm{op}}\right\} r \sum_{m,m'=1}^{m_*}\frac{1}{2^{\frac{m-1}{2}}}\frac{1}{2^{\frac{m'-1}{2}}}\epsilon_{m,m'}.\end{aligned}$$

By definition of $\epsilon_{m,m'}$, we have for any $m, m' \in [m_*]$

$$\frac{1}{2^{\frac{m-1}{2}}}\frac{1}{2^{\frac{m'-1}{2}}}\epsilon_{m,m'} \lesssim \frac{1+\log(4NT\delta^{-1})}{\sqrt{n}} + \frac{(1+\log(4NT\delta^{-1}))^2}{n}.$$

Combining the last two displays with the fact that $m_* \lesssim \log(d)$, we get for any $\delta \in (0,1)$ with probability at least $1-\delta$, for any $n \in [N]$

$$\frac{1}{n}\left\|\mathbf{M}_n\right\|_{\mathrm{op},\mathcal{C}_{\mathrm{r}}} \leq C \max_{1\leq k\leq K}\left\{\left\|\boldsymbol{\Sigma}_k\right\|_{\mathrm{op}}\right\} r\left(\frac{(1+\log(4NT\delta^{-1}))\log^2(d)}{\sqrt{n}} + \frac{(1+\log(4NT\delta^{-1}))^2\log^2(d)}{n}\right).$$

□

C.2 RSC condition and Matrix perturbation

The next lemma guarantees that if the $\|\cdot\|_{\mathrm{op},\mathcal{C}_r}$ metric between two matrices is small enough and if one of these two matrices satisfies the RSC condition, then the other matrix also satisfied the RSC condition. Specifically, the next lemma links the RSC constant $\kappa(\overline{\boldsymbol{\Sigma}})$ associated to the multi-task theoretical matrix, to the adapted one $\kappa(\overline{\boldsymbol{\Sigma}}_n)$.

Lemma 3 (RSC condition and Random Matrices)**.** *Let $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ be two $dT \times dT$ block-diagonal matrices (with blocks matrices of size $d \times d$). Suppose that the RSC condition is met for the covariance matrix $\boldsymbol{\Sigma}_0$ with constant $\kappa(\boldsymbol{\Sigma}_0)$ and that $\left\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\right\|_{\mathrm{op},\mathcal{C}_r} \leq \widetilde{\lambda}$, where*

$$\widetilde{\lambda} \leq \kappa(\boldsymbol{\Sigma}_0).$$

Then, the RSC condition is also met for the covariance matrix $\boldsymbol{\Sigma}_1$ with constant $\kappa(\boldsymbol{\Sigma}_1) \geq \kappa(\boldsymbol{\Sigma}_0)/2$. Moreover, for all $\boldsymbol{\Delta} \in \mathbb{R}^{d\times T}$ such that $\boldsymbol{\Delta} \in \mathcal{C}(r)$ (see equation 8), we have

$$\left\|\mathrm{Vec}(\boldsymbol{\Delta})\right\|_{\boldsymbol{\Sigma}_0}^2 \leq 2\left\|\mathrm{Vec}(\boldsymbol{\Delta})\right\|_{\boldsymbol{\Sigma}_1}^2 \leq 3\left\|\mathrm{Vec}(\boldsymbol{\Delta})\right\|_{\boldsymbol{\Sigma}_0}^2. \tag{43}$$

The proof combines Definition 1 and the $\|\cdot\|_{\mathrm{op},\mathcal{C}_r}$ metric closeness $\left\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\right\|_{\mathrm{op},\mathcal{C}_r} \leq \widetilde{\lambda}$.

Proof. For any $\boldsymbol{\Delta} \in \mathcal{C}_r$, we have

$$\left| \|\mathrm{Vec}(\boldsymbol{\Delta})\|_{\boldsymbol{\Sigma}_0}^2 - \|\mathrm{Vec}(\boldsymbol{\Delta})\|_{\boldsymbol{\Sigma}_1}^2 \right| \leq \|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_{\mathrm{op},\mathcal{C}_r} \|\mathrm{Vec}(\boldsymbol{\Delta})\|_2^2 \leq \widetilde{\lambda} \|\mathrm{Vec}(\boldsymbol{\Delta})\|_2^2 .$$

According to Definition 1, for any $\boldsymbol{\Delta\Delta} \in \mathcal{C}_r$, the RSC condition gives

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 = \|\mathrm{Vec}(\boldsymbol{\Delta})\|_2^2 \leq \frac{\|\mathrm{Vec}(\boldsymbol{\Delta})\|_{\Sigma_0}^2}{2\kappa(\boldsymbol{\Sigma}_0)}.$$

Combining these two last results we have

$$\left| \|\mathrm{Vec}(\boldsymbol{\Delta})\|_{\boldsymbol{\Sigma}_0}^2 - \|\mathrm{Vec}(\boldsymbol{\Delta})\|_{\boldsymbol{\Sigma}_1}^2 \right| \leq \widetilde{\lambda} \frac{\|\mathrm{Vec}(\boldsymbol{\Delta})\|_{\boldsymbol{\Sigma}_0}^2}{2\kappa(\boldsymbol{\Sigma}_0)}.$$

Equivalently, the following result holds

$$\left| \frac{\|\mathrm{Vec}(\boldsymbol{\Delta})\|_{\boldsymbol{\Sigma}_1}^2}{\|\mathrm{Vec}(\boldsymbol{\Delta})\|_{\boldsymbol{\Sigma}_0}^2} - 1 \right| \leq \frac{\widetilde{\lambda}}{2\kappa(\boldsymbol{\Sigma}_0)}.$$

The statement follows directly by the selected $\widetilde{\lambda}$. □

C.3 Combining everything

We can now prove Proposition 2.

Proof. In the RSC condition (see Definition 1) the considered covariance matrix $\overline{\boldsymbol{\Sigma}} \in \mathbb{R}^{dT \times dT}$ is block-diagonal consisting of T blocks, one for each different task. Let us consider now the single-task adapted matrix

$$\boldsymbol{\Sigma}_{t,n} = \frac{1}{n} \sum_{s=1}^{n} \mathbb{E}\left[\mathbf{x}_{t,s} \mathbf{x}_{t,s}^\top \middle| \overline{\mathcal{F}}_{s-1} \right]$$

and the multi-task one $\overline{\boldsymbol{\Sigma}}_n = \mathrm{diag}(\boldsymbol{\Sigma}_{1,n}, ..., \boldsymbol{\Sigma}_{T,n}) \in \mathbb{R}^{dT \times dT}$. Relying on (Oh et al., 2021, Lemma 10), under Assumption 2 we have

$$\overline{\boldsymbol{\Sigma}}_n = \frac{1}{n} \sum_{s=1}^{n} \mathbb{E}\left[\overline{\mathbf{x}}_s \overline{\mathbf{x}}_s^\top \middle| \overline{\mathcal{F}}_{s-1} \right] \succeq (2\nu\omega_{\mathcal{X}})^{-1} \overline{\boldsymbol{\Sigma}}, \tag{44}$$

where $\overline{\mathbf{x}}_s = [\mathbf{x}_{1,s}^\top, \cdots, \mathbf{x}_{T,s}^\top]^\top \in \mathbb{R}^{dT}$. Now, let us denote $\widetilde{\mathbf{w}}_n = \arg\min_{\overline{\mathbf{w}} \in \mathcal{C}(r)} \frac{\mathrm{Vec}(\overline{\mathbf{w}}) \overline{\boldsymbol{\Sigma}}_n \mathrm{Vec}(\overline{\mathbf{w}})}{2\|\mathrm{Vec}(\overline{\mathbf{w}})\|_2^2}$.

Relying on the RSC condition and thanks to the previous display, we obtain

$$\frac{\mathrm{Vec}(\widetilde{\mathbf{w}})_n^\top \overline{\boldsymbol{\Sigma}}_n \mathrm{Vec}(\widetilde{\mathbf{w}})_n}{\|\mathrm{Vec}(\widetilde{\mathbf{w}})_n\|_2^2} \geq \frac{\mathrm{Vec}(\widetilde{\mathbf{w}})_n^\top \overline{\boldsymbol{\Sigma}} \mathrm{Vec}(\widetilde{\mathbf{w}})_n}{2\nu\omega_{\mathcal{X}} \|\mathrm{Vec}(\widetilde{\mathbf{w}})_n\|_2^2} \geq \frac{\kappa(\overline{\boldsymbol{\Sigma}})}{2\nu\omega_{\mathcal{X}}}.$$

Hence, $\overline{\boldsymbol{\Sigma}}_n$ satisfies the RSC condition with constant $\kappa\left(\overline{\boldsymbol{\Sigma}}_n\right) = \kappa\left(\overline{\boldsymbol{\Sigma}}\right) / (2\nu\omega_{\mathcal{X}})$.

Next we consider the operator norm deviation of the multi-task adapted matrix $\boldsymbol{\Sigma}_0 = \overline{\boldsymbol{\Sigma}}_n$ from the multi-task empirical matrix $\boldsymbol{\Sigma}_1 = \widehat{\overline{\boldsymbol{\Sigma}}}_n$. In view of (33), the quantity $\left\|\widehat{\overline{\boldsymbol{\Sigma}}}_n - \overline{\boldsymbol{\Sigma}}_n\right\|_{\mathrm{op},\mathcal{C}_r}$ becomes smaller as n increases. This means that if we take n large enough, then we have whp that $\left\|\widehat{\overline{\boldsymbol{\Sigma}}}_n - \overline{\boldsymbol{\Sigma}}_n\right\|_{\mathrm{op},\mathcal{C}_r} \leq \kappa\left(\overline{\boldsymbol{\Sigma}}\right) / (4\nu\omega_{\mathcal{X}}) = \kappa(\overline{\boldsymbol{\Sigma}}_n)/2$. We formalize this argument below.

For a fixed $\delta \in (0, 1)$, recall the definition of $N_0(\delta)$ in (12). In view of Lemma 2, if $n \geq N_0(\delta)$ then with probability at least $1 - \delta$,

$$\left\|\widehat{\overline{\boldsymbol{\Sigma}}}_n - \overline{\boldsymbol{\Sigma}}_n\right\|_{\mathrm{op},\mathcal{C}_r} \leq \frac{\kappa\left(\overline{\boldsymbol{\Sigma}}\right)}{4\nu\omega_{\mathcal{X}}} \leq \frac{\kappa\left(\overline{\boldsymbol{\Sigma}}_n\right)}{2}.$$

Then Lemma 3 yields the result. □

C.4 Technical Lemma

Consider linear subspaces $\mathbf{U}, \mathbf{V}$ of $\mathbb{R}^d$ of dimension $d_{\mathbf{U}}$ and $d_{\mathbf{V}}$ respectively. Denote by $P_{\mathbf{U}}$ and $P_{\mathbf{V}}$ the orthogonal projections onto $\mathbf{U}$ and $\mathbf{V}$ respectively. We have for any $n \in [N]$

$$\max_{\mathbf{u} \in B_{\mathbf{U}}(0,1), \mathbf{v} \in B_{\mathbf{V}}(0,1)} \langle \mathbf{M}_{t,n} \mathbf{u}, \mathbf{v} \rangle \leq \| P_{\mathbf{U}} \mathbf{M}_{t,n} P_{\mathbf{V}} \|_{\mathrm{op}} .$$

Lemma 4. *Let Assumption 1 be satisfied. For any $\delta \in (0,1)$ and any $t \in [T]$, with probability at least $1-\delta$,*

$$\max_{n \in [N]} \left\{ \frac{1}{n} \| P_{\mathbf{U}} \mathbf{M}_{t,n} P_{\mathbf{V}} \|_{\mathrm{op}} \right\} \leq C \max_{k \in [K]} \{ \| \mathbf{\Sigma}_k \|_{\mathrm{op}} \} \Bigg(\sqrt{\frac{\left(d_{\mathbf{U}} \vee d_{\mathbf{V}} + \log(4N\delta^{-1}) \right) \log(4N\delta^{-1} d)}{n}} \\ + \frac{\sqrt{\left(d_{\mathbf{U}} + \log(4N\delta^{-1}) \right) \left(d_{\mathbf{V}} + \log(4N\delta^{-1} d) \right)} \log(4N\delta^{-1})}{n} \Bigg), \tag{45}$$

where $C = C(C_{\mathbf{z}}) > 0$ may depend only on $C_{\mathbf{z}}$.

Proof. We have for any $n \geq 1$

$$\| P_{\mathbf{U}} (\mathbf{M}_{t,n} - \mathbf{M}_{t,n-1}) P_{\mathbf{V}} \|_{\mathrm{op}} = \left\| P_{\mathbf{U}}(\mathbf{x}_{t,n}) P_{\mathbf{V}}(\mathbf{x}_{t,n})^\top - \mathbb{E} \left[P_{\mathbf{U}}(\mathbf{x}_{t,n}) P_{\mathbf{V}}(\mathbf{x}_{t,n})^\top \middle| \overline{\mathcal{F}}_{n-1} \right] \right\|_{\mathrm{op}} \\ \leq \| P_{\mathbf{U}}(\mathbf{x}_{t,n}) \| \, \| P_{\mathbf{V}}(\mathbf{x}_{t,n}) \| + \max_{k \in [K]} \{ \| \mathbf{\Sigma}_k \|_{\mathrm{op}} \}. \tag{46}$$

In view of Assumption 1, we have $\| P_{\mathbf{U}}(\mathbf{x}_{t,n}) \|_2^2 = \mathbf{z}_{t,n}^\top P_{\mathbf{U}} \mathbf{\Sigma}_k P_{\mathbf{U}} \mathbf{z}_{t,n}$ for some $k \in [K]$. We apply now Hanson-Wright's inequality conditionally on $\overline{\mathcal{F}}_{n-1}$ to get for any $x > 0$

$$\mathbb{P} \left(\mathbf{z}_{t,n}^\top P_{\mathbf{U}} \mathbf{\Sigma}_k P_{\mathbf{U}} \mathbf{z}_{t,n} \leq \mathbb{E} \left[\mathbf{z}_{t,n}^\top P_{\mathbf{U}} \mathbf{\Sigma}_k P_{\mathbf{U}} \mathbf{z}_{t,n} \middle| \overline{\mathcal{F}}_{n-1} \right] + C \, C_{\mathbf{z}}^2 \left(\sqrt{ \| P_{\mathbf{U}} \mathbf{\Sigma}_k P_{\mathbf{U}} \|_{\mathrm{F}}^2 x } + \| P_{\mathbf{U}} \mathbf{\Sigma}_k P_{\mathbf{U}} \|_{\mathrm{op}} \ x \right) \middle| \overline{\mathcal{F}}_{n-1} \right) \geq 1 - e^{-x}.$$

Note that $\mathbb{E} \left[\mathbf{z}_{t,n}^\top P_{\mathbf{U}} \mathbf{\Sigma}_k P_{\mathbf{U}} \mathbf{z}_{t,n} \middle| \overline{\mathcal{F}}_{n-1} \right] = \mathrm{trace}(P_{\mathbf{U}} \mathbf{\Sigma}_k P_{\mathbf{U}}) \leq \max_{k \in [K]} \{ \| \mathbf{\Sigma}_k \|_{\mathrm{op}} \} d_{\mathbf{U}}$ since $\mathbf{z}_k$ is a zero mean isotropic random vector. We define the following event in $\overline{\mathcal{F}}_n$

$$\Omega_n(\mathbf{U}) := \left\{ \| P_{\mathbf{U}}(\mathbf{x}_{t,n}) \|^2 \leq \max_{k \in [K]} \left\{ \| \mathbf{\Sigma}_k \|_{\mathrm{op}} \right\} \left(d_{\mathbf{U}} + C \, C_{\mathbf{z}}^2 \left(\sqrt{ d_{\mathbf{U}} \ \log(4N\delta^{-1}) } + \log(4N\delta^{-1}) \right) \right) \right\}, \tag{47}$$

and

$$\Omega_n(\mathbf{V}) := \left\{ \| P_{\mathbf{V}}(\mathbf{x}_{t,n}) \|^2 \leq \max_{k \in [K]} \left\{ \| \mathbf{\Sigma}_k \|_{\mathrm{op}} \right\} \left(d_{\mathbf{V}} + C \, C_{\mathbf{z}}^2 \left(\sqrt{ d_{\mathbf{V}} \ \log(4N\delta^{-1}) } + \log(4N\delta^{-1}) \right) \right) \right\}. \tag{48}$$

Combining the last three displays and taking $x = \log(\delta^{-1})$ immediately implies for any $n \in [N]$

$$\mathbb{P} \left(\Omega_n(\mathbf{U}) \middle| \overline{\mathcal{F}}_{n-1} \right) \geq 1 - \frac{\delta}{4N}, \quad \mathbb{P} \left(\Omega_n(\mathbf{V}) \middle| \overline{\mathcal{F}}_{n-1} \right) \geq 1 - \frac{\delta}{4N} \tag{49}$$

We define, for any $n \in [N]$

$$\bar{\Omega}_n = \bigcap_{l=0}^{n} \Omega_l(\mathbf{U}) \cap \Omega_l(\mathbf{V}) \in \overline{\mathcal{F}}_n, \tag{50}$$

with $\Omega_0(\mathbf{U}) = \Omega_0(\mathbf{V})$ being the whole sample space. The Bayes rule combined with (49) gives

$$\mathbb{P} \left(\bar{\Omega}_N \right) = \prod_{n=1}^{N} \mathbb{P} \left(\Omega_n(\mathbf{U}) \cap \Omega_n(\mathbf{V}) \middle| \bigcap_{k=0}^{n-1} \Omega_k(\mathbf{U}) \cap \Omega_k(\mathbf{V}) \right) \geq \left(1 - \frac{\delta}{2N} \right)^N . \tag{51}$$

Bernoulli's inequality ($(1+x)^n \geq 1 + nx$ for any $x > -1$ and integer $n \geq 1$) gives

$$\mathbb{P} \left(\bar{\Omega}_N \right) = \prod_{n=1}^{N} \mathbb{P} \left(\Omega_n(\mathbf{U}) \cap \Omega_n(\mathbf{V}) \middle| \bigcap_{k=0}^{n-1} \Omega_k(\mathbf{U}) \cap \Omega_k(\mathbf{V}) \right) \geq 1 - \frac{\delta}{2}. \tag{52}$$

Define the stopping times $\tau(\mathbf{U})$ and $\tau(\mathbf{V})$ as follows

$$\tau(\mathbf{U}) = \inf\left\{n \geq 0 \,:\, \|P_{\mathbf{U}}(\mathbf{x}_{t,n})\|^2 \geq \max_{k\in[K]}\left\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\right\}\left(d_{\mathbf{U}} + C\,C_{\mathbf{z}}^2\left(\sqrt{d_{\mathbf{U}}\,\log(4N\delta^{-1})} + \log(4N\delta^{-1})\right)\right)\right\}, \quad (53)$$

and

$$\tau(\mathbf{V}) = \inf\left\{n \geq 0 \,:\, \|P_{\mathbf{V}}(\mathbf{x}_{t,n})\|^2 \geq \max_{k\in[K]}\left\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\right\}\left(d_{\mathbf{V}} + C\,C_{\mathbf{z}}^2\left(\sqrt{d_{\mathbf{V}}\,\log(4N\delta^{-1})} + \log(4N\delta^{-1})\right)\right)\right\}, \quad (54)$$

Again by definition of the Trace-Norm bandit, $\tau(\mathbf{U})$ and $\tau(\mathbf{V})$ are both stopping times relative to the filtration $\{\overline{\mathcal{F}}_n\}_{n\geq 0}$, so is $\tau = \tau(\mathbf{U}) \wedge \tau(\mathbf{V})$. Hence the stopped process $\{\overline{\mathbf{M}}_n^\tau\}_{n\geq 0}$ defined as follows

$$\overline{\mathbf{M}}_n^\tau := P_{\mathbf{U}}\mathbf{M}_{t,n\wedge\tau}P_{\mathbf{V}}, \quad \forall n \geq 0,$$

is also a martingale adapted to the filtration $\{\overline{\mathcal{F}}_n\}_{n\geq 0}$. Furthermore $\{\overline{\mathbf{M}}_n^\tau\}_{n\geq 0}$ is a bounded martingale. Hence we can apply the Freedman inequality to it.

By definition of τ and (52), we have $\{\tau > N\} = \overline{\Omega}_N$. Hence for any $x > 0$

$$\begin{aligned}\mathbb{P}\left(\|P_{\mathbf{U}}\mathbf{M}_{t,n}P_{\mathbf{V}}\|_{\mathrm{op}} \geq x\right) &= \mathbb{P}\left(\{P_{\mathbf{U}}\mathbf{M}_{t,n}P_{\mathbf{V}} \geq x\} \cap \{\tau > N\}\right) + \mathbb{P}\left(\{\|P_{\mathbf{U}}\mathbf{M}_{t,n}P_{\mathbf{V}}\|_{\mathrm{op}} \geq x\} \cap \{\tau \leq N\}\right)\\ &\leq \mathbb{P}\left(\left\|\overline{\mathbf{M}}_n^\tau\right\|_{\mathrm{op}} \geq x\right) + \mathbb{P}\left(\{\tau \leq N\}\right) \leq \mathbb{P}\left(\left\|\overline{\mathbf{M}}_n^\tau\right\|_{\mathrm{op}} \geq x\right) + \delta/2, \end{aligned} \quad (55)$$

where we have used (52) in the last line.

Deviation bound for $\left\|\overline{\mathbf{M}}_n^\tau\right\|_{\mathrm{op}}$. We will use again (Tropp, 2011, Corollary 1.3).

In view of (46), (47) and (48), we have

$$\left\|\overline{\mathbf{M}}_n^\tau - \overline{\mathbf{M}}_{n-1}^\tau\right\|_{\mathrm{op}} \leq \max_{k\in[K]}\left\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\right\}\left(1 + (1 + C\,C_{\mathbf{z}}^2/2)\sqrt{(d_{\mathbf{U}} + \log(4N\delta^{-1}))\,(d_{\mathbf{V}} + \log(4N\delta^{-1}))}\right). \quad (56)$$

Next, we set $Y_{t,n} = \overline{\mathbf{M}}_n^\tau - \overline{\mathbf{M}}_{n-1}^\tau$. Elementary computations give

$$\boldsymbol{W}_{\mathrm{col}} = \sum_{s=1}^n \mathbb{E}\left[Y_{t,n}Y_{t,n}^\top|\overline{\mathcal{F}}_{s-1}\right] \leq (1 + C\,C_{\mathbf{z}}^2/2)^2 n\big(d_{\mathbf{V}} + \log(4N\delta^{-1})\big)\max_{k\in[K]}\{\|P_{\mathbf{V}}\mathbf{\Sigma}_k P_{\mathbf{V}}\|_{\mathrm{op}}\}\max_{k\in[K]}\{\|P_{\mathbf{U}}\mathbf{\Sigma}_k P_{\mathbf{U}}\|_{\mathrm{op}}\},$$

and

$$\boldsymbol{W}_{\mathrm{row}} = \sum_{s=1}^n \mathbb{E}\left[Y_{t,n}^\top Y_{t,n}|\overline{\mathcal{F}}_{s-1}\right] \leq (1 + C\,C_{\mathbf{z}}^2/2)^2 n\big(d_{\mathbf{U}} + \log(4N\delta^{-1})\big)\max_{k\in[K]}\{\|P_{\mathbf{V}}\mathbf{\Sigma}_k P_{\mathbf{V}}\|_{\mathrm{op}}\}\max_{k\in[K]}\{\|P_{\mathbf{U}}\mathbf{\Sigma}_k P_{\mathbf{U}}\|_{\mathrm{op}}\}.$$

Hence

$$\boldsymbol{W}_{\mathrm{row}} \vee \boldsymbol{W}_{\mathrm{col}} \leq (1 + C\,C_{\mathbf{z}}^2/2)^2 n\big(d_{\mathbf{U}} \vee d_{\mathbf{V}} + \log(4N\delta^{-1})\big)\max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}^2\}. \quad (57)$$

Applying (Tropp, 2011, Corollary 1.3), we get for any $\delta \in (0,1)$ with probability at least $1 - \delta/(2N)$ that

$$\begin{aligned}\left\|\overline{\mathbf{M}}_n^\tau\right\|_{\mathrm{op}} \leq (1 + C\,C_{\mathbf{z}}^2/2)\max_{k\in[K]}\{\|\mathbf{\Sigma}_k\|_{\mathrm{op}}\}\Bigg(&\sqrt{n\left(d_{\mathbf{U}} \vee d_{\mathbf{V}} + \log(4N\delta^{-1})\right)\log(4N\delta^{-1}d)}\\ &+ 2\sqrt{(d_{\mathbf{U}} + \log(4N\delta^{-1}))\,(d_{\mathbf{V}} + \log(4N\delta^{-1}))}\log(4N\delta^{-1}d)\Bigg).\end{aligned} \quad (58)$$

Dividing by n and a trivial union bound gives the result.

□

D PROOF OF THEOREM 1

We recall that $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_T]$ is of rank smaller than r. Hence $\mathbf{W}$ admits the following singular value decomposition

$$\mathbf{W} = \sum_{j=1}^{r} \sigma_j(\mathbf{W}) u_j(\mathbf{W}) \otimes v_j(\mathbf{W}),$$

with singular values $\sigma_1(\mathbf{W}) \geq \cdots \geq \sigma_r(\mathbf{W}) > 0$ and corresponding left and right orthonormal singular vectors $u_1(\mathbf{W}), \dots, u_r(\mathbf{W})$ of $\mathbb{R}^d$ and $v_1(\mathbf{W}), \dots, v_r(\mathbf{W})$ of $\mathbb{R}^T$. We denote by P_U the orthogonal projection onto $U = \text{l.s.}(u_1(\mathbf{W}), \dots, u_r(\mathbf{W}))$. By definition of the SVD, we obviously have $P_U(\mathbf{w}_t) = \mathbf{w}_t$ for any $1 \leq t \leq T$.

We consider now the instantaneous multi-task regret

$$\overline{R}_n = \sum_{t=1}^{T} R_{t,n} = \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}^*, \mathbf{w}_t \rangle - \langle \mathbf{x}_{t,n}, \mathbf{w}_t \rangle \leq 2 \sum_{t=1}^{T} \max_{\mathbf{x} \in \mathcal{D}_{t,n}} |\langle \mathbf{x}, \mathbf{w}_t \rangle|,$$

where $\mathbf{x}_{t,n}^* = \arg\max_{\mathbf{x} \in \mathcal{D}_{t,n}} \langle \mathbf{x}, \mathbf{w}_t \rangle$.

Assumption 1 guarantees that $\frac{\langle \mathbf{x}, \mathbf{w}_t \rangle}{\|\mathbf{w}_t\|}$ is $C_{\mathbf{x}}$-subgaussian and consequently, for any $\delta \in (0,1)$, with probability at least $1-\delta$

$$|\langle \mathbf{x}, \mathbf{w}_t \rangle| \leq C_{\mathbf{x}} \|\mathbf{w}_t\| \sqrt{\log(e\delta^{-1})} \leq C_{\mathbf{z}} \max_{k \in [K]} \left\{ \|\Sigma_k\|_{\text{op}}^{1/2} \right\} L \sqrt{\log(e\delta^{-1})}.$$

An union bound over $t \in [T]$, $n \in [N]$, $\mathbf{x} \in \mathcal{D}_{t,n}$ gives with probability at least $1 - \delta/2$

$$\max_{\mathbf{x} \in \mathcal{D}_{t,n}} |\langle \mathbf{x}, \mathbf{w}_t \rangle| \leq C_{\mathbf{z}} \max_{k \in [K]} \left\{ \|\Sigma_k\|_{\text{op}}^{1/2} \right\} L \sqrt{\log(2eTKN\delta^{-1})}, \quad \forall t \in [T],\ \forall k \in [K], \forall n \in [N].$$

Hence with the same probability, $\forall n \in [N]$

$$\overline{R}_n \leq C_{\mathbf{z}} T \max_{k \in [K]} \left\{ \|\Sigma_k\|_{\text{op}}^{1/2} \right\} L \sqrt{\log(2eTKN\delta^{-1})}$$

We will now split the regret into two phases. During the first N_0 interactions the collected data will not meet the RSC condition. Hence, the incurred regret will be of order $TN_0\sqrt{\log(2eTKN\delta^{-1})}$. Conversely, from the N_0+1-th round up to the N-th one, the RSC condition will be met allowing the result of Lemma 1.

We have

$$\begin{aligned}
\sum_{n=\lceil N_0 \rceil}^{N} \overline{R}_n &= \sum_{n=\lceil N_0 \rceil}^{N} \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}^*, \mathbf{w}_t - \widehat{\mathbf{w}}_{t,n} \rangle + \sum_{n=\lceil N_0 \rceil}^{N} \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}^* - \mathbf{x}_{t,n}, \widehat{\mathbf{w}}_{t,n} \rangle + \sum_{n=\lceil N_0 \rceil}^{N} \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}, \widehat{\mathbf{w}}_{t,n} - \mathbf{w}_t \rangle \\
&\leq \sum_{n=\lceil N_0 \rceil}^{N} \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}^*, \mathbf{w}_t - \widehat{\mathbf{w}}_{t,n} \rangle + \sum_{n=\lceil N_0 \rceil}^{N} \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}, \widehat{\mathbf{w}}_{t,n} - \mathbf{w}_t \rangle = \sum_{n=\lceil N_0 \rceil}^{N} \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}^* - \mathbf{x}_{t,n}, \mathbf{w}_t - \widehat{\mathbf{w}}_{t,n} \rangle
\end{aligned} \tag{59}$$

where we have used that for any $1 \leq n \leq N$, $1 \leq t \leq T$, $\langle \mathbf{x}_{t,n}^* - \mathbf{x}_{t,n}, \widehat{\mathbf{w}}_{t,n} \rangle \leq 0$ a.s. by definition of our policy.

Next, we have

$$\left| \sum_{t=1}^{T} \langle \mathbf{x}_{t,n}^* - \mathbf{x}_{t,n}, \mathbf{w}_t - \widehat{\mathbf{w}}_{t,n} \rangle \right| \leq 2 \sum_{t=1}^{T} \max_{\mathbf{x} \in \mathcal{D}_{t,n}} |\langle \mathbf{x}, \mathbf{w}_t - \widehat{\mathbf{w}}_{t,n} \rangle|. \tag{60}$$

In view of (4), we note that $\widehat{\mathbf{W}}_n \in \overline{\mathcal{F}}_{n-1}$. Thus, conditionally on $\overline{\mathcal{F}}_{n-1}$, Assumption 1 guarantees that $\frac{\langle \mathbf{x}, \mathbf{w}_t - \widehat{\mathbf{w}}_{t,n} \rangle}{\|\mathbf{w}_t - \widehat{\mathbf{w}}_{t,n}\|}$ is $C_{\mathbf{x}}$-subgaussian and consequently, for any $\delta \in (0,1)$, with probability at least $1-\delta$

$$|\langle \mathbf{x}, \mathbf{w}_t - \widehat{\mathbf{w}}_{t,n} \rangle| \leq C_{\mathbf{x}} \|\mathbf{w}_t - \widehat{\mathbf{w}}_{t,n}\| \sqrt{\log(e\delta^{-1})} \leq C_{\mathbf{z}} \max_{k \in [K]} \left\{ \|\Sigma_k\|_{\text{op}}^{1/2} \right\} \|\mathbf{w}_t - \widehat{\mathbf{w}}_{t,n}\| \sqrt{\log(e\delta^{-1})}.$$

An union bound over t, n, k gives with probability at least $1-\delta$

$$\max_{t\in[T],n\in[N]}\max_{\mathbf{x}\in\mathcal{D}_{t,n}}|\langle\mathbf{x},\mathbf{w}_t-\widehat{\mathbf{w}}_{t,n}\rangle| \leq C_{\mathbf{z}}\max_{k\in[K]}\left\{\|\Sigma_k\|_{\mathrm{op}}^{1/2}\right\}\|\mathbf{w}_t-\widehat{\mathbf{w}}_{t,n}\|\sqrt{\log(eTKN\delta^{-1})}. \tag{61}$$

Combining the last two displays, we get with probability at least $1-\delta$, for any $t\in[T]$, $n\in[N]$,

$$\begin{aligned}\left|\sum_{t=1}^{T}\langle\mathbf{x}_{t,n}^{*}-\mathbf{x}_{t,n},\mathbf{w}_t-\widehat{\mathbf{w}}_{t,n}\rangle\right| &\leq 2C_{\mathbf{z}}\max_{k\in[K]}\left\{\|\Sigma_k\|_{\mathrm{op}}^{1/2}\right\}\left(\sum_{t=1}^{T}\|\mathbf{w}_t-\widehat{\mathbf{w}}_{t,n}\|\right)\sqrt{\log(eTKN\delta^{-1})}\\ &\leq 2C_{\mathbf{z}}\max_{k\in[K]}\left\{\|\Sigma_k\|_{\mathrm{op}}^{1/2}\right\}\sqrt{T}\left\|\mathbf{W}-\widehat{\mathbf{W}}_n\right\|_F\sqrt{\log(eTKN\delta^{-1})}.\end{aligned}$$

Recall that Lemmas 1 and 2 combined with Propositions 1 and 2 gives, for any $\lceil N_0\rceil\leq n\leq N$ and any $\delta\in(0,1)$, with probability at least $1-\frac{\delta}{2N}$,

$$\left\|\widehat{\mathbf{W}}_n-\mathbf{W}\right\|_{\mathrm{F}}\lesssim_{c_\eta,\sigma,C_{\mathbf{x}},\kappa(\overline{\mathbf{\Sigma}}),\max_{1\leq k\leq K}\left\{\|\Sigma_k\|_{\mathrm{op}}^{1/2}\right\}}\sqrt{\frac{r}{n}}\left(\sqrt{T+d+\log\left(\frac{4N(d+T)}{\delta}\right)}\log^{3/2}\left(\frac{8N(d+T)}{\delta}\right)\right). \tag{62}$$

Combining the last two displays, we obtain with probability at least $1-\delta$

$$\left|\sum_{t=1}^{T}\langle\mathbf{x}_{t,n}^{*}-\mathbf{x}_{t,n},\mathbf{w}_t-\widehat{\mathbf{w}}_{t,n+1}\rangle\right|\lesssim\sqrt{\frac{rT}{n}}\left(\sqrt{T+d+\log\left(\frac{4N(d+T)}{\delta}\right)}\log^{3/2}\left(\frac{8N(d+T)}{\delta}\right)\right)\sqrt{\log(2e\delta^{-1}TKN)}. \tag{63}$$

Summing over n in the previous display along with an union bound argument and (59) gives with probability at least $1-\delta$

$$\sum_{n=\lceil N_0\rceil}^{N}\overline{R}_s\leq C\sqrt{rTN}\left(\sqrt{T+d+\log\left(\frac{4N(d+T)}{\delta}\right)}\log^{3/2}\left(\frac{8N(d+T)}{\delta}\right)\right)\sqrt{\log(2e\delta^{-1}TKN)}, \tag{64}$$

where $C=C\left(\eta,\sigma,C_{\mathbf{z}},\kappa(\overline{\mathbf{\Sigma}}),\max_{1\leq k\leq K}\left\{\|\Sigma_k\|_{\mathrm{op}}^{1/2}\right\}\right)$ is a finite constant under our assumptions.

E NUMERICAL EXPERIMENTS

We perform here some additional experiments on the impact of the rank r, the dimension d and the number of tasks T on the performance of our trace norm bandit policy (Algorithm 1). We also compare its performance to the MLingreedy policy of Yang et al. (2020).

Comparison to MLingreedy (Yang et al., 2020). We compare the performance of the trace norm bandit (Alg. 1) to that of the MLingreedy policy and independent task learning (ITL). We recall that the MLingreedy policy use a rank parameter $\overline{r}$. This parameter is set equal to the true rank r of $\mathbf{W}$ in Yang et al. (2020) to derive the theoretical properties of this policy.

In our experiments we will use different values for the rank parameter $\overline{r}$, either the true rank r or an overestimated rank $\overline{r}=\min(2r,d,T)$ or an underestimated rank $\overline{r}=\max(\lfloor r/2\rfloor,1)$. At each epoch of the the MLingreedy policy, we used stochastic gradient descent to solve the matrix factorization step. We perform 5 repetitions of our experiments for several values of dimension d, T, the parameter $\overline{r}$ of MLingreedy and the regularization parameter $l\in\{l_1,l_2,l_3\}$ of the trace norm policy.

In Figures 3 and 4, we implement MLingreedy with the true value of rank r (most favorable case for this policy in theory) and compare its performance to that of our trace norm bandit and ITL. Trace norm bandit performs consistently above ITL and MLingreedy with true parameter for all considered configurations of T and d. Noticeably MLingreedy with true parameter performs significantly worse than ITL for $d=40$ for almost all values of N.

In Figure 5, we explore in more details the impact of parameters l and $\overline{r}$ on the performance of the trace norm bandit and MLingreedy, respectively.

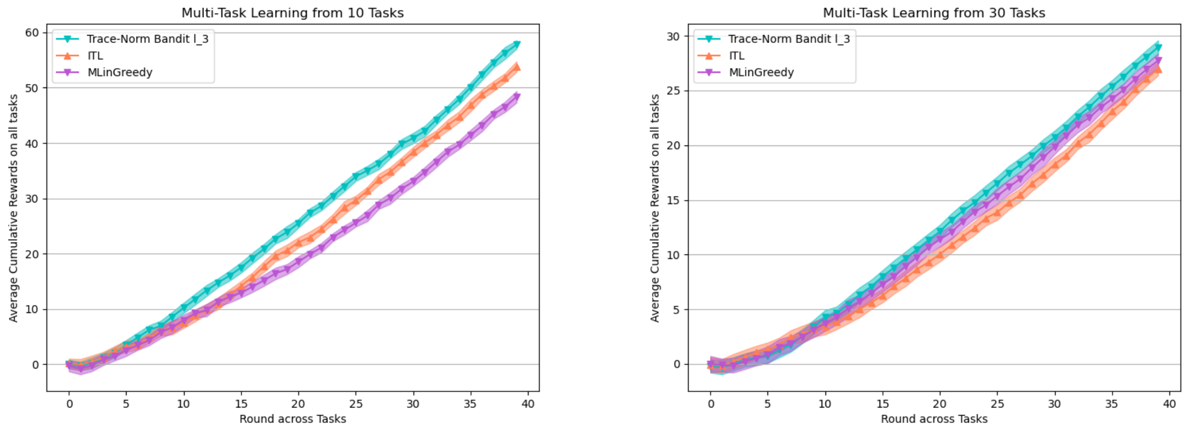

Figure 3: Averaged cumulative reward over all tasks with $d = 10$ features, $T = 10$ tasks (left) and $T = 30$ (right). Each task lasts for $N = 40$ rounds, has $K = 10$ arms, noise variance $\sigma^2 = 1$ and true rank $r = 5$.

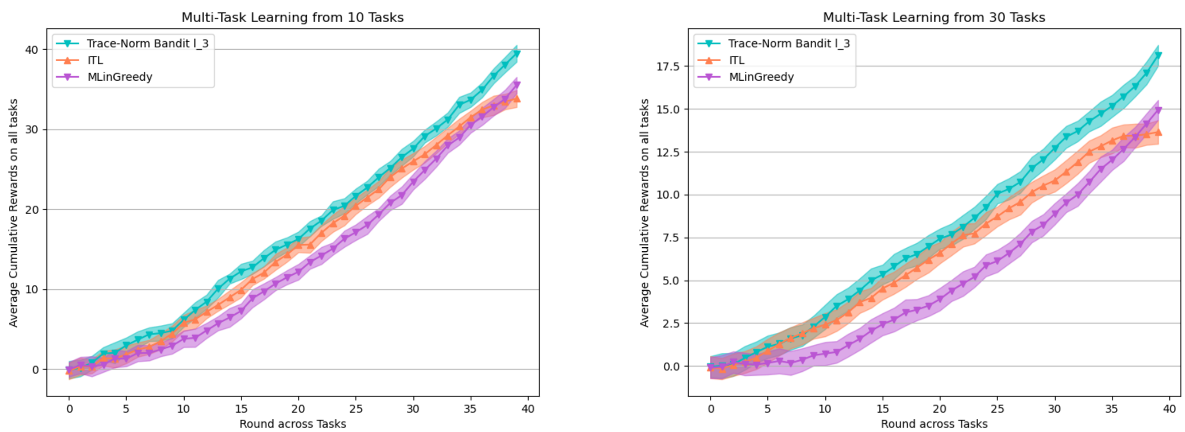

Figure 4: Averaged cumulative reward over all tasks with $d = 40$ features, $T = 10$ tasks (left) and $T = 30$ (right). Each task lasts for $N = 40$ rounds, has $K = 10$ arms, noise variance $\sigma^2 = 1$ and true rank $r = 5$.

- In higher dimension $d = 40$ and $T = 10$, the trace norm bandit performs significantly better than MLingreedy. Actually MLingreedy performance is significantly worse than ITL even when using the true rank ($\bar{r} = r$) until N becomes larger than 38 whereas the trace norm policy performs significantly better than ITL even when N is small.
- For $d = 40$ and $T = 30$, the trace norm bandit performs better than ITL for any $N \geq 25$ whereas MLingreedy with overestimated rank and true rank are below ITL until $N \geq 35$ and $N \geq 38$ respectively. MLingreedy with underestimated rank is significantly worse than ITL for all N.

In summary, the trace norm bandit performs consistently better than ITL uniformly for most values of N, for all values of d and T and for several choices of the regularization parameter. Conversely, MLingreedy is far more sensitive to the choice of rank parameter $\bar{r}$ and the values of d and T. MLingreedy with underestimated rank performs far worse than ITL in all configurations. MLingreedy with overestimated rank performs sometimes better than MLingreedy with true rank. Overall, its performance are worse than those of trace norm bandit.

Impact of rank r. In Figure 6, we note that Alg. 1 performs uniformly better than the ITL policy uniformly over all values of rank $r \in [d]$. The performance of the oracle tends to that of ITL as r tends to d. This is expected as the oracle working in a r-dimensional space loses the benefit of working in a small dimensional space when $r \approx d$. Interestingly, the

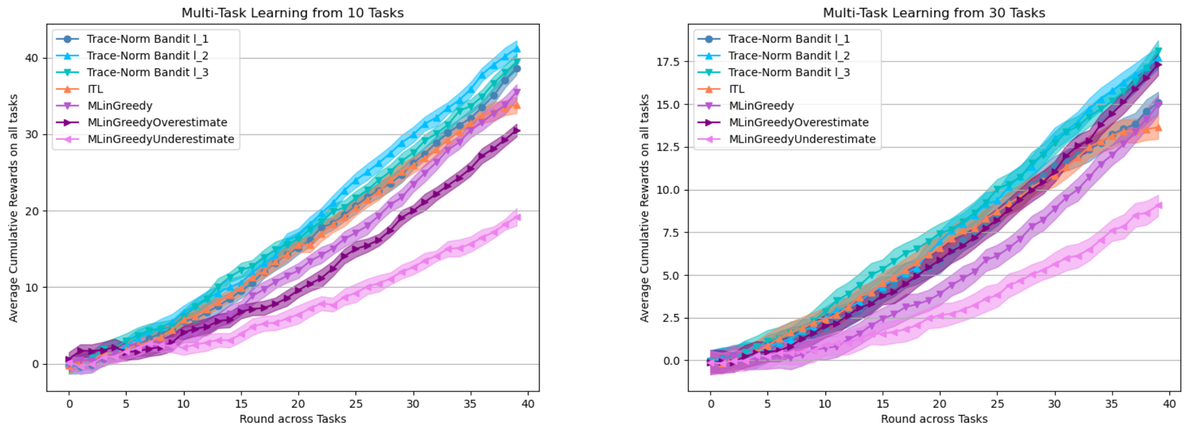

Figure 5: Averaged cumulative reward over all tasks with $d = 40$ features, $T = 10$ tasks (top) and $T = 30$ (bottom). Each task lasts for $N = 40$ rounds, has $K = 10$ arms, noise variance $\sigma^2 = 1$ and true rank $r = 5$.

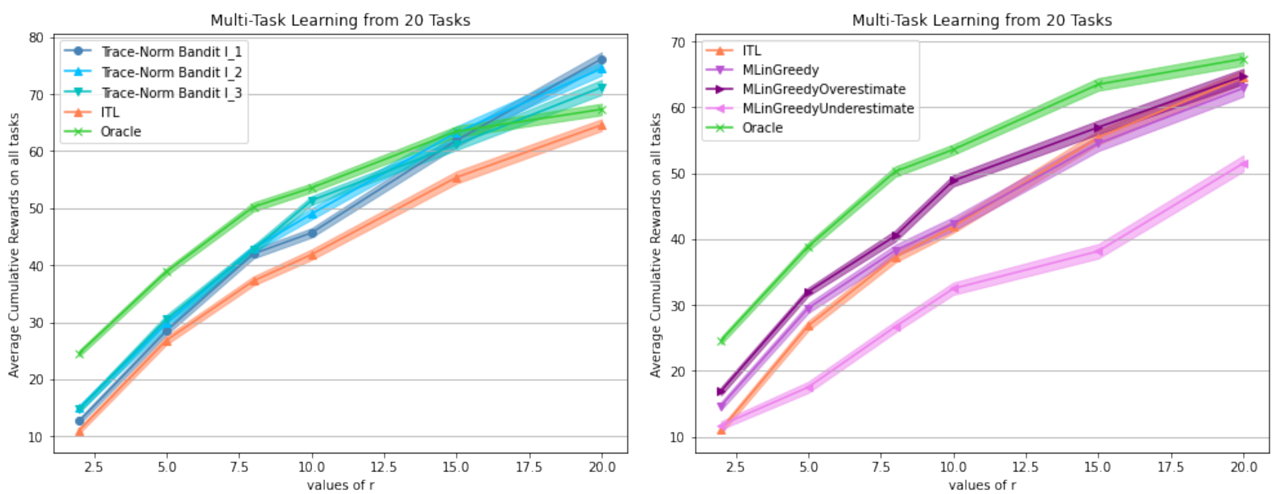

Figure 6: Averaged cumulative reward over all tasks for $d = 20$, $T = 20$, $\sigma^2 = 1$ as a function of the rank r.

performance of the trace norm bandit becomes superior to that of the oracle as r gets close to d. A plausible explanation is that nuclear norm regularization can still bring some benefit even when $r \approx d$ as it can still perform dimension reduction.