Meta-learning with Stochastic Linear Bandits

Leonardo Cella [1 2] **Alessandro Lazaric** [3] **Massimiliano Pontil** [2]

Abstract

We investigate meta-learning procedures in the setting of stochastic linear bandits tasks. The goal is to select a learning algorithm which works well on average over a class of bandits tasks, that are sampled from a task-distribution. Inspired by recent work on learning-to-learn linear regression, we consider a class of bandit algorithms that implement a regularized version of the well-known OFUL algorithm, where the regularization is a square euclidean distance to a bias vector. We first study the benefit of the biased OFUL algorithm in terms of regret minimization. We then propose two strategies to estimate the bias within the learning-to-learn setting. We show both theoretically and experimentally, that when the number of tasks grows and the variance of the task-distribution is small, our strategies have a significant advantage over learning the tasks in isolation.

1. Introdution

The multi-armed bandit MAB (Lattimore & Szepesvári, 2020; Auer et al., 2002; Siegmund, 2003; Robbins, 1952; Cesa-Bianchi, 2016; Bubeck et al., 2012) is a simple framework formalizing the online learning problem constrained to partial feedback. In the last decades it has receiving increasing attention due to its wide practical importance and the theoretical challenges in designing principled and efficient learning algorithms. In particular, applications range from recommender systems (Li et al., 2010; Cella & Cesa-Bianchi, 2019; Bogers, 2010), to clinical trials (Villar et al., 2015), and to adaptive routing (Awerbuch & Kleinberg, 2008), among others.

In this paper, we are concerned with linear bandits (Abbasi-Yadkori et al., 2011; Chu et al., 2011; Auer, 2003), a consolidated MAB setting in which each arm is associated with a vector of features and the arm payoff function is modeled by a (unknown) linear regression of the arm feature vector. Our study builds upon the OFUL algorithm introduced in (Abbasi-Yadkori et al., 2011), which in turned improved the theoretical analysis initially investigated in (Chu et al., 2011; Auer, 2003). Nonetheless, it may still require a long exploration in order to estimate well the unknown linear regression vector. An appealing approach to solve this bottleneck is to leverage already completed tasks by transferring the previously collected experience to speedup the learning process. This framework finds its most common application in the recommendation system domain, where we wish to recommend contents to a new user by matching his preference. Our objective is to rely on past interactions corresponding to navigation of different users to speedup the learning process.

Previous Work During the past decade, there have been numerous theoretical investigation of transfer learning, with a particular attention to the problems of multi-task (MTL) (Ando & Zhang, 2005; Maurer & Pontil, 2013; Maurer et al., 2013; 2016; Cavallanti et al., 2010) and learning-to-learn (LTL) or meta-learning (Baxter, 2000; Alquier et al., 2017; Denevi et al., 2018a;b; 2019; Pentina & Urner, 2016). The main difference between these two settings is that MTL aims to solve the problem of learning well on a prescribed set of tasks (the learned model is tested on the same tasks used during training), whereas LTL studies the problem of selecting a learning algorithm that works well on tasks from a common environment (i.e. sampled from a prescribe distribution), relying on already completed tasks from the same environment (Pentina & Urner, 2016; Balcan et al., 2019; Denevi et al., 2018a; 2019). In either case the base tasks considered have always been supervised learning ones. Recently, the MTL setting has been extended to a class of bandit tasks, with encouraging results empirically and theoretically (Azar et al., 2013; Calandriello et al., 2014; Zhang & Bareinboim, 2017; Deshmukh et al., 2017; Liu et al., 2018), as well as the case where tasks belong to a (social) graph, a setting that is usually referred to as *collaborative linear bandit* (Cesa-Bianchi et al., 2013; Soare et al., 2014; Gentile et al., 2014; 2017). Differently from these works, the principal goal of this paper is to investigate the adoption of the meta-learning framework, which has been

[1]University of Milan [2]Istituto Italiano di Tecnologia [3]Facebook AI Research. Correspondence to: Leonardo Cella <leonardocella@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

successfully considered within the supervised setting, to the setting of linear stochastic bandits.

Contributions Our contribution is threefold. First, we introduce in Section 3 a variant of the OFUL algorithm in which the regularization term is modified by introducing a bias vector, analyzing the impact of the bias in terms of regret minimization. Second, and more importantly, in Sections 4 and 5 we propose two alternative approaches to estimate the bias, within the meta-learning setting. We establish theoretical results on the regret of these methods, highlighting that, when the task-distribution has a small variance and the number of tasks grows, adopting the proposed meta-learning methods lead a substantial benefit in comparison to using the standard OFUL algorithm. Finally, in Section 6 we compare experimentally the proposed methods with respect to the standard OFUL algorithm on both synthetic and real data.

2. Learning Foundations

In this section we start by briefly recalling the standard stochastic linear bandit framework and we then present the considered LTL setting.

2.1. Linear Stochastic Bandits

Let T be a positive integers and let $[T] = \{1, \dots, T\}$. A Linear Stochastic MAB is defined by a sequence of T interactions between the agent and the environment. At each round $t \in [T]$, the learner is given a decision set $\mathcal{D}_t \subseteq \mathbb{R}^d$ from which it has to pick an arm $\mathbf{x}_t \in \mathcal{D}_t$. Subsequently, it observes the corresponding reward $y_t = \mathbf{x}_t^\top \mathbf{w}^* + \eta_t$ which is defined by a linear relation with respect to an unknown parameter $\mathbf{w}^* \in \mathbb{R}^d$ combined with a sub-gaussian random noise term η_t. Thanks to the knowledge of the true parameter $\mathbf{w}^*$, at each round t the optimal policy picks the arm $\mathbf{x}_t^* = \arg\max_{\mathbf{x} \in D_t} \mathbf{x}^\top \mathbf{w}^*$, maximizing the instantaneous reward. The learning objective is to maximize the cumulative reward, or equivalently, to minimize the *pseudo-regret*

$$R(T, \mathbf{w}^*) = \sum_{t=1}^{T} (\mathbf{x}_t^* - \mathbf{x}_t)^\top \mathbf{w}^*.$$

As learning algorithm we consider OFUL (Abbasi-Yadkori et al., 2011). At each round $t \in [T]$, it estimates $\mathbf{w}^*$ by ridge-regression over the observed arm reward pairs, that is,

$$\widehat{\mathbf{w}}_{t+1}^{\lambda} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{X}_t \mathbf{w} - \mathbf{y}_t\|_2^2 + \lambda \|\mathbf{w}\|_2^2 \qquad (1)$$

where $\mathbf{X}_t$ is the matrix whose rows are $\mathbf{x}_1^\top, \dots, \mathbf{x}_t^\top$, $\mathbf{I}$ is the $d \times d$ identity matrix and $\mathbf{y}_t = (y_1, \dots, y_t)^\top$. A key insight behind OFUL is to update online a confidence interval $\mathcal{C}_t$ containing the true parameter $\mathbf{w}^*$ with high probability and centered in $\widehat{\mathbf{w}}_t^{\lambda}$. According to Theorem 2 of (Abbasi-Yadkori et al., 2011), assuming $\|\mathbf{w}^*\|_2 \le S$ and $\|\mathbf{x}\|_2 \le L$ $\forall \mathbf{x} \in \cup_{s=1}^{t} \mathcal{D}_s$, then for any $\delta > 0$, w.p. $\ge 1 - \delta$, $\forall t \ge 0$, $\mathbf{w}^*$ lies in

$$\mathcal{C}_t(\delta) = \Bigg\{ \mathbf{w} \in \mathbb{R}^d : \left\|\widehat{\mathbf{w}}_t^{\lambda} - \mathbf{w}\right\|_{\mathbf{V}_t^{\lambda}} \le R\sqrt{d \log \frac{1 + tL^2/\lambda}{\delta}} + \lambda^{\frac{1}{2}} S =: \beta_t^{\lambda}(\delta) \Bigg\} \qquad (2)$$

where $\mathbf{V}_t^{\lambda} = \lambda \mathbf{I} + \mathbf{X}_t^\top \mathbf{X}_t$. According to the optimism in the face of uncertainty principle, at each round t OFUL picks the arm $\mathbf{x}_t$ by solving the following optimization problem:

$$\mathbf{x}_t = \arg\max_{\mathbf{x} \in \mathcal{D}_t} \max_{\widetilde{\mathbf{w}}_t^{\lambda} \in \mathcal{C}_t} \mathbf{x}^\top \widetilde{\mathbf{w}}_t. \qquad (3)$$

As was proved in Lemma 5 of (Kuzborskij et al., 2019), this corresponds to pick:

$$\mathbf{x}_t \in \arg\max_{\mathbf{x} \in \mathcal{D}_t} \left\{ \mathbf{x}^\top \widehat{\mathbf{w}}_{t-1}^{\lambda} + \beta_{t-1}(\delta) \|\mathbf{x}\|_{\left(\mathbf{V}_{t-1}^{\lambda}\right)^{-1}} \right\}. \quad (4)$$

Finally, with probability at least $1 - \delta$, OFUL satisfies (see Theorem 3 of (Abbasi-Yadkori et al., 2011)):

$$R(T, \mathbf{w}^*) \le 4\sqrt{Td \log\left(1 + \frac{TL}{\lambda d}\right)} \Bigg(\lambda^{\frac{1}{2}} S + R\sqrt{2 \log(1/\delta) + d \log(1 + TL/(\lambda d))} \Bigg).$$

We can now formally introduce the considered LTL learning framework for the family of tasks we analyze in this work: biased regularized linear stochastic bandits.

2.2. LTL with Linear Stochastic Bandits.

We assume that each learning task $\mathbf{w} \in \mathbb{R}^d$ representing a linear bandit, is sampled from a task-distribution ρ of bounded support in $\mathbb{R}^d$. The objective is to design a meta-learning algorithm which is well suited to the environment. Specifically, we assume to receive a sequence of tasks $\mathbf{w}_1, \dots, \mathbf{w}_N, \dots$ which are independently sampled from the task-distribution (*environment*) ρ. Due to the interactive nature of the bandit setting, we do not have any prior information related to a new task; we collect information about it along the interaction with the environment. After completing the j-th task, we store the whole interaction in a dataset Z_j which is formed by T entries $(\mathbf{x}_{j,t}, y_{j,t})_{t=1}^{T}$. Clearly, the dataset entries are not i.i.d sampled from a given distribution, but each dataset Z_j corresponds to the recording of the learning policy in terms of the arm $\mathbf{x}_{j,t}$ picked from the decision set $\mathcal{D}_t^j$ and its corresponding reward $y_{j,t}$ while facing the task specified by the unknown vector $\mathbf{w}_j$. Starting from these datasets, we wish to design an algorithm $\mathcal{A}$ which suffers a low regret on a new task $\mathbf{w}_{N+1} \sim \rho$. This can be stated into requiring that $\mathcal{A}$ trained over N datasets has small *transfer-regret*:

$$\mathcal{R}(T, \rho) = \mathbb{E}_{\mathbf{w} \sim \rho}\Big[\mathbb{E}\big[R(T, \mathbf{w})\big]\Big]$$

where the inner expectation is with respect to rewards realizations due to their noisy components.

3. Biased Regularized OFUL

We now introduce BIAS-OFUL, a biased version of OFUL, which is instrumental for our meta-learning setting. Although not feasible, the proposed algorithm it serves as a basis to study the theoretical properties of meta-learning with stochastic linear bandit tasks. In Section 6 we will present a more practical version of it.

Regularized Confidence Sets The idea of following a bias in a specific family of learning algorithms is not new in the LTL literature (Denevi et al., 2018a; 2019; 2018b). Inspired by (Denevi et al., 2019) we modify the regularization in the computation of the confidence set centroid $\widehat{\mathbf{w}}_t^\lambda$, where the regularization is now defined as a square euclidean distance to the bias parameter $\mathbf{h} \in \mathbb{R}^d$. Given a fixed vector $\mathbf{h}$, at each round $t \in [T]$ BIAS-OFUL estimates the regularized centroid of the confidence ellipsoid as

$$\widehat{\mathbf{w}}_t^{\mathbf{h}} = \arg\min_{\mathbf{w}} \left\|\mathbf{X}_t^\top \mathbf{w} - \mathbf{Y}_t\right\|_2^2 + \lambda \left\|\mathbf{w} - \mathbf{h}\right\|_2^2$$

whose solution is given by

$$\widehat{\mathbf{w}}_t^{\mathbf{h}} = \left(\mathbf{V}_t^\lambda\right)^{-1} \mathbf{X}_t^\top (\mathbf{Y}_t - \mathbf{X}_t \mathbf{h}) + \mathbf{h}. \tag{5}$$

This result follows directly from the standard ridge-regression by using the substitution $\mathbf{v} = \mathbf{w} - \mathbf{h}$.
As we have mentioned in the previous section, at each round t OFUL keeps also updated a confidence interval $\mathcal{C}_t$ (see Equation 2) centered in $\widehat{\mathbf{w}}_t^\lambda$ which contains $\mathbf{w}^*$ with high probability. We now derive a confidence set for the biased regularized estimate $\widehat{\mathbf{w}}_t^{\mathbf{h}}$, assuming that we have access to an oracle to compute the distance $\|\mathbf{h} - \mathbf{w}^*\|_2$. This seems quite restrictive, however later in the paper we will show how levering similar related tasks we can exploit this bound to take advantage of the bias version of OFUL, without having to know the above distance a-priori.

Theorem 1. *Assuming $\|\mathbf{h}\|_2 \leq S$, $\|\mathbf{w}^*\|_2 \leq S$ and $\|\mathbf{x}\|_2 \leq L \; \forall \; \mathbf{x} \in \cup_{s=1}^t \mathcal{D}_s$, then for any $\delta > 0$, with probability at least $1-\delta$, $\forall t \geq 0$, $\mathbf{w}^*$ lies in the set*

$$\mathcal{C}_t^{\mathbf{h}}(\delta) = \Bigg\{ \mathbf{w} \in \mathbb{R}^d : \left\|\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}\right\|_{\mathbf{V}_t^\lambda} \leq \lambda^{\frac{1}{2}} \left\|\mathbf{h} - \mathbf{w}^*\right\|_2 + \\ + R\sqrt{2\log\left(\frac{\det\left(\mathbf{V}_t^\lambda\right)^{1/2}}{\det\left(\lambda I\right)^{1/2}\delta}\right)} = \beta_t^{\mathbf{h}}(\delta) \Bigg\}. \tag{6}$$

The proof can be found in the appendix material. We will now study the impact of the bias $\mathbf{h}$ in terms of regret.

3.1. Regret Analysis with Fixed Bias

Given the confidence set defined in Theorem 1 and the *optimism principle* translated into selecting the next arm according to Equation 4, we can analyze the expected pseudo-regret depending on the value of $\mathbf{h}$.

Lemma 1. *(BIAS-OFUL Expected Regret) Under the same assumptions of Theorem 1, if in addition, for all t and all $\mathbf{x} \in \mathcal{D}_t$, $\mathbf{x}^\top \mathbf{w}^* \in [-1, 1]$, and considering $\lambda \geq 1$, we have*

$$\overline{R}(T, \mathbf{w}^*) = \mathbb{E}\left[R(T, \mathbf{w}^*)\right] \\ \leq C\sqrt{Td\log\left(1 + \frac{TL}{\lambda d}\right)}\Bigg(\lambda^{\frac{1}{2}} \left\|\mathbf{w}^* - \mathbf{h}\right\|_2 + \\ + R\sqrt{d\log(T + T^2L/(\lambda d))}\Bigg)$$

where the expectation is respect to the reward generation and $C > 0$ is a constant factor.

We now analyze the regret for two different values of $\mathbf{h}$. In particular we wish to highlight how setting a good bias can speedup the process of learning with respect to using the standard OFUL approach (Abbasi-Yadkori et al., 2011).

Corollary 1. *Under the conditions of Lemma 1, the following bounds on the expected regret of BIAS-OFUL holds:*

(i) Independent Task Learning (ITL)*, given by setting $\mathbf{h} = \mathbf{0}$ satisfies the following expected regret bound*

$$\overline{R}(T, \mathbf{w}^*) \leq C\sqrt{Td\log\left(1 + \frac{TL}{\lambda d}\right)}\Bigg(\lambda^{\frac{1}{2}} S + \\ + R\sqrt{d\log(T + T^2L/(\lambda d))}\Bigg)$$

which is of order $\mathcal{O}(d\sqrt{T})$ for any $\lambda \geq 1$.

(ii) The Oracle*, given by setting $\mathbf{h} = \mathbf{w}^*$ satisfies*

$$\overline{R}(T, \mathbf{w}^*) \leq C\sqrt{Td\log\left(1 + \frac{TL}{\lambda d}\right)} \\ \left(R\sqrt{d\log(T + T^2L/(\lambda d))}\right)$$

which is 0 as $\lambda \to \infty$.

The proofs can be found in the supplementary material. The main intuition is that, as long as we can set $\mathbf{h} = \mathbf{w}^*$, the bigger the the regularization parameter λ is, the more the Oracle policy tends to select the arm only based on $\mathbf{w}^*$, thereby becoming equivalent to the optimal policy.

3.2. Transfer Regret Analysis with Fixed Bias

Following the above analysis for the single task case, we now study the impact of the bias in the transfer regret bound. To this end, we introduce the variance and the mean absolute distance of a bias vector $\mathbf{h}$ relative to the environment of task,

$$\text{Var}_{\mathbf{h}} = \mathbb{E}_{\mathbf{w}\sim\rho}\big[\,\|\mathbf{w}-\mathbf{h}\|_2^2\,\big], \quad \text{Mar}_{\mathbf{h}} = \mathbb{E}_{\mathbf{w}\sim\rho}\big[\,\|\mathbf{w}-\mathbf{h}\|_2\,\big]$$

and we observe that $\overline{\mathbf{w}} = \mathbb{E}_{\mathbf{w}\sim\rho}\mathbf{w} = \arg\min_{\mathbf{h}\in\mathbb{R}^d} \text{Var}_{\mathbf{h}}$ and $\mathbf{m} = \arg\min_{\mathbf{h}\in\mathbb{R}^d} \text{Mar}_{\mathbf{h}}$. With this in hand, we can now analyze how the transfer regret can be upper bounded as a function of the introduced terms.

Lemma 2. *(Transfer Regret Bound) Under the same conditions in Theorem 1 and Lemma 1, the expected transfer regret of BIAS-OFUL can be upper bounded as:*

$$\begin{aligned}\mathcal{R}(T,\rho) &\leq C\sqrt{Td\lambda\log\left(1+\frac{TL}{\lambda d}\right)}\text{Mar}_{\mathbf{h}} + \\ &\quad + RCd\sqrt{T\log\left(T+\frac{T^2L}{\lambda d}\right)\log\left(1+\frac{TL}{\lambda d}\right)} \\ &\leq C\sqrt{Td\lambda\log\left(1+\frac{TL}{\lambda d}\right)}\textit{Var}_{\mathbf{h}} + \\ &\quad + RCd\sqrt{T\log\left(T+\frac{T^2L}{\lambda d}\right)\log\left(1+\frac{TL}{\lambda d}\right)}\end{aligned}$$

Proof. The first statement is the expectation with respect to the *task*-distribution ρ applied to Lemma 1, while the second follows by applying Jensen's inequality. □

We can now replicate what we have done in Corollary 1 and consider the transfer regret bound for two different values of the hyper-parameter $\mathbf{h}$. The main difference is that here, there is not an a-priori correct value for $\mathbf{h}$ as it depends on the task-distribution ρ.

Corollary 2. *Under the same assumptions in Theorem 1 and Lemma 1, and setting $\lambda = \frac{1}{T\text{Var}_{\mathbf{h}}}$, the following bounds on the transfer regret hold*

(i) Independent Task Learning (ITL)*, given by setting the bias hypeparameter $\mathbf{h}$ equal to $\mathbf{0}$, satisfies*

$$\mathcal{R}(T,\rho) \leq \left[1+\sqrt{Td\log\left(T+\frac{T^3L\textit{Var}_{\mathbf{0}}}{d}\right)}\right] C\sqrt{d\log\left(1+\frac{T^2L\textit{Var}_{\mathbf{0}}}{d}\right)}$$

(ii) The Oracle*, given by setting the bias hyperparameter $\mathbf{h}$ equal to the mean task $\overline{\mathbf{w}}$, satisfies*

$$\mathcal{R}(T,\rho) \leq \left[1+\sqrt{Td\log\left(T+\frac{T^3L\textit{Var}_{\rho}}{d}\right)}\right] C\sqrt{d\log\left(1+\frac{T^2L\textit{Var}_{\rho}}{d}\right)}$$

where $\text{Var}_{\rho} = \text{Var}_{\overline{\mathbf{w}}}$.

Proof. These results directly follow from Lemma 2. We have picked $\lambda = \frac{1}{T\text{Var}_{\mathbf{h}}}$ in order to highlight the multiplicative term $\log(1+\text{Var}_{\mathbf{h}})$ which tends to zero as the the variance $\text{Var}_{\mathbf{h}}$ goes to zero. □

Therefore, running BIAS-OFUL with bias $\mathbf{h}$ equal to $\overline{\mathbf{w}}$ brings a substantial benefit with respect to the unbiased case when the second moment of the task-distribution ρ is much bigger than its variance. Specifically, we introduce the following assumption.

Assumption 1. *(Low Biased Variance)*

$$\textit{Var}_{\rho} = \mathbb{E}_{\mathbf{w}\sim\rho}\|\mathbf{w}-\overline{\mathbf{w}}\|_2^2 \ll \mathbb{E}_{\mathbf{w}\sim\rho}\|\mathbf{w}\|_2^2 = \textit{Var}_{\mathbf{0}}. \quad (7)$$

Notice also that the choice $\lambda = 1/(T\text{Var}_{\mathbf{h}})$, implies that, as $\text{Var}_{\overline{\mathbf{w}}}$ tends to 0, the regret upper bound of the oracle case tends to zero too reflecting the result of Corollary 1. More in general, we can state that when the environment (i.e. the task-distribution ρ) satisfies Assumption 1, leveraging on tasks similarity would gives a substantial benefit compared to learning each task separately. Since in practice the mean task parameter $\overline{\mathbf{w}}$ is unknown, in the following sections we propose two alternative approaches to estimate $\overline{\mathbf{w}}$.

Algorithm 1 Within Task Algorithm: BIAS-OFUL

Require: $\lambda > 0, \widehat{\mathbf{h}}_0 \in \mathbb{R}^d$
$\widehat{\mathbf{w}}_0^{\mathbf{h}} = \widehat{\mathbf{h}}_0, \mathbf{V}_0^{-1} = \frac{1}{\lambda}\mathbf{I}$.
for $t = 1$ **to** T **do**
  GET decision set D_t
  SELECT $\mathbf{x}_t \in D_t$ with bias $\mathbf{h} = \widehat{\mathbf{h}}_{j,t}^{\lambda}$
  OBSERVE reward y_t
  UPDATE $\mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top$
  UPDATE $\widehat{\mathbf{h}}_t$ according to the meta-algorithm
  UPDATE $\widehat{\mathbf{w}}_t^{\mathbf{h}}$ using Equation 5
end for

Algorithm 2 Meta-Algorithm: Estimating $\widehat{\mathbf{h}}^{\lambda}$

for $j = 1$ **to** N **do**
  SAMPLE new task $\mathbf{w}_j \sim \rho$
  SET $\widehat{\mathbf{h}}_{j,0}^{\lambda}$
  RUN Algorithm 1 with parameter $\widehat{\mathbf{h}}_{j,0}^{\lambda}$
end for

4. A High Variance Solution

In this section, we present our first meta-learning method. We begin by introducing some additional notation. We

let $\mathbf{x}^{\mathbf{h}}_{j,t}$ be the arm pulled by the BIAS-OFUL algorithm (Algorithm 1) at round t-th of the j-th task. We denote by $\mathbf{V}_{j,T} = \sum_{s=1}^{T} \mathbf{x}^{\mathbf{h}}_{j,s} \mathbf{x}^{\mathbf{h}\top}_{j,s}$ the design matrix computed with the T arms picked during the j-th task. For each terminated task $j \in [N]$ we also define $\mathbf{b}_{j,T} = \mathbf{X}_{j,T}^{\top} \mathbf{Y}_{j,T}$. Finally, we introduce the *mean estimation error*

$$\epsilon_{N,t}(\rho) = \left\| \overline{\mathbf{w}} - \widehat{\mathbf{h}}^{\lambda}_{N,t} \right\|_2^2$$

which is the error of our estimate $\widehat{\mathbf{h}}^{\lambda}_{N,t}$ with respect to the true mean task $\overline{\mathbf{w}}$, at round t of the $N+1$-th task.

4.1. Averaging the Estimated Task Parameters

An intuitive solution to bound the estimation error $\epsilon_{N,t}$ is to simply average of the estimated task parameters $\widehat{\mathbf{w}}^{\lambda}_{j}$ computed according to Equation 1 on the dataset Z_j without considering any bias.

$$\widehat{\mathbf{h}}^{\lambda}_{N,t+1} = \frac{1}{NT+t} \left(\sum_{j=1}^{N} T \widehat{\mathbf{w}}^{\lambda}_{j,T} + t \widehat{\mathbf{w}}^{\lambda}_{N+1,t} \right). \quad (8)$$

By adopting this approach, we have the following bound on the transfer regret.

Theorem 2. *(Transfer Regret Bound). Let the assumptions of Lemma 2 hold and let $\widehat{\mathbf{h}}^{\lambda}_{N,t}$ be defined as in Equation* (8)*. Then, it hold that*

$$\mathcal{R}(T,\rho) \leq dC \sqrt{T \log \left(1 + \frac{T^2 L \left(\mathrm{Var}_{\overline{\mathbf{w}}} + \epsilon_{N,T}(\rho) \right)}{d} \right)}$$

where the mean estimation error can be bound as

$$\sqrt{\epsilon_{N,T}(\rho)} \leq H_{\rho}(N+1, \overline{\mathbf{w}}) + \max_{j=1,\dots,N} \frac{\beta^{\lambda}_{j}\left(1/T\right)}{\lambda^{1/2}_{\min}(\mathbf{V}^{\lambda}_{j,T})}.$$

Here, $\beta^{\lambda}_{j}\left(\frac{1}{T}\right)$ refers to the confidence interval computed with OFUL (see Equation 2) and $H_{\rho}(N+1, \overline{\mathbf{w}}) = \left\| \overline{\mathbf{w}} - \overline{\mathbf{h}}_{N,t} \right\|_2$ with $\overline{\mathbf{h}}_{N,t+1} = \frac{1}{NT+t}\left(\sum_{j=1}^{N} T \mathbf{w}_j + t \mathbf{w}_{N+1} \right)$.

Proof. We follow the reasoning in Corollary 2, this time setting $\mathbf{h} = \widehat{\mathbf{h}}^{\lambda}_{N,T}$, and then observe that

$$\begin{aligned}
\sqrt{\epsilon_{N,T}(\rho)} &= \left\| \overline{\mathbf{w}} - \widehat{\mathbf{h}}^{\lambda}_{N,T} \right\|_2 \\
&\leq \left\| \overline{\mathbf{w}} - \overline{\mathbf{h}}_{N,T} \right\|_2 + \left\| \overline{\mathbf{h}}_{N,T} - \widehat{\mathbf{h}}^{\lambda}_{N,T} \right\|_2 \\
&= H_{\rho}(N+1, \overline{\mathbf{w}}) + \left\| \overline{\mathbf{h}}_{N,T} - \widehat{\mathbf{h}}^{\lambda}_{N,T} \right\|_2 \\
&\leq H_{\rho}(N+1, \overline{\mathbf{w}}) + \max_{1 \leq j \leq N+1} \left\| \mathbf{w}_j - \widehat{\mathbf{w}}^{\lambda}_{j,T} \right\|_2 \\
&\leq H_{\rho}(N+1, \overline{\mathbf{w}}) + \max_{1 \leq j \leq N+1} \frac{\left\| \mathbf{w}_j - \widehat{\mathbf{w}}^{\lambda}_{j,T} \right\|_{\mathbf{V}^{\lambda}_{j,T}}}{\lambda^{1/2}_{\min}(\mathbf{V}^{\lambda}_{j,T})} \\
&\leq H_{\rho}(N+1, \overline{\mathbf{w}}) + \max_{1 \leq j \leq N+1} \frac{\beta^{\lambda}_{j}\left(1/T\right)}{\lambda^{1/2}_{\min}(\mathbf{V}^{\lambda}_{j,T})}.
\end{aligned}$$

□

The term $H_{\rho}(N+1, \overline{\mathbf{w}})$ denotes the estimation error of the empirical mean computed from the $N+1$ tasks vectors $(\mathbf{w}_j)_{j=1}^{N+1}$, relative to the true mean $\overline{\mathbf{w}}$. Since the $\mathbf{w}_j$ are independent random d-dimensional vectors drawn from ρ we can apply the following vectorial version of the Bennett's inequality (see, e.g., Smale & Zhou, 2007, Lemma 2).

Lemma 3. *Let $\mathbf{w}_1, \dots, \mathbf{w}_N$ be N independent random vectors with values in $\mathbb{R}^d$ sampled from the task-distribution ρ. Assuming that $\forall j \in [N] : \|\mathbf{w}_j\| \leq S$, then for any $0 < \delta < 1$, it holds, with probability at least $1-\delta$*

$$H(N, \overline{\mathbf{w}}) \leq \frac{2 \, \log(2/\delta) \, S}{N} + \sqrt{\frac{2 \, \log(2/\delta) \, \mathrm{Var}_{\rho}}{N}}.$$

The above lemma says that the error $H_{\rho}(N, \overline{\mathbf{w}})$ goes to zero as N grows to infinity. Therefore the estimation error $\epsilon_{N,t}(\rho)$ is dominated by the "variance" term $\max_{1 \leq j \leq N} \beta^{\lambda}_{j}\left(1/T\right) \lambda^{-1/2}_{\min}(\mathbf{V}^{\lambda}_{j,T})$, associated with the worst past task. By relying on linear regression results (Lai & Wei, 1982) we have that $\lambda_{\min}(\mathbf{V}_j) \geq \log T$. Moreover, as $\lambda_{\min}(\mathbf{V}^{\lambda}_{j}) \geq \lambda + \lambda_{\min}(\mathbf{V}_j)$, we observe an increasing sensitivity of the incurred variance to the λ parameter for small value of T. Finally, according to our choice of $\lambda = 1/T\mathrm{Var}_{\widehat{\mathbf{h}}^{\lambda}}$, the suffered variance increases with the variance of our estimator. The latter in turns increases with the variance of the distribution ρ, which corresponds to the case in which Assumption 1 tends to be violated.

5. A High Bias Solution

In this section we will present an alternative estimator of the true mean $\overline{\mathbf{w}}$, which is inspired by the existing multi-task bandit literature (Gentile et al., 2014; 2017; Soare et al., 2014). This estimator exploits together all the samples associated to the past tasks $Z_1, \dots, Z_N$, with the aim of reducing the variance. This is unlike the previous estimator which separately considers the ridge-regression estimates $\widehat{\mathbf{w}}^{\lambda}_{1}, \dots, \widehat{\mathbf{w}}^{\lambda}_{N}$ in Equation 8. As we will see, this approach will reduce the variance but it will introduce an extra-bias. Before presenting this second approach we require some more notation. We let $\widetilde{\mathbf{V}}_{N,t} = \sum_{j=1}^{N} \mathbf{V}_{N,T} + \mathbf{V}_{N+1,t}$ the global design matrix containing the design matrices associated to past tasks $\mathbf{V}_{1,T}, \dots, \mathbf{V}_{N,T}$ and the current design matrix $\mathbf{V}_{N+1,t}$. Analogously $\widetilde{\mathbf{b}}_{N,t} = \sum_{j=1}^{N} \mathbf{b}_{j,T} + \mathbf{b}_{N+1,t}$ refers to global counterpart of $\mathbf{b}_{j,t}$. We denote

with $|A| = \sup\{\|A\mathbf{x}\| : \mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| = 1\}$ the norm of matrix A induced by the norm $\|\cdot\|$, which if no specified is the Euclidean norm. Finally, we denote with $\sigma_{\max}(\mathbf{A})$ the biggest singular value associated with matrix $\mathbf{A}$.

5.1. Global Ridge Regression

In order to reduce the variance, our second approach estimates, at each round t of the new sampled task $N+1$, the mean task $\overline{\mathbf{w}}$ as a *global ridge regression* computed over all the available samples as

$$\widehat{\mathbf{h}}^{\lambda}_{N,t} = \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t-1}\right)^{-1} \widetilde{\mathbf{b}}_{N,t-1}. \tag{9}$$

Our next result provides a bound on the transfer regret of this proposed strategy.

Theorem 3. *(Transfer Regret Bound). Let the assumptions of Lemma 2 hold and let $\widehat{\mathbf{h}}^{\lambda}_{N,t}$ be defined as in Equation* (9)*. Then, the following upper bound holds*

$$\mathcal{R}(T,\rho) \leq dC\sqrt{T\log\left(1+\frac{T^2L\left(\mathrm{Var}_{\overline{\mathbf{w}}}+\epsilon_{N,t}(\rho)\right)}{d}\right)}$$

where the mean estimation error can be bound as

$$\sqrt{\epsilon_{N,T}(\rho)} \leq \frac{S}{\lambda+\nu_{\min}} + 2(N+1)\max_{1\leq j\leq N+1}\widetilde{H}(N+1,\mathbf{w}_j) + R\sqrt{\frac{2}{\lambda+\nu_{\min}}\log\left(T\left(1+\frac{NTL^2}{\lambda d}\right)\right)} + H_{\rho}(N+1,\overline{\mathbf{w}})$$

and defined $\nu_{\min} = \lambda_{\min}(\widetilde{\mathbf{V}}_{N,T})$ and we introduced

$$\widetilde{H}(N,\mathbf{w}_j) = H_{\rho}(j,\mathbf{w}_j)\sigma_{\max}\left(\mathbf{V}_{j,T}\widetilde{\mathbf{V}}^{-1}_{N,T}\right)$$

which is a weighted form of the estimation error $H_{\rho}(j,\mathbf{w}_j)$ towards the current task vector $\mathbf{w}_j$, where the weights are defined in terms of tasks misalignment $\sigma_{\max}\left(\mathbf{V}_{j,T}\widetilde{\mathbf{V}}^{-1}_{N,T}\right)$.

The proof is presented in Section D of the appendix. The previous variance term $\frac{\beta^{\lambda}_j(1/T)}{\lambda_{\min}(\mathbf{V}^{\lambda}_{j,T})}$ has been now replaced by $\frac{\beta^{\lambda}(1/NT)}{\lambda+\nu_{\min}}$. It should be easy to observe that $\nu_{\min} \geq \frac{N}{d}\lambda_{\min}(\mathbf{V}_j)\ \ \forall j \in [N]$ which leads a reduction of factor d/N to the variance, which goes to zero as N goes to infinity. This gain does not come for free, in fact this approach introduces a potentially high bias: $2(N+1)\max_{j=1,\dots,N+1}\widetilde{H}(N+1,\mathbf{w}_j)$ which increases with the tasks misalignment $\sigma_{\max}\left(\mathbf{V}_{j,T}\widetilde{\mathbf{V}}^{-1}_{N,T}\right)$.

5.2. Tasks Misalignment

We now analyze the tasks misalignment factors appearing in Theorem 3, namely, the quanitities $\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right)$ and $\widetilde{H}(N,\mathbf{w}_j)$. For this purpose, we consider two opposite environments of tasks.

In the first case we assume that all the tasks parameters are equal to each other and far from the zero d-dimensional vector. This scenario, which corresponds to put all the mass of the task-distribution ρ on a single task parameter $\overline{\mathbf{w}}$, is clearly in agreement with Assumption 1. We expect this to be the most favorable scenario, since after completing a task, we face exactly the same task again and again. In this case, independently on the covariance matrices, whose construction also depends on the decision sets available in the different tasks, it is simple to observe that we are not suffering any bias, that is, $\widetilde{H}(N,\mathbf{w}_j) = 0 \quad$ for every $j \in [N]$ as all the task parameters are equal to each other.

The second environment is characterized by a task distribution ρ that is unform on finitely many orthogonal tasks. For instance, this is the scenario when ρ is uniform distributed over the standard basis vectors $\{(S,0,\dots,0),\dots,(0,\dots,0,S)\} \in \mathbb{R}^d$. Differently from the previous scenario, here after completing a task we will probably face an orthogonal task. It should be quite natural to see that this is the most unfavorable case and to expect to not have transfer learning between tasks. This is confirmed by the regret bound due to the misalignment expressed by the covariance matrices $\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right)$. Indeed, since we can have at most d misaligned arms, we have the following upper bound $\frac{d}{N}$ to the term $\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right)$. Based on these observations we can conclude that the bigger the cardinality of the set of basis induced by the distribution ρ, the larger the number of completed tasks required to have a proper transfer. We will now focus on an intermediate case satisfying Assumption 1. In order to control the term $\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right)$ and to give the possibility to generate aligned matrices when dealing with similar tasks, we introduce an additional mild assumption:

Assumption 2. *(Shared Induced Basis) The decision sets are shared among all the tasks and tasks sampled according to Assumption* 1 *induces that the covariance matrices generated by running the BIAS-OFUL algorithm (Algorithm 1) share the same basis:*

$$\mathbf{V}_i = \mathbf{P}\Sigma_i\mathbf{P}^*, \quad \forall i \in [N]. \tag{10}$$

This assumption is quite mild as it just states that similar tasks share the same pulled arms with no restrictions on the pulling frequency. This is the case when the decision set is fixed among different rounds and tasks, that is, $\mathcal{D}_{j,t} = \mathcal{D}\ \forall j \in [N]$ and $\forall t \in [T]$, and consists of d orthogonal arms. If Assumption 2 is satisfied, then we can obtain the following bound: $\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right) \leq 1$. Furthermore, if we denote by M the number of tasks necessary to achieve a stationary behavior of the BIAS-OFUL policy in terms of covariance matrices, then $\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right) \leq 1/(N-M)$.

5.3. Smallest Global Eigenvalue $\nu_{\min}$

It only remains to analyze the term $\nu_{\min}$. We observe that it satisfies the lower bound

$$\nu_{\min} = \lambda_{\min}\left(\sum_{j=1}^{N+1} \mathbf{V}_{j,T}\right) \geq \sum_{j=1}^{N+1} \lambda_{\min}(\mathbf{V}_{j,T}) \geq (N+1)\log T$$

where in the last step we have relied on linear regression result from (Lai & Wei, 1982) which shows that the condition $\mathcal{O}(\lambda_{\min}) = \log(\lambda_{\max})$ is required to guarantee asymptotic consistency, necessary to have sublinear anytime regret. Since $\min_{j\in[N]} \lambda_{\max}(\mathbf{V}_j) = \mathcal{O}(T)$, this condition implies that $\min_{j\in[N]} \lambda_{\min}(\mathbf{V}_j) \geq \log T$.

6. Experiments

In this section we test the real effectiveness of the proposed approaches. The theoretical results stated that the method presented in Section 4 does not introduce any bias but it may incur an additional variance according to the variance of the task-distribution Var_ρ. On the contrary, the solution proposed in Section 5 which massively uses all the observed samples together, reduces the variance (at least) by a factor d/N, at the price of an extra bias term.

As it was mentioned in Section 3, the parameter $\mathbf{w}^*$ associated to each single task is unknown, therefore we cannot compute the gap $\|\widehat{\mathbf{h}}^\lambda - \mathbf{w}^*\|_2$ defining the term $\beta_t^{\mathbf{h}}(1/T)$. The main issue is that according to Equation 4, in order to pick the next arm, it seems that the algorithm needs to compute its exact value. However, we can simply split the norm and rely on the assumption that $\|\mathbf{w}^*\| \leq S$, so to remove the dependency on $\mathbf{w}^*$. Indeed, it is important to emphasize that the real knowledge transfer happens in terms of $\mathbf{w}^{\mathbf{h}}$, see Equation 5. This can be noticed by observing that the gap $\left\|\widehat{\mathbf{h}}^\lambda - \mathbf{w}^*\right\|$ equally affects all the available arms.

6.1. Experimental Results

In all the presented experiments the policy OPT knows the parameter $\mathbf{w}_j$ associated to task j and picks the next arm as $\mathbf{x}_{j,t} = \arg\max_{\mathbf{x}\in D_{j,t}} \mathbf{x}^\top \mathbf{w}_j$. The policies AVG-OFUL and RR-OFUL implement Algorithms 1 and 2 and estimate $\widehat{\mathbf{h}}$ as per Equations 8 and Equation 9, respectively. The Oracle policy knows the mean task parameter $\overline{\mathbf{w}}$ and uses it as the bias $\mathbf{h}$ in BIAS-OFUL (Corollary 2 (ii)). Analogously, the ITL policy consists of BIAS-OFUL with bias set equal to $\mathbf{0}$, see Corollary 2 (i). The regularization hyper-parameter λ was selected over a logarithmic scale. We will start by considering a pair of synthetic experiments in which we show how the hyper-parameter λ affects the performance. We then present experiments on two real datasets. We will denote with K the size of the decision set $\mathcal{D}$.

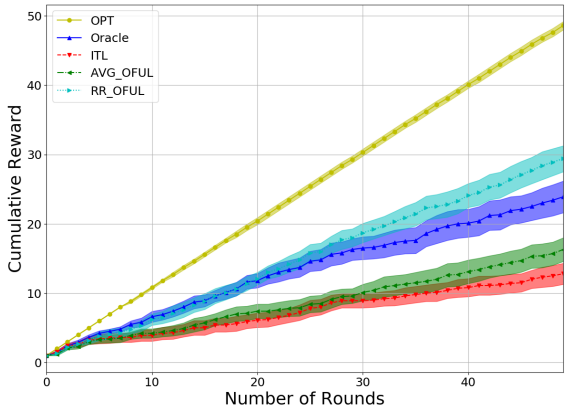

Figure 1. Cumulative reward measured after $N = 10$ tasks and averaged over 10 independent test tasks, with $\lambda = 1$.

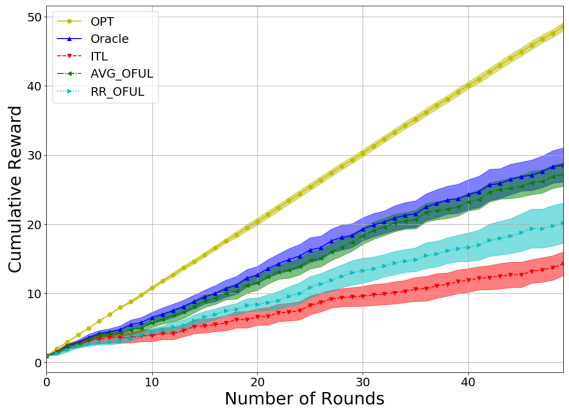

Figure 2. Cumulative reward measured after $N = 10$ tasks and averaged over 10 independent test tasks, with $\lambda = 100$.

Synthetic Data Similarly to what was done in (Denevi et al., 2019), we first generated an environment of tasks in which running the Oracle policy is expected to outperform the ITL approach. In agreement with Assumption 1, we sample the task vectors from a distribution characterized by a much smaller variance than its second moment. That is, each task parameter $\mathbf{w}_j$ is sampled from a Gaussian distribution with mean $\overline{\mathbf{w}}$ given by the vector in $\mathbb{R}^d$ with all components equal to 1 and $\mathrm{Var}_\rho = 1$. As far as the decision set concerns, we first generate a random square matrix $\mathbf{P}$ with size d and then compute its qr factorization $\mathbf{P} = \mathbf{QR}$, where $\mathbf{Q}$ is a matrix with orthonormal columns and $\mathbf{R}$ is an upper-triangular matrix. We then associate to each base arm the direction associated to a column of the matrix $\mathbf{Q}$. This will guarantee having arms that are almost orthogonal each other. Finally, at each round $t \in [T]$ the decision set $\mathcal{D}_t$ is initialized as a set of K random vector that are first shifted towards the respective arm base direction and then normalized. Notice that by following this generation mechanism we avoid any inductive bias between the task vectors and the arms ones, as they are actually

independent. Each task consists of $T = 50$ rounds, in which we have $K = 5$ arms of size $d = 20$. In order to generate the rewards, we first compute the inner product between the user (task) vector and the arm (input) vector, we shift the resulting output interval $[0, 1]$ and then add to a Gaussian noise $\mathcal{N}(0.5, 1)$, to compute the rewards. Finally, we assigned reward 1 to the arm having the maximum final reward, 0 to the others. In Figures 1 and 2, we report the results generated with $\lambda = 1$ and $\lambda = 100$, respectively. It is easy to observe that the stronger the regularization, the more the AVG-OFUL tends to the Oracle. Conversely, RR-OFUL get penalized with the increasing of λ, due to its bias.

LastFM Data The first dataset we considered is extracted from the music streaming service Last.fm (Cantador et al., 2011). It contains 1892 possible users and 17632 artists. This dataset contains information about the artists listened by a given user, and we used this information to define the payoff function. We first removed from the set of items those with less than 30 ratings and then we repeat the same procedure for the users. This operation yields an user rating matrix of size 741 x 538. Starting from this reduced matrix we derived the arms and the users vectors by computing an SVD decomposition where we kept only the first $d = 10$ features associated to the users and to the items. In order to consider tasks satisfying Assumption 1, we randomly pick an user and compute the set of its $N = 20$ most similar users according to the l2-distance between their vectors. Each task lasts $T = 5$ rounds and consists of $K = 5$ arms. At each round t, the decision set consists of one arm whose rating was at least equal to 4 and $K - 1$ arms whose ratings were at most equal to 3. The rewards were then generated analogously to the synthetic case. The Oracle policy knows $\overline{\mathbf{w}}$ which is computed as the average between the $N = 20$ considered user vectors. In Figure 3 (and Figure 4) we displayed the cumulative regret suffered with respect to the optimal policy, which during each task $j \in [N]$ knows the true user parameter $\mathbf{w}_j$. The vertical yellow lines indicate the end of each task. From the presented results we can observe that both the proposed policies AVG-OFUL and RR-OFUL outperform the ITL approach, while the Oracle policy is consistent with Corollary 2 and Assumption 1.

Movielens Here we consider the Movielens data (Harper & Konstan, 2015). It contains 1M anonymous ratings of approximately 3900 movies made by 6040 users. As before we first removed from the set of movies those with less than 500 ratings, and from the set of users those with less than 200 rated movies. This preprocessing procedure yields an user rating matrix of size 847 x 618. Unlike the Last.fm case, here adopting SVD to generate the arm/user vectors seems not appropriate. Indeed, by exploring the retrieved singular values, we could not find a subspace which provides a good approximation of the real ratings unless we keep all the latent features. Therefore, in order to find a set of similar

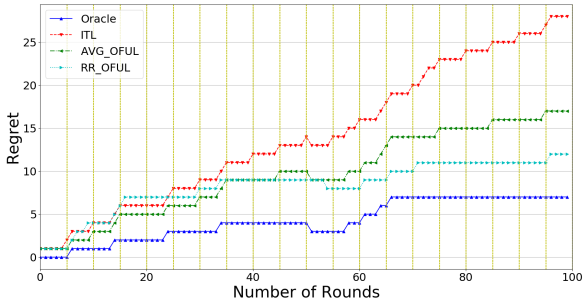

Figure 3. Empirical Transfer regret associated with Lastfm.

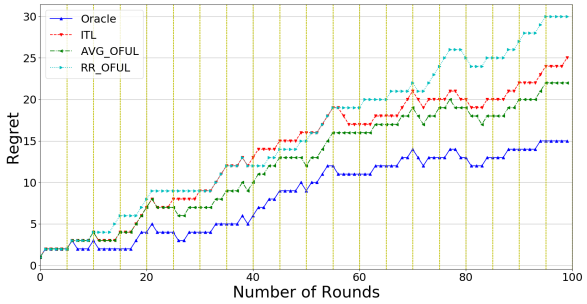

Figure 4. Empirical Transfer regret associated with Movielens.

users we observe better results by using the KMeans clustering algorithm over the user vectors. The results displayed in Figure 4 were generated by running KMeans with $C = 20$ clusters with user vectors of size $d = 10$. We then picked all the resulting clusters by filtering out the clusterings with a silhouette value lower than 0.15 and for each cluster of the clustering we have discarded those with less than 20 users. Furthermore, in order to let the tasks be simpler, we reduced the variance of the noisy components affecting rewards to 0.1. The difficulty in finding a valid set of similar tasks yields a high task misalignment, which is confirmed by the fact that the best performance occur for small value of λ. Indeed, Figure 4 considers $\lambda = 1$. Here the AVG-OFUL policy behaves almost equally to the ITL approach, conversely, the task misalignment caused bad performances to the RR-OFUL policy, confirming its higher sensitivity to task dissimilarity (see Theorem 3).

7. Conclusions and Future Work

In this work we studied a meta-learning framework with stochastic linear bandit tasks. We have first introduced a novel regularized version of OFUL, where the regularization depends on the Euclidean distance to a bias vector. We showed that setting appropriately the bias leads a substantial improvement compared to learning each task in isolation. This observation motivated two alternative approaches to estimate this bias: while the first one may suffer a potentially

high variance, the second might incur a strong bias.

In the future, it would be valuable to investigate the existence of unbiased estimators which do not suffer any variance. Furthermore, while in our analysis we set $\lambda = 1/T\mathrm{Var}_{\mathbf{h}}$, in the future it would be also interesting to learn its value as part of the learning problem. Experimentally, we observed that when Assumption 1 is satisfied, adopting the unbiased estimator yields better results than the second one, which is biased. One more direction of future research would be to extend other meta-learning approaches, such as those based on feature sharing, to the banding setting. Finally, a problem which remains to be studied is the combination of meta-learning with non-stochastic bandits.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 2312–2320, USA, 2011. Curran Associates Inc. ISBN 978-1-61839-599-3. URL http://dl.acm.org/citation.cfm?id=2986459.2986717.

Alquier, P., Mai, T. T., and Pontil, M. Regret Bounds for Lifelong Learning. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on rtificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 261–269, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL http://proceedings.mlr.press/v54/alquier17a.html.

Ando, R. K. and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853, December 2005. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1046920.1194905.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944941.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

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

Azar, M. G., Lazaric, A., and Brunskill, E. Sequential transfer in multi-armed bandit with finite set of models. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2220–2228, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999860.

Balcan, M.-F., Khodak, M., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pp. 424–433, 2019.

Baxter, J. A model of inductive bias learning. *J. Artif. Int. Res.*, 12(1):149–198, March 2000. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=1622248.1622254.

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

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

Calandriello, D., Lazaric, A., and Restelli, M. Sparse Multi-task Reinforcement Learning. In *NIPS - Advances in Neural Information Processing Systems 26*, Montreal, Canada, December 2014. URL https://hal.inria.fr/hal-01073513.

Cantador, I., Brusilovsky, P., and Kuflik, T. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. Linear algorithms for online multitask classification. *J. Mach. Learn. Res.*, 11:2901–2934, December 2010. ISSN 1532-4435.

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

Cesa-Bianchi, N. *Multi-armed Bandit Problem*, pp. 1356–1359. Springer New York, New York, NY, 2016. ISBN 978-1-4939-2864-4. doi: 10.1007/978-1-4939-2864-4_768. URL https://doi.org/10.1007/978-1-4939-2864-4_768.

Cesa-Bianchi, N., Gentile, C., and Zappella, G. A gang of bandits. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pp. 737–745, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999611.2999694.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL http://proceedings.mlr.press/v15/chu11a.html.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. Learning to learn around a common mean. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 10169–10179. Curran Associates, Inc., 2018a.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. Incremental learning-to-learn with statistical guarantees. *arXiv preprint arXiv:1803.08089*, 2018b.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. Learning-to-learn stochastic gradient descent with biased regularization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1566–1575, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/denevi19a.html.

Deshmukh, A. A., Dogan, U., and Scott, C. Multi-task learning for contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 4848–4856, 2017.

Gentile, C., Li, S., and Zappella, G. Online clustering of bandits. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–757–II–765. JMLR.org, 2014. URL http://dl.acm.org/citation.cfm?id=3044805.3044977.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. On context-dependent clustering of bandits. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1253–1262, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/gentile17a.html.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), December 2015. ISSN 2160-6455. doi: 10.1145/2827872.

Kuzborskij, I., Cella, L., and Cesa-Bianchi, N. Efficient linear bandits through matrix sketching. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 177–185. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/kuzborskij19a.html.

Lai, T. L. and Wei, C. Z. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10:154–166, 1982.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Liu, B., Wei, Y., Y., Z., Yan, Z., and Yang, Q. Transferable contextual bandit for cross-domain recommendation. In *In Thirty-Second AAAI Conference on Artificial Intelligence.*, 2018.

Maurer, A. and Pontil, M. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pp. 55–76, 2013.

Maurer, A., Pontil, M., and Romera-Paredes, B. Sparse coding for multitask and transfer learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. II–343–II–351. JMLR.org, 2013. URL http://dl.acm.org/citation.cfm?id=3042817.3042932.

Maurer, A., Pontil, M., and Romera-Paredes, B. The benefit of multitask representation learning. *J. Mach. Learn. Res.*, 17(1):2853–2884, January 2016. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2946645.3007034.

Pentina, A. and Urner, R. Lifelong learning with weighted majority votes. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3612–3620. Curran Associates, Inc., 2016.

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

Siegmund, D. Herbert robbins and sequential analysis. *Annals of statistics*, pp. 349–365, 2003.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Soare, M., Alsharif, O., Lazaric, A., and Pineau, J. Multi-task linear bandits. In *NIPS'14 Workshop on Transfer and Multi-task Learning*, 2014.

Villar, S. S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Zhang, J. and Bareinboim, E. Transfer learning in multi-armed bandits: A causal approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 1340–1346. AAAI Press, 2017. ISBN 978-0-9992411-0-3.