Multi-Task and Meta-Learning with Sparse Linear Bandits Supplementary material

Leonardo Cella [*1] **Massimiliano Pontil** [†1,2]

[1]Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, 16163, Genoa, Italy
[2]Department of Computer Science, University College London, London, WC1E 6BT, UK

Abstract

Motivated by recent developments on meta-learning with linear contextual bandit tasks, we study the benefit of feature learning in both the multi-task and meta-learning settings. We focus on the case that the task weight vectors are *jointly sparse*, i.e. they share the same small set of predictive features. Starting from previous work on standard linear regression with the group-lasso estimator we provide novel oracle-inequalities for this estimator when samples are collected by a bandit policy. Subsequently, building on a recent lasso-bandit policy, we investigate its group-lasso variant and analyze its regret bound. We specialize the proposed policy to the multi-task and meta-learning settings, demonstrating its theoretical advantage. We also point out a deficiency in the state-of-the-art lower bound and observe that our method has a smaller upper bound. Preliminary experiments confirm the effectiveness of our approach in practice.

1 INTRODUCTION

Stochastic bandits [see Lattimore and Szepesvári, 2020, Auer et al., 2002, Siegmund, 2003, Robbins, 1952, Cesa-Bianchi, 2016, Bubeck and Cesa-Bianchi, 2012, and references therein] are effective approaches to solve the online learning problem constrained to partial feedback. In the last decade they have receiving increasing attention thanks to both their wide practical usage and the challenge of designing efficient and theoretically grounded algorithms. This methodology has been applied to a variety areas, ranging from recommender systems [Cella and Cesa-Bianchi, 2019, Bogers, 2010], to online auctions [Weed et al., 2016, Nedelec et al., 2020] and to adaptive routing [Awerbuch and Kleinberg, 2008], among others.

In this work we focus on linear stochastic bandits [Abbasi-Yadkori et al., 2011, Li et al., 2010, Chu et al., 2011, Auer, 2003, Bastani and Bayati, 2020, Kim and Paik, 2019], a well studied setting in which each arm is associated with a vector of features, and the arm payoff is given by a linear regression of the feature vector. A difficulty behind this problem is that often each arm is described by many features, thus requiring a long exploration in order to obtain an accurate estimate of the unknown regression vector. Two main approaches have been adopted to speedup the learning process: meta-learning [Cella et al., 2020, Boutilier et al., 2020, Kveton et al., 2020] and feature learning [Kim and Paik, 2019, Bastani and Bayati, 2020, Abbasi-Yadkori et al., 2012, Gopalan et al., 2016]. The former approach aims to compensate the need for large samples by leveraging relationships between multiple tasks (e.g. those corresponding to similar users). The latter solution investigates the existence of a low-dimensional space still satisfying the linearity assumption but requiring much less samples than those needed when considering all the features. Very recently, Yang et al. [2020] combined these approaches and showed how feature learning could improve the efficiency of linear bandits in the multi-task framework.

Research Objectives and Challenges. Similarly to [Yang et al., 2020] in this work we investigate the multi-task framework assuming that the tasks are sparse and share the same sparsity pattern. We will also analyze the meta-learning setting and show that both problems can be solved through a group-lasso bandit policy.
Since we deal with a bandit framework, noisy components affecting observations are not i.i.d. but satisfy a martingale condition [see Abbasi-Yadkori et al., 2012]. Similarly, samples are collected sequentially by a bandit policy that varies over time, according to the already observed data. Consequently, we cannot employ existing oracle inequalities for the group-lasso [Lounici et al., 2011, Bühlmann and Van

[*] leonardocella@gmail.com
[†] massimiliano.pontil@iit.it

De Geer, 2011], which rely on the i.i.d. assumption. Moreover, recent concentration results investigating lasso-bandit policies [Bastani and Bayati, 2020] do not apply to our setting, since we need to handle the norm of a vector of sub-Gaussian random variables rather than their sum.

Previous Work. Transfer learning across bandit tasks has been investigated in [Cella et al., 2020, Kveton et al., 2020, Boutilier et al., 2020, Gentile et al., 2017, 2014, Deshmukh et al., 2017, Soare, 2015, Azar et al., 2013]. Works like [Cella et al., 2020, Soare, 2015, Gentile et al., 2014, 2017] considered task similarities to be proportional to the euclidean distance between their unknown regression vectors. Differently, in [Kveton et al., 2020] the authors proposed a gradient based approach to estimate the regression parameter.

The problem of learning with high-dimensional context vectors has caught recent attention. In [Kim and Paik, 2019, Bastani and Bayati, 2020], they consider the true regression vector to be specified by only a sparse and small subset of the original set of features. Since the ordinary least square estimator is not expected to work well in this scenario, authors propose two different solutions to embed the standard lasso estimator within a bandit policy. Interestingly, Yang et al. [2020] showed that when the sparsity assumption holds *jointly* across different tasks, this can be leveraged for learning a common representation and speeding up the exploration phase in each task. A similar joint sparse assumption across tasks has been already investigated in the reinforcement-learning framework by Calandriello et al. [2014]. The main difference here is that, while they could assume to have access to a generative model of the MDP, in our bandit setting such assumption cannot be evaluated, since samples are collected while learning.

Contributions and Organization. Our contributions are threefold. Firstly, in Section 3 we give oracle inequalities for the group-lasso estimator when the training data are not independently sampled, but rather they are collected by a bandit policy and their noisy components satisfy a certain martingale-condition [Abbasi-Yadkori et al., 2012]. Secondly, relying on the proved oracle inequalities, in Section 4 we propose a group-lasso bandit policy that is inspired by the doubly-robust lasso bandit policy [Kim and Paik, 2019]. Thirdly, In Section 5 we show how to use the introduced policy in the multi-task and meta-learning settings. In the latter case our upper bound is novel and highlight the benefit of the proposed policy. In the former case, our approach outperforms the state of the art solution of [Yang et al., 2020]. We further complement our analysis by providing a lower bound that point out a weaknesses of the existing result in [Yang et al., 2020]. Finally, in Section 6 we present preliminary numerical experiments which corroborate our theoretical findings.

2 PRELIMINARIES

We begin by introducing the linear contextual bandit setting and then present its multi-task extension, focusing on jointly sparse regression vectors.

2.1 LINEAR CONTEXTUAL BANDITS

Let N be a positive integer and let $[N] = \{1, \ldots, N\}$. A linear contextual bandit problem consists of a sequence of N interactions between a learning policy π and an environment. At each round $n \in [N]$, the policy has to pick one arm $\mathbf{x}_n \in \mathcal{K}_n$ from a given decision set $\mathcal{K}_n \subseteq \mathbb{R}^M$ s.t. $|\mathcal{K}_n| = K$. Subsequently, only the reward associated to the chosen arm $\mathbf{x}_n$ will be observed, and no feedback will be available for the remaining (not selected) arms. In the linear bandit setting, the observed (instantaneous) reward satisfies $y_n = \mathbf{x}_n^\top \mathbf{w} + \eta_n$, that is a linear relation with respect to an unknown parameter $\mathbf{w} \in \mathbb{R}^M$ and subject to additive zero-mean noise η_n which we assume throughout to be conditionally sub-Gaussian.

We assume the optimal policy π^* to know the true parameter $\mathbf{w}$, hence at each round $n \in [N]$, π^* picks the arm $\mathbf{x}_n^* = \arg\max_{\mathbf{x} \in \mathcal{K}_n} \mathbf{x}^\top \mathbf{w}$ maximizing the instantaneous reward. The objective is to minimize the *pseudo-regret*

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

which measures the gap incurred with respect to the optimal policy. If no further assumptions hold, a standard algorithm to face the linear bandit problem is OFUL [Abbasi-Yadkori et al., 2011]. At each round $n \in [N]$, this method estimates $\mathbf{w}$ by ridge-regression over the observed arm reward pairs, that is $\widehat{\mathbf{w}}_{n+1}^{\lambda} = \arg\min_{\mathbf{w} \in \mathbb{R}^M} \|\mathbf{X}_n \mathbf{w} - \mathbf{y}_n\|_2^2 + \lambda \|\mathbf{w}\|_2^2$ where $\mathbf{X}_n$ is the matrix whose rows are $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$, $\mathbf{I}$ is the $M \times M$ identity matrix and $\mathbf{y}_n = (y_1, \ldots, y_n)^\top$. Subsequently, OFUL updates a confidence interval $\mathcal{C}_n$ centered in $\widehat{\mathbf{w}}_n^{\lambda}$ and containing the real parameter $\mathbf{w}$ with high probability. It then picks the arm $\mathbf{x}_n$ as

$$\mathbf{x}_n = \arg\max \left\{ \mathbf{x}^\top \mathbf{v} \ : \ \mathbf{x} \in \mathcal{K}_n, \ \mathbf{v} \in \mathcal{C}_n \right\}. \qquad (1)$$

Finally, similarly to [Kim and Paik, 2019, Assumption 2], we make the following assumption on the sets $\mathcal{K}_1, \ldots, \mathcal{K}_N$.

Assumption A (i.i.d. Arms)**.** *Let* $\mathcal{K}_n = \{\mathbf{x}_{1,n}, \ldots, \mathbf{x}_{K,n}\}$. *There exist* K *distributions* $\mathcal{P}_1, \ldots, \mathcal{P}_K$ *with support on the unit sphere* $\{\mathbf{x} \in \mathbb{R}^M \ : \ \|\mathbf{x}\|^2 \leq 1\}$ *such that, for every* $k \in [K]$,

$$\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,n} \overset{\text{i.i.d.}}{\sim} \mathcal{P}_k$$

.

2.2 SPARSE LINEAR BANDITS AND TRANSFER LEARNING

In this work we are interested in the case where we have to solve $T > 1$ different linear bandit tasks. Each task $t \in [T]$ is fully specified by its regression vector $\mathbf{w}_t \in \mathbb{R}^d$. Notice that in order to have a clear notation, differently from the above subsection we now consider d-dimensional weight vectors. Let us introduce the following additional notation. Let $J(\mathbf{w}) = \{j \in [d] : w_j \neq 0\}$ be the *sparsity pattern* of w, that is the subset of $s = |J(\mathbf{w})|$ non-zero components of vector $\mathbf{w} \in \mathbb{R}^d$. Differently from existing multi-task and meta-learning bandit works [Cella et al., 2020, Soare, 2015, Kveton et al., 2020], here we study the advantage of the group sparsity assumption for transfer learning among tasks, an idea which was originally investigated in the standard supervised learning setting [see Lounici et al., 2011, and references therein]. As we shall see, a key difficulty when moving to the bandit setting is that samples are not i.i.d. anymore. Indeed, they are not given in advance, but rather incrementally collected by a policy while it is learning (hence potentially incurring a regret). Our main assumption is that the considered tasks are jointly sparse.

Assumption B (Jointly Sparse Tasks)**.** *The task parameters $\{\mathbf{w}_1, \dots, \mathbf{w}_T\} \subseteq \mathbb{R}^d$ share a common sparsity pattern of small cardinality. That is, there is a set $J \subseteq [d]$ such that*

$$J(\mathbf{w}_t) \subseteq J \quad \forall t \in [T]$$

and, letting $s = |J|$, we have that $s \ll d$.

In the ideal scenario in which the set of relevant features are known *a-priori*, we could consider the OFUL policy introduced in Section 2.1. That is, we would run OFUL on the set J of s relevant features, obtaining a per task regret bound $O(s\sqrt{N})$ and computational costs of order $O(s^2)$. In particular, when restricting to the finite-action linear contextual bandit setting [Auer, 2003, Chu et al., 2011] (decision sets are fixed with finite size $|\mathcal{K}| = K$) a regret bound of $O(\sqrt{KNs})$ would outperform the original $O(\sqrt{KNd})$ result. Since in practice the set of s relevant features is not known a-priori, we will refer to the above policy as Oracle.

In Section 5 we will show how both the multi-task and meta-learning problems can be reformulated as an instance of a single bandit setting with specific group sparsity structure. Thus, in the next two sections we will consider the single task bandit problem under group sparsity constraints. First, in Section 3 we investigate oracle inequalities for the corresponding group-lasso estimator considering data to be collected in a non i.i.d. fashion. Then, using such inequalities, in Section 4 we will present our group-lasso policy.

3 GROUP-LASSO INEQUALITIES WITH MARTINGALE NOISE

Let us consider $G \leq M$ and let the sets $\mathcal{G}_1, \dots, \mathcal{G}_G$ form a partition of the set $[M] = \{1, \dots, M\}$[1]. That is, $\forall i \neq j$ $\mathcal{G}_i \cap \mathcal{G}_j = \varnothing$ and $\cup_{j \in [G]} \mathcal{G}_j = [M]$. For every $j \in [G]$, we denote by $|\mathcal{G}_j| = M_j$ the number of features indexed by the subset $\mathcal{G}_j$. Given the matrix $\mathbf{X}_n = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times M}$ we define the Gram matrix $\mathbf{V}_n = \mathbf{X}_n^\top \mathbf{X}_n \in \mathbb{R}^{M \times M}$. Analogously, we denote by $\mathbf{X}_{n,\mathcal{G}_j}$ the $n \times M_j$ submatrix of $\mathbf{X}$ whose columns contain only the original features indexed by $\mathcal{G}_j$, and let $\mathbf{V}_{n,\mathcal{G}_j} \in \mathbb{R}^{M_j \times M_j}$ be the corresponding Gram matrix.
For every vector $\mathbf{w} \in \mathbb{R}^M$, we denote by $\mathbf{w}^j = (w_i : i \in \mathcal{G}_j)$. Now, given a subset $J \subseteq [G]$, let $\mathbf{w}_J = (\mathbf{w}^j \mathbb{I}\{j \in J\} : j \in J)$. Finally, analogously to the previous notation let us use $J(\mathbf{w}) = \{j : \mathbf{w}^j \neq 0, j \in [G]\}$ and $M(\mathbf{w}) = |J(\mathbf{w})|$. Hence, $J(\mathbf{w})$ contains the indices of the relevant groups and $M(\mathbf{w})$ is the number of such groups. Note that, if $N = M$, each group is a singleton, indexing the corresponding feature. In [Kim and Paik, 2019, Bastani and Bayati, 2020] authors assumed the linear regression vector to be sparse. Differently we make a group-sparsity assumption.

Assumption C (Group Sparsity)**.** *We assume that the vector $\mathbf{w} \in \mathbb{R}^M$ associated to a linear bandit task is group-sparse, that is, there exists a value $s \leq M$ such that*

$$M(\mathbf{w}) \ll s.$$

As in the standard supervised-learning setting, at each round $n \in [N]$ we consider the group-lasso estimator $\widehat{\mathbf{w}}_n$ [Yuan and Lin, 2006] which is defined as the solution of the following optimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^M} \left\{ \frac{1}{n} \|\mathbf{X}_n \mathbf{w} - \mathbf{y}_n\|^2 + 2 \sum_{j \in [G]} \lambda_{n,j}(\delta) \left\|\mathbf{w}^j\right\| \right\}, \tag{2}$$

where $\lambda_{n,1}(\delta), \dots, \lambda_{n,G}(\delta)$ are positive real parameters that will be specified later.
The first result we present is a generalization of [Lemma 11.2 Van De Geer and Bühlmann, 2009] and [Lemma 3.1 Lounici et al., 2011] for the group-lasso case when the noise are not i.i.d. but form a martingale difference sequence.

As we shall see, differently from [Kim and Paik, 2019] we cannot rely on Proposition 1 of Bastani and Bayati [2020], indeed their analysis refers to the standard lasso estimator which in turn can build on standard sub-Gaussian concentration arguments. Below we present a complete characterization of the oracle inequalities relying on the concentration of the euclidean norm of a sub-Gaussian random vector.

[1] As we shall see in Section 5, this setting includes the jointly sparse multitask setting by letting $M = dT$, the vector $\mathbf{w}$ be the concatenation of vectors $\mathbf{w}_1, \dots, \mathbf{w}_T \in \mathbb{R}^d$, and each of the $G = d$ groups contain the same component across the T tasks.

Lemma 1 (Group-Lasso Oracle Inequality Without i.i.d. Data)**.** *Let $\mathcal{F}_n$ denote the filtration up to round $n-1$*

$$\mathcal{F}_{n-1} = \{\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n\}.$$

For any $n \in [N]$ let $\mathbf{x}_n \in \mathbb{R}^M$ and $y_n = \mathbf{x}_n^\top \mathbf{w} + \eta_n \in \mathbb{R}$, where η_n is zero-mean conditionally 1-sub-Gaussian w.r.t. $\mathcal{F}_{n-1}$. For any $j \in [G]$ choose $\lambda_{n,j}(\delta) \geq \sqrt{\frac{M_j}{n}\log\left(\frac{G}{\delta}\right)}$, where $\delta \in (0,1)$. Then with probability at least $1-\delta$

$$\frac{1}{n}\left\|\mathbf{X}_n\left(\widehat{\mathbf{w}}_n - \mathbf{w}\right)\right\|^2 \leq 4\sum_{j\in[G]}\lambda_{n,j}(\delta)\left\|(\mathbf{w}-\widehat{\mathbf{w}}_n)^j\right\|. \tag{3}$$

When considering the lasso estimator additional assumptions are usually required in order to guarantee a fast rate of convergence towards the true vector $\mathbf{w} \in \mathbb{R}^M$ [Van De Geer and Bühlmann, 2009, Bühlmann and Van De Geer, 2011]. In the literature two alternatives emerged among others, the restricted eigenvalues (RE) assumption [Bickel et al., 2009, Koltchinskii et al., 2009] and the compatibility conditions (CC) [Van De Geer and Bühlmann, 2009]. In order to get fast rates, in [Lounici et al., 2011] authors analyze the group-lasso estimator convergence properties under the RE assumption in the fixed-design setting (considering $\mathbf{X}_N$ not to be a random quantity). However, when considering the bandit framework, $\mathbf{X}_N \in \mathbb{R}^{N\times M}$ consists of $\mathbf{x}_n \in \mathbb{R}^M$ $\mathcal{F}_n$-measurable random vectors. Notably, asking a bandit policy to collect samples satisfying the RE assumption is too demanding. Hence, here we begin by recalling the group lasso CC [Section 8.3.3 Bühlmann and Van De Geer, 2011] which are weaker than the RE assumption [Van De Geer and Bühlmann, 2009].

Definition 1 (Σ-Compatibility Conditions)**.** *Let $\Sigma \in \mathbb{R}^{M\times M}$ be symmetric and positive semi-definite, and let $J \subset [G]$. We say that J satisfies the Σ-group lasso compatibility conditions at round $n \in [N]$, with constant $\phi_\Sigma(J) > 0$, if $\forall\ \Delta \in \mathbb{R}^M \setminus \{\mathbf{0}\}$ such that $\sum_{j\in J^c}\lambda_{n,j}(\delta)\left\|\Delta^j\right\| \leq 3\sum_{j\in J}\lambda_{n,j}(\delta)\left\|\Delta^j\right\|$, it holds*

$$\left\|\Delta_J\right\|^2 \leq \frac{\|\Delta\|_\Sigma^2}{n\phi_\Sigma^2(J)},$$

where J^c denotes the complement of the set of indices J.

We are now ready to present the main result of this section.

Theorem 1 (Fast-rate Group Lasso Oracle Inequalities)**.** *Let $\mathcal{F}_n$ denote the filtration up to round $t-1$*

$$\mathcal{F}_{n-1} = \{\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{n-1}, y_{n-1}, \mathbf{x}_n\}.$$

For any $n \in [N]$ let $\mathbf{x}_n \in \mathbb{R}^M$ and $y_n = \mathbf{x}_n^\top \mathbf{w} + \eta_n \in \mathbb{R}$ with η_n be zero-mean conditionally 1-sub-Gaussian w.r.t. $\mathcal{F}_{n-1}$. Assume that the index set $J := J(\mathbf{w}) \subset [G]$ satisfies the Σ-Compatibility conditions with matrix $\Sigma = \mathbf{X}_n^\top \mathbf{X}_n$ and constant $\phi_\Sigma(J)$. For any $j \in [G]$ choose $\lambda_{n,j}(\delta) \geq \sqrt{\frac{M_j}{n}\log\left(\frac{G}{\delta}\right)}$ where $\delta \in (0,1)$, then with probability at least $1-\delta$ it holds that

$$\frac{1}{n}\left\|\mathbf{X}_n\left(\widehat{\mathbf{w}}_n - \mathbf{w}\right)\right\|^2 \leq \frac{16}{\phi_\Sigma^2(J)}\sum_{j\in J}\lambda_{n,j}^2(\delta), \tag{4}$$

$$\sum_{j\in[G]}\left\|\left(\widehat{\mathbf{w}}_n - \mathbf{w}\right)^j\right\| \leq \frac{16}{\phi_\Sigma(J)}\sum_{j\in J}\frac{\lambda_{n,j}^2(\delta)}{\min_{j\in J}\lambda_{n,j}}. \tag{5}$$

Building on this result, remarkably on the bound of Equation (5), in the next section we will present our group-lasso variant of the doubly-robust lasso policy of [Kim and Paik, 2019].

4 GROUP-LASSO BANDIT POLICY

While in the standard supervised regression problem the Σ-Compatibility conditions (Definition 1) refers to the Gram matrix $\Sigma = \mathbf{X}^\top\mathbf{X}$ computed over all the i.i.d. samples, the same does not occur in the stochastic bandit setting (see Section 2.1). Specifically, the Gram matrix $\mathbf{V}_n = \mathbf{X}_n^\top\mathbf{X}_n$ is not built to satisfy the compatibility conditions, hence we cannot directly use the results of Theorem 1. Indeed, since a bandit policy aims at choosing arms yielding maximum reward, the collected samples will tend to span only a small region of the whole context/input space. To overcome this obstacle, in this section we will follow the approach introduced by Kim and Paik [2019], which differently from Bastani and Bayati [2020] does not require to force sampling all arms every $O(\log N)$ rounds in order to collect i.i.d. observations. Following the result in [Bastani and Bayati, 2020, Corollary 3.4], in the next lemma we show that when $\mathbf{x}_n$ are i.i.d. and $\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^\top]$ satisfies the Σ-CC with constant $\phi_\Sigma(J)$, the same property would hold when considering matrix $\widehat{\Sigma}_n = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top/n$.

Lemma 2 (CC with Random Matrices)**.** *Let $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n \in \mathbb{R}^M$ be sampled independently from a distribution $\mathcal{P}$ with covariance Σ and supported on the unit sphere, and let $\widehat{\Sigma}_n = \sum_{i=1}^n \widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i^\top/n$ be the empirical covariance. Suppose that the set J satisfies the Σ-CC in Definition 1 with constant $\phi_\Sigma(J)$ and let $c = \min\left(\frac{1}{2}, \frac{\phi_\Sigma(J)^2}{256M(\mathbf{w})}\right)$. If*

$$n \geq z_N \equiv \max\left(\frac{3}{c^2}\log G, \frac{1}{c^2}\log\frac{N^2}{\delta}\right)$$

then, with probability at least $1-\delta/n^2$, the set J satisfies the $\widehat{\Sigma}_n$-CC with constant $\frac{\phi_\Sigma(J)}{\sqrt{2}}$.

Following the reasoning in [Kim and Paik, 2019] we can now develop our doubly-robust group-lasso policy π_{DR}. At each round $n \in [N]$ and for any arm $\mathbf{x} \in \mathcal{K}_n$, we denote by

$$\pi_{\mathbf{x}}^{DR}(n) = \mathbb{P}[\mathbf{x}_n = \mathbf{x} \mid \overline{\mathcal{F}}_{n-1}]$$

the conditional probability of selecting arm $\mathbf{x}$ associated to policy π_{DR} conditioned on

$$\overline{\mathcal{F}}_{n-1} = \Big\{ \{\mathbf{x}_i(1)\}_{i\in\mathcal{K}_1}, y_1, \dots, \\ \dots, \{\mathbf{x}_i(n-1)\}_{i\in\mathcal{K}_{n-1}}, y_{n-1}, \{\mathbf{x}_i(n)\}_{i\in\mathcal{K}_n}\Big\}.$$

Notice that differently from $\mathcal{F}_{n-1}$, for each round $n \in [N]$ $\overline{\mathcal{F}}_{n-1}$ contains all the different arms of $\mathcal{K}_n$ and not only the one chosen $\mathbf{x}_n$. As done in [Kim and Paik, 2019] we can now construct the doubly-robust pseudo reward

$$\widehat{y}_n = \frac{1}{K}\left[\sum_{\mathbf{x}(i)\in\mathcal{K}_n} \mathbf{x}(i)^\top \widehat{\mathbf{w}}_n + \frac{y_n - \mathbf{x}_n^\top \widehat{\mathbf{w}}_n}{\pi^{DR}_{\mathbf{x}_n}(n)}\right], \qquad (6)$$

where $\widehat{\mathbf{w}}_n$ is the estimate of true parameter $\mathbf{w}$ given $\overline{\mathcal{F}}_{n-1}$. The defined estimator satisfies $\mathbb{E}[\widehat{y}_n|\overline{\mathcal{F}}_{n-1}] = \widehat{\mathbf{x}}_n^\top \mathbf{w}$ independently on the adopted $\widehat{\mathbf{w}}_n$ where $\widehat{\mathbf{x}}_n = \frac{1}{K}\sum_{\mathbf{x}\in\mathcal{K}_n} \mathbf{x}$. Let us now introduce our last assumption.

Assumption D (Assumption 3 in [Kim and Paik, 2019]). *For each $n \in [N]$, the matrix*

$$\Sigma = \mathbb{E}\left[\widehat{\mathbf{x}}_n \widehat{\mathbf{x}}_n^\top\right]$$

satisfies the CC (Definition 1) with constant $\phi_\Sigma(J)$.

Hence, thanks to Assumptions (A) and (D), when considering the pair $\left(\widehat{\mathbf{X}}_n, \widehat{\mathbf{y}}_n\right) = \left([\widehat{\mathbf{x}}_1, ..., \widehat{\mathbf{x}}_n]^\top, [\widehat{y}_1, ..., \widehat{y}_n]^\top\right)$ and applying on it the group-lasso estimator (Equation (2)) we can leverage the Σ-CC of Lemma 2. Thus, the group-lasso estimator $\widehat{\mathbf{w}}_n$ would satisfy the fast-rate results of Theorem 1.

Following the result of [Kim and Paik, 2019], with the next result we show that the considered doubly-robust pseudo reward has constant variance, conditional on $\overline{\mathcal{F}}_{n-1}$.

Proposition 1. *Let $\mathrm{Var}_{\widehat{y}_n}$ be the conditional variance of $\widehat{y}_n$ given $\overline{\mathcal{F}}_{n-1}$. Then $\mathrm{Var}_{\widehat{y}}$ can be upper bounded by a constant if the following conditions are met*

$$\pi^{DR}_{\mathbf{x}}(n) = \begin{cases} \frac{1}{K} & \text{if } n \leq z_N \\ \frac{32}{K}\frac{M}{\phi_\Sigma(J)\sqrt{M_{min}}}\sqrt{\frac{\log G}{n}} & \text{if } n > z_N \end{cases}$$

where $M_{min} = \min_{j\in J} M_j$ and z_N is defined in Lemma 2.

Following the same construction adopted in Kim and Paik [2019] for the standard lasso estimator we propose the DR group-lasso bandit policy π^{DR} (Algorithm 1). Basically, π^{DR} selects the next arm $\mathbf{x}_n$ according to the discrete uniform distribution $\mathcal{U}(\mathcal{K}_n)$ with support $\mathcal{K}_n$ as long as $n \leq z_N$. When $n > z_N$ it randomizes the arm selection guaranteeing the satisfiability of Proposition 1. To do so, it first generates a random sample m_n according to a Bernoulli distribution $\mathcal{B}(\widetilde{\lambda}_n)$ with mean

$$\widetilde{\lambda}_n = \widetilde{\lambda}\frac{M}{\phi_\Sigma(J)\sqrt{M_{\min}}}\sqrt{\frac{\log G}{n}} \qquad (7)$$

Algorithm 1 Doubly Robust (DR) Group-Lasso Bandit

```
Require: $\widetilde{\lambda}, \delta, z_N$
Set $\widehat{\mathbf{w}}_0 = \mathbf{0}$.
for $n = 1$ to $N$ do
    if $n \leq z_N$ then
        Select $\mathbf{x}_n \sim \mathcal{U}(\mathcal{K}_n)$
    else
        Sample $m_n \sim \mathcal{B}(\widetilde{\lambda}_n)$ with $\widetilde{\lambda}_n$ as in Eq. (7)
        if $m_n = 1$ then
            Select $\mathbf{x}_n \sim \mathcal{U}(\mathcal{K}_n)$
        else
            Select $\mathbf{x}_n = \arg\max_{\mathbf{x}\in\mathcal{K}_n} \mathbf{x}^\top \widehat{\mathbf{w}}_n$
        end if
    end if
    Observe feedback $y_n$
    Compute $\widehat{y}_n$ according to Eq. (6) with
    $\pi^{DR}_{\mathbf{x}}(n) = \frac{\widetilde{\lambda}_{1,n}}{K} + (1 - \widetilde{\lambda}_{1,n})\mathbb{I}\{m_n = 0\}$
    Update $\widehat{\mathbf{w}}_n$ according to Eq. (2) with parameters
    $\lambda_{n,j}(\delta) = \sqrt{\frac{M_j}{n}\log\left(\frac{G}{\delta}\right)} \quad \forall j \in [G]$
end for
```

where $\widetilde{\lambda} > 0$ is a tuning parameter. Then, if $m_n = 1$ it picks the next arm with probability $1/K$, otherwise $\mathbf{x}_n = \arg\max_{\mathbf{x}\in\mathcal{K}_n} \mathbf{x}^\top \widehat{\mathbf{w}}_n$. Recall that $\widehat{\mathbf{w}}_n$ is computed relying on the pair $(\widehat{\mathbf{X}}_n, \widehat{\mathbf{y}}_n)$. We can now present the main result of this section that is a high-probability regret bound for the DR group-lasso policy.

Theorem 2. *Fixing an horizon $N > 0$, let Assumptions A and D hold true. Given the sets of indices $\mathcal{G}_1, \dots, \mathcal{G}_G$ satisfying Assumption C. Considering the linear bandit setting specified in Section 2.1, running the policy π^{DR} yields the following regret upper bound with probability at least $1 - 2\delta$*

$$R(N, \mathbf{w}) \leq O\left(\frac{M(\mathbf{w}) M_{max}\sqrt{N}}{\sqrt{M_{min}}}\right)$$

where we denoted $M_{max} = \max_{j\in[G]} M_j$.

The complete statement can be found in Section B of the supplementary material. Differently from [Kim and Paik, 2019] it relies on the group-lasso oracle inequality proved in Theorem 1 and not on the vanilla-lasso variant introduced in [Bastani and Bayati, 2020].

Remark 1. *Notice that by considering $G = M$ groups satisfying $\mathcal{G}_j = j \; \forall j \in [M]$, the regret bound would be of order $O(s\sqrt{N})$. This result is coherent with the bound obtained in [Kim and Paik, 2019] for the simple lasso estimator.*

Result Discussion. If compared to the standard ridge-regression based bandit policy [Abbasi-Yadkori et al., 2011], we got a reduction in terms of regret of order $\sqrt{M_{\min}}$.
As we will discuss in the next section, when considering the multi-task and meta-learning frameworks groups

would be homogeneous (i.e. $M_j = M/T \ \forall j \in [G]$). In this special case, the regret bound would be of order $O(M(\mathbf{w})\sqrt{NM/T})$.

5 MULTI-TASK AND META-LEARNING

We now turn to the main goal of this paper, which is to present a suitable solution to speedup the learning rate when dealing with multiple tasks with common sparsity patterns. We discuss both the setting of multitask and meta-learning.

5.1 BOUNDING THE MULTITASK REGRET

In multitask learning, we consider the problem of concurrently learning T linear bandit tasks each lasting for N_0 rounds and specified by regression vectors $\mathbf{w}_1, \dots, \mathbf{w}_T \in \mathbb{R}^d$ that satisfy the joint sparsity condition in Assumption B. At each round $n \in [N_0]$ the learner will sequentially face all the T tasks. When considering the generic task $t \in [T]$ our objective is to pick the next arm $\mathbf{x}_{t,n} \in \mathbb{R}^d$ from a finite-set of alternatives $\mathcal{K}^t_n \subset \mathbb{R}^d$ associated to that task. The learning objective is then to minimize the so called multi-task regret:

$$\overline{R}(T, N_0) = \sum_{t=1}^{T} R(N_0, \mathbf{w}_t) = \sum_{t=1}^{T}\sum_{n=1}^{N_0} (\mathbf{x}^*_{t,n} - \mathbf{x}_{t,n})^\top \mathbf{w}_t, \tag{8}$$

where $\mathbf{x}^*_{t,n} = \arg\max_{\mathbf{x}\in\mathcal{K}^t_n} \mathbf{x}^\top \mathbf{w}_t$.
Our solution relies on the doubly-robust group-Lasso policy π^{DR} defined in Section 4. Hence, at each round $n \in [N_0]$ we might want to solve the following set of learning problems

$$\begin{cases} \widehat{\mathbf{y}}_{1,n} & = \widehat{\mathbf{X}}_{1,n}\mathbf{w}_1 + \boldsymbol{\eta}_{1,n} \\ & \vdots \\ \widehat{\mathbf{y}}_{T,n} & = \widehat{\mathbf{X}}_{T,n}\mathbf{w}_T + \boldsymbol{\eta}_{T,n}. \end{cases} \tag{9}$$

There, for each task $t \in [T]$ and round $n \in [N_0]$ we have $\widehat{\mathbf{X}}_{t,n} \in \mathbb{R}^{n\times d}$ to be a matrix of i.i.d. components, $\mathbf{w}_t \in \mathbb{R}^d$ and $\widehat{\mathbf{y}}_{t,n} \in \mathbb{R}^n$. Finally, $\boldsymbol{\eta}_{t,n} \in \mathbb{R}^n$ is a random vector with conditionally independent components $\eta_{t,n}$ with respect to the sigma algebra

$$\overline{\mathcal{F}}^t_{n-1} = \Big\{ \{\mathbf{x}_j(1)\}_{j\in\mathcal{K}^t_1}, y_1, \dots \\ \dots, \{\mathbf{x}_j(n-1)\}_{j\in\mathcal{K}^t_{n-1}}, y_{n-1}, \{\mathbf{x}_j(n)\}_{j\in\mathcal{K}^t_n} \Big\}.$$

Inspired by the work of [Lounici et al., 2011] we now show how using our bandit policy based on the group-lasso estimator (see Algorithm 1) can boost the performance in terms of multi-task and transfer regrets (see Equation (8)) above, and Equation (12) below, respectively).

Reduction Scheme. Let us set the number of features $M = dT$ and denote with $\overline{\mathbf{w}} \in \mathbb{R}^M$ the vector obtained by stacking together $\mathbf{w}_1, \dots, \mathbf{w}_T \in \mathbb{R}^d$. Similarly, $\widehat{\overline{\mathbf{y}}}_n$ represents the $N = nT$ dimensional vector obtained by stacking $\widehat{\mathbf{y}}_{1,n}, \dots, \widehat{\mathbf{y}}_{T,n}$. We also consider $\widehat{\overline{\mathbf{X}}} \in \mathbb{R}^{N\times M}$ to be a block diagonal matrix consisting of T blocks, where each block $t \in [T]$ corresponds to the design matrix $\widehat{\mathbf{X}}_{t,n} \in \mathbb{R}^{n\times d}$ associated to task t.
In the proposed reduction, Algorithm 1 will play for a total of $N = TN_0$ rounds. Equivalently, at each round $n \in [N]$ we can assume Algorithm 1 to observe all T tasks sequentially.
Finally, we will organize the M features in the following $\mathcal{G}_1, \dots, \mathcal{G}_G$ groups, where $G = d$. Specifically, for any $j \in [d]$ the group $\mathcal{G}_j$ contains the same number T of features referring to the same feature j over the different T tasks. Hence, the multi-task learning problem reduces to minimize the multi-task regret in Equation (8), using the group-lasso bandit policy π^{DR} to estimate the concatenated vector $\overline{\mathbf{w}} \in \mathbb{R}^M$, considering the proposed groups and matrices definitions.
We now give an equivalent version of the Σ-(CC) adapted to the multi-task case.

Definition 2. *Let $\overline{\Sigma} \in \mathbb{R}^{M\times M}$ be symmetric and positive semi-definite. Given the index set $\overline{J} \subset [M]$, at round $n \in [N_0]$ the $\overline{\Sigma}$-group lasso compatibility conditions are met for the set $\overline{J}$ with constant $\phi_{\overline{\Sigma}}(\overline{J}) > 0$, if $\forall\ \Delta \in \mathbb{R}^M \setminus \{\mathbf{0}\}$ such that $\sum_{j\in J^c} \lambda_{n,j}(\delta)\, \|\Delta_j\|_{2,1} \leq 3\sum_{j\in J} \lambda_{n,j}(\delta)\, \|\Delta_j\|_{2,1}$, it holds*

$$\|\Delta_J\|^2 \leq \frac{\|\Delta\|^2_{\overline{\Sigma}}}{nT\phi^2_{\overline{\Sigma}}(J)},$$

where $\|\Delta\|_{2,1} = \sum_{j\in[G]} \|\Delta^j\|$.

Now, if considering the optimization problem stated in Equation (2) with regularization parameters $\lambda_{n,j}(\delta) = \lambda_n \geq \sqrt{\frac{T}{n}\log\frac{d}{\delta}}\ \forall j \in [d]$ we could still leverage the regret upper bound of Theorem 2 relative to the doubly-robust group-Lasso bandit policy π^{DR}. Yet, given the stated assumption and the proposed reduction scheme, we can adopt the group-lasso bandit policy when considering the multi-task problem. Hence, the following corollary of Theorem 2 holds.

Corollary 1. *Let us consider the multi-task bandit model, aimed to minimize the multi-task regret in Equation (8). Choose the groups $\mathcal{G}_1, \dots, \mathcal{G}_G$ as in the multitask setting described above and let Assumptions A, B and D hold with respect to Definition 2. Then, the multi-task regret associated to the policy π^{DR} satisfies*

$$\overline{R}(T, N_0) \leq \widetilde{O}\left(s\sqrt{TN_0}\right) \tag{10}$$

with probability at least $1-2\delta$, where $\widetilde{O}(\cdot)$ hides logarithmic factors.

Result Discussion. If compared to the proof of Theorem 2, the only difference results in the application of the oracle inequalities. Indeed, at each round $n \in [N]$ the obtained convergence rate for Equation (5) scales with respect to the sparsity level s associated to each task weight vector. Additionally, as we mentioned at the end of the last section, all groups have the same cardinality T, which yields a bound of order $\sqrt{T}$.

The benefit given by the proposed solution can be observed by comparing to the regret one would incur by separately running T independent lasso-bandit policies [Kim and Paik, 2019], that is, one policy per task. In this case, the regret bound would be of order $\overline{R}(T, n) \leq O\left(sT\sqrt{n}\right)$ which is bigger than the RHS of (10) by a factor $\sqrt{T}$.

To the best of our knowledge, [Yang et al., 2020] is the only work which considered the same multi-task bandit setting. Their regret bound is of order $\overline{R}(T, N_0) \leq O\left(T\sqrt{sN_0} + \sqrt{dsTN_0}\right)$. The second term of their bound is never smaller than the RHS of Equation (10) as d is assumed to satisfy $d \gg s$. The relation between the first term and our regret bound depends on the relation between the sparsity constant s and the number of tasks T. Yet, our bound is worse only if s is bigger than T, which goes against their lower-bound assumption. To conclude, we can state that our regret bound is smaller by a multiplicative factor of order $O(\sqrt{d/s})$ than the bound in [Yang et al., 2020].

We also point out that in [Yang et al., 2020] the authors claim that their bound is optimal up to poly-logarithmic factors. This would mean that either our claim is false or that their lower bound argument is vacuous. In the next result we present a different and simple lower bound argument for the multi-task setting. In the proof we also highlight two weaknesses of the lower bound demonstration used in [Yang et al., 2020].

Theorem 3 (Lower Bound MTL)**.** *Let us consider the multi-task setting described in Section 2.2, where each task $t \in [T]$ satisfies Assumptions A and B. Then the multi-task regret (see Equation (8)) can be lower bounded as*

$$\overline{R}(T, N_0) \geq O\left(s\sqrt{TN_0}\right). \tag{11}$$

Proof. We consider the simpler scenario where all T sparse tasks (i.e. $s \ll d$) are equal to each other (which clearly satisfy Assumption B). This is equivalent to consider a single bandit task which lasts for TN_0 rounds instead of N_0. The known cumulative regret lower bound for this single task is of order $O\left(\sqrt{sdN_0}\right) > O\left(s\sqrt{N_0}\right)$ [see Theorem 24.3 Lattimore and Szepesvári, 2020]. The claimed lower bound directly follows by considering $N_0 = TN_0$. The underling reasoning characterizing our bound specifically refers to the multi-task setting where tasks cannot be assumed to be independent. Indeed, we consider the extreme case where they are all equal. □

Remark 2. *In the proof presented in [Theorem 2 Yang et al., 2020] we found two vacuous steps. The first one is in the contradiction argument used to prove Lemma 6 therein. Indeed, in the contradiction they compare to the lower bound associated to the single-task problem of [Han et al., 2020]. The problem is that, this argument would hold only if tasks would be completely independent to each other. This is not the case when you have multiple-tasks satisfying assumptions like Assumption B. Indeed, the referred single-task lower bound considers an harder scenario where less data are available, those associated to a single task (no data for additional tasks are given). The second deficiency appears in the proof of Lemma 7 in [Han et al., 2020], in a way analogous to the incorrect proof of their Lemma 6. The difference is in the application of Lemma 8 which is now adopted to lower bound the regret incurred in subgroups of tasks. As before, the implicit assumption is that the considered groups are independent to each other.*

Remark 3. *We want to remark that due to Assumptions A and D the setting considered in Corollary 1 is slightly simpler than the one of Theorem 3. Hence, even if the obtained bound matches the result of Theorem 3 up to logarithmic factors, we cannot claim the optimality of our solution.*

5.2 BOUNDING THE TRANSFER REGRET

We now move on to the meta-learning problem. We can adopt a reasoning similar to the one used for the multi-task setting. Specifically, we consider the set of tasks not to be observed in parallel, but to be given in sequence. After having solved T linear bandit tasks with parameters $\mathbf{w}_1, \dots, \mathbf{w}_T \in \mathbb{R}^d$ satisfying Aassumption B, our objective is to perform well on the next $(T+1)$-th task which will still satisfy the same assumption. Following the reasoning in [Cella et al., 2020], this problem can be formulated as the minimization of the transfer regret which is defined as

$$\widetilde{R}(T, N_0) = R(N_0, \mathbf{w}_{T+1}) \tag{12}$$

having observed $\cup_{t=1}^{T} \overline{\mathcal{F}}^{t}_{N_0}$, which corresponds to the T $\sigma-$algebras associated to the already completed tasks. At each round $n \in [N_0]$ associated to task $T+1$, we could indeed assume to have to solve $T+1$ tasks in parallel all considered at the same round n. Hence, the following holds.

Corollary 2. *Let us consider the meta-learning bandit model, that is, the minimization of the transfer regret in Equation (12). Choose the groups $\mathcal{G}_1, \dots, \mathcal{G}_G$ as in the multitask setting described above and let Assumptions A, B and D hold with respect to Definition 2. Then, the transfer regret $\widetilde{R}(T, N_0)$ associated to the policy π^{DR} satisfies*

$$\widetilde{R}(T, N_0) \leq \widetilde{O}\left(\frac{s\sqrt{N_0}}{\sqrt{T}}\right) \tag{13}$$

with probability at least $1-2\delta$, where $\widetilde{O}(\cdot)$ hides logarithmic factors.

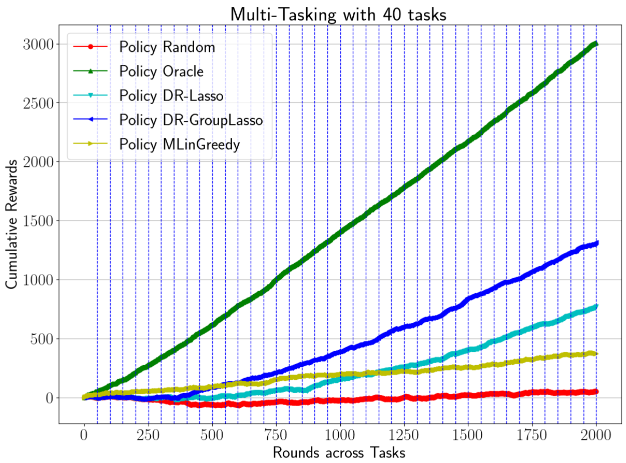

Figure 1: Cumulative reward measured over $T = 40$ tasks $N_0 = 50$ rounds long. Each task has $K = 50$ arms with $d = 50$ features and sparsity constant $s = 5$.

Result Discussion. As before, in order to see the benefit of the proposed solution we could compare with the strategy adopting the original lasso policy [Kim and Paik, 2019] independently on each task, whose regret would satisfy $\widetilde{R}(T, N_0) = R(N_0, \mathbf{w}_{T+1}) \leq O\left(s\sqrt{N_0}\right)$. Hence, coherently with the result of Corollary 1, the proposed solution would obtain a regret reduction of a factor $\sqrt{T}$.

At last we wish to point out that, while in this work we have investigated the benefit of jointly sparse multi-task bandits starting from the lasso bandit policy presented in [Kim and Paik, 2019], starting from the general result of Theorem 1, we conjecture that a similar generalization to group-lasso would hold with alternative lasso-bandit policies [see Bastani and Bayati, 2020, Hao et al., 2020, Ariu et al., 2020].

6 EXPERIMENTAL RESULTS

In this section we test the effectiveness of the approach proposed in Section 5 when solving the multi-task problem under group-sparsity. The theoretical bound stated in Corollary 1 showed a reduction of the incurred regret of order $\sqrt{T}$ compared to running the vanilla lasso-bandit policy [Kim and Paik, 2019] independently on the T tasks.

6.1 SYNTHETIC EXPERIMENTS

In the experiment displayed in Figure 1 we compare five different policies. All the parameters associated to the compared policies have been carefully selected over a logarithmic scale. The Oracle policy consists of separately running an instance of the OFUL strategy [Abbasi-Yadkori et al., 2011] on each task where the features space is restricted only to the shared active dimensions. The random policy simply chooses the arm to be pulled randomly. Considering this policy is necessary just to ensure the designed experiment instance not to be too simple. Then, we have the DR-lasso bandit policy [Kim and Paik, 2019] and our group-lasso variant proposed in Section 4. We want to highlight that the main focus of our experiment is to compare these two lasts policies. Remarkably the benefit brought by the considered group-lasso structure.
At last, we compare with the MLinGreedy strategy proposed in [Yang et al., 2020, Algorithm 1]. Particularly, referring to the notation in the above paper, at the end of each of the $m \in [\log_2 \log_2 N_0]$ phases, we separately run a ridge-regression scheme on each task considering only samples collected during the last phase. Subsequently, we combine the obtained estimators in a matrix $Z_n \in \mathbb{R}^{d \times n}$ whose columns consist of the previously estimated vectors. Finally, matrices B and W in [Yang et al., 2020, Algorithm 1] are calculated via a QR decomposition of matrix Z_n truncated at s features.

Synthetic Data. We generated an environment of tasks in agreement with Assumption B. Differently from the single-task environment used in [Kim and Paik, 2019], we ran $T = 40$ tasks in parallel, each lasting for 50 rounds. We considered each task to be a linear bandit problem with a 50-dimensional feature space, where only $s = 5$ features contribute the reward definition. The s sparse features are sampled according to the uniform distribution over the d original features. As done in [Kim and Paik, 2019], we consider $K = 50$ arm vectors which are sampled from a zero mean Gaussian distribution with covariance matrix Σ satisfying $\Sigma_{i,i} = 1 \ \forall i \in [d]$ and $\Sigma_{i,j} = 0.7 \ \forall i \neq j$. The noisy components characterizing rewards are drawn from a Gaussian distribution with 0 mean and 0.05 standard deviation. The vertical blue-lines indicates the end of the rounds associated to each task.

Result Discussion. In Figure 1 we can observe that the more the tasks, the higher the cumulative-reward collected by the group-lasso policy if compared to its vanilla lasso counterpart. The MLinGreedy strategy of [Yang et al., 2020] collects an higher cumulative reward than the random policy but his performance are far from being close to the one of our DR group-lasso solution. Finally, from a comparison with the Oracle policy we can observe a significant gap. This points out a practical limitation of the considered base strategy [Kim and Paik, 2019] which is independent on the multi-task setting considered in this paper. Hence, in the future it would be interesting to investigate a better performing algorithm. Notice that this does not seem to be a simple task, indeed, as remarked in [Kim and Paik, 2019] the considered policy already outperforms alternative strategies proposed for the

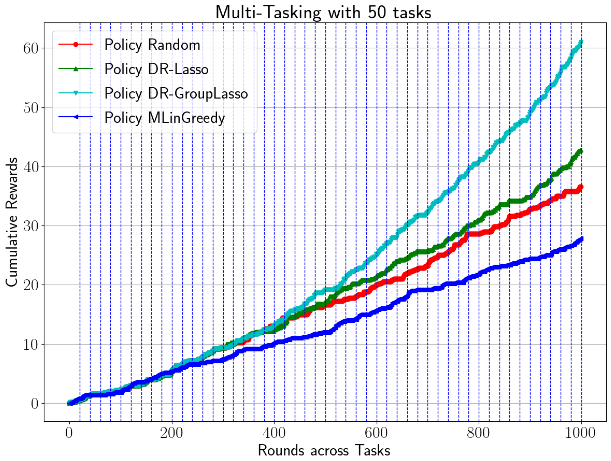

Figure 2: Cumulative reward measured over $T = 50$ tasks $N_0 = 20$ rounds long and averaged over 5 repetitions. Each task has $K = 30$ arms with $d = 20$ features.

sparse bandit setting.

6.2 REAL-DATA EXPERIMENTS

LastFM Data. We consider the Last.FM[2] dataset consisting of 92800 artist listening records from 1892 users. First, we apply an SVD transformation and consider the $d = 20$ most-relevant features for the users and the songs. Then, we randomly pick $T = 50$ users/tasks, each having $N_0 = 20$ rounds. In each round we randomly pick 29 arms (items) whose rating was < 4, and only one whose rating was ≥ 4. Finally, a positive binary reward is given only to the single arm having the highest score among the $K = 30$ available (scores are computed as the inner product between the user-song features).

Result Discussion. Coherently to the results observed over the synthetic data, in Figure 2 we can observe that the gap between the proposed policy and the competitors increase the more the number of tasks. This can be observed by comparing either to the vanilla DR-Lasso policy or to the MLinGreedy one. It is important to remark that this behavior was not always met overall the conducted experiments as it requires Assumption B to hold. Secondly, we can observe that when considering real data, the MLinGreedy policy seems to collect poor performance.

7 CONCLUSIONS

In this work we have investigated the benefit of the group-sparsity assumption in the linear bandit setting. Building on an existing lasso-bandit policy, we have generalized it to the group-lasso estimator. We provided novel group-lasso oracle inequalities suited for bandit collected samples, and discussed its application to the above setting. We then applied the group lasso bandit policy to multi-task and meta-learning linear bandits problems under joint sparsity assumptions on the task weight vectors. Specifically, in the multi-task setting we proposed a novel and simple lower bound which highlights some weaknesses of the existing state-of-the-art result. Finally, we corroborate our theoretical results with synthetic experiments. The poor practical performance of the vanilla policy also affected our group-lasso generalization. Hence, we are now considering alternative base policies.
In the future it would also be valuable to study optimal policies for the group lasso and multi-task learning setting in terms of minimax regret bound considering linear contextual bandit tasks.

[2] https://grouplens.org/datasets/hetrec-2011/

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 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.

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*, 3:397–422, 2003.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

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

Mohammad Gheshlaghi Azar, Alessandro Lazaric, and Emma Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 2220–2228, 2013.

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

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Toine Bogers. Movie recommendation using random walks over the contextual graph. In *Proc. of the 2nd Intl. Workshop on Context-Aware Recommender Systems*, 2010.

Craig Boutilier, Chih-Wei Hsu, Branislav Kveton, Martin Mladenov, Csaba Szepesvari, and Manzil Zaheer. Differentiable bandit exploration. *arXiv preprint arXiv:2002.06772*, 2020.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

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 *Advances in Neural Information Processing Systems 26*, 2014.

Laurent Cavalier, G. K. Golubev, Dominique Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *The Annals of Statistics*, 30(3):843–874, 2002.

Leonardo Cella and Nicolò Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs. *arXiv preprint arXiv:1910.02757*, 2019.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In *Proc. 37th International Conference on Machine Learning*, volume 119, pages 1360–1370, 2020.

Nicolò Cesa-Bianchi. *Multi-armed Bandit Problem*, pages 1356–1359. Springer New York, New York, NY, 2016.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 208–214, 2011.

Aniket An Deshmukh, Urun Dogan, and Clay Scott. Multi-task learning for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 4848–4856, 2017.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 2014.

Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, and Evans Etrue. On context-dependent clustering of bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1253–1262, 2017.

Aditya Gopalan, Odalric-Ambrym Maillard, and Mohammadi Zaki. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*, 2016.

Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. *arXiv preprint arXiv:2011.04020*, 2020.

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

Vladimir Koltchinskii et al. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

Branislav Kveton, Martin Mladenov, Chih-Wei Hsu, Manzil Zaheer, Csaba Szepesvari, and Craig Boutilier. Differentiable meta-learning in contextual bandits. *arXiv preprint arXiv:2006.05094*, 2020.

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.

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.

Thomas Nedelec, Clément Calauzènes, Noureddine El Karoui, and Vianney Perchet. Learning in repeated auctions. *arXiv preprint arXiv:2011.09365*, 2020.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

David Siegmund. Herbert robbins and sequential analysis. *Annals of Statistics*, pages 349–365, 2003.

Marta Soare. *Sequential Resource Allocation in Linear Stochastic Bandits*. PhD thesis, Lille University of Science and Technology, 2015.

Sara A. Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583, 2016.

Jiaqi Yang, Wei Hu, Jason D. Lee, and Simon S. Du. Provable benefits of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):49–67, 2006.

A PROOFS OF LEMMA 1 AND THEOREM 1

Proof of Lemma 1. Differently from [Lounici et al., 2011] we couldn't rely on [Cavalier et al., 2002, Equation (27)] yet we have to rely on martingale theory. When estimating the true parameter $\mathbf{w} \in \mathbb{R}^d$, according to Equation (2)

$$\frac{1}{n}\left\|\mathbf{X}_n\widehat{\mathbf{w}}_n - \mathbf{y}_n\right\|^2 + 2\sum_{j=1}^{G}\lambda_{n,j}\left\|\widehat{\mathbf{w}}_n^j\right\| \leq \frac{1}{n}\left\|\mathbf{X}_n\mathbf{w} - \mathbf{y}_n\right\|^2 + 2\sum_{j=1}^{G}\lambda_{n,j}\left\|\mathbf{w}^j\right\|.$$

Thanks to the reward definition $\mathbf{y}_n = \mathbf{X}_n\mathbf{w} + \boldsymbol{\eta}_n$ where $\boldsymbol{\eta}_n = [\eta_1, \dots, \eta_n]^\top$, the following holds

$$\frac{1}{n}\left\|\mathbf{X}_n\left(\widehat{\mathbf{w}}_n - \mathbf{w}\right) - \boldsymbol{\eta}_n\right\|^2 + 2\sum_{j=1}^{G}\lambda_{n,j}\left\|\widehat{\mathbf{w}}_n^j\right\| \leq \frac{1}{n}\left\|\boldsymbol{\eta}_n\right\|^2 + 2\sum_{j=1}^{G}\lambda_{n,j}\left\|\mathbf{w}^j\right\| \tag{14}$$

$$\implies \frac{1}{n}\left\|\mathbf{X}_n\left(\widehat{\mathbf{w}}_n - \mathbf{w}\right)\right\|^2 \leq 2\sum_{j=1}^{G}\lambda_{n,j}\left(\left\|\mathbf{w}^j\right\| - \left\|\widehat{\mathbf{w}}_n^j\right\|\right) + \frac{2}{n}\boldsymbol{\eta}_n^\top\mathbf{X}_n\left(\mathbf{w} - \widehat{\mathbf{w}}_n\right).$$

Applying Cauchy-Schwarz inequality to the last term on the RHS gives

$$\frac{1}{n}\left\|\mathbf{X}_n\left(\widehat{\mathbf{w}}_n - \mathbf{w}\right)\right\|^2 \leq 2\sum_{j=1}^{G}\lambda_{n,j}\left(\left\|\mathbf{w}^j\right\| - \left\|\widehat{\mathbf{w}}_n^j\right\|\right) + \sum_{j=1}^{G}\frac{2}{n}\left\|\left(\boldsymbol{\eta}_n^\top\mathbf{X}_n\right)^j\right\|\left\|\mathbf{w}^j - \widehat{\mathbf{w}}_n^j\right\|. \tag{15}$$

Now, for each group $\mathcal{G}_j \in \{\mathcal{G}_1, \dots, \mathcal{G}_G\}$ of the partition of $[d]$ we define $\mathcal{A}_{n,j} = \left\{\frac{2}{n}\left\|\left(\boldsymbol{\eta}_n^\top\mathbf{X}_n\right)^j\right\| \leq \lambda_{n,j}\right\}$. The objective is to guarantee that the previous event holds with probability at least $1-\delta$ overall the G groups.

$$\mathbb{P}\left[\mathcal{A}_n\right] = \mathbb{P}\left[\bigcap_{j=1}^{G}\mathcal{A}_{n,j}\right] \geq 1 - \bigcup_{j=1}^{G}\mathbb{P}\left[\mathcal{A}_{n,j}^c\right] \geq 1 - \sum_{j=1}^{G}\mathbb{P}\left[\mathcal{A}_{n,j}^c\right]$$

where $\mathcal{A}_{n,j}^c$ denotes the complement of event $\mathcal{A}_{n,j}$. The random variable $\left(\boldsymbol{\eta}_n^\top\mathbf{X}_n\right)^j$ consists of M_j terms, each of which is defined as $\sum_{s=1}^{n}\mathcal{D}_{s,i}\ \forall i \in \mathcal{G}_j$ where $\mathcal{D}_{s,i} = \eta_s\mathbf{x}_{s,i}$. We can observe that $\mathbb{E}\left[\mathcal{D}_{s,i}|\mathcal{F}_{s-1}\right] = 0$, however since our objective is to control the norm of a vector of independent sub-Gaussian random variables, differently from [Bastani and Bayati, 2020] we cannot rely on known sub-Gaussian tail inequalities. Indeed, when controlling $\left\|\left(\boldsymbol{\eta}_n^\top\mathbf{X}_n\right)^j\right\|$ we have to handle terms of the form η_s^2, which follow sub-exponential tails.

Let us now denote the i-th column of X_n by $\mathbf{X}_n^{(i)}$. It should be simple to observe that $\boldsymbol{\eta}_n^\top\mathbf{X}_n^{(i)}$ is distributed as a sub-Gaussian random variable with parameter n, indeed it satisfies

$$\mathbb{E}\left[\exp\left(\lambda\boldsymbol{\eta}_n^\top\mathbf{X}_n^{(i)}\right)\right] = \mathbb{E}\left[\exp\left(\lambda\eta_1 + \dots + \eta_{n-1}\right)\mathbb{E}\left[\exp\left(\lambda\eta_n\right)|\mathcal{F}_n\right]\right] \leq \exp\left(\lambda^2 n\right).$$

Hence vector $\left(\boldsymbol{\eta}_n^\top\mathbf{X}_n\right)^j$ consists of M_j sub-Gaussian independent random variables each with parameter n.

It follows that it is a sub-Gaussian vector with parameter $t\ M_j$, indeed thanks to the independence between the M_j components, we have

$$\mathbb{E}\left[\exp\left(\lambda\mathbf{u}^\top\left(\boldsymbol{\eta}_n^\top\mathbf{X}_n\right)^j\right)\right] \leq \exp\left(\lambda^2 M_j n\right) \quad \forall\mathbf{u} \in \mathbb{S}_{M_j} = \left\{\mathbf{u} \in \mathbb{R}^{M_j} : \|\mathbf{u}\| = 1\right\}.$$

Finally, relying on concentration results for sub-Gaussian random vectors [see Vershynin, 2018, Theorem 3.1.1], we have with probability at least $1 - \delta/N$

$$\left\|\left(\boldsymbol{\eta}_n^\top\mathbf{X}_n\right)^j\right\| \leq 4\sqrt{nM_j} + 2\sqrt{n}\sqrt{\log\left(\frac{G}{\delta}\right)} = \widetilde{O}\left(\sqrt{nM_j}\right) \tag{16}$$

where $\widetilde{O}(\cdot)$ hides $\log$ factors. Notice that in statement of the lemma we used $\lambda_{n,j} = 4\sqrt{\frac{M_j}{n}} + 2\sqrt{\log\left(\frac{1}{n}\right)\frac{G}{\delta}}$. Assuming event $\mathcal{A}_n$ to hold and starting from Equation (15) we have

$$\frac{1}{n}\left\|\mathbf{X}_n\left(\widehat{\mathbf{w}}_n - \mathbf{w}\right)\right\|^2 \leq 4\sum_{j\in J(\mathbf{w})}\lambda_{n,j}\min\left(\left\|\mathbf{w}^j\right\|, \left\|(\mathbf{w}-\widehat{\mathbf{w}}_n)^j\right\|\right) \leq 4\sum_{j\in J(\mathbf{w})}\lambda_{n,j}\left\|(\mathbf{w}-\widehat{\mathbf{w}}_n)^j\right\|$$

where we first used the definition of $\mathcal{A}_n$ and subsequently $\|A-B\| + \|A\| - \|B\| \leq 2\min(\|A\|, \|A-B\|)$. □

Proof of Theorem 1. Let us call $\Delta = \widehat{\mathbf{w}}_n - \mathbf{w}$ and let $J = J(\mathbf{w}) = \{\mathcal{G}_j : \mathbf{w}^j \neq 0 \in \mathbb{R}^{M_j}\}$. Starting from the result of Lemma 1 we have

$$\frac{1}{n}\left\|\mathbf{X}_n\Delta\right\|^2 \leq 4\sum_{j\in J(\mathbf{w})}\lambda_{n,j}\left\|\Delta^j\right\| \leq 4\sqrt{\sum_{j\in J(\mathbf{w})}\lambda_{n,j}^2}\left\|\Delta_J\right\|. \tag{17}$$

At the same time, when event $\mathcal{A}_n$ holds we have $\sum_{j\in J^c}\lambda_{j,n}\left\|\Delta^j\right\| \leq 3\sum_{j\in J}\lambda_{j,n}\left\|\Delta^j\right\|$. We can then leverage the Σ-compatibility conditions specified in Definition 1 with $\Sigma = \mathbf{X}_n^\top\mathbf{X}_n$ ensuring

$$\left\|\Delta_J\right\|^2 \leq \frac{\|\Delta\|_\Sigma^2}{n\,\phi_\Sigma^2(J)} = \frac{\|\mathbf{X}_n\Delta\|^2}{n\,\phi_\Sigma^2(J)}.$$

We can then obtain the following fast-rate oracle inequality

$$\frac{1}{n}\left\|\mathbf{X}_n\Delta\right\|^2 \leq 4\frac{\|\mathbf{X}_n\Delta\|}{\phi_\Sigma(J)}\sqrt{\frac{1}{n}\sum_{j\in J(\mathbf{w})}\lambda_{n,j}^2} \Longrightarrow \frac{1}{n}\left\|\mathbf{X}_n\Delta\right\|^2 \leq \frac{16}{\phi_\Sigma^2(J)}\sum_{j\in J(\mathbf{w})}\lambda_{n,j}^2. \tag{18}$$

In order to prove the second statement, thanks to the analysis presented in the previous lemma we have

$$\sum_{j\in[G]}\lambda_{n,j}\left\|\Delta^j\right\| \leq 4\sum_{j\in J}\lambda_{n,j}\left\|\Delta^j\right\|.$$

We can now use Cauchy-Schwarz and the CC to upper bound $\|\Delta_J\|$ obtaining

$$\sum_{j\in[G]}\lambda_{n,j}\left\|\Delta^j\right\| \leq 4\sqrt{\sum_{j\in J}\lambda_{n,j}^2}\left\|\Delta_J\right\| \leq 4\sqrt{\frac{1}{n}\sum_{j\in J}\lambda_{n,j}^2}\frac{\|\mathbf{X}_n\Delta\|}{\phi_\Sigma(J)}.$$

Finally, by using $\sum_{j\in[N]}\left\|\Delta^j\right\| \leq \sum_{j\in[N]}\frac{\lambda_{j,t}(\gamma)}{\min_{j\in[N]}\lambda_{j,t}(\gamma)}\left\|\Delta^j\right\|$ and Equation (18) we have

$$\sum_{j\in[G]}\left\|\Delta^j\right\| \leq \frac{16}{\phi_\Sigma(J)}\sum_{j\in J}\frac{\lambda_{n,j}^2}{\min_{j\in J}\lambda_{n,j}}.$$

□

B PROOFS OF LEMMA 2, PROPOSITION 1 AND THEOREM 2

Proof of Lemma 2. Defining the family of real-valued functions taking as input the random sample $\mathbf{x} \in \mathbb{R}^d$ as

$$g_{i,j,k}(\mathbf{x}) = \frac{\mathbf{x}_i\mathbf{x}_j\mathbb{I}\{i,j\in\mathcal{G}_k\} - \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_j\mathbb{I}\{i,j\in\mathcal{G}_k\}\right]}{2G}. \tag{19}$$

Adapting [Bastani and Bayati, 2020, Lemma EC.4] to the group-lasso case we have for all $w > 0$

$$\mathbb{P}\left[\left\|\Sigma - \widehat{\Sigma}\right\|_\infty \geq 2G\left(2 + \sqrt{2w} + \sqrt{\frac{2\log(2G)}{n}} + \frac{\log(2G)}{n}\right)\right] \leq \exp(-nw) \tag{20}$$

where Σ^j and $\widehat{\Sigma}^j$ are respectively the Gram matrix and the empirical Gram matrix with all the non $\mathcal{G}_j$ components set to zero, additionally we used $\frac{\|\Sigma-\widehat{\Sigma}\|_\infty}{2G} = \max_{k\in[G]} \max_{i,j\in\mathcal{G}_k} \sum_{s=1}^{n} |g_j(\mathbf{x}_s)|/t$.
Finally, suppose the Σ-CC holds for a set $J \subseteq [G]$ with constant $\phi_\Sigma(J)$ and $\left\|\widehat{\Sigma}-\Sigma\right\|^2 \leq \frac{\phi^2_\Sigma(J)}{32|J|}$, then the set J satisfies the $\widehat{\Sigma}$-CC with constant $\phi_\Sigma(J)/\sqrt{2}$. The proof follows directly by the combination of [Bühlmann and Van De Geer, 2011, Corollary 6.8] with the reasoning used in [Kim and Paik, 2019, Corollary 3.4]. □

Proof of Proposition 1. Starting from the conditional variance definition, the following holds

$$\begin{aligned}\mathrm{Var}_{\widehat{y}_n} &= \mathbb{E}\left[\left(\widehat{y}_n - \overline{\mathbf{x}}_n^\top \mathbf{w}\right)^2 \middle| \overline{\mathcal{F}}_{n-1}\right] = \mathbb{E}\left[\left(\overline{\mathbf{x}}_n^\top(\widehat{\mathbf{w}}_n - \mathbf{w}) + \frac{\eta_n}{K\pi_{\mathbf{x}}^{DR}(n)} + \frac{\mathbf{x}_n^\top(\mathbf{w}-\widehat{\mathbf{w}}_n)}{K\pi_{\mathbf{x}}^{DR}(n)}\right)^2 \middle| \overline{\mathcal{F}}_{t-1}\right] \\ &\leq \mathbb{E}\left[\left(\sum_{j\in J}\left\|\overline{\mathbf{x}}_n^j\right\|\left\|(\widehat{\mathbf{w}}_n-\mathbf{w})^j\right\| + \frac{\eta_n}{K\pi_{\mathbf{x}}^{DR}(n)} + \sum_{j\in J}\frac{\left\|\mathbf{x}_n^j\right\|\left\|(\mathbf{w}-\widehat{\mathbf{w}}_n)^j\right\|}{K\pi_{\mathbf{x}}^{DR}(n)}\right)^2 \middle| \overline{\mathcal{F}}_{n-1}\right],\end{aligned}$$

where the inequality follows from Cauchy-Schwarz.
Now, the second term in the square can be controlled thanks to the conditionally sub-Gaussianity assumption. Finally, the first and the third can be bounded by a constant as long as the following holds

$$\pi_{\mathbf{x}}^{DR}(n) = \begin{cases} \frac{1}{K} & \text{if } n \leq z_N \\ \frac{32}{K}\frac{d}{\phi_\Sigma(J)\sqrt{\min_{j\in J} M_j}}\sqrt{\frac{\log G}{n}} & \text{if } n > z_N. \end{cases}$$

□

Proof of Theorem 2. The proof follows from the one of [Kim and Paik, 2019, Theorem 4.1] with the only difference being in the considered parameters. The first component of the regret $R(N,a)$ is relative to the first z_N rounds where arms are selected according to the uniform distribution. The second component $R(N,b)$ considers only the cases where $m_N = 1$ and $t \geq z_N$. This term can be controlled with [Kim and Paik, 2019, Lemma 4.2] which gives

$$R(T,b) \leq \sum_{n=z_N}^{N} \widetilde{\lambda}_n \frac{d}{\sqrt{d_{\min}}}\sqrt{\frac{\log G}{n}} + \sqrt{\frac{N}{2}\log\left(\frac{1}{\delta}\right)} \leq \frac{d\sqrt{N\log N}}{\sqrt{d_{\min}}} + \sqrt{\frac{N}{2}\log\left(\frac{1}{\delta}\right)}.$$

Finally, the last term $R(T,c)$ can be controlled with [Kim and Paik, 2019, Lemma 4.3] which ensures

$$\mathbb{P}\left[\sum_{j\in J}\left\|(\widehat{\mathbf{w}}_n - \mathbf{w})^j\right\| \geq d_n\right] \leq \delta \qquad \text{for any } n \geq z_N$$

where

$$d_n = 32\frac{\max_{j\in J}\lambda^2_{n,j}(\gamma)}{\min_{j\in J}\lambda_{n,j}(\gamma)}\frac{M(\mathbf{w})}{\phi^2_\Sigma(J)} = 32\frac{M(\mathbf{w})\max_{j\in J} M_j}{\phi_\Sigma(J)\sqrt{\min_{j\in J} M_j}}\sqrt{\frac{\log N}{n}}.$$

We can finally control the last term considering the CC to be satisfied with constant $\phi/\sqrt{2}$ when considering the empirical gram matrix $\widehat{\Sigma}$. If we also consider the last result to hold, it follows that with probability at least $1-2\delta$

$$\begin{aligned}\sum_{n=1}^{N}\sum_{j\in J}\left\|(\widehat{\mathbf{w}}_n-\mathbf{w})^j\right\| \leq \sum_{n=1}^{N} d_n &\leq 32\frac{M(\mathbf{w})\max_{j\in J} M_j}{\phi_\Sigma(J)\sqrt{\min_{j\in J} M_j}}\sqrt{\log G}\sum_{n=1}^{N}\frac{1}{\sqrt{n}} \\ &\leq 32\frac{M(\mathbf{w})\max_{j\in J} M_j}{\phi_\Sigma(J)\sqrt{\min_{j\in J} M_j}}\sqrt{\log G}\sqrt{N\log N}.\end{aligned}$$

The obtained regret bound is the sum of these three terms. □