Università degli Studi di Milano

Ph.D. Program in Computer Science

(XXXIII Cycle)

Department of Computer Science

A thesis submitted for the degree of

Doctor of Philosophy

Efficiency and Realism in Stochastic Bandits

Author
Leonardo Cella

Supervisor
Nicolò Cesa-Bianchi

PhD Coordinator
Paolo Boldi

Academic Year 2019–2020

Contents

Abstract

This manuscript is dedicated to the analysis of the application of stochastic bandits to the recommender systems domain. Here a learning agent sequentially recommends one item from a catalog of available alternatives. Consequently, the environment returns a reward that is a noisy observation of the rating associated to the suggested item. The peculiarity of the bandit setting is that no information is given about not recommended products, and the collected rewards are the only information available to the learning agent. By relying on them the learner adapts his strategy towards reaching its learning objective, that is, maximizing the cumulative reward collected over all the interactions.

In this dissertation we cover the investigation of two main research directions: the development of efficient learning algorithms and the introduction of a more realistic learning setting. In addressing the former objective we propose two approaches to speedup the learning process. The first solution aims to reduce the computational costs associated to the learning procedure, while the second's goal is to boost the learning phase by relying on data corresponding to terminated recommendation sessions. Regarding the latter research line, we propose a novel setting representing use-cases that do not fit in the standard bandit model.

Estratto

Questo manoscritto è dedicato all'analisi dell'applicazione dei banditi stocastici ai sistemi di raccomandazione. Qui un apprenditore raccomanda in maniera sequenziale un oggetto estratto da un catalogo contenente varie alternative. Di conseguenza, l'ambiente restituisce una ricompensa definita come un'osservazione rumorosa della valutazione attribuita all'oggetto suggerito. La particolarità dell'ambiente bandits è che non si ha alcuna informazione relativa agli oggetti non raccomandati, e le ricompense raccolte sono le uniche informazioni disponibili all'apprenditore. Facendo affidamento su di esse, l'apprenditore adatta la sua strategia col fine di soddisfare la propria funzione obiettivo, ovvero la massimizzazione della ricompensa cumulativa raccolta su tutte le interazioni. In questa dissertazione trattiamo l'indagine di due direzioni di ricerca principali: lo sviluppo di algoritmi di apprendimento efficienti e l'introduzione di un ambiente di apprendimento più realistico. Nell'affrontare il primo obiettivo proponiamo due approcci per accelerare il processo di apprendimento. La prima soluzione mira a ridurre i costi computazionali associati alla procedura di apprendimento, mentre l'obiettivo della seconda proposta è quello di potenziare la fase di apprendimento facendo affidamento sui dati corrispondenti alle sessioni di raccomandazione precedenti. Infine, proponiamo e analizziamo un nuovo modello di apprendimento che rappresenta casi d'uso non inclusi nel modello classico dei banditi stocastici.

Acknowledgments

The manuscript you are reading is merely the result of the time and the energies spent during an unrepeatable three years long experience. Besides the achieved theoretical guarantees, the real outcome consists of the valuable connections built in and out of the research community. Focusing on the connections built out of the research field, a big thanks has to be given to my friends, from the oldest to the newest. I cannot even forget the dreamers of our social cooperative and the members of the *giovedei* football championship. Thanks you all guys. More importantly, thanks to my family for the continuous and persistent support, the right suggestions and the precious listening, thanks Elena, Valerio, Giovanna and Dina.

Within the research community a big thank to my advisor Nicolò Cesa-Bianchi. It is useless to say that without him this would not be possible. Thanks for having given me this valuable opportunity, the patience and the various teachings. Thanks also to the scientist I had the pleasure to work with: Massimiliano Pontil, Alessandro Lazaric, Ilja Kuzborskij, Claire Vernade and Giovanni Zappella.

Chapter 1

Introduction

Overview of the Studied Problems. This manuscript is dedicated to the theoretical analysis of efficient and effective solutions to the sequential recommendation problem. With the wide-spreading of the World-Wide-Web, people around the world are spending more and more time surfing the internet and looking for new and personalized content. This process produces large amount of data regarding the users (such as location, date, gender, age etc.) and at the same time provides feedback to the services providers regarding their recommendations. The high frequency of interactions has caused traditional batch recommendation systems to be unsuitable to this highly-interactive scenario. Indeed, the considered recommendation problem becomes how to leverage past feedback and user side information, to decide what to (quickly) recommend to get the customer satisfied. Formally, we represent the user as a stochastic agent who samples item ratings from a stochastic function which is parameterized by the recommended products and represents the customer preferences. The learner's objective is to maximize the cumulative reward collected during the whole user navigation by matching his interests. In order to achieve this goal at each new choice the learner can only rely on the information collected during the past interactions. This lack of information arises a trade-off, at each new choice the learner can either *exploit* its estimates and take what he thinks to be the best available recommendation, or *explore* potentially suboptimal products in order to acquire more information on the customer (environment).

Theoretical Settings. Multi-armed bandits are the most used way to formalize the described learning problem and to analyze the trade-off between following an optimality criteria and gaining more information on the environment. Throughot this dissertation we will focus on two well known bandit settings: the stochastic bandit and the stochastic linear bandit frameworks. In the former case we assume the available items (arms) are adimensional and their ratings (value) are sampled from unknown distributions which they are associated to. In the latter scenario, we assume the arms to be d-dimensional vectors representing their features and their value are given by a (unknown) linear regression of the arm features representing the user preference. In both cases, the learning objective is to learn the customer tastes while minimizing the regret incurred by making recommendation which do not match his interests. In Chapter 2 we formally define the stochastic bandit and the linear stochastic bandit settings along with their most popular learning algorithms.

Contributions In this dissertation we tackle three main complexities in the literature of the sequential recommendation problem that suggest that standard bandit algorithms are not an optimal solution to the considered problem.
Firstly, the biggest limitation affecting both standard (adimensional) and linear stochastic bandits is that ratings are assumed to be invariant to the provided recommendations. Indeed, if we consider the ideal case where the learning agent knows the user preference beforehand, than it would be natural to let it always recommend the most liked product. Even if this may be a winning strategy in the short term, it overlooks the user's preferences which in fact are not *static*. For instance, the user may get bored of receiving the same recommendation over and over. Starting from these motivations, in Chapter 3 we introduce a novel non-stationary bandit setting. There, the payoff of an arm is not static anymore and grows with the time since the arm was last played. We also show that the optimal policy do no stick recommending the best arm but rather varies its recommendations following unknown combinatorial patterns.
Moving to the stochastic linear bandit setting, firstly we observe that while their most popular algorithm exhibit good theoretical and empirical performances, they require potentially high time to update their model after each interaction. In Chapter 4 our objective is to propose an alternative learning scheme that significantly reduces the update time while preserving good quality recommendations. The second fragility suffered by existing linear bandit algorithms is the curse-of-dimensionality. With the increasing of the number of arms features (d), existing models may require many *explorative* interactions that can let the user quit his navigation session. In Chapter 5 we propose an approach to speedup the learning process that takes advantage on terminated navigation sessions to acquire information about the user and then reducing the number of *explorative* interactions.

Outline

The remainder of this manuscript is organized as follows:

- In Chapter 2 we introduce some preliminary notations, definitions and the existing core results in the stochastic bandits literature.
- In Chapter 3 we formalize a novel nonstationary stochastic bandit setting and propose a suitable learning strategy. This work has been published as Cella and Cesa-Bianchi [2020] and benefit from the supervision of Nicolò Cesa-Bianchi.
- In Chapter 4 we investigate the adoption of a sketching technique to speedup the update time of linear bandit policies. This chapter refers to a joint work with Ilja Kuzborskij and Nicolò Cesa-Bianchi which has been published as Kuzborskij et al. [2019].
- Finally, in Chapter 5 we investigate two transfer learning strategies applied to linear bandit tasks. This problem has been published as Cella et al. [2020] and was a joint work with Alessandro Lazaric and Massimiliano Pontil.
- The manuscript ends with Chapter 6. There, we provide closing remarks and propose candidate future research directions.

Chapter 2

Notation and Preliminaries

In this chapter we introduce the adopted notation and the state of the art approaches for both the stochastic and the linear stochastic bandits. We present here the main concepts and proof techniques that we will recall in the next chapters. We begin by formally defining the stochastic bandit setting and presenting the upper confidence bound principle that was formally analyzed in Auer et al. [2002]. We will then formalize the linear bandit setting and present two state of the art learning algorithms: OFUL [Abbasi-Yadkori et al., 2011] and Thompson Sampling [Abeille and Lazaric, 2017].

2.1 Stochastic Bandits

Multi-armed bandits is a very powerful interactive framework for algorithms that make decisions over time and under uncertainty [Lattimore and Szepesvári, 2018, Auer et al., 2002, Cesa-Bianchi and Lugosi, 2006, Siegmund, 2003, Robbins, 1952, Cesa-Bianchi, 2016, Bubeck et al., 2012]. The name comes from the slot machines that can be found in a casino. There we have many slot machines (a.k.a. one-armed bandits) and the gambler would like to play the most profitable one, that is the one yielding the higher monetary reward if played.

2.1.1 Setting Formalization

We start by considering the stochastic bandit setting that was first formalized in Robbins [1952]. Here, an algorithms sequentially interacts with an environemnt for T rounds. In each round $t \in [T] = 1, \ldots, T$, the environment provides a decision set $\mathcal{K} = \{x_1, \ldots, x_k\}$ consisting of K possible actions (a.k.a. arms) to choose from. The learner picks one of the available arm $X_t \in \mathcal{K}$ and then the environment samples a reward $Y_t \in \mathbb{R}$ from a fixed but unknown distribution associated with the pulled arm X_t with mean μ_{X_t}. The generated reward will be the only feedback provided to the learning agent at round t and we call this a *bandit feedback*. As in the casino the objective of the gambler is to collect as much reward as possible thorugh its sequence of interactions, here the objective of the learner is to maximize the cumulative reward collected during the T rounds. In achieving this objective the learner can only rely on the collected feedbacks. This arises a natural trade-off:

Algorithm 1 The UCB1 Algorithm.

```
Input: number of rounds T, decision set $\mathcal{K}$
PULL each arm once
for $t = K + 1, 2, \dots, T$ do
    UPDATE arm indices according to Equation 2.2
    SELECT arm $x_t = \arg\max_{i \in \mathcal{K}} ucb(i)$
    OBSERVE reward $Y_t$
end for
```

from one side he needs to *explore* the environment in order to acquire information on the available arms to identify the most profitable one, from the other it has to *exploit* the arm which seems to be better according to the acquired information to collect more rewards. The learning algorithms that we will present in this dissertation find smart ways to deal with this exploration/exploitation trade-off. The reward maximization objective of the learner can be equivalently stated in terms of regret incurred with respect to the optimal strategy which collects the biggest cumulative reward possible. Formally, if we denote with $i^* \in \mathcal{K}$ the arm whose distribution has the largest mean $\mu^* = \max_{i=1,\dots,k} \mu_i$ then we can define the incurred expected regret as

$$R(T) = \mathbb{E}\Big[T\mu^* - \sum_{t=1}^{T} Y_t\Big] = T\mu^* - \sum_{t=1}^{T} \mu_{X_t}, \tag{2.1}$$

where the expectation is with respect to the randomness in the reward generation.

2.1.2 Algorithms and Results

Bandits found their first application in clinincal trials (Thompson [1933], Gittins [1979], Villar et al. [2015]). Here the goal is to identify the most effective drug out of a finite number of alternatives with unknown effects. In order to achieve this objective, the learning agent can sequentially select one of the drugs and administers it to the current patient. Once it has observed the effects induced by the drug, the learner updates its estimate hoping to do better with the next patient. In the last decades bandit algorithm expanded their applications to web-oriented scenarios like web search Radlinski et al. [2008], news recommendation Li et al. [2010], music playlist construction Cella and Cesa-Bianchi [2020] and recommendation systems (Li et al. [2011], Bresler et al. [2014], Gentile et al. [2014], Bresler et al. [2016], Gentile et al. [2017]). All the proposed models are inspired by the breakthrough that have been firstly introduced in Auer et al. [2002]. There, relying on the Chernoff-Hoeffding inequality, authors developed a finite-time analysis of the UCB1 algorithm (see Algorithm 1) which maintains arms indices defined by a high-probability upper bound on the expected value associated to the arm. More in detail, at each round $t \in [T]$ UCB1 associates to each arm $i \in \mathcal{K}$ an index $ucb(i)$ defined as

$$ucb(i) = \widehat{\mu}_i\big(T(i,t)\big) + \sqrt{\frac{2\log t}{T(i,t)}} \tag{2.2}$$

where $T(i,t)$ denotes the number of times arm $i \in \mathcal{K}$ has been pulled up to round t and $\widehat{\mu}_i\big(T(i,t)\big)$ is the empirical mean computed over the $T(i,t)$ rewards associated to arm i. These indices represent the largest statistically plausible true mean values associated to the arms based on the available observations. As shown in Algorithm 1 UCB1 follows the *Optimism in the Face of Uncertainty* (OFU) principle and pulls the arm with highest index.

Theorem 1. *For all $K > 1$, the expected regret of the UCB1 strategy after any number T of rounds is upper bounded by*

$$\left[8 \sum_{i \in \mathcal{K} \setminus \{i^*\}} \frac{\log T}{\Delta_i}\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{i \in \mathcal{K}} \Delta_i\right) \tag{2.3}$$

where Δ_i represents the suboptimality mean gap $\mu^* - \mu_i$.

Proof Sketch First, we can notice that the regret can be equivalently rewritten as

$$R(T) = \sum_{i \in \mathcal{K}} \Delta_i \mathbb{E}\big[T(i,T)\big]. \tag{2.4}$$

Therefore, our objective is to minimize the number of pulls associated to the set of suboptimal arms $\{i \in \mathcal{K} : \mu_i < \mu^*\}$. According to the UCB1 algorithm, the following holds:

$$\begin{aligned}
T(i,T) &= 1 + \sum_{t=K+1}^{T} \mathbb{I}\{X_t = i\} \\
&\leq l + \sum_{t=K+1}^{T} \mathbb{I}\{X_t = i, T(i,t-1) \geq l\} \\
&\leq l + \sum_{t=K+1}^{T} \mathbb{I}\{\widehat{\mu}_i\big(T(i,t-1)\big) + c_{t-1,T(i,t-1)} \geq \widehat{\mu}_{i^*}\big(T(i^*,t-1)\big) + \\
&\quad + c_{t-1,T(i^*,t-1)}, T(i,t-1) \geq l\} \\
&\leq l + \sum_{t=K+1}^{T} \mathbb{I}\left\{\min_{0<s_i<t} \widehat{\mu}_i(s_i) + c_{t-1,s_i} \geq \max_{l \leq s \leq t} \widehat{\mu}_{i^*}(s) + c_{t-1,s}\right\} \\
&\leq l + \sum_{t=K+1}^{T} \sum_{s=1}^{t-1} \sum_{s_i=l}^{t-1} \mathbb{I}\{\widehat{\mu}_i(s_i) + c_{t,s_i} \geq \widehat{\mu}_{i^*}(s) + c_{t,s}\}
\end{aligned}$$

where l is an arbitrary positive integer and $c_{t,s} = \sqrt{\frac{2\log t}{s}}$. In order to control the argument $\widehat{\mu}_i(s_i) + c_{t-1,s_i} \geq \widehat{\mu}_{i^*}(s) + c_{t-1,s}$ we can add and subtract the optimal true mean μ^*, the considered suboptimal true mean μ_i and the confidence bound associated to the suboptimal arm c_{t,s_i}. We can then observe that when the previous argument is satisfied, than at least one of the following conditions must hold:

$$\widehat{\mu}_i(s_i) \geq \mu_i + c_{t,s_i} \tag{2.5}$$

$$\widehat{\mu}_{i^*}(s) \leq \mu^* - c_{t,s} \tag{2.6}$$

$$\mu^* - \mu_i \leq 2c_{t,s_i} \tag{2.7}$$

Intuitively, this means that the suboptimal pull is justified either by an overestimation of the mean associated to the suboptimal arm (Equation 2.5), or an understimation of the mean associated to the optimal arm (Equation 2.6) or a still too large confidence bound (Equation 2.7) compared to the suboptimality gap Δ_i. The first two cases can be directly controlled by applying the Chernoff-Hoeffding inequality (Proposition 4 in the Appendix material) and cause the $(1 + \frac{\pi^2}{3})$ term. Finally, the last case is false up to $l = \lceil \frac{8 \log T}{\Delta_i^2} \rceil$ rounds and results in the $\left[8 \sum_{i \in \mathcal{K} \setminus \{i^*\}} \frac{\log T}{\Delta_i} \right]$ term.

2.2 Linear Stochastic Bandits

An interesting alternative to the stochastic bandit problem is given by the linear bandits which considers actions as a subset of $\mathbb{R}^d$. Here, the observed reward has an expected value which is an unknown linear function of the action. It is simple to observe that this setting generalizes the previous one by taking actions as the standard orthonormal basis. Before presenting the learning setting and the algorithms we need to introduce some additional notation. Let $\mathcal{B}(\mathbf{z}, r) \subset \mathbb{R}^d$ be the Euclidean ball of center $\mathbf{z}$ and radius $r > 0$ and let $\mathcal{B}(r) = \mathcal{B}(\mathbf{0}, r)$. Given a positive definite $d \times d$ matrix $\mathbf{A}$, we define the inner product $\langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{A}} = \mathbf{x}^\top \mathbf{A} \mathbf{z}$ and the induced norm $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A} \mathbf{x}}$, for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, if not specified $\|\cdot\|$ is the Euclidean norm. Throughout the dissertation, we write $f \overset{\widetilde{\mathcal{O}}}{=} g$ to denote $f = \widetilde{\mathcal{O}}(g)$.

2.2.1 Setting Formalization

We describe the linear bandit protocol in Algorithm 2. Here, at each round $t \in [T]$, the learner has to select one arm $\mathbf{x}_t$ from a set of alternatives $D_t \subset \mathbb{R}^d$. The observed reward corresponding to the taken arm has expected value satisfying

$$Y_t = \mathbf{x}_t^\top \mathbf{w}^\star + \eta_t \tag{2.8}$$

where $\mathbf{w}^\star \in \mathbb{R}^d$ is an unknown parameter and η_t is a random noise satisfying some constraints that we will specified soon. This learning problem is particularly relevant in cases where the number of arms is very large. The main intuition is that, given the assumed reward structure, each pull gives information on the unknown parameter $\mathbf{w}^\star$ which indirectly, gives information about the value of not pulled arms. It is than natural to see that here the objective is to estimate the d-dimensional feature vector $\mathbf{w}^\star$. Similarly to what we have done for the simpler stochastic case, we can introduce the objective function in terms of regret minimization. Thanks to the knowledge of the true parameter $\mathbf{w}^\star$, 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 then to maximize the cumulative reward, or equivalently, to minimize the *pseudo-regret*

$$R(T, \mathbf{w}^\star) = \sum_{t=1}^{T} (\mathbf{x}_t^* - \mathbf{x}_t)^\top \mathbf{w}^\star. \tag{2.9}$$

Algorithm 2 (Linear Bandit)

```
for $t = 1, 2, \dots$ do
    GET decision set $D_t \subset \mathbb{R}^d$
    Use current policy to SELECT action $\mathbf{x}_t \in D_t$
    OBSERVE reward $Y_t \in \mathbb{R}$
    UPDATE the current policy using pair $(\mathbf{x}_t, Y_t)$
end for
```

We introduce some standard assumptions for the linear stochastic bandit setting. At any round $t = 1, 2, \dots, T$ the decision set $D_t \subset \mathbb{R}^d$ is finite and such that $\|\mathbf{x}\|_2 \le L$ for all $\mathbf{x} \in D_t$ and for all $t \ge 1$. The noise sequence $\eta_1, \eta_2 \dots, \eta_T$ is conditionally R-subgaussian for some fixed constant $R \ge 0$. Formally, for all $t \ge 1$ and all $\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda \eta_t} \,\middle|\, \eta_1, \dots, \eta_{t-1}\right] \le \exp\left(\lambda^2 R^2/2\right)$. Note that this implies $\mathbb{E}[\eta_t \mid \eta_1, \dots, \eta_{t-1}] = 0$ and $\mathrm{Var}[\eta_t \mid \eta_1, \dots, \eta_{t-1}] \le R^2$. Finally, we assume that a known upper bound S on $\|\mathbf{w}^\star\|$ is available.

2.2.2 Algorithms and Results

Both OFUL and Linear TS operate by computing a confidence ellipsoid to which $\mathbf{w}^\star$ belongs with high probability. Let $\mathbf{X}_t = [\mathbf{x}_1, \dots, \mathbf{x}_t]^\top$ be the $t \times d$ matrix of all actions selected up to round t by an arbitrary policy for linear contextual bandits. For $\lambda > 0$, define the regularized correlation matrix of actions $\mathbf{V}_t^\lambda$ and the regularized least squares (RLS) estimate $\widehat{\mathbf{w}}_t$ as

$$\mathbf{V}_t^\lambda = \mathbf{X}_t^\top \mathbf{X}_t + \lambda \mathbf{I} \quad \text{and} \quad \widehat{\mathbf{w}}_t = \left(\mathbf{V}_t^\lambda\right)^{-1} \sum_{s=1}^{t} \mathbf{x}_s Y_s \,. \tag{2.10}$$

The following theorem [Abbasi-Yadkori et al., 2011, Theorem 2] bounds in probability the distance, in terms of the norm $\|\cdot\|_{\mathbf{V}_t^\lambda}$, between the optimal parameter $\mathbf{w}^\star$ and the RLS estimate $\widehat{\mathbf{w}}_t$.

Theorem 2 (Confidence Ellipsoid)**.** *Let $\widehat{\mathbf{w}}_t$ be the RLS estimate constructed by an arbitrary policy for linear contextual bandits after t rounds of play. For any $\delta \in (0,1)$, the optimal parameter $\mathbf{w}^\star$ belongs to the set $C_t \equiv \left\{\mathbf{w} \in \mathbb{R}^d \;:\; \|\mathbf{w} - \widehat{\mathbf{w}}_t\|_{\mathbf{V}_t^\lambda} \le \beta_t(\delta)\right\}$ with probability at least $1 - \delta$, where*

$$\beta_t(\delta) = R\sqrt{d \ln\left(1 + \frac{tL^2}{\lambda d}\right) + 2\ln\left(\frac{1}{\delta}\right)} + S\sqrt{\lambda} \,. \tag{2.11}$$

OFUL. The actions selected by OFUL are solutions to the following constrained optimization problem

$$\begin{aligned} &\mathbf{x}_t = \arg\max_{\mathbf{x} \in D_t} \max_{\mathbf{w} \in \mathbb{R}^d} \mathbf{x}^\top \mathbf{w} \\ &\text{such that} \quad \|\mathbf{w} - \widehat{\mathbf{w}}_{t-1}\|_{\mathbf{V}_{t-1}^\lambda} \le \beta_{t-1}(\delta) \,. \end{aligned} \tag{2.12}$$

Algorithm 3 (OFUL)

```
Input: $\delta, \lambda > 0$
$\widehat{\mathbf{w}}_0 = \mathbf{0}$, $\left(\mathbf{V}_0^\lambda\right)^{-1} = \frac{1}{\lambda}\mathbf{I}$.
for $t = 1, 2, \dots$ do
    GET decision set $D_t$
    SELECT $\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x} \in D_t} \left\{ \widehat{\mathbf{w}}_{t-1}^\top \mathbf{x} + \beta_{t-1}(\delta) \left\|\mathbf{x}\right\|_{\left(\mathbf{V}_t^\lambda\right)^{-1}} \right\}$
    OBSERVE reward $Y_t$
    UPDATE $\left(\mathbf{V}_t^\lambda\right)^{-1}$ and $\widehat{\mathbf{w}}_t$ according to (2.10)
end for
```

OFUL can be formulated as Algorithm 3. Note that $\mathbf{x}_t$ maximizes the sum of an *exploitation* term consisting on the expected reward estimate $\widehat{\mathbf{w}}_{t-1}^\top \mathbf{x}$ plus an *exploration* term $\beta_{t-1}(\delta) \left\|\mathbf{x}\right\|_{\left(\mathbf{V}_t^\lambda\right)^{-1}}$ that provides an upper confidence bound for the RLS estimate in the direction of $\mathbf{x}$. More in detail, the more the arm $\mathbf{x} \in \mathbb{R}^d$ is correlated to the design matrix $\mathbf{V}$, the more its norm weighted by the inverse of the same matrix will be small. This means that the more an arm has been pulled during past rounds, the more accurate will be our estimates on it. The next theorem states an upper bound on the regret incurred by the OFUL algorithm (see Theorem 3 of Abbasi-Yadkori et al. [2011]).

Theorem 3. *Assume that for all $t \in [T]$ and all $\mathbf{x} \in D_t, \mathbf{x}^\top \mathbf{w}^\star \in [-1, 1]$. Then, with probability at least $1 - \delta$, the regret incurred by OFUL satisfies:*

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

Linear TS. The linear Thompson Sampling algorithm of Agrawal and Goyal [2013] is Bayesian in nature: the selected actions and the observed rewards are used to update a Gaussian prior over the parameter space. Each action $\mathbf{x}_t$ is selected by maximizing $\mathbf{x}^\top \widehat{\mathbf{w}}_t^{\text{TS}}$ over $\mathbf{x} \in D_t$, where $\widehat{\mathbf{w}}_t^{\text{TS}}$ is a random vector drawn from the posterior. As shown by Abeille and Lazaric [2017], linear TS can be equivalently defined as a randomized algorithm based on the RLS estimate (see Algorithm 4). The random vectors $\mathbf{Z}_t$ are drawn i.i.d. from a suitable multivariate distribution $\mathcal{D}^{\text{TS}}$ that need not be related to the posterior. In order to prove regret bounds, it is sufficient that the law of $\mathbf{Z}_t$ satisfies certain properties.

Definition 1 (TS-sampling distribution)**.** *A multivariate distribution $\mathcal{D}^{\text{TS}}$ on $\mathbb{R}^d$, absolutely continuous w.r.t. the Lebesgue measure, is TS-sampling if it satisfies the following two properties:*

- *(Anti-concentration) There exists $p > 0$ such that for any $\mathbf{u}$ with $\|\mathbf{u}\| = 1$, $\mathbb{P}\big(\mathbf{u}^\top \mathbf{Z} \geq 1\big) \geq p$.*
- *(Concentration) There exist $c, c' > 0$ such that for all $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\|\mathbf{Z}\| \leq \sqrt{cd \ln\left(\frac{c'd}{\delta}\right)}\right) \geq 1 - \delta\,.$$

Algorithm 4 Linear TS

```
Input: $\delta, \lambda > 0, m \in \{1, \dots, d-1\}, \mathcal{D}^{\text{TS}}$ (sampling distribution)
$\widehat{\mathbf{w}}_0 = \mathbf{0}, \left(\mathbf{V}_0^\lambda\right)^{-1} = \frac{1}{\lambda}\mathbf{I}_{d \times d}, \delta' = \delta/(4T)$
for $t = 1, 2, \dots$ do
    GET decision set $D_t$
    SAMPLE $\mathbf{Z}_t \sim \mathcal{D}^{\text{TS}}$
    SELECT $\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x} \in D_t} \mathbf{x}^\top \left(\widehat{\mathbf{w}}_{t-1} + \widetilde{\beta}_t(\delta') \left(\mathbf{V}_t^\lambda\right)^{-\frac{1}{2}} \mathbf{Z}_t\right)$
    OBSERVE reward $Y_t$
    UPDATE $\left(\mathbf{V}_t^\lambda\right)^{-\frac{1}{2}}$ and $\widehat{\mathbf{w}}_t$ using Equation (2.10)
end for
```

Similarly to OFUL, linear TS uses the notion of confidence ellipsoid. However, due to the properties of the sampling distribution $\mathcal{D}^{\text{TS}}$, the ellipsoid used by linear TS is larger by a factor of order $\sqrt{d}$ than the ellipsoid used by OFUL. This causes an extra factor of $\sqrt{d}$ in the regret bound, whose result is formally presented in the next theorem.

Theorem 4. *Under the same assumptions holding for Theorem 3, with probability at least* $1 - \delta$*, the regret of linear TS satisfies*

$$R(T, \mathbf{w}^\star) \leq \left[\beta_T(\delta') + \gamma_T(\delta')(1 + \frac{4}{p})\right] \sqrt{2Td \log\left(1 + \frac{T}{\lambda}\right)} + \frac{4\gamma_T(\delta')}{p} \sqrt{\frac{8T}{\lambda} \log \frac{4}{\delta}} \tag{2.13}$$

where $\gamma(\delta) = \beta(\delta)\sqrt{cd \log(c'c/\delta)}$ *and* $\delta' = \frac{\delta}{4T}$.

Note that both OFUL and linear TS need to maintain $\left(\mathbf{V}_t^\lambda\right)^{-1}$ $\left(\text{or } \left(\mathbf{V}_t^\lambda\right)^{-\frac{1}{2}}\right)$, which requires time $\mathcal{O}\left(d^2\right)$ to update.

Chapter 3

Stochastic Bandits with Delay-Dependent Payoffs

We dedicate this chapter to the analysis of a non-stationary stochastic bandit problem. The introduced setting is motivated by recommendation problems in music streaming platforms and in education. Here, the expected reward of an arm depends on the number of rounds that have passed since the arm was last pulled. We begin by proving that finding an optimal policy is NP-hard even when all model parameters are known. Then, we introduce a class of ranking policies provably approximating, to within a constant factor, the expected reward of the optimal policy. We show an algorithm whose regret with respect to the best ranking policy is bounded by $\widetilde{\mathcal{O}}\big(\sqrt{kT}\big)$, where k is the number of arms and T is time. Our algorithm uses only $\mathcal{O}\big(k \ln\ln T\big)$ switches, which helps when switching between policies is costly. As constructing the class of learning policies requires ordering the arms according to their expectations, we also bound the number of pulls required to do so. Finally, we run experiments to compare our algorithm against UCB on different problem instances.

3.1 Introduction

As introduced in Section 2.1, in the simplest stochastic bandit framework Lai and Robbins [1985] rewards are realizations of i.i.d. draws from fixed and unknown distributions associated to each arm. In that setting the optimal policy is to consistently recommend the arm with the highest reward expectation. On the other hand, in scenarios like song recommendation, users may grow tired of listening to the same music genre over and over. Here, playlists typically consists of different music genres interleaved according to certain patterns. This is naturally formalized as a nonstationary bandit setting, where the payoff of an arm grows with the time since the arm was last played. In this case policies consistently recommending the same arm are seldom optimal. E-learning applications, where arms corresponds to questions that students have to answer, are other natural examples of the same phenomenon, as asking again immediately the same question that the student has just answered is not very effective. In the remaining of the chapter we introduce a simple nonstationary stochastic bandit model, B2DEP, in which the expected reward $\mu_i(\tau)$ of an arm i is a bounded nondecreasing function of the number τ of rounds that have passed since the arm was last selected by the policy. More specifically, we assume each arm i has an unknown baseline payoff expectation μ_i (equal to the expected payoff when the arm is pulled for the first time) and an unknown delay parameter $d_i > 0$. If the arm was pulled recently (that is, $1 \leq \tau \leq d_i$), then the expected payoff may be smaller that its baseline value: $\mu_i(\tau) \leq \mu_i$. Vice versa, if $\tau > d_i$, then $\mu_i(\tau)$ is guaranteed to match the baseline value μ_i. In the song recommendation example, the delays d_i model the extent to which listening to a song of genre i affects how much a user is willing to listen to more songs of that same genre.

Since τ can be viewed as a notion of state for arm i, our model can be compared to nonstationary models, such as rested bandits Gittins [1979] and restless bandits Whittle [1988] —see also Tekin and Liu [2012]. In restless bandits the reward distribution of an arm changes irrespective of the policy being used, whereas in rested bandits the distribution changes only when the arm is selected by the policy. Our setting is neither rested nor restless, as our reward distributions change differently according to whether the arm is selected by the policy or not.

Optimal strategies for restless bandits are notoriously hard to compute, or even approximate —see, e.g., Guha et al. [2010]. In Section 3.4 we make a reduction to the Periodic Maintenance Scheduling Problem Bar-Noy et al. [2002] to prove that the optimization problem of finding an optimal periodic policy in our setting is NP-Hard. In order to circumvent the hardness of computing the optimal periodic policy, in Section 3.5 we identify a simple class of periodic policies that are efficiently learnable, and whose expected reward is provably to within a constant factor of that of the optimal policy. Our approximating class is pretty natural: it contains all *ranking policies* that cycle over the r best arms (where r is the parameter to optimize) according to the unknown ordering based on the arms' baseline payoff expectations. Note that a top-r ranking policy pulls each of the first r arms with a delay exactly equal to r. As it turns out, learning the best ranking policy can be formulated in terms of minimizing the standard notion of regret. This is unlike the problem of learning the best periodic policy, which instead requires minimizing the harder notion of policy regret Arora et al. [2012].

Consider the task of learning the best ranking policy. In our music streaming example, a ranking policy is a playlist for the user. As changing the playlist streamed to the user may be costly in practice, we also introduce a *switching cost* for selecting a different ranking policy. Controlling the number of switches could also have a good effect in our nonstationary setting, when the expected reward of a ranking policy may depend on which other ranking polices were played earlier. The learning agent should ensure that a ranking policy is played many times consecutively (i.e., infrequent switches), so that estimates are calibrated (i.e., computed in the same context of past plays).

A standard bandit strategy like UCB Auer et al. [2002], which guarantees a regret of $\mathcal{O}\big(\sqrt{kT \ln T}\big)$ irrespective of the size of the suboptimality gaps between the expected reward of the optimal ranking policy and that of the other policies, performs a number of switches growing with the squared inverse of these gaps. In Section 3.6 we show how to learn the best ranking policy using a simple variant of a learning algorithm based on action elimination proposed in Cesa-Bianchi et al. [2013a]. Similarly to UCB, this algorithm has a distribution-free regret bounded by $\sqrt{kT}$. However, a bound $\mathcal{O}\big(k \ln \ln T\big)$ on the number of switches is also guaranteed irrespective of the size of the gaps.

In Section 3.7 we turn to the problem of constructing the class of ranking policies, which amounts to learning the ordering of the arms according to their baseline payoff expectations $\mu_1, \ldots, \mu_k$. Assuming $\mu_1 > \cdots > \mu_k$, this can be reduced to the problem of learning the ordering of reward expectations in a standard stochastic bandit with i.i.d. rewards. We show that this is possible with a number of pulls bounded by $\sum_i 1/\Delta_i^2$ (ignoring logarithmic factors), where Δ_i is the smallest gap between $\mu_{i-1} - \mu_i$ and $\mu_i - \mu_{i-1}$. Note that this bound is not significantly improvable, because $1\big/\Delta_i^2$ samples of each arm i are needed to verify that $\mu_{i-1} < \mu_i < \mu_{i+1}$.

Finally, in Section 3.8 we describe experiments comparing our low-switch algorithm against UCB in both large-gap and in small-gap settings.

3.2 Related works

Our setting is a variant of the model introduced by Kleinberg and Immorlica [2018]. In that work, $\mu_i(\tau)$ are concave, nondecreasing functions satisfying $\mu_i(\tau) \leq \mu_i(\tau - 1) + 1$. Note that this setting and ours are incomparable. Indeed, unlike Kleinberg and Immorlica [2018] we assume a specific parametric form for the functions $\mu_i(\cdot)$, which are nondecreasing and bounded by 1. On the other hand, we do not assume concavity, which plays a key role in their analysis.

Pike-Burke and Grunewalder [2018] consider a setting in which the expected reward functions $\mu_i(\cdot)$ are sampled from a Gaussian Process with known kernel. The main result is a bound of order $\sqrt{kT}$ on the Bayesian d-step lookahead regret, where d is a user-defined parameter. This notion of regret is defined by dividing time in length-d blocks, and then summing the regret in each block against the greedy algorithm optimizing the next d pulls given the agent's current configuration of delays (i.e., how long ago each arm was last pulled). Similarly to Pike-Burke and Grunewalder [2018], we also compete against a greedy block strategy. However, in our case the block length is unknown, and the greedy strategy is not defined in terms of the agent's delay configuration.

A special case of our model is investigated in the very recent work by Basu et al. [2019]. Unlike B2DEP, they assume $\mu_i(\tau) = 0$ for all $\tau \leq d_i$ and complete knowledge of the delays d_i. In fact, they even assume that every arm i cannot be selected in the next d_i time steps after a pull. Their main result is a regret bound for a variant of UCB competing against the greedy policy. They also show NP-hardness of finding the optimal policy through a reduction similar to ours. It is not clear how their learning approach could be extended to prove results in our more general setting, where $\mu_i(\tau)$ could be positive even when $\tau \leq d_i$ and the delays d_i are unknown.

A different approach to nonstationary bandits in recommender systems considers expected reward functions that depend on the number of times the arm was played so far (Levine et al. [2017], Cortes et al. [2017], Bouneffouf and Féraud [2016], Heidari et al. [2016], Seznec et al. [2019], Warlop et al. [2018]). These cases correspond to a rested bandit model, where each arm's expected reward can only change when the arm is played.

The fact that we learn ranking strategies is reminiscent of stochastic combinatorial semi-bandits Kveton et al. [2015b], where the number of arms in the schedule is a parameter of the learning problem. In particular, similarly to (Radlinski et al. [2008], Kveton et al. [2015a], Katariya et al. [2016]) our strategies learn rankings of the actions, but unlike those approaches in our case the optimal number of elements in the ranking must be learned too.

3.3 The B2DEP setting

As presented in Section 2.1, in the classical stochastic multi-armed bandit model, at each round $t = 1, 2, \ldots$ the agent pulls an arm from $\mathcal{K} = \{1, \ldots, k\}$ and receives the associated payoff, which is a $[0, 1]$-valued random variable independently drawn from the (fixed but unknown) probability distribution associated with the pulled arm. The payoff is the only feedback revealed to the agent at each round. The agent's goal is to maximize the expected cumulative payoff over any number of rounds.

In the B2DEP (Bandits with DElay DEpendend Payoff) variant introduced here, at each round $t \in [T]$, the learner picks an arm $X_t \in \mathcal{K}$ and observes the realization Y_t of a reward random variable whose (conditional) expectation $\mu_{X_t}(\tau)$ is an increasing function of τ, the number of rounds since the last time arm $X_t \in \mathcal{K}$ has been pulled. Specifically, for any $i \in \mathcal{K}$ and $t \in [T]$, let us define by $\mathcal{F}_t$ the σ-algebra generated by the past history of pulls and observed reward random variables $X_1, Y_1, \ldots, X_{t-1}, Y_{t-1}$. Given a time horizon T, a learning policy π is a function that maps at each round $t \in [T]$ the observed history $X_1, Y_1 \ldots, X_{t-1}, Y_{t-1}$ to the next action $X_t \in \mathcal{K}$. For any $i \in \mathcal{K}$, $t \in [T]$ we can then define

$$\mu_{i,t} = \mathbb{E}_{\mathcal{F}_t}[Y_t] = (1 - f(\tau)\mathbb{I}\{0 < \tau \leq d_i\})\,\mu_i \tag{3.1}$$

where μ_i is the unknown *baseline reward expectation* for arm i, $f : \mathbb{N} \to [0, 1]$ is an unknown nonincreasing function, and τ is the number of rounds that have passed since that arm was last pulled (conventionally, $\tau = 0$ means that an arm is pulled for the first time). From now on our interest will focus on the expected value $\mu_{i,t} = \mu_i(\tau)$ associated to an arm-delay pair, where the dependency on t is fully captured by the τ random variable. When f is identically zero, B2DEP reduces to the standard stochastic bandit model with payoff expectations $\mu_1, \ldots, \mu_k$. The unknown arm-dependent delay parameters $d_i > 0$ control the number of rounds after which the arm's expected payoff is guaranteed to return to its baseline value μ_i.

Let $g_t(\pi)$ be the payoff collected by policy π at round t. Given an instance of B2DEP, the optimal policy π^* maximizes, over all policies π, the long term expected average payoff

$$\lim_{T\to\infty} \frac{G_T(\pi)}{T} \quad \text{where} \quad G_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} g_t(\pi)\right].$$

Note that, the payoff conditional expectations at any time step t can be computed given the current (not random) *delay vector* $\tau(t) = \big(\tau_1(t), \ldots, \tau_k(t)\big)$, where each integer $0 \leq \tau_i(t) \leq d_i$ counts how many rounds have passed since $i \in \mathcal{K}$ was last pulled (setting $\tau_i(t) = 0$ if i was never pulled or if it was last pulled more than d_i steps ago). Hence, any delay-based policy —e.g., any deterministic function of the current delay vector— is eventually periodic, meaning that $\pi\big(\tau(t)\big) = \pi\big(\tau(t + P)\big)$ for all $t_0 \leq t \leq T$, where P is the period and t_0 is the length of the transient.

Consider the greedy policy π_{greedy} defined as follows: At each round t, π_{greedy} pulls the arm $i \in \mathcal{K}$ with the highest expected reward according to current delays

$$\pi_{\text{greedy}}\big(\tau(t)\big) = \arg\max_{i \in \mathcal{K}} \mu_i\big(\tau_i(t)\big) \tag{3.2}$$

where $\tau_i(t) = 0$ if i was never pulled before. It is easy to see that π_{greedy} is not always optimal. For example consider the following instance of $B2DEP$ with $k = 2$: $f(\tau) = \frac{1}{2}$ for all τ, $\mu_1 = 1$, $\mu_2 = \frac{1}{2} - \varepsilon$, $d_1 = d_2 = 1$. Then π_{greedy} always pulls arm 1 and achieves $G_t(\pi_{\text{greedy}}) = 1 + \frac{T-1}{2}$, whereas $G_T(\pi^*) = 1 + \frac{T-1}{2}\big(\frac{3}{2} - \varepsilon\big)$ where π^* alternates between arm 1 and arm 2. Hence $G_T(\pi_{\text{greedy}}) \le \frac{2}{3} G_T(\pi^*)$.
In the next section we show that the problem of finding the optimal periodic policy for B2DEP is intractable.

3.4 Hardness results

We show that the optimization problem of finding an optimal policy for B2DEP is NP-hard, even when all the instance parameters are known. Our proof relies on the NP-completeness of the Periodic Maintenance Scheduling Problem (PMSP) shown by Bar-Noy et al. [2002]. Although a very similar result can also be proven using the reduction of Basu et al. [2019], introduced for a special case of our B2DEP setting, we give our proof for completeness.
A maintenance schedule on n machines $\{1, \dots, n\}$ is any infinite sequence over $\{0, 1, \dots, n\}$, where 0 indicates that no machine is scheduled for service at that time. An instance of the PMSP decision problem is given by integer service intervals $\ell_1, \dots, \ell_n$ such that $\sum_{i=1}^{n} \frac{1}{\ell_i} \le 1$. The question is whether there exists a maintenance schedule such that the consecutive service times of each machine i are exactly ℓ_i times apart. The following result holds (proof in the supplementary material).

Theorem 5. *It is NP-hard to decide whether an instance of B2DEP has a periodic policy π achieving*

$$\lim_{T \to \infty} \frac{G_T(\pi)}{T} \ge \sum_{i=1}^{k} \frac{\mu_i}{d_i + 1} .$$

3.5 Approximating the optimal policy

In order circumvent the computational problem of finding the best periodic policy, we introduce a simple class $\Pi_{\mathcal{K}}$ of periodic *ranking policies* whose best element π_{ghost} has a cumulative expected payoff not too far from that of π^*. Without loss of generality, assume that $\mu_1 > \cdots > \mu_k$. Let $\Pi_{\mathcal{K}} \equiv \{\pi_m : m \in \mathcal{K}\}$, where each policy π_m cycles over the arm sequence $1, \dots, m$ fixing the random variable τ to be equal to m. For a fixed rank m, the expected average reward $g(m)$ of policy π_m is defined by

$$g(m) = \frac{1}{m} \sum_{j=1}^{m} \mu_j(m) .$$

Since π_{ghost} maximizes $g(m)$ over $m \in \mathcal{K}$, we have $\pi_{\text{ghost}} \equiv \pi_{r^\star}$ where

$$r^\star \in \operatorname*{arg\,max}_{m=1,\dots,k} \frac{1}{m} \sum_{j=1}^{m} \mu_j(m) \tag{3.3}$$

We now study the expected approximation error incurred by introducing policy π_{ghost}. Specifically, we can bound $G_T(\pi_{\text{ghost}})$ in terms of $G_T(\pi^*)$ as follows.

Theorem 6.

$$G_T(\pi_{\text{ghost}}) \geq \big(1 - f(r_0)\big) G_T(\pi^*) + O(1)$$

where r_0 is the largest arm index r such that

$$\mu_i > \max_{j=1,\dots,i-1} \mu_j(i-j) \quad i = 2, \dots, r$$

and $r_0 = 1$ if $\mu_2 \leq \mu_1(1)$.

The definition of r_0 is better understood in the context of the more intuitive delay-based policy π_{greedy}. Note indeed that $r_0 + 1$ is the first round in which π_{greedy} prefers to pull one of the arms that were played in the first r_0 rounds rather than the next arm $r_0 + 1$.

Proof. Since $r^\star$ maximizes (3.3),

$$\begin{aligned} G_T(\pi_{\text{ghost}}) &= \frac{T}{r^\star} \sum_{i=1}^{r^\star} \mu_i(r^\star) + \mathcal{O}(1) \\ &\geq \frac{T}{r_0} \sum_{i=1}^{r_0} \mu_i(r_0) + \mathcal{O}(1) \\ &\geq \frac{T}{r_0} \sum_{i=1}^{r_0} \big(1 - f(r_0)\big)\mu_i + \mathcal{O}(1) \end{aligned}$$

where the $\mathcal{O}(1)$ term takes into account that $r^\star$ may not divide T, and the fact that in the first $r^\star$ rounds the expected reward is $\mu_1 + \cdots + \mu_{r^\star}$ instead of $\mu_1(r^\star) + \cdots + \mu_{r^\star}(r^\star)$. Now split the T time steps in blocks of length r_0. Because r_0 is —by definition— the largest expected reward any policy can achieve in r_0 consecutive steps, the expected reward of π^* in any of these blocks is at most $\mu_1 + \cdots + \mu_{r_0}$. Therefore

$$G_T(\pi^*) \leq \frac{T}{r_0} \sum_{i=1}^{r_0} \mu_i + \mathcal{O}(1)$$

where, as before, the $\mathcal{O}(1)$ term takes into account that r_0 may not divide T. This concludes the proof. □

The proof of Theorem 6 actually shows that both $r^\star$ and r_0 achieve the claimed approximation. However, by definition $G_T(\pi_{\text{ghost}})$ is bigger than the total reward of the policy that cycles over $1, \dots, r_0$. Also, learning π_{ghost} is relatively easy, as we show in Section 3.6.
It is easy to see that $g(m)$ is not monotone due to the presence of the coefficients d_i. For example, consider the B2DEP instance defined by $k = 3$, $\mu_1 = 1$, $\mu_2 = \frac{2}{3}$, $\mu_3 = \frac{1}{2}$, $d_1 = d_2 = d_3 = 2$, and $f(\tau) = 2^{-\tau}$. Then $g(2) < g(1) < g(3)$.

Algorithm 5 (π_{low})

```
Input: Policy set $\Pi_{\mathcal{K}}$, confidence $\delta \in (0,1)$, horizon $T$
1: Let $\mathcal{A}_1 \equiv \mathcal{K}$ be the initial set of active policies
2: repeat                                         ▷ $s$ indexes the stage number
3:     for $m \in \mathcal{A}_s$ do
4:         SELECT $\pi_m$ for $T_s/\big(m|\mathcal{A}_s|\big) + 1$ times
5:         UPDATE $\widehat{g}_s(m)$ discarding the first play
6:     end for
7:     Let $\widehat{m}_s = \arg\max_{m \in \mathcal{A}_s} \widehat{g}_s(\widehat{m}_s)$
8:     UPDATE $\mathcal{A}_{s+1} = \{m \in \mathcal{A}_s \,:\, \widehat{g}_s(m) \geq \widehat{g}_s(\widehat{m}_s) - 2C_s\}$
9: until overall number of pulls exceeds $T$
```

3.6 Learning the ghost policy

In this section we deal with the problem of learning $r^\star$ assuming the correct ordering $1, \dots, k$ of the arms (such that $\mu_1 > \dots > \mu_k$) is known. In the next section, we consider the problem of learning this ordering.

Our search space is the set of ranking policies $\Pi_{\mathcal{K}} \equiv \{\pi_m \,:\, m \in \mathcal{K}\}$, where each policy π_m cycles over the arm sequence $1, \dots, m$. Note that, by definition, $\pi_{\text{ghost}} \equiv \pi_{r^\star}$. The average reward $g(m)$ of policy π_m is defined by $g(m) = \big(\mu_1(m) + \dots + \mu_m(m)\big)/m$. Note that every time the learning algorithm chooses to play a different policy $\pi_m \in \Pi_{\mathcal{K}}$, an extra cost is incurred due to the need of calibrating the estimates for $g(m)$. In fact, if we played a policy different from π_m in the previous round, the reward expectation associated with the play of π_m in the current round is potentially different from $g(m)$. This is due to the fact that we cannot guarantee that each arm in the schedule used by π_m was pulled exactly m steps earlier. This implies that we need to play each newly selected policy more than once, as the first play cannot be used to reliably estimate $g(m)$.

We now introduce the policy π_{low} (Algorithm 5), a simple variant of a learning algorithm based on action elimination proposed in Cesa-Bianchi et al. [2013a]. This policy has a regret bound similar to UCB while guaranteeing a bound $\mathcal{O}\big(k \ln \ln T\big)$ on the number of switches, irrespective of the size of the gaps. In Section 3.8 we compare $\pi_{\text{low}}$with UCB.

In each stage s, algorithm π_{low} plays each policy π_m in the active set $\mathcal{A}_s$ for $T_s/\big(m|\mathcal{A}_s|\big) + 1$ times, where $T_s = T^{1-2^{-s}}$. Then, the algorithm computes the sample average reward $\widehat{g}_s(m)$ based on these plays, excluding the first one because of calibration (lines 3–6). After that, the empirically best policy is selected (7). Finally, the active set is recomputed (line 8) excluding all policies whose sample average reward is significantly smaller than that of the empirically best policy. The quantity C_s is derived from a standard Chernoff-Hoeffding bound (see Proposition 4 in the Appendix material) and is equal to $\sqrt{\frac{k}{2T_s} \ln \frac{2kS}{\delta}}$ where

$$S = \min\left\{ j \in \mathbb{N} \,:\, \sum_{s=1}^{j} \big(|\mathcal{A}_s| + T_s\big) \geq T \right\}$$

implying $S = \mathcal{O}\big(\ln\ln T\big)$. The terms $|\mathcal{A}_s|$ account for the extra calibration pull each time we switch to a new policy in $\Pi_{\mathcal{K}}$. We can prove the following bound on the regret of π_{low} with respect to π_{ghost}.

Theorem 7. *When run on an instance of B2DEP with parameters δ and T, with probability at least $1-\delta$ Algorithm 5 guarantees*

$$\begin{aligned} &G_T(\pi_{\text{ghost}}) - G_T(\pi_{\text{low}}) \\ &\quad = \mathcal{O}\left(k^2 \ln\ln T + \sqrt{kT\left(\ln\frac{k}{\delta} + \ln\ln\ln T\right)}\right) \end{aligned} \tag{3.4}$$

with probability at least $1-\delta$.

Note that this bound is distribution-free. That is, it does not depend on the gaps $g(r^\star) - g(m)$ (which in general could be arbitrarily small). The rate $\sqrt{T}$, as opposed to the $\ln T$ rate of distribution-dependent bounds, cannot be improved upon in general Bubeck et al. [2012].

Proof. The proof is an adaptation of [Cesa-Bianchi et al., 2013a, Theorem 6]. Note that $\mathcal{A}_S \subseteq \cdots \subseteq \mathcal{A}_1$ by construction. Also, our choice of C_s and Chernoff-Hoeffding bound (Proposition 4 in the Appendix material) implies that

$$\max_{m\in\mathcal{A}_s} \big|\widehat{g}_s(m) - g(m)\big| \le C_s \tag{3.5}$$

simultaneously for all $s = 1, \dots, S$ with probability at least $1-\delta$. To see this, note that in every stage s the estimates $\widehat{g}_s(m)$ are computed using $T_s/\big(m|\mathcal{A}_s|\big)$ plays. Since a play of π_m consists of $m \le k$ pulls, we have that each $g(m)$ is estimated using $T_s/|\mathcal{A}_s| \ge T_s/k$ realizations of a sequence of random variables whose expectations have average exactly equal to $g(m)$.

We now claim that, with probability at least $1-\delta$, $r^\star \in \bigcap_{s=1}^S A_s$ and $0 \le \widehat{g}_s(\widehat{m}_s) - \widehat{g}_s(r^\star) \le 2C_s$ for all $s = 1, \dots, S$.

We prove the claim by induction on $s = 1, \dots, S$. We first show that the base case $s = 1$ holds with probability at least $1 - \delta/S$. Then we show that if the claim holds for $s-1$, then it holds for s with probability at least $1-\delta/S$ over all random events in stage s. Therefore, using a union bound over $s = 1, \dots, S$ we get that the claim holds simultaneously for all s with probability at least $1-\delta$.

For the base case $s = 1$ note that $r^\star \in A_1$ by definition, and thus $0 \le \widehat{g}_1(\widehat{m}_1) - \widehat{g}_1(r^\star)$ holds. Moreover: $\widehat{g}_1(\widehat{m}_1) - g(\widehat{m}_1) \le C_1$, $g(r^\star) - \widehat{g}_1(r^\star) \le C_1$, and $g(\widehat{m}_1) - g(r^\star) \le 0$, where the two first inequalities hold with probability at least $1-\delta$ because of (3.5). This implies $0 \le \widehat{g}_1(\widehat{m}_1) - \widehat{g}_1(r^\star) \le 2C_1$ as required. We now prove the claim for $s > 1$. The inductive assumption

$r^\star \in \mathcal{A}_{s-1}$ and $0 \le \widehat{g}_{s-1}(\widehat{m}_{s-1}) - \widehat{g}_{s-1}(r^\star) \le 2C_{s-1}$

directly implies that $r^\star \in \mathcal{A}_s$. Thus we have $0 \le \widehat{g}_s(\widehat{m}_s) - \widehat{g}_s(r^\star)$, because $\widehat{m}_s$ maximizes $\widehat{g}_s$ over a set that contains $r^\star$. The rest of the proof of the claim closely follows that of the base case $s = 1$.

We now return to the proof of the theorem. For any $s = 1, \dots, S$ and for any $m \in \mathcal{A}_s$ we have

$$\begin{aligned}
g(r^\star) - g(m) &\le g(r^\star) - \widehat{g}_{s-1}(m) + C_{s-1} \quad \text{by (3.5)} \\
&\le g(r^\star) - \widehat{g}_{s-1}(\widehat{m}_{s-1}) + 3C_{s-1} \\
&\quad \text{by definition of } \mathcal{A}_{s-1}, \text{ since } m \in \mathcal{A}_s \subseteq \mathcal{A}_{s-1} \\
&\le g(r^\star) - \widehat{g}_{s-1}(r^\star) + 3C_{s-1} \\
&\quad \text{since } \widehat{m}_{s-1} \text{ maximizes } \widehat{g}_{s-1} \text{ in } \mathcal{A}_{s-1} \\
&\le 4C_{s-1} \quad \text{by (3.5)}
\end{aligned}$$

holds with probability at least $1 - \delta/S$. Hence, recalling that the number of switches between two different policies in $\Pi_{\mathcal{K}}$ is deterministically bounded by kS, the regret of the player can be bounded as follows,

$$\begin{aligned}
&G_T(\pi_{\text{ghost}}) - G_T(\pi_{\text{low}}) \\
&\quad = k^2 S + \sum_{s=1}^{S} \frac{T_s}{|\mathcal{A}_s|} \sum_{m \in \mathcal{A}_s} \Big(g(r^\star) - g(m) \Big) \\
&\quad = k^2 S + T_1 + \sum_{s=2}^{S} \frac{T_s}{|\mathcal{A}_s|} \sum_{m \in \mathcal{A}_s} \Big(g(r^\star) - g(m) \Big) \\
&\quad \le k^2 S + T_1 \; + \sum_{i=2}^{S} 4 T_s \sqrt{\frac{k}{2T_{s-1}} \ln \frac{2kS}{\delta}} \\
&\quad = k^2 S + T_1 + 4\sqrt{k \ln \frac{2kS}{\delta}} \sum_{s=2}^{S} \frac{T_s}{\sqrt{T_{s-1}}}
\end{aligned}$$

where the $k^2 S$ term accounts for the regret suffered in the kS plays where we switched between two policies in $\Pi_{\mathcal{K}}$ and paid maximum regret due to calibration for at most k steps (as each policy in $\Pi_{\mathcal{K}}$ is implemented with at most k pulls). Now, since $T_1 = \sqrt{T}$, $T_s/\sqrt{T_{s-1}} = \sqrt{T}$ and $S = \mathcal{O}\big(\ln \ln T\big)$, we obtain that with probability at least $1 - \delta$ the regret is at most of order $k^2 \ln \ln T + \sqrt{T} + \sqrt{kT \left(\ln \frac{k}{\delta} + \ln \ln \ln T \right)}$ as desired. □

3.7 Learning the ordering of the arms

In this section we show how to recover, with high probability, the correct ordering $\mu_1 > \cdots > \mu_k$ of the arms. Initially, we ignore the problem of calibration, and focus on the task of learning the arm ordering when each pulls of arm i returns a sample from the true baseline reward distribution with expectation μ_i.

`BanditRanker` (Algorithm 6) is an action elimination procedure. The arms in the set $\mathcal{A}_r$ of active arms are sampled once each (line 3), and their average rewards are kept sorted in decreasing order (line 4). We use $\widehat{\mu}_{i,r}$ to denote the sample average of rewards

```
Algorithm 6 (BanditRanker)
Input: Confidence $\delta \in (0,1)$
Output: A permutation $[1], \dots, [k]$ of $\mathcal{K}$.
 1: Let $\mathcal{A}_1 \equiv \mathcal{K}$ be the initial set of active arms
 2: repeat                                              ▷ $r$ indexes the round number
 3:     SAMPLE once all arms in $\mathcal{A}_r$                           ▷ sampling round
 4:     SORT the empirical means $\widehat{\mu}_{[1],r} \geq \cdots \geq \widehat{\mu}_{[n],r}$
 5:     for $i = 1$ to $|\mathcal{A}|$ do
 6:         if $\widehat{\mu}_{[i],r} + 2\varepsilon_r < \min_{j \in \mathcal{K}^+_{[i],r}} \widehat{\mu}_{j,r}$ then
 7:             if $\widehat{\mu}_{[i],r} - 2\varepsilon_r > \max_{j \in \mathcal{K}^-_{[i],r}} \widehat{\mu}_{[s],r}$ then
 8:                 REMOVE $[i]$ from $\mathcal{A}_r$
 9:                 RANK before $[i]$ all arms in $\mathcal{K}^+_{[i],r}$
10:                 RANK after $[i]$ all arms in $\mathcal{K}^-_{[i],r}$
11:             end if
12:         end if
13:     end for
14: until $|\mathcal{A}_t| \leq 1$
```

obtained from arm i after r sampling rounds, and define the indexing $[1], \dots, [k]$ be such that $\widehat{\mu}_{[1],r} \geq \cdots \geq \widehat{\mu}_{[k],r}$, where ties are broken according to the original arm indexing.
When the confidence interval around the average reward of an arm $[i]$ is not overlapping anymore with the confidence intervals of the other arms (lines 6–7), $[i]$ is removed from $\mathcal{A}_r$ and not sampled anymore (line 8). Moreover, the set $\mathcal{K}^+_{[i],r}$ of all arms $[b] \in \mathcal{A}_r$ such that $\widehat{\mu}_{[b],r} \geq \widehat{\mu}_{[i],r}$ (if any) is ranked before $[i]$ (line 9). Similarly, the set let $\mathcal{K}^-_{[i],r}$ of all arms $[s] \in \mathcal{A}_r$ such that $\widehat{\mu}_{[s],r} \leq \widehat{\mu}_{[i],r}$ (if any) is ranked after $[i]$ (line 10). The algorithm ends when all arms are removed (line 14).
The parameter ε_r determining the confidence interval after r sampling rounds is defined by

$$\varepsilon_r = \sqrt{\frac{1}{2r} \ln \frac{2kr(r+1)}{\delta}} \,. \tag{3.6}$$

The sequence of removed arms can be stored in a binary tree whose root is the first removed arm and whose left (resp., right) leaf contain all arms whose average reward was bigger (resp., smaller) when the first arm was removed. When a new arm is removed, the leaf to which it belongs is split using the same logic that we used for the root. Eventually, all nodes contain a single arm and the in-order traversal of the tree provides the desired ordering.
We introduce the following quantity, measuring the suboptimality gaps between arm that are adjacent in the correct ordering,

$$\Delta_i = \begin{cases} \Delta_{1,2} & \text{if } i = 1 \\ \min\left\{\Delta_{i-1,i}, \Delta_{i,i+1}\right\} & \text{if } 1 < i < k \\ \Delta_{k-1,k} & \text{if } i = k \end{cases}$$

where $\Delta_{i,j} = \mu_i - \mu_j$.
We are now ready to state and prove the main result of this section.

Theorem 8. *If Algorithm 6 is run with parameter δ on a k-armed stochastic bandit problem, the correct ordering $\mu_1 > \cdots > \mu_k$ of the arms is returned with probability at least $1 - \delta$ after a number of pulls of order*

$$\sum_{i=1}^{k-1} \frac{1}{\Delta_i^2} \ln \frac{1}{\delta \Delta_i} \; . \tag{3.7}$$

Note that, up to logarithmic factors, the bound stated in Theorem 8 is of the same order as the sample used by an ideal procedure that knows $\Delta_1, \dots, \Delta_k$ and uses the optimal order $1/\Delta_i^2$ of samples to determine the position of each arm i in the correct ordering.

Proof. The proof is an adaptation of Even-Dar et al. [2006, Theorem 8]. Using Chernoff-Hoeffding bounds (see Proposition 4 in the Appendix), the choice of ε_r ensures that

$$\begin{aligned} \Pr\left(\exists\, r \geq 1 \; \exists i \in \mathcal{K} \; \left|\widehat{\mu}_{i,r} - \mu_i\right| > \varepsilon_r\right) &\leq 2k \sum_{r \geq 1} e^{-2\varepsilon_t^2 r} \\ &\leq \delta \; . \end{aligned} \tag{3.8}$$

If an action $[i]$ is eliminated after r sampling rounds, then it must be that $\widehat{\mu}_{[b],r} - 2\varepsilon_r > \widehat{\mu}_{[i],r} > \widehat{\mu}_{[s],r} + 2\varepsilon_r$ for all $[b] \in \mathcal{K}^+_{[i],r}$ and all $[s] \in \mathcal{K}^-_{[i],r}$. Condition (3.8) then ensures that, with probability at least $1 - \delta$, $\mu_{[b]} > \mu_{[i]} > \mu_{[s]}$ for all such b and s. This implies that the current ordering of $\mu_{[j],r}$ for $j \in \mathcal{A}_r$ is correct with respect to $[i]$. Since $\varepsilon_r \to 0$, every action is eventually eliminated. Therefore, with probability at least $1 - \delta$ the sequence of eliminated arms i and their corresponding sets $\mathcal{K}^+_{[i],r}, \mathcal{K}^-_{[i],r}$ provide the correct arm ordering.
We now proceed to bounding the number of samples. Under condition (3.8), for all $b < i < s$,

$$\Delta_{b,i} - 2\varepsilon_r = \big(\mu_b - \varepsilon_r\big) - \big(\mu_i + \varepsilon_r\big) \leq \widehat{\mu}_{b,r} - \widehat{\mu}_{i,r} \; .$$

Therefore, if $\widehat{\mu}_{b,r} - \widehat{\mu}_{i,r} \leq 2\varepsilon_r$, then $\Delta_{b,i} \leq 4\varepsilon_r$. Recalling the definition (3.6) of ε_r and solving by $r = r(b,i)$ we get

$$r(b,i) = \mathcal{O}\left(\frac{1}{\Delta_{b,i}^2} \ln \frac{1}{\delta \Delta_{b,i}}\right) \; .$$

Thus, after $r(b,i)$ sampling rounds, $\widehat{\mu}_{b,r(b,i)} - \widehat{\mu}_{i,r(b,i)} > 2\varepsilon_{r(b,i)}$ with probability at least $1 - \delta$. Similarly, after $r(i,s)$ sampling rounds, $\widehat{\mu}_{i,r(i,s)} - \widehat{\mu}_{s,r(i,s)} > 2\varepsilon_{r(i,s)}$ with probability at least $1 - \delta$.
This further implies that after $N_i = \mathcal{O}\left(\frac{1}{\Delta_i^2} \ln \frac{1}{\delta \Delta_i}\right)$ many sampling rounds, action i is eliminated and not sampled any more.
Re-define the indexing $[1], \dots, [k]$ so that $\Delta_{[1]} > \cdots > \Delta_{[k]}$. Hence $N_{[1]} < \cdots < N_{[k]}$ by definition. We now compute a bound on the overall number of pulls based on our bound on the number of sampling rounds. With probability at least $1 - \delta$, we have that: $kN_{[1]}$ pulls are

needed to eliminate arm $[1]$, $(k-1)\big(N_{[2]} - N_{[1]}\big)$ pulls are needed to eliminate arm $[2]$, and so on. Hence, the total number of pulls needed to eliminate all arms is

$$
\begin{aligned}
&\sum_{i=0}^{k-2}(k-i)\big(N_{[i+1]} - N_{[i]}\big) \\
&\quad = kN_{[k-1]} - \sum_{i=0}^{k-2} i\big(N_{[i+1]} - N_{[i]}\big) \\
&\quad = kN_{[k-1]} - (k-1)N_{[k-1]} + \sum_{i=0}^{k-2} N_{[i+1]} \\
&\quad = N_{[k-1]} + \sum_{i=1}^{k-1} N_{[i]}
\end{aligned}
$$

with probability at least $1-\delta$ where we set conventionally $N_{[0]} = 0$. □

In order to apply `BanditRanker` to an instance of B2DEP, we assume that an upper bound $d_0 > \max_i d_i$ be available in advance to the algorithm. This ensures that $\mu_i(d_0) = \mu_i$ for all $i \in \mathcal{K}$. In each sampling round r, we partition the arms in $\mathcal{A}_r$ in groups of size d_0 and make $2d_0$ pulls for each group by cycling twice over the arms in an arbitrary order. Then, the first d_0 pulls in each group are discarded, while the last d_0 pulls are used to estimate the expectations μ_i (when d_0 does not divide $|\mathcal{A}_r|$ we can add to $\mathcal{A}_r$ arms that were already removed, or arms from previous groups, just for the purpose of calibrating). The sample size bound (3.7) remains of the same order (because the extra pulls only add a factor of two).

We could have used a naive approach that continuously pulls all arms until no overlaps occur. This approach would pull each arm $1/(\min_{i\in[k]}\Delta_i)^2$ times requiring than much more samples compared to (3.7).

Finally, notice that a symmetric argument would hold in case the delay upper bound d_0 would not be available. Indeed, the same logic holds considering $d_0 = 1$, that represents the minimum delay for which all arms incurs the same relative discount $1 - f(d_0)$. It should be easy to observe that this would require many more samples as gaps Δ_i would be smaller.

3.8 Experiments

In this section we present an empirical evaluation of our policy π_{low} in a synthetic environment with Bernoulli rewards. In order to study the impact the switching cost on ranking policies when the suboptimality gap is small, we also define a setting in which there are two distinct ranking policies that are both optimal —see Figure 3.1.

We plot regrets against the policy π_{ghost}. Our policy π_{low} is run without any specific tuning (other than the knowledge of the horizon T) and with δ set to 0.1 in all experiments. The benchmark π_{ucb} consists of running UCB1 (summarized in Algorithm 1 in Section 2.1) —with the same scaling factor as in the original article by Auer et al. [2002]— over the class $\Pi_{\mathcal{K}}$ of ranking policies, where calibration is addressed by rolling out twice each ranking

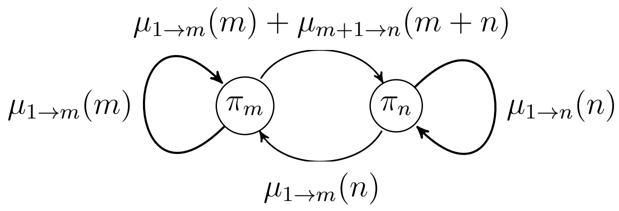

Figure 3.1: Transitions between policies π_m and π_n assuming $n > m$, where the notation $\mu_{m\to n}(d)$ stands for $\mu_m(d) + \cdots + \mu_n(d)$. The expected reward obtaining by switching between policies is different from the one obtained by cycling over the same policy.

policy selected by UCB1 and using only the second roll-out to compute reward estimates.

Since both π_{low} and π_{ghost} are run over $\Pi_{\mathcal{K}}$, we implicitly assume that `BanditRanker` successfully ranked the arms in a preliminary stage.

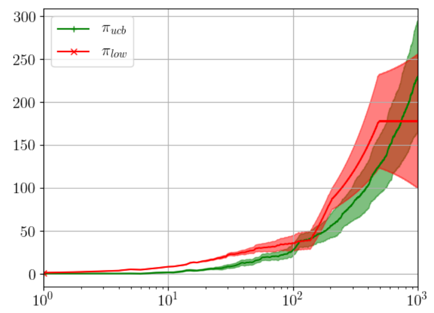

Figure 3.2: Comparing regrets of π_{low} and π_{ucb} against π_{ghost} with 7 arms and baseline expectations $0, 1/3, 2/3, 4/5, 13/15, 14/15, 1$ and $f(\tau) = (0.999)^\tau$. A unit cost is charged for switching between ranking policies. Curves are averages of 5 runs each using a different sample of delays $d_1, \ldots, d_7$ uniformly drawn from $\{1, \ldots, 6\}$. We plot expectations of sampled arms rather than realized rewards.

Figure 3.2 shows that when the gap between the best and the second best ranking policy is not too small (0.1 on average in these experiments), then π_{ucb} is competitive against π_{low} even in the presence of unit switching costs. This happens because, in order to minimize the number of switches, π_{low} samples a suboptimal policy more frequently than π_{ucb}. Although this oversampling does not affect the distribution-free regret bound of π_{low}, it hurts performance unless the suboptimality gap is small enough to cause the switching costs to prevail, a case which is addressed next. Note also that π_{low} eventually stops exploration because all policies but one have been eliminated, while π_{ucb} keeps on exploring, albeit at a logarithmic rate.
In the second experiment we consider two arms with $\mu_1 = 1$, $f(1) = 0.3$, $f(2) = 0.25$, $d_1 = d_2 = 2$, and μ_2 chosen so that $g(1) = g(2)$ to simulate a vanishing suboptimality gap between π_1 and π_2. Figure 3.3 (upper part) shows that π_{low} performs better than π_{ucb} due to its low switch regime. On the other hand, Figure 3.3 (lower part) shows that when the switching cost is zero, switching between two good policies becomes more advantageous

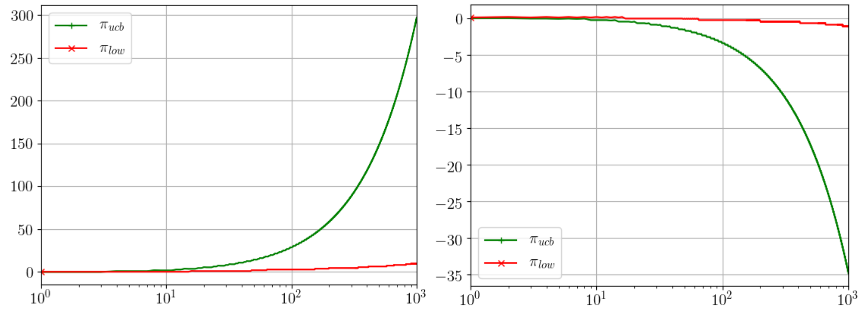

Figure 3.3: Comparing regrets of π_{low} and π_{ucb} against π_{ghost} with 2 arms such that $g(1) = g(2)$ with unit cost charged for switching between the two policies (upper part) and without any cost for switching (lower part).

than using a single good policy, and the regret of both π_{ucb} and π_{low} becomes negative (in this case π_{ucb}, which has no control over the number of switches, outperforms π_{low}). The reason for this advantage is explained by Fact 1 below (proof in the supplementary material), see also Figure 3.1.

Fact 1. *If an instance of B2DEP admits two optimal ranking policies, then consistently switching between these two policies achieves an average expected reward higher than sticking to either one.*

To summarize, the experiments confirm that, in the presence of switching costs, π_{low} works better than π_{ucb} only when the suboptimalty gap is very small. The advantage of π_{low} over π_{ucb} is however reduced by the fact that switching between two good policies is better than consistently playing either one of the two (Fact 1). Note also that π_{low} stops exploring because T is known. This preliminary knowledge can be dispensed with using a doubling trick, or some more sophisticated method. Also, it would be interesting to design a method that achieves the best between the performance of π_{ucb} and π_{low}, according to the size of the suboptimality gap.

3.9 Conclusions

Motivated by music recommendation in streaming platforms, we introduced a new stochastic bandit model with nonstationary reward distributions. To cope with the NP-hardness of learning the optimal policy caused by nonstationarity, we introduced a restricted class of ranking policies approximating the optimal performance. We then proved sample and regret bounds on the problem of learning the best ranking policy in this class. One of the main problems left open by our work is that of deriving more practical learning algorithms, able to simultaneously learn the ranking of the arms and the best cutoff value $r^\star$, while minimizing their regret with respect to the best ranking policy.

3.10 Proofs

3.10.1 Proof of the Hardness Result (Theorem 5)

Proof. Given an instance $\ell_1, \dots, \ell_n$ of PMSP, we construct a B2DEP instance with $|\mathcal{K}| = n+1$ arms such that $d_i = \ell_i - 1$ and $\mu_i = 1$ for all $i = 1, \dots, n$, $\mu_{n+1} = 0$, and $f \equiv 1$. The long-term average reward for a periodic policy in this setting is

$$\sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\mathbb{I}\{t_{i,j} > d_i\}}{t_{i,j}}$$

where N_i is the number of times the policy plays arm i in a period and $t_{i,j}$ is the number of time steps between when arm i was played for the j-th time in the cycle and the last time it was played (in the same cycle or in the previous cycle, excluding the transient). Clearly, if the PMSP instance has a feasible schedule, then we can design a bandit policy that replicates that schedule (playing arm $n+1$ at all time steps where no machines are scheduled for maintenance). The long-term average reward of this policy is at most $\sum_{i=1}^{n} \frac{1}{d_i+1}$. Moreover, if we have a periodic bandit policy with long-term average reward exactly equal to $\sum_{i=1}^{n} \frac{1}{d_i+1}$, this means that each arm $i = 1, \dots, n$ is eventually played after exactly $d_i + 1 = \ell_i$ rounds. Indeed, the only way to have

$$\frac{1}{N_i} \sum_{n=1}^{N_i} \frac{\mathbb{I}\{t_{i,j} > d_i\}}{t_{i,j}} \ge \frac{1}{d_i + 1}$$

is by setting $t_{i,j} = d_i + 1$ for all $j = 1, \dots, N_i$. □

3.10.2 Proof of the Swtiching Result (Fact 1)

Proof. We use the following notation: $\mu_{m\to n}(d)$, where $n > m$, stands for $\mu_m(d) + \dots + \mu_n(d)$. Consider two optimal ranking policies π_m and π_n with $n > m$. Then $g(m) = g(n)$, where $g(n) = \frac{1}{n}\mu_{1\to n}(n)$ and similarly for $g(m)$. The expected total reward of playing π_m after π_n is $\mu_{1\to m}(n)$, and the expected total reward of playing π_n after π_m is $\mu_{1\to m}(m) + \mu_{m+1\to n}(m+n)$. We want to prove

$$\frac{\mu_{1\to m}(n) + \mu_{1\to m}(m) + \mu_{m+1\to n}(m+n)}{m+n} \ge \frac{\mu_{1\to m}(m)}{m} .$$

Rearranging gives $\mu_{1\to m}(n) + \mu_{m+1\to n}(m+n) \ge \frac{n}{m}\mu_{1\to m}(m)$. Since $\frac{1}{n}\mu_{1\to n}(n) = \frac{1}{m}\mu_{1\to m}(m)$, we have

$$\mu_{1\to m}(n) + \mu_{m+1\to n}(m+n) \ge \mu_{1\to n}(n) .$$

Observing that $\mu_{1\to n}(n) = \mu_{1\to m}(n) + \mu_{m+1\to n}(n)$, the above is equivalent to

$$\mu_{m+1\to n}(m+n) \ge \mu_{m+1\to n}(n)$$

which is always true since in our model expected rewards are non-decreasing with delays. □

Chapter 4

Efficient Linear Bandits through Matrix Sketching

We dedicate this chapter to the design of efficient strategies for the linear bandit learning problem. We consider the popular OFUL and Thompson Sampling algorithms that have been presented in Section 2.2 of Chapter 2. As it was highlighted, they both share an update time of order $\mathcal{O}(d^2)$, which could be potentially expensive when dealing with arms represented by a large number of features d. We show how they can be made efficient using Frequent Directions, a deterministic online sketching technique. More precisely, we show that a sketch of size m allows a $\mathcal{O}(md)$ update time for both algorithms. This computational speedup is accompanied by regret bounds of order $(1+\varepsilon_m)^{3/2}d\sqrt{T}$ for OFUL and of order $\big((1+\varepsilon_m)d\big)^{3/2}\sqrt{T}$ for Thompson Sampling, where ε_m is bounded by the sum of the tail eigenvalues not covered by the sketch. In particular, when the selected contexts span a subspace of dimension at most m, our algorithms have a regret bound matching that of their slower, non-sketched counterparts. Experiments on real-world datasets corroborate our theoretical results.

4.1 Introduction

We consider two of the most popular algorithms for stochastic linear bandits: OFUL Abbasi-Yadkori et al. [2011] and linear Thompson Sampling Agrawal and Goyal [2013] (linear TS for short). As we shown in Section 2.2, while exhibiting good theoretical and empirical performances, both algorithms require $\Omega\big(d^2\big)$ time to update their model after each round. In this Chapter we investigate whether it is possible to significantly reduce this update time while ensuring that the regret remains nicely bounded.

The quadratic dependence on d is due to the computation of the inverse correlation matrix of past actions (a cubic dependence is avoided because each new inverse is a rank-one perturbation of the previous inverse). The occurrence of this matrix is caused by the linear nature of rewards: to compute their decisions, both algorithms essentially solve a regularized least squares problem at every round. In order to improve the running time, we sketch the correlation matrix using a specific technique —Frequent Directions, Ghashami et al. [2016]—

that works well in a sequential learning setting. While matrix sketching is a well-known approach Woodruff [2014], to the best of our knowledge this is the first work that applies sketching to linear contextual bandits while providing rigorous performance guarantees. With a constant sketch size of m, a rank-one update of the correlation matrix takes only time $\mathcal{O}(md)$, which is linear in d for a constant sketch size. However, this speed-up comes at a price, as sketching reduces the matrix rank causing a loss of information which —in turn— affects the least squares estimates used by the algorithms. Our main technical contribution shows that when OFUL and linear TS are run with a sketched correlation matrix, their regret blows up by a factor which is controlled by the spectral decay of the correlation matrix of selected actions. More precisely, we show that the sketched variant of OFUL, called SOFUL, achieves a regret bounded by

$$R(T, \mathbf{w}^\star) \overset{\widetilde{\mathcal{O}}}{=} \left(1+\varepsilon_m\right)^{\frac{3}{2}} \Big(m + d \ln\left(1+\varepsilon_m\right)\Big)\sqrt{T} \tag{4.1}$$

where m is the sketch size and ε_m is upper bounded by the spectral tail (sum of the last $d-m+1$ eigenvalues) of the correlation matrix for all T rounds. In the special case when the selected actions span a number of dimensions equal or smaller than the sketch size, then $\varepsilon_m = 0$ implying a regret of order $m\sqrt{T}$. Thus, we have a regret bound matching that of the slower, non-sketched counterpart.[1] When the correlation matrix has rank larger than the sketch size, the regret of SOFUL remains small to the extent the spectral tail of the matrix grows slowly with T. In the worst case of a spectrum with heavy tails, SOFUL may incur linear regret. In this respect, sketching is only justified when the computational cost of running OFUL cannot be afforded. Similarly, we prove that the efficient sketched formulation of linear TS enjoys a regret bound of order

$$R(T, \mathbf{w}^\star) \overset{\widetilde{\mathcal{O}}}{=} \Big(m + d \ln(1+\varepsilon_m)\Big)\left(1+\varepsilon_m\right)^{\frac{3}{2}} \sqrt{dT}\ . \tag{4.2}$$

Once again, for $\varepsilon_m = 0$ our bound is of order $m\sqrt{dT}$, which matches the regret bound for linear TS. When the rank of the correlation matrix is larger than the sketch size, the bound for linear TS behaves similarly to the bound for SOFUL.

Finally, we show a problem-dependent regret bound for SOFUL. This bound, which exhibits a logarithmic dependence on T, depends on the smallest gap Δ between the expected reward of the best and the second best action across the T rounds,

$$R(T, \mathbf{w}^\star) \overset{\widetilde{\mathcal{O}}}{=} \frac{1}{\Delta}\left(1+\varepsilon_m\right)^{3} \Big(m + d \ln\left(1+\varepsilon_m\right)\Big)^{2} (\ln T)^2\ . \tag{4.3}$$

When $\varepsilon_m(T) = 0$ this bound is of order $\frac{m^2}{\Delta}(\ln T)^2$ which matches the corresponding bound for OFUL. Experiments on six real-world datasets support our theoretical results.

Additional related work. For an introduction to contextual bandits, we refer the reader to the recent monograph of Lattimore and Szepesvári [2018]. The idea of applying sketching

[1] The regret bound of OFUL in Abbasi-Yadkori et al. [2011, Theorem 3] is stated as $\mathcal{O}(d\sqrt{T})$, however, it can be improved for low-rank problems by using the "log-det" formulation of the confidence ellipsoid.

techniques to linear contextual bandits was also investigated by Yu et al. [2017], where they used random projections to preliminarly draw a random m-dimensional subspace which is then used in every round of play. However, the per-step computation time of their algorithm is cubic in m rather than quadratic like ours. Morover, random projection introduces an additive error ε in the instantaneous regret which becomes of order $m^{-1/2}$ for any value of the confidence parameter δ bounded away from 1. A different notion of compression in contextual bandits is explored by Jun et al. [2017], where they use hashing algorithms to obtain a computation time sublinear in the number K of actions. An application of sketching (including Frequent Directions) to speed up 2nd order algorithms for online learning is studied by Luo et al. [2016], in a RKHS setting by Calandriello et al. [2017], and in stochastic optimization by Gonen et al. [2016].

4.2 Sketching the correlation matrix

The idea of sketching is to maintain an approximation of $\mathbf{X}_t$, denoted by $\mathbf{S}_t \in \mathbb{R}^{m \times d}$, where $m \ll d$ is a small constant called the sketch size. If we choose m such that $\mathbf{S}_t^\top \mathbf{S}_t$ approximates $\mathbf{X}_t^\top \mathbf{X}_t$ well, we could use $\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}$ in place of $\mathbf{V}_t^\lambda$. In the following we use the notation $\widetilde{\mathbf{V}}_t = \mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}$ to denote the sketched regularized correlation matrix. The RLS estimate based upon it is denoted by

$$\widetilde{\mathbf{w}}_t = \widetilde{\mathbf{V}}_t^{-1} \sum_{s=1}^{t} \mathbf{x}_s Y_s \ . \tag{4.4}$$

A trivial replacement of $\mathbf{V}^\lambda$ with $\widetilde{\mathbf{V}}$ does not yield an efficient algorithm. On the other hand, using the Woodbury identity we may write

$$\widetilde{\mathbf{V}}_t^{-1} = \frac{1}{\lambda}\left(\mathbf{I}_{d \times d} - \mathbf{S}_t^\top \mathbf{H}_t \mathbf{S}_t\right)$$

where $\mathbf{H}_t = \left(\mathbf{S}_t \mathbf{S}_t^\top + \lambda \mathbf{I}_{m \times m}\right)^{-1}$. Here matrix-vector multiplications involving $\mathbf{S}_t$ require time $\mathcal{O}(md)$, while matrix-matrix multiplications involving $\mathbf{H}_t$ require time $\mathcal{O}(m^2)$. So, as long as $\mathbf{S}_t$ and $\mathbf{H}_t$ can be efficiently maintained, we obtain an algorithm for linear stochastic bandits where $\widetilde{\mathbf{V}}_t^{-1}$ can be updated in time $\mathcal{O}(md + m^2)$. Next, we focus on a concrete sketching algorithm that ensures efficient updates of $\mathbf{S}_t$ and $\mathbf{H}_t$.

Frequent Directions. Frequent Directions (FD) Ghashami et al. [2016] is a deterministic sketching algorithm that maintains a matrix $\mathbf{S}_t$ whose last row is invariably $\mathbf{0}$. On each round, we insert $\mathbf{x}_t^\top$ into the last row of $\mathbf{S}_{t-1}$, perform an eigendecomposition $\mathbf{S}_{t-1}^\top \mathbf{S}_{t-1} + \mathbf{x}_t \mathbf{x}_t^\top = \mathbf{U}_t \, \boldsymbol{\Sigma}_t \, \mathbf{U}_t^\top$, and then set $\mathbf{S}_t = \left(\boldsymbol{\Sigma}_t - \rho_t \mathbf{I}_{m \times m}\right)^{\frac{1}{2}} \mathbf{U}_t$, where ρ_t is the smallest eigenvalue of $\mathbf{S}_t^\top \mathbf{S}_t$. Observe that the rows of $\mathbf{S}_t$ form an orthogonal basis, and therefore $\mathbf{H}_t$ is a diagonal matrix which can be updated and stored efficiently. Now, the only step in question is an eigendecomposition, which can also be done in time $\mathcal{O}(md)$ —see [Ghashami et al., 2016, Section 3.2]. Hence, the total update time per round is $\mathcal{O}(md)$. The updates of matrices $\mathbf{S}_t$ and $\mathbf{H}_t$ are summarized in Algorithm 7.

Algorithm 7 (FD Sketching)

```
Input: $\mathbf{S}_{t-1} \in \mathbb{R}^{m \times d}, \mathbf{x}_t \in \mathbb{R}^d, \lambda > 0$
Compute eigendecomposition $\mathbf{U}^\top \mathrm{diag}\{\rho_1, \dots, \rho_m\}\,\mathbf{U} = \mathbf{S}_{t-1}^\top \mathbf{S}_{t-1} + \mathbf{x}_t \mathbf{x}_t^\top$
$\mathbf{S}_t \leftarrow \mathrm{diag}\{\sqrt{\rho_1 - \rho_m}, \dots, \sqrt{\rho_{m-1} - \rho_m}, 0\}\,\mathbf{U}$
$\mathbf{H}_t \leftarrow \mathrm{diag}\left\{\frac{1}{\rho_1 - \rho_m + \lambda}, \dots, \frac{1}{\lambda}\right\}$
Output: $\mathbf{S}_t, \mathbf{H}_t$
```

It is not hard to see that FD sketching sequentially identifies the top-m eigenvectors of the matrix $\mathbf{V}_T = \mathbf{X}_T^\top \mathbf{X}_T$. Thus, whenever we use a sketched estimate, we lose a part of the spectrum tail. This loss is captured by the following notion of *spectral error*,

$$\varepsilon_m = \min_{k=0,\dots,m-1} \frac{\lambda_{d-k} + \lambda_{d-k+1} + \dots + \lambda_d}{\lambda(m-k)} \tag{4.5}$$

where $\lambda_1 \geq \ldots \geq \lambda_d$ are the eigenvalues of the correlation matrix $\mathbf{V}_T$ ($\mathbf{V}_T^\lambda$ with $\lambda = 0$). Note that $\varepsilon_m \leq (\lambda_m + \dots + \lambda_d)/\lambda$. For matrices with low rank or light-tailed spectra we expect this spectral error to be small. In the following, we use $\widetilde{m}$ to denote the quantity $m + d\ln(1 + \varepsilon_m)$ which occurs often in our bounds involving sketching. Note that $\widetilde{m} \geq m$ and $\widetilde{m} \to m$ as the spectral error vanishes.
Since the matrix $\mathbf{V}_t^\lambda$ is used to compute both the RLS estimate $\widehat{\mathbf{w}}_t$ and the norm $\|\cdot\|_{\mathbf{V}_t^\lambda}$, the sketching of $\mathbf{V}_t^\lambda$ clearly affects the confidence ellipsoid. The next theorem quantifies how much the confidence ellipsoid must be blown up in order to compensate for the sketching error. Let ρ_t be the smallest eigenvalue of the FD-sketched correlation matrix $\mathbf{S}_t^\top \mathbf{S}_t$ and let $\bar{\rho}_t = \rho_1 + \dots + \rho_t$. The following proposition due to Ghashami et al. [2016] (see the proof of Thm. 3.1, bound on Δ) relates $\bar{\rho}_t$ to ε_m (Equation 4.5).

Proposition 1. *For any $t = 0, \dots, T$, any $\lambda > 0$, and any sketch size $m = 1, \dots, d$, it holds that $\bar{\rho}_t/\lambda \leq \varepsilon_m$.*

A key lemma in the analysis of regret is the following sketched version of [Abbasi-Yadkori et al., 2011, Lemma 11], which bounds the sum of the ridge leverage scores. Although sketching introduces the spectral error ε_m, it also improves the dependence on the dimension from d to m whenever ε_m is sufficiently small.

Lemma 1 (Sketched leverage scores)**.**

$$\sum_{t=1}^{T} \min\left\{1, \|\mathbf{x}_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} \leq 2\left(1 + \varepsilon_m\right)\left(\widetilde{m} + m\ln\left(1 + \frac{TL^2}{m\lambda}\right)\right) . \tag{4.6}$$

We can now state the main result of this section.

Theorem 9 (Sketched confidence ellipsoid)**.** *Let $\widetilde{\mathbf{w}}_t$ be the RLS estimate constructed by an arbitrary policy for linear contextual bandits after t rounds of play. For any $\delta \in (0, 1)$,* *the optimal parameter* $\mathbf{w}^\star$ *belongs to the set* $\widetilde{C}_t \equiv \left\{\mathbf{w} \in \mathbb{R}^d \ : \ \|\mathbf{w} - \widetilde{\mathbf{w}}_t\|_{\widetilde{\mathbf{V}}_t} \le \widetilde{\beta}_t(\delta)\right\}$ *with probability at least* $1-\delta$*, where*

$$\widetilde{\beta}_t(\delta) = R\sqrt{m \ln\left(1 + \frac{tL^2}{m\lambda}\right) + 2\ln\frac{1}{\delta} + d\ln\left(1 + \frac{\bar{\rho}_t}{\lambda}\right)} \cdot \sqrt{1 + \frac{\bar{\rho}_t}{\lambda}} + S\sqrt{\lambda}\left(1 + \frac{\bar{\rho}_t}{\lambda}\right) \tag{4.7}$$

$$\overset{\widetilde{\mathcal{O}}}{=} R\sqrt{\widetilde{m}\,(1+\varepsilon_m)} + S\sqrt{\lambda}\,(1+\varepsilon_m) \ . \tag{4.8}$$

Note that (4.8) is larger than its non-sketched counterpart (2.11) due to the factors $1 + \varepsilon_m$. However, when the spectral error ε_m vanishes, $\widetilde{\beta}_t(\delta)$ becomes of order $R\sqrt{m} + S\sqrt{\lambda}$, which improves upon (2.11) since we replace the dependence on the ambient space dimension d with the dependence on the sketch size m. In the following, we use the abbreviation $M_\lambda = \max\left\{1, 1/\sqrt{\lambda}\right\}$.

4.3 Sketched OFUL

Equipped with the sketched confidence ellipsoid and the sketched RLS estimate, we can now introduce SOFUL (Algorithm 8), the sketched version of OFUL. SOFUL enjoys the following regret bound, characterized in terms of the spectral error.

Algorithm 8 (SOFUL)

```
Input: $\delta, \lambda > 0, m \in \{1, \dots, d-1\}$
$\widetilde{\mathbf{w}}_0 = \mathbf{0}, \widetilde{\mathbf{V}}_0^{-1} = \frac{1}{\lambda}\mathbf{I}_{d\times d}, \mathbf{S}_0 = \mathbf{0}_{m\times d}$
for $t = 1, 2, \dots$ do
    GET decision set $D_t$
    SELECT $\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x}\in D_t} \left\{\widetilde{\mathbf{w}}_{t-1}^\top \mathbf{x} + \widetilde{\beta}_{t-1}(\delta)\,\|\mathbf{x}\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\}$
    OBSERVE reward $Y_t$
    UPDATE $\mathbf{S}_t, \mathbf{H}_t$ using Alg. 7 given $\mathbf{S}_{t-1}, \mathbf{x}_t$
    UPDATE $\widetilde{\mathbf{V}}_t^{-1} \leftarrow \frac{1}{\lambda}\left(\mathbf{I}_{d\times d} - \mathbf{S}_t^\top \mathbf{H}_t \mathbf{S}_t\right)$
    UPDATE $\widetilde{\mathbf{w}}_t$ using (4.4)
end for
```

Theorem 10. *The regret of SOFUL with FD-sketching of size* m *w.h.p. satisfies*

$$R(T, \mathbf{w}^\star) \overset{\widetilde{\mathcal{O}}}{=} M_\lambda \big(1 + \varepsilon_m\big)^{\frac{3}{2}} \widetilde{m}\left(R + S\sqrt{\lambda}\right)\sqrt{T} \, .$$

Similarly to Abbasi-Yadkori et al. [2011], we also prove a distribution dependent regret bound for SOFUL. This bound is polylogarithmic in time and depends on the smallest difference Δ between the rewards of the best and the second best action in the decision sets,

$$\Delta = \min_{t=1,\dots,T} \ \max_{\mathbf{x}\in D_t \setminus \{\mathbf{x}_t^\star\}} \left(\mathbf{x}_t^\star - \mathbf{x}\right)^\top \mathbf{w}^\star \, .$$

Theorem 11. *The regret of SOFUL with FD-sketching of size m w.h.p. satisfies*

$$R(T, \mathbf{w}^\star) \overset{\widetilde{\mathcal{O}}}{=} M_\lambda \left(1 + \varepsilon_m\right)^3 \widetilde{m}^2 \left(R^2 + S^2\lambda\right) \frac{(\ln T)^2}{\Delta} .$$

Proofs of the regret bounds appear in the supplementary material (Section 4.5.2).

4.4 Sketched linear TS

In this section we introduce a variant of linear TS (Algorithm 4) based on FD-sketching. Similarly to SOFUL, sketched linear TS (see Algorithm 9) uses the FD-sketched approximation $\widetilde{\mathbf{V}}_{t-1}$ of the correlation matrix $\mathbf{V}_{t-1}$ in order to select the action $\mathbf{x}_t$. Note that, in this case, we need both $\widetilde{\mathbf{V}}_{t-1}^{-1}$ and $\widetilde{\mathbf{V}}_{t-1}^{-\frac{1}{2}}$ to compute $\mathbf{x}_t$. Using the generalized Woodbury identity (Corollary 1 in Section 4.5.1 for proofs), we can write

Algorithm 9 (Sketched linear TS)

```
Input: $\delta, \lambda > 0, m \in \{1, \dots, d-1\}$, $\mathcal{D}^{\text{TS}}$ (TS-sampling distribution)
$\widetilde{\mathbf{w}}_0 = \mathbf{0}, \widetilde{\mathbf{V}}_0^{-1} = \frac{1}{\lambda}\mathbf{I}_{d\times d}, \mathbf{S}_0 = \mathbf{0}_{m\times d}, \delta' = \delta/(4T)$
for $t = 1, 2, \dots$ do
    GET decision set $D_t$
    SAMPLE $\mathbf{Z}_t \sim \mathcal{D}^{\text{TS}}$
    SELECT $\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x}\in D_t} \mathbf{x}^\top \left(\widetilde{\mathbf{w}}_{t-1} + \widetilde{\beta}_t(\delta')\widetilde{\mathbf{V}}_{t-1}^{-\frac{1}{2}}\mathbf{Z}_t\right)$
    OBSERVE reward $Y_t$
    UPDATE $\mathbf{S}_t, \mathbf{H}_t$ using Algorithm 7 given $\mathbf{S}_{t-1}, X_t$
    UPDATE $\widetilde{\mathbf{V}}_t^{-1} \leftarrow \frac{1}{\lambda}\left(\mathbf{I}_{d\times d} - \mathbf{S}_t^\top \mathbf{H}_t \mathbf{S}_t\right)$
    UPDATE $\widetilde{\mathbf{w}}_t$ using (4.4)
end for
```

$$\widetilde{\mathbf{V}}_t^{-\frac{1}{2}} = \mathbf{S}_t'^\top \left(\mathbf{S}_t'\mathbf{S}_t'^\top\right)^{-1} \left(\frac{\lambda}{2}\mathbf{I} + \mathbf{S}_t'\mathbf{S}_t'^\top\right)^{-\frac{1}{2}} \mathbf{S}_t'$$

where

$$\mathbf{S}_t' = \left(\boldsymbol{\Sigma}_t + \left(\frac{\lambda}{2} - \rho_t\right)\mathbf{I}_{m\times m}\right)^{\frac{1}{2}} \mathbf{U}_t .$$

Note that $\widetilde{\mathbf{V}}_t^{-\frac{1}{2}}$ can still be computed in time $\mathcal{O}\big(md + m^2\big)$ because $\mathbf{S}_t'\mathbf{S}_t'^\top$ is a diagonal matrix.

The confidence ellipsoid stated in Theorem 9 applies to any contextual bandit policy, and so also to the $\widetilde{\mathbf{w}}_t$ constructed by sketched linear TS. However, as shown by Abeille and Lazaric [2017], the analysis needs a confidence ellipsoid larger by a factor equal to the bound on $\|\mathbf{Z}\|$ appearing in the concentration property of the TS-sampling distribution. More precisely, the *TS-confidence ellipsoid* is defined by

$$\widetilde{C}_t^{\text{TS}} \equiv \left\{\mathbf{w} \in \mathbb{R}^d \ : \ \|\mathbf{w} - \widetilde{\mathbf{w}}_t\|_{\widetilde{\mathbf{V}}_t} \le \widetilde{\gamma}_t\big(\delta/(4T)\big)\right\}$$

where

$$\widetilde{\gamma}_t(\delta) = \widetilde{\beta}_t(\delta)\sqrt{cd\ln\left(\frac{c'd}{\delta}\right)}\,. \tag{4.9}$$

The quantity $\widetilde{\beta}_t(\delta)$ is defined in (4.7) and c, c' are the concentration constants of the TS-sampling distribution (Definition 1). We are now ready to prove a bound on the regret of linear TS with FD-sketching.

Theorem 12. *The regret of FD-sketched linear TS, run with sketch size m w.h.p. satisfies*

$$R(T, \mathbf{w}^\star) \overset{\widetilde{\mathcal{O}}}{=} M_\lambda\,(1+\varepsilon_m)^{\frac{3}{2}}\,\widetilde{m}\left(R + S\sqrt{\lambda}\right)\sqrt{dT}\,.$$

The proof of Theorem 12 closely follows the analysis of Abeille and Lazaric [2017] with some key modifications due to the sketching operations. For completeness, we include the proof in Section 4.5.3.

4.5 Proofs

4.5.1 Linear algebra and sketching tools

We start by introducting a basic relationship between the correlation matrix of actions $\mathbf{X}_s^\top\mathbf{X}_s$ and its FD-sketched estimate $\mathbf{S}_t^\top\mathbf{S}_t$ with sketch size $m \le d$. Recall that ρ_t is the smallest eigenvalue of $\mathbf{S}_t^\top\mathbf{S}_t$ for $t = 1, \dots, T$ and $\bar{\rho}_t = \rho_1 + \dots + \rho_t$. Recall also that $\widetilde{\mathbf{V}} = \mathbf{S}_t^\top\mathbf{S}_t + \lambda\mathbf{I}$.

Proposition 2. *Let $\mathbf{S}_s$ be the matrix computed by FD-sketching at time step $s = 1, \dots, t$ (where $\mathbf{S}_0 = \mathbf{0}$). Then $\mathbf{V}_s = \mathbf{X}_s^\top\mathbf{X}_s = \mathbf{S}_s^\top\mathbf{S}_s + \bar{\rho}_s\mathbf{I}$.*

Proof. By construction, $\mathbf{S}_{s-1}^\top\mathbf{S}_{s-1} + \mathbf{x}_s\mathbf{x}_s^\top = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{U}_s^\top$ where $\mathbf{S}_s = \left(\mathbf{\Sigma}_s - \rho_s\mathbf{I}_{m\times m}\right)^{\frac{1}{2}}\mathbf{U}_s$. Thus,

$$\mathbf{S}_s^\top\mathbf{S}_s = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{U}_s^\top - \rho_s\mathbf{I} = \mathbf{S}_{s-1}^\top\mathbf{S}_{s-1} + \mathbf{x}_s\mathbf{x}_s^\top - \rho_s\mathbf{I}$$

Summing both sides of the above over $s = 1, \dots, t$ we get

$$\mathbf{S}_t^\top\mathbf{S}_t = \sum_{s=1}^t \mathbf{x}_s\mathbf{x}_s^\top - \sum_{s=1}^t \rho_s\mathbf{I}$$

which implies the desired result. □

In the following lemma, we show a sketch-specific version of the determinant-trace inequality (Lemma 17). When the spectral error is small, the right-hand side of the inequality depends on the sketch size m rather than the ambient dimension d.

Lemma 2.

$$\ln\left(\frac{\det(\mathbf{V}_t^\lambda)}{\det(\lambda\mathbf{I})}\right) \le d\ln\left(1+\frac{\bar{\rho}}{\lambda}\right) + m\ln\left(1+\frac{tL^2}{m\lambda}\right)\,.$$

Proof. Let $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_d \geq 0$ be the eigenvalues of $\mathbf{S}_t^\top \mathbf{S}_t$. We start by looking at the ratio of determinants. Using Proposition 2 we can write

$$\begin{aligned} \frac{\det(\mathbf{V}_t^\lambda)}{\det(\lambda \mathbf{I})} = \frac{\det\big(\mathbf{S}_s^\top \mathbf{S}_s + \bar{\rho}_s \mathbf{I} + \lambda \mathbf{I}\big)}{\det(\lambda \mathbf{I})} &= \prod_{i=1}^{d} \left(\frac{\widetilde{\lambda}_i}{\lambda} + 1 + \frac{\bar{\rho}}{\lambda} \right) \\ &= \left(1 + \frac{\bar{\rho}}{\lambda}\right)^{d-m} \prod_{i=1}^{m} \left(\frac{\widetilde{\lambda}_i}{\lambda} + 1 + \frac{\bar{\rho}}{\lambda} \right) \end{aligned} \tag{4.10}$$

since $\widetilde{\lambda}_{m+1} = \cdots = \widetilde{\lambda}_d = 0$ because $\mathbf{S}_t^\top \mathbf{S}_t$ has rank at most m. Using the AM-GM inequality (Lemma 16 in the Appendix), the product in (4.10) can be bounded as

$$\begin{aligned} \prod_{i=1}^{m} \left(\frac{\widetilde{\lambda}_i}{\lambda} + 1 + \frac{\bar{\rho}}{\lambda} \right) &\leq \left(1 + \frac{\bar{\rho}}{\lambda} + \frac{1}{m\lambda} \sum_{i=1}^{m} \widetilde{\lambda}_i \right)^m = \left(1 + \frac{\bar{\rho}}{\lambda} + \frac{\operatorname{tr}(\mathbf{S}_t^\top \mathbf{S}_t)}{m\lambda} \right)^m \\ &\leq \left(1 + \frac{\bar{\rho}}{\lambda} + \frac{tL^2}{m\lambda} \right)^m \end{aligned} \tag{4.11}$$

where the last inequality holds because

$$\begin{aligned} \operatorname{tr}(\mathbf{S}_t^\top \mathbf{S}_t) = \operatorname{tr}\left(\widetilde{\mathbf{V}}_t - \lambda \mathbf{I} \right) &\leq \operatorname{tr}\left(\mathbf{V}_t^\lambda - \lambda \mathbf{I} \right) && \text{(by Proposition 2)} \\ &= \sum_{s=1}^{t} \operatorname{tr}(\mathbf{x}_s \mathbf{x}_s^\top) \leq tL^2 \,. && \text{(by definition of } \mathbf{V}_t\text{)} \end{aligned}$$

Finally, substituting (4.11) into (4.10) and taking logs on both sides gives

$$\begin{aligned} \ln\left(\frac{\det(\mathbf{V}_t^\lambda)}{\det(\lambda \mathbf{I})} \right) &\leq (d-m) \ln\left(1 + \frac{\bar{\rho}}{\lambda}\right) + m \ln\left(1 + \frac{\bar{\rho}}{\lambda} + \frac{tL^2}{m\lambda} \right) \\ &= d \ln\left(1 + \frac{\bar{\rho}}{\lambda}\right) + m \ln\left(1 + \frac{\frac{tL^2}{m\lambda}}{1 + \frac{\bar{\rho}}{\lambda}} \right) \\ &\leq d \ln\left(1 + \frac{\bar{\rho}}{\lambda}\right) + m \ln\left(1 + \frac{tL^2}{m\lambda} \right) \end{aligned}$$

concluding the proof. □

We start with the proof of a simple lemma that is used in the definition of OFUL (see Algorithm 3).

Lemma 3. *For any positive definite $d \times d$ matrix $\mathbf{A}$, for any $\mathbf{w}_0, \mathbf{x} \in \mathbb{R}^d$ and $c > 0$, the solution of*

$$\begin{aligned} \max_{\mathbf{w} \in \mathbb{R}^d} \quad & \mathbf{w}^\top \mathbf{x} \\ \textit{s.t.} \quad & \|\mathbf{w} - \mathbf{w}_0\|_{\mathbf{A}} \leq c \end{aligned}$$

has value $\mathbf{w}_0^\top \mathbf{x} + c \, \|\mathbf{x}\|_{\mathbf{A}^{-1}}$.

Proof. Let $\mathbf{u} = \mathbf{A}^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}_0)$ so that $\mathbf{w} = \mathbf{A}^{-\frac{1}{2}}\mathbf{u} + \mathbf{w}_0$. Then the optimization problem can be equivalently rewritten as

$$\begin{aligned} \max_{\mathbf{w} \in \mathbb{R}^d} \quad & \mathbf{u}^\top \mathbf{A}^{-\frac{1}{2}} \mathbf{x} + \mathbf{w}_0^\top \mathbf{x} \\ \text{s.t.} \quad & \|\mathbf{u}\| \le c \end{aligned}$$

Then the solution is clearly $\mathbf{u} = c\,\mathbf{A}^{-\frac{1}{2}}\mathbf{x} / \|\mathbf{x}\|_{\mathbf{A}^{-1}}$, which achieves the claimed value. □

Our regret analyses follow Abbasi-Yadkori et al. [2011], Abeille and Lazaric [2017] and related works. However, due to the sketching of the correlation matrix, some key components of the proofs now depend on the spectral error (4.5). In Section 4.5.1, we present tools specific to the analysis of linear bandits with FD-sketching. These tools are used to bound the instantaneous regret $\left(\mathbf{x}^\star - \mathbf{x}_t\right)^\top \mathbf{w}^\star$ in terms of the norm $\|\mathbf{w}^\star - \widetilde{\mathbf{w}}_t\|_{\widetilde{\mathbf{V}}_{t-1}}$ and the ridge leverage scores $\sum_{t=1}^T \|\mathbf{x}_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}$. Armed with these results, we then prove our regret bounds in Sections 4.5.2 and 4.5.3.

Next, we recall some standard tools from the analysis of linear bandits. All results in Section 7.2 are by Abbasi-Yadkori et al. [2011]. The next lemma is similar to Abbasi-Yadkori et al. [2011, Lemma 11]. However, now the statement depends on the sketched matrix $\widetilde{\mathbf{V}}_{t-1}$ instead of $\mathbf{V}_{t-1}^\lambda$. Although we pay in terms of the spectral error ε_m, we also improve the dependence on the dimension from d to m whenever ε_m is sufficiently small.

Lemma 4 (Sketched leverage scores)**.**

$$\sum_{t=1}^T \min\left\{1, \|\mathbf{x}_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} \le 2\left(1+\varepsilon_m\right)\left(d \ln\left(1+\varepsilon_m\right) + m \ln\left(1 + \frac{TL^2}{m\lambda}\right)\right) . \tag{4.12}$$

Proof. Throughout the proof, unless stated explicitly, we drop the subscripts containing t. Therefore, $\mathbf{V} = \mathbf{V}_{t-1}^\lambda$, $\widetilde{\mathbf{V}} = \widetilde{\mathbf{V}}_{t-1}$, $\mathbf{x} = \mathbf{x}_t$, and $\bar{\rho} = \bar{\rho}_{t-1}$. Now suppose that $(\widetilde{\lambda}_i + \lambda, \widetilde{\mathbf{u}}_i)$ is an i-th eigenpair of $\widetilde{\mathbf{V}}$. Then, Proposition 2 implies that a corresponding eigenpair of $\mathbf{V}$ is $(\widetilde{\lambda}_i + \lambda + \bar{\rho}, \widetilde{\mathbf{u}}_i)$. Using this fact we have that

$$\begin{aligned} \|\mathbf{x}\|^2_{\mathbf{V}^{-1}} &= \mathbf{x}^\top \widetilde{\mathbf{V}} \widetilde{\mathbf{V}}^{-1} \mathbf{V}^{-1} \mathbf{x} \\ &= \mathbf{x}^\top \left(\sum_{i=1}^d \widetilde{\mathbf{u}}_i \widetilde{\mathbf{u}}_i^\top \frac{1}{\widetilde{\lambda}_i + \lambda} \frac{\widetilde{\lambda}_i + \lambda}{\widetilde{\lambda}_i + \lambda + \bar{\rho}} \right) \mathbf{x} \\ &\ge \frac{\lambda}{\lambda + \bar{\rho}} \mathbf{x}^\top \left(\sum_{i=1}^d \widetilde{\mathbf{u}}_i \widetilde{\mathbf{u}}_i^\top \frac{1}{\widetilde{\lambda}_i + \lambda} \right) \mathbf{x} = \frac{\lambda}{\lambda + \bar{\rho}} \|\mathbf{x}\|^2_{\widetilde{\mathbf{V}}^{-1}} . \end{aligned}$$

Furthermore, this implies that

$$\begin{aligned} & \min\left\{1, \frac{\lambda}{\lambda + \bar{\rho}} \|\mathbf{x}\|^2_{\widetilde{\mathbf{V}}^{-1}}\right\} \le \min\left\{1, \|\mathbf{x}\|^2_{\mathbf{V}^{-1}}\right\} \\ \implies \quad & \min\left\{1 + \frac{\bar{\rho}}{\lambda}, \|\mathbf{x}\|^2_{\widetilde{\mathbf{V}}^{-1}}\right\} \le \left(1 + \frac{\bar{\rho}}{\lambda}\right) \min\left\{1, \|\mathbf{x}\|^2_{\mathbf{V}^{-1}}\right\} \\ & \qquad\qquad \text{(multiply both sides by } 1 + \tfrac{\bar{\rho}}{\lambda}\text{)} \\ \implies \quad & \min\left\{1, \|\mathbf{x}\|^2_{\widetilde{\mathbf{V}}^{-1}}\right\} \le \left(1 + \frac{\bar{\rho}}{\lambda}\right) \min\left\{1, \|\mathbf{x}\|^2_{\mathbf{V}^{-1}}\right\} . \end{aligned}$$

Finally, combining the above with Lemma 18, equation (7.3), and using the fact that $\bar{\rho}_{t-1} \leq \bar{\rho}_T$, we obtain

$$
\begin{aligned}
\sum_{t=1}^{T} \min\left\{1, \|\mathbf{x}_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} &\leq 2\left(1+\frac{\bar{\rho}_T}{\lambda}\right) \ln\left(\frac{\det(\mathbf{V}_T^{\lambda})}{\det(\lambda I)}\right) \\
&\leq 2\left(1+\frac{\bar{\rho}_T}{\lambda}\right)\left(d \ln\left(1+\frac{\bar{\rho}_T}{\lambda}\right) + m \ln\left(1+\frac{TL^2}{m\lambda}\right)\right) \\
&\qquad \text{(by Lemma 2)} \\
&\leq 2\left(1+\varepsilon_m\right)\left(d \ln\left(1+\varepsilon_m\right) + m \ln\left(1+\frac{TL^2}{m\lambda}\right)\right)
\end{aligned}
$$

where the last inequality follows from Proposition 1. □

Now we prove Theorem 9, characterizing the confidence ellipsoid generated by the sketched estimate.

Theorem 9 (Sketched confidence ellipsoid – restated). *For any $\delta \in (0,1)$, the optimal parameter $\mathbf{w}^\star$ belongs to the set*

$$
\widetilde{C}_t \equiv \left\{\mathbf{w} \in \mathbb{R}^d \ : \ \|\mathbf{w} - \widetilde{\mathbf{w}}_t\|_{\widetilde{\mathbf{V}}_t} \leq \widetilde{\beta}_t(\delta)\right\}
$$

with probability at least $1-\delta$, where

$$
\begin{aligned}
\widetilde{\beta}_t(\delta) &= R\sqrt{m \ln\left(1+\frac{tL^2}{m\lambda}\right) + 2\ln\left(\frac{1}{\delta}\right) + d \ln\left(1+\frac{\bar{\rho}_t}{\lambda}\right)}\sqrt{1+\frac{\bar{\rho}_t}{\lambda}} + S\sqrt{\lambda}\left(1+\frac{\bar{\rho}_t}{\lambda}\right) \\
&\overset{\widetilde{\mathcal{O}}}{=} R\sqrt{\left(m + d\ln(1+\varepsilon_m)\right)\left(1+\varepsilon_m\right)} + S\sqrt{\lambda}\left(1+\varepsilon_m\right) \ .
\end{aligned}
$$

Proof. Throughout the proof we frequently use Proposition 2, implying $\mathbf{X}_t^\top \mathbf{X}_t = \mathbf{S}_t^\top \mathbf{S}_t + \bar{\rho}_t \mathbf{I}$. For brevity, in the following we drop subscripts containing t in matrices. Let $\eta_t = (\eta_1, \eta_2 \ldots, \eta_t)$, and by definition of the sketched estimate we have that

$$
\begin{aligned}
\widetilde{\mathbf{w}}_t &= \left(\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}\right)^{-1} \mathbf{X}_t^\top \left(\mathbf{X}_t \mathbf{w}^\star + \eta_t\right) \\
&= \left(\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}\right)^{-1} \mathbf{X}_t^\top \eta_t + \left(\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}\right)^{-1} \mathbf{X}_t^\top \mathbf{X}_t \mathbf{w}^\star \\
&= \left(\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}\right)^{-1} \mathbf{X}_t^\top \eta_t \\
&\quad + \left(\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}\right)^{-1} \left(\mathbf{X}_t^\top \mathbf{X}_t + (\lambda - \bar{\rho}_t)\mathbf{I}\right) \mathbf{w}^\star - (\lambda - \bar{\rho}_t)\left(\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}\right)^{-1} \mathbf{w}^\star \\
&= \left(\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}\right)^{-1} \mathbf{X}_t^\top \eta_t + \mathbf{w}^\star - (\lambda - \bar{\rho}_t)\left(\mathbf{S}_t^\top \mathbf{S}_t + \lambda \mathbf{I}\right)^{-1} \mathbf{w}^\star \\
&= \widetilde{\mathbf{V}}_t^{-1} \mathbf{X}_t^\top \eta_t + \mathbf{w}^\star - (\lambda - \bar{\rho}_t)\widetilde{\mathbf{V}}_t^{-1} \mathbf{w}^\star \ .
\end{aligned} \tag{4.13}
$$

Then, by (4.13), for any $\mathbf{x} \in \mathbb{R}^d$ we have that

$$\begin{aligned}
\mathbf{x}^\top \left(\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\right) &= \mathbf{x}^\top \widetilde{\mathbf{V}}_t^{-1} \mathbf{X}_t^\top \eta_t - (\lambda - \bar{\rho}_t) \mathbf{x}^\top \widetilde{\mathbf{V}}_t^{-1} \mathbf{w}^\star \qquad (4.14) \\
&\leq \left\|\mathbf{x}^\top \widetilde{\mathbf{V}}_t^{-1}\right\|_{\mathbf{V}_t^\lambda} \|\mathbf{X}_t^\top \eta_t\|_{(\mathbf{V}_t^\lambda)^{-1}} - (\lambda - \bar{\rho}_t) \left\langle \mathbf{x}, \mathbf{w}^\star \right\rangle_{\widetilde{\mathbf{V}}_t^{-1}} \\
&\qquad \text{(by Cauchy-Schwartz)} \\
&\leq \left\|\mathbf{x}^\top \widetilde{\mathbf{V}}_t^{-1}\right\|_{\mathbf{V}_t^\lambda} \|\mathbf{X}_t^\top \eta_t\|_{(\mathbf{V}_t^\lambda)^{-1}} + |\lambda + \bar{\rho}_t| \left| \left\langle \mathbf{x}, \mathbf{w}^\star \right\rangle_{\widetilde{\mathbf{V}}_t^{-1}} \right| \\
&\qquad \text{(by the triangle inequality.)}
\end{aligned}$$

We now choose $\mathbf{x} = \widetilde{\mathbf{V}}_t (\widetilde{\mathbf{w}}_t - \mathbf{w}^\star)$ and proceed by bounding terms in the above. By the choice of $\mathbf{x}$, we have that $\mathbf{x}^\top \left(\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\right) = \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}^2$, $\left\|\mathbf{x}^\top \widetilde{\mathbf{V}}_t^{-1}\right\|_{\mathbf{V}_t^\lambda} = \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\mathbf{V}_t^\lambda}$ and

$$\begin{aligned}
\left\langle \mathbf{x}, \mathbf{w}^\star \right\rangle_{\widetilde{\mathbf{V}}_t^{-1}} &= \left(\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\right)^\top \mathbf{w}^\star \leq \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_2 \, \|\mathbf{w}^\star\|_2 \qquad \text{(by Cauchy-Schwartz)} \\
&\leq \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_2 \, S \, .
\end{aligned}$$

Finally, by Lemma 19, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\|\mathbf{X}^\top \eta_t\|_{(\mathbf{V}_t^\lambda)^{-1}} \leq \sqrt{B_t(\delta)} \qquad \forall t \geq 0 \, .$$

The left-hand side of (4.14) can now upper bounded as

$$\begin{aligned}
& \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}^2 \leq \sqrt{B_t(\delta)} \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\mathbf{V}_t^\lambda} + S(\lambda + \bar{\rho}_t) \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_2 \\
\implies \quad & \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t} \leq \sqrt{B_t(\delta)} \frac{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\mathbf{V}_t^\lambda}}{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}} + S(\lambda + \bar{\rho}_t) \frac{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_2}{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}} \, . \qquad (4.15)
\end{aligned}$$

Now we handle the ratios of norms in the right-hand side of (4.15). First,

$$\begin{aligned}
\frac{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\mathbf{V}_t^\lambda}}{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}} &= \sqrt{\frac{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}^2 + \bar{\rho}_t \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_2^2}{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}^2}} \\
&= \sqrt{1 + \bar{\rho}_t \frac{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_2^2}{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}^2}} \leq \sqrt{1 + \frac{\bar{\rho}_t}{\lambda}}
\end{aligned}$$

since $\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}^2 \geq \lambda \|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_2^2$ and, using the same reasoning,

$$\frac{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_2}{\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t}} \leq \frac{1}{\sqrt{\lambda}} \, .$$

Substituting these into (4.15) gives

$$\|\widetilde{\mathbf{w}}_t - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_t} \leq \sqrt{B_t(\delta) \left(1 + \frac{\bar{\rho}_t}{\lambda}\right)} + S\sqrt{\lambda} \left(1 + \frac{\bar{\rho}_t}{\lambda}\right) \, .$$

Now we provide a deterministic bound on $B_t(\delta)$. Using Lemma 2 we have

$$\begin{aligned}\sqrt{B_t(\delta)} &= R\sqrt{2\ln\left(\frac{1}{\delta}\det\left(\mathbf{V}_t^{\lambda}\right)^{\frac{1}{2}}\det\left(\lambda\mathbf{I}\right)^{-\frac{1}{2}}\right)} \\ &\le R\sqrt{d\ln\left(1+\frac{\bar{\rho}_t}{\lambda}\right)+m\ln\left(1+\frac{tL^2}{m\lambda}\right)+2\ln\left(\frac{1}{\delta}\right)}\ .\end{aligned}$$

This proves the first statement (4.7). Finally, (4.8) follows by Proposition 1. □

We close this section by computing a closed form for $\widetilde{\mathbf{V}}_t^{-\frac{1}{2}}$, the square root of the inverse of the sketched correlation matrix. This is used by sketched linear TS for selecting actions. We make use of the generalized Woodbury matrix identity (see Lemma 15) to prove the following:

Corollary 1. *For $\lambda > 0$, let*

$$\mathbf{S}'_t = \left(\mathbf{\Sigma}_t + \left(\frac{\lambda}{2}-\rho_t\right)\mathbf{I}_{m\times m}\right)^{\frac{1}{2}}\mathbf{U}_t\ .$$

Then

$$\widetilde{\mathbf{V}}_t^{-\frac{1}{2}} = \mathbf{S}_t'^{\top}\left(\mathbf{S}'_t\mathbf{S}_t'^{\top}\right)^{-1}\left(\frac{\lambda}{2}\mathbf{I}+\mathbf{S}'_t\mathbf{S}_t'^{\top}\right)^{-\frac{1}{2}}\mathbf{S}'_t\ .$$

Proof. We apply Lemma 15 with $f(\widetilde{\mathbf{V}}) = \widetilde{\mathbf{V}}_t^{-\frac{1}{2}}$. However, since $\mathbf{S}_t\mathbf{S}_t^{\top}$ is singular by design, we apply the theorem with $\mathbf{B}$ set the non-singular proxy matrix $\mathbf{S}'_t$, $\mathbf{A}$ set to $\mathbf{S}_t'^{\top}$, and α set to $\lambda/2$. Thus $\widetilde{\mathbf{V}}_t = \mathbf{S}_t'^{\top}\mathbf{S}'_t + \frac{\lambda}{2}\mathbf{I}_{d\times d}$ and

$$\begin{aligned}\left(\mathbf{S}_t'^{\top}\mathbf{S}'_t+\frac{\lambda}{2}\mathbf{I}_{d\times d}\right)^{-\frac{1}{2}} &= \sqrt{\frac{2}{\lambda}}\mathbf{I}_{m\times m}+\mathbf{S}_t'^{\top}\left(\mathbf{S}'_t\mathbf{S}_t'^{\top}\right)^{-1}\left(\left(\frac{\lambda}{2}\mathbf{I}_{m\times m}+\mathbf{S}'_t\mathbf{S}_t'^{\top}\right)^{-\frac{1}{2}}-\sqrt{\frac{2}{\lambda}}\mathbf{I}_{m\times m}\right)\mathbf{S}'_t \\ &= \mathbf{S}_t'^{\top}\left(\mathbf{S}'_t\mathbf{S}_t'^{\top}\right)^{-1}\left(\frac{\lambda}{2}\mathbf{I}_{m\times m}+\mathbf{S}'_t\mathbf{S}_t'^{\top}\right)^{-\frac{1}{2}}\mathbf{S}'_t \qquad (4.16)\end{aligned}$$

where (4.16) follows since $\mathbf{S}_t'^{\top}\left(\mathbf{S}'_t\mathbf{S}_t'^{\top}\right)^{-1}\mathbf{S}'_t = \mathbf{I}_{m\times m}$. □

4.5.2 Proof of the regret bound for SOFUL (Theorem 10)

We start with a preliminary lemma.

Lemma 5. *For any $\delta > 0$, the instantaneous regret of SOFUL satisfies*

$$(\mathbf{x}_t^{\star}-\mathbf{x}_t)^{\top}\mathbf{w}^{\star} \le 2\widetilde{\beta}_{t-1}(\delta)\|\mathbf{x}_t\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}} \qquad t=1,\dots,T\ .$$

Proof. Let $\widetilde{\mathbf{w}}^{\mathrm{so}}_{t-1}$ be the FD-sketched RLS estimate of OFUL (Algorithm 8). Recall that the optimal action at time t is $\mathbf{x}^\star_t = \arg\max_{\mathbf{x}\in D_t} \mathbf{x}^\top \mathbf{w}^\star$, whereas

$$\left(\mathbf{x}_t, \widetilde{\mathbf{w}}^{\mathrm{so}}_{t-1}\right) = \mathop{\arg\max}_{(\mathbf{x},\mathbf{w})\in D_t\times\widetilde{C}_{t-1}} \mathbf{x}^\top \mathbf{w} \ .$$

We use these facts to bound the instantaneous regret,

$$\begin{aligned}
\left(\mathbf{x}^\star_t - \mathbf{x}_t\right)^\top \mathbf{w}^\star &\le \mathbf{x}_t^\top \widetilde{\mathbf{w}}^{\mathrm{so}}_{t-1} - \mathbf{x}_t^\top \mathbf{w}^\star \\
&= \mathbf{x}_t^\top \left(\widetilde{\mathbf{w}}^{\mathrm{so}}_{t-1} - \mathbf{w}^\star\right) \\
&= \mathbf{x}_t^\top \left(\widetilde{\mathbf{w}}^{\mathrm{so}}_{t-1} - \widetilde{\mathbf{w}}_{t-1}\right) + \mathbf{x}_t^\top \left(\widetilde{\mathbf{w}}_{t-1} - \mathbf{w}^\star\right) \\
&\le \|\mathbf{x}_t\|_{\widetilde{\mathbf{V}}^{-1}_{t-1}} \left(\|\widetilde{\mathbf{w}}^{\mathrm{so}}_{t-1} - \widetilde{\mathbf{w}}_{t-1}\|_{\widetilde{\mathbf{V}}_{t-1}} + \|\widetilde{\mathbf{w}}_{t-1} - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_{t-1}}\right) \\
&\qquad\qquad \text{(by Cauchy-Schwartz)} \\
&\le 2\widetilde{\beta}_{t-1}(\delta)\|\mathbf{x}_t\|_{\widetilde{\mathbf{V}}^{-1}_{t-1}} \qquad \text{(by Theorem 9)}
\end{aligned}$$

concluding the proof. □

Now we are ready to prove the regret bound.

Proof of Theorem 10. Bounding the regret using Lemma 5 gives

$$\begin{aligned}
R(T, \mathbf{w}^\star) &= \sum_{t=1}^T \left(\mathbf{x}^\star_t - \mathbf{x}_t\right)^\top \mathbf{w}^\star \\
&\le 2\sum_{t=1}^T \min\left\{LS, \widetilde{\beta}_{t-1}(\delta)\|\mathbf{x}_t\|_{\widetilde{\mathbf{V}}^{-1}_{t-1}}\right\} \\
&\qquad \left(\text{since } \max_{t=1,\dots,T} \max_{\mathbf{x}\in D_t} |\mathbf{x}^\top \mathbf{w}^\star| \le LS \text{ by Cauchy-Schwartz}\right) \\
&\le 2\sum_{t=1}^T \widetilde{\beta}_{t-1}(\delta) \min\left\{\frac{L}{\sqrt{\lambda}}, \|\mathbf{x}_t\|_{\widetilde{\mathbf{V}}^{-1}_{t-1}}\right\} \quad \left(\text{since } \min_{t=0,\dots,T-1} \min_{\delta\in[0,1]} \widetilde{\beta}_t(\delta) \ge S\sqrt{\lambda}\right) \\
&\le 2\left(\max_{t=0,\dots,T-1} \widetilde{\beta}_t(\delta)\right) \sum_{t=1}^T \min\left\{\frac{L}{\sqrt{\lambda}}, \|\mathbf{x}_t\|_{\widetilde{\mathbf{V}}^{-1}_{t-1}}\right\} \\
&\le 2\max\left\{1, \frac{L}{\sqrt{\lambda}}\right\} \left(\max_{t=0,\dots,T-1} \widetilde{\beta}_t(\delta)\right) \sum_{t=1}^T \min\left\{1, \|\mathbf{x}_t\|_{\widetilde{\mathbf{V}}^{-1}_{t-1}}\right\} \\
&\le 2\max\left\{1, \frac{L}{\sqrt{\lambda}}\right\} \left(\max_{t=0,\dots,T-1} \widetilde{\beta}_t(\delta)\right) \sqrt{T\sum_{t=1}^T \min\left\{1, \|\mathbf{x}_t\|^2_{\widetilde{\mathbf{V}}^{-1}_{t-1}}\right\}} \ . \\
&\qquad\qquad \text{(by Cauchy-Schwartz)}
\end{aligned}$$

Now we finish by further bounding the terms in the above. In particular, we bound $\widetilde{\beta}_t(\delta)$ by (4.8)

$$\max_{t=0,\dots,T-1} \widetilde{\beta}_t(\delta) \overset{\widetilde{\mathcal{O}}}{=} R\sqrt{\left(m + d\ln(1+\varepsilon_m)\right)(1+\varepsilon_m)} + S\sqrt{\lambda}\,(1+\varepsilon_m)$$

while the bound on the summation term uses Lemma 4,

$$\sqrt{\sum_{t=1}^{T} \min\left\{1, \|X_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\}} \overset{\widetilde{\mathcal{O}}}{=} \sqrt{(1+\varepsilon_m)\left(d \ln(1+\varepsilon_m)+m\right)}\ .$$

Then, using $M_\lambda = \max\left\{1, \frac{L}{\sqrt{\lambda}}\right\}$ and $\widetilde{m} = m + d\ln(1+\varepsilon_m)$,

$$\begin{aligned} R(T, \mathbf{w}^\star) &\overset{\widetilde{\mathcal{O}}}{=} M_\lambda \sqrt{T}\left(R\sqrt{\widetilde{m}\,(1+\varepsilon_m)} + S\sqrt{\lambda}\,(1+\varepsilon_m)\right)\sqrt{\widetilde{m}\,(1+\varepsilon_m)} \\ &\overset{\widetilde{\mathcal{O}}}{=} M_\lambda \sqrt{T}\Big(R\,\widetilde{m}\,(1+\varepsilon_m) + S\sqrt{\lambda}\,(1+\varepsilon_m)^{\frac{3}{2}}\sqrt{\widetilde{m}}\Big) \\ &\overset{\widetilde{\mathcal{O}}}{=} M_\lambda\,(1+\varepsilon_m)^{\frac{3}{2}}\,\widetilde{m}\left(R + S\sqrt{\lambda}\right)\sqrt{T} \end{aligned}$$

which completes the proof. □

Proof of Theorem 11. Recall that

$$\Delta \leq \min_{t=1,\ldots,T} \left(\mathbf{x}_t^\star - \mathbf{x}_t\right)^\top \mathbf{w}^\star\ .$$

Similarly to the proof of Theorem 10, we use Lemma 5 to bound the instantaneous regret. However, we first use the gap assumption to bound the regret in terms of the sum of squared instantaneous regrets,

$$\begin{aligned} R(T, \mathbf{w}^\star) &= \sum_{t=1}^{T} \left(\mathbf{x}_t^\star - \mathbf{x}_t\right)^\top \mathbf{w}^\star \\ &\leq \frac{1}{\Delta}\sum_{t=1}^{T}\left(\left(\mathbf{x}_t^\star - \mathbf{x}_t\right)^\top \mathbf{w}^\star\right)^2 \\ &\leq \frac{2}{\Delta}\sum_{t=1}^{T}\min\left\{2L^2S^2, \widetilde{\beta}_{t-1}(\delta)^2\|\mathbf{x}_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} &\quad (4.17) \\ &\leq \frac{2}{\Delta}\left(\max_{t=0,\ldots,T-1}\widetilde{\beta}_t(\delta)^2\right)\sum_{t=1}^{T}\min\left\{\frac{2L^2}{\lambda}, \|\mathbf{x}_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} &\quad (4.18) \\ &\leq \frac{2}{\Delta}\max\left\{1, \frac{2L^2}{\lambda}\right\}\left(\max_{t=0,\ldots,T-1}\widetilde{\beta}_t(\delta)^2\right)\sum_{t=1}^{T}\min\left\{1, \|\mathbf{x}_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} &\quad (4.19) \end{aligned}$$

where (4.18) holds because $\min_t \min_\delta \widetilde{\beta}_t(\delta)^2 \geq S^2\lambda$. Inequality (4.17) holds because

$$\begin{aligned} \left(\left(\mathbf{x}_t^\star - \mathbf{x}_t\right)^\top \mathbf{w}^\star\right)^2 &\leq 2(\mathbf{x}_t^{\star\top}\mathbf{w}^\star)^2 + 2(\mathbf{x}_t^\top \mathbf{w}^\star)^2 \\ &\leq 4L^2S^2 &\quad \text{(by Cauchy-Schwartz)} \end{aligned}$$

and because of Lemma 5.

We now finish bounding the regret by further bounding the individual terms in (4.19). In particular, we use (4.8) to bound $\widetilde{\beta}_t(\delta)$ as follows

$$\begin{aligned}
\max_{t=0,\dots,T-1} \widetilde{\beta}_t(\delta)^2 &\overset{\widetilde{\mathcal{O}}}{=} R^2 \left(\sqrt{\big(m + d\ln(1+\varepsilon_m)\big)\,(1+\varepsilon_m)} + S\sqrt{\lambda}\,(1+\varepsilon_m) \right)^2 \\
&\overset{\widetilde{\mathcal{O}}}{=} R^2\big(m + d\ln(1+\varepsilon_m)\big)\,(1+\varepsilon_m) + S^2\lambda\,(1+\varepsilon_m)^2 \ .
\end{aligned}$$

Lemma 4 gives

$$\sum_{t=1}^{T} \min\left\{1, \|X_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} \overset{\widetilde{\mathcal{O}}}{=} (1+\varepsilon_m)\,\Big(m\ln(T) + d\ln\left(1+\varepsilon_m\right)\Big) \ .$$

Then, using again $M_\lambda = \max\left\{1, \frac{L}{\sqrt{\lambda}}\right\}$ and $\widetilde{m} = m + d\ln(1+\varepsilon_m)$,

$$\begin{aligned}
R(T,\mathbf{w}^\star) &\overset{\widetilde{\mathcal{O}}}{=} \frac{M_\lambda^2}{\Delta}\left(R^2\widetilde{m}\,(1+\varepsilon_m) + S^2\lambda\,(1+\varepsilon_m)^2\right)(1+\varepsilon_m)\,\widetilde{m} \\
&\overset{\widetilde{\mathcal{O}}}{=} \frac{M_\lambda^2}{\Delta}\left(\widetilde{m}R^2 + S^2\lambda\right)(1+\varepsilon_m)^3\,\widetilde{m} \overset{\widetilde{\mathcal{O}}}{=} \frac{M_\lambda^2}{\Delta}\left(R^2 + S^2\lambda\right)(1+\varepsilon_m)^3\,\widetilde{m}^2
\end{aligned}$$

concluding the proof. □

4.5.3 Proof of the regret bound for Sketched Linear TS (Theorem 12)

Here $\widetilde{\mathbf{w}}^{\text{TS}}_{t-1}$ is used to denote the FD-sketched RLS estimate of linear TS (Algorithm 9). As in Abeille and Lazaric [2017], we split the regret as follows

$$\begin{aligned}
R(T,\mathbf{w}^\star) &= \sum_{t=1}^{T}\left(\mathbf{x}^\star_t - \mathbf{x}_t\right)^\top\mathbf{w}^\star = \sum_{t=1}^{T}\left(\mathbf{x}_t^{\star\top}\mathbf{w}^\star - \mathbf{x}_t^\top\widetilde{\mathbf{w}}^{\text{TS}}_{t-1}\right) + \sum_{t=1}^{T}\left(\mathbf{x}_t^\top\widetilde{\mathbf{w}}^{\text{TS}}_{t-1} - \mathbf{x}_t^\top\mathbf{w}^\star\right) \\
&= \sum_{t=1}^{T}\left(J_t(\mathbf{w}^\star) - J_t(\widetilde{\mathbf{w}}^{\text{TS}}_{t-1})\right) + \sum_{t=1}^{T}\left(\mathbf{x}_t^\top\widetilde{\mathbf{w}}^{\text{TS}}_{t-1} - \mathbf{x}_t^\top\mathbf{w}^\star\right) \qquad (4.20)
\end{aligned}$$

where

$$J_t(\mathbf{w}) = \max_{\mathbf{x}\in D_t}\mathbf{x}^\top\mathbf{w}$$

is an "optimistic" reward function. Most of the proof is concerned with bounding the first term in (4.20). The second term is instead obtained in way similar to the analysis of OFUL. Fix any $\delta \in (0,1)$, let $\delta' = \frac{\delta}{4T}$, and introduce events

$$\begin{aligned}
\widetilde{E}_t &\equiv \left\{\|\widetilde{\mathbf{w}}_s - \mathbf{w}^\star\| \le \widetilde{\beta}_s(\delta'),\ s = 1,\dots,t\right\} \\
\widetilde{E}^{\text{TS}}_t &\equiv \left\{\|\widetilde{\mathbf{w}}^{\text{TS}}_s - \widetilde{\mathbf{w}}_s\| \le \widetilde{\gamma}_s(\delta'),\ s = 1,\dots,t\right\}
\end{aligned}$$

and $E_t \equiv \widetilde{E}_t \cap \widetilde{E}^{\text{TS}}_t$. Observe that, by definition,

$$\widetilde{E}_T \subset \cdots \subset \widetilde{E}_1 \qquad \text{and} \qquad \widetilde{E}^{\text{TS}}_T \subset \cdots \subset \widetilde{E}^{\text{TS}}_1 \qquad (4.21)$$

We also use the following lower bound on the probability of E_T.

Lemma 6. $\mathbb{P}\left(E_T\right) \geq 1 - \frac{\delta}{2}$.

Proof. The proof is identical to the proof of Abeille and Lazaric [2017, Lemma 1], the only difference being that we use the confidence ellipsoid defined in Theorem 9. □

We study the regret when E_T occurs,

$$\begin{aligned}
\mathbb{I}\{E_T\}\, R(T, \mathbf{w}^\star) &= \sum_{t=1}^{T} \mathbb{I}\{E_T\} \left(J_t(\mathbf{w}^\star) - J_t(\widetilde{\mathbf{w}}_{t-1}^{\text{TS}})\right) + \sum_{t=1}^{T} \mathbb{I}\{E_T\} \left(\mathbf{x}_t^\top \widetilde{\mathbf{w}}_{t-1}^{\text{TS}} - \mathbf{x}_t^\top \mathbf{w}^\star\right) \\
&\leq \sum_{t=1}^{T} \mathbb{I}\{E_{t-1}\} \left(J_t(\mathbf{w}^\star) - J_t(\widetilde{\mathbf{w}}_{t-1}^{\text{TS}})\right) + \sum_{t=1}^{T} \mathbb{I}\{E_{t-1}\} \left(\mathbf{x}_t^\top \widetilde{\mathbf{w}}_{t-1}^{\text{TS}} - \mathbf{x}_t^\top \mathbf{w}^\star\right) \\
&\qquad \text{(using (4.21))} \\
&= \sum_{t=1}^{T} r_t^{\text{TS}} + \sum_{t=1}^{T} r_t^{\text{RLS}} \qquad (4.22)
\end{aligned}$$

where we introduced the notation

$$r_t^{\text{TS}} = \mathbb{I}\{E_{t-1}\} \left(J_t(\mathbf{w}^\star) - J_t(\widetilde{\mathbf{w}}_{t-1}^{\text{TS}})\right) \qquad \text{and} \qquad r_t^{\text{RLS}} = \mathbb{I}\{E_{t-1}\} \left(\mathbf{x}_t^\top \widetilde{\mathbf{w}}_{t-1}^{\text{TS}} - \mathbf{x}_t^\top \mathbf{w}^\star\right) .$$

First we focus on r_t^{TS}, and get that

$$\begin{aligned}
r_t^{\text{TS}} &= \left(J_t(\mathbf{w}^\star) - J_t(\widetilde{\mathbf{w}}_{t-1}^{\text{TS}})\right) \mathbb{I}\{E_{t-1}\} \\
&\leq \left(J_t(\mathbf{w}^\star) - \inf_{\mathbf{w} \in \widetilde{C}_{t-1}^{\text{TS}}} J_t(\mathbf{w})\right) \mathbb{I}\{E_{t-1}\} \qquad \text{(because } E_{t-1} \text{ implies } \widetilde{\mathbf{w}}_{t-1}^{\text{TS}} \in \widetilde{C}_{t-1}^{\text{TS}}) \\
&\leq \left(J_t(\mathbf{w}^\star) - \inf_{\mathbf{w} \in \widetilde{C}_{t-1}^{\text{TS}}} J_t(\mathbf{w})\right) \mathbb{I}\left\{\widetilde{E}_{t-1}\right\} . \qquad \text{(using (4.21))}
\end{aligned}$$

Consider the following set of "optimistic" coefficients $\mathbf{w}$ such that $J_t(\mathbf{w}^\star) \leq J_t(\mathbf{w})$ and, moreover, $\mathbf{w}$ belongs to the sketched TS confidence ellipsoid,

$$W_t^{\text{OPT-TS}} \equiv \left\{\mathbf{w} \in \mathbb{R}^d \;:\; J_t(\mathbf{w}^\star) \leq J_t(\mathbf{w})\right\} \cap \widetilde{C}_t^{\text{TS}} .$$

Then, for $\widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}}$

$$r_t^{\text{TS}} \leq \left(J_t(\widetilde{\mathbf{w}}^{\text{TS}}) - \inf_{\mathbf{w} \in \widetilde{C}_{t-1}^{\text{TS}}} J_t(\mathbf{w})\right) \mathbb{I}\left\{\widetilde{E}_{t-1}\right\} . \qquad (4.23)$$

We now use Abeille and Lazaric [2017, Proposition 3 and Lemma 2] (restated below here for convenience) to argue about the convexity of J and relate its gradient to the chosen action.

Proposition 3. *For any finite set D of actions $\mathbf{x}$ such that $\|\mathbf{x}\| \leq 1$, $\max_{\mathbf{x} \in D} \mathbf{x}^\top \mathbf{w}$ is convex on $\mathbb{R}^d$. Moreover, it is continuous with continuous first derivatives (except for a zero-measure set w.r.t. the Lebesgue measure).*

Lemma 7. *For any* $\mathbf{w} \in \mathbb{R}^d$*, we have*

$$\nabla\Big(\max_{\mathbf{x}\in D} \mathbf{x}^\top \mathbf{w}\Big) = \arg\max_{\mathbf{x}\in D} \mathbf{x}^\top \mathbf{w}$$

(except for a zero-measure w.r.t. the Lebesgue measure).

Relying on the two results above, we can proceed as follows. Introduce $J_t^{/L}(\mathbf{w}) = J_t(\mathbf{w})/L = \max_{\mathbf{x}\in D_t}(\mathbf{x}/L)^\top \mathbf{w}$. Then by Proposition 3, $J_t^{/L}(\mathbf{w})$ is convex for $\mathbf{w} \in \mathbb{R}^d$ since $\|\mathbf{x}/L\| \le 1$. Then, by letting $\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}}) = \nabla J_t(\widetilde{\mathbf{w}}^{\text{TS}})$, for any $\widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}}$ we have

$$\begin{aligned}
J_t(\widetilde{\mathbf{w}}^{\text{TS}}) - \inf_{\mathbf{w}\in\widetilde{C}_{t-1}^{\text{TS}}} J_t(\mathbf{w}) &= L\left(J_t^{/L}(\widetilde{\mathbf{w}}^{\text{TS}}) - \inf_{\mathbf{w}\in\widetilde{C}_{t-1}^{\text{TS}}} J_t^{/L}(\mathbf{w})\right)\\
&\le L \sup_{\mathbf{w}\in\widetilde{C}_{t-1}^{\text{TS}}} \left\{\nabla J_t^{/L}(\widetilde{\mathbf{w}}^{\text{TS}})^\top \left(\widetilde{\mathbf{w}}^{\text{TS}} - \mathbf{w}\right)\right\}\\
&= L \sup_{\mathbf{w}\in\widetilde{C}_{t-1}^{\text{TS}}} \left\{\left(\frac{\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}})}{L}\right)^\top \left(\widetilde{\mathbf{w}}^{\text{TS}} - \mathbf{w}\right)\right\}\\
&\le \|\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}})\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}} \sup_{\mathbf{w}\in\widetilde{C}_{t-1}^{\text{TS}}} \|\widetilde{\mathbf{w}}^{\text{TS}} - \mathbf{w}\|_{\widetilde{\mathbf{V}}_{t-1}} \quad \text{(by Cauchy-Schwartz)}\\
&\le 2\widetilde{\gamma}_{t-1}(\delta')\|\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}})\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}
\end{aligned}$$

where the last inequality holds for all $\widetilde{\mathbf{w}}^{\text{TS}} \in \widetilde{C}_{t-1}^{\text{TS}}$ and by the triangle inequality. Substituting this into (4.23), and taking expectation with respect to $\widetilde{\mathbf{w}}^{\text{TS}}$ yields

$$r_t^{\text{TS}} \le 2\widetilde{\gamma}_{t-1}(\delta')\,\mathbb{E}\left[\|\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}})\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\mathbb{I}\left\{\widetilde{E}_{t-1}\right\} \,\middle|\, \widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}},\, \mathcal{F}_{t-1}\right] . \tag{4.24}$$

where we use $\mathcal{F}_t$ to denote the σ-algebra generated by the random variables $\eta_1, \mathbf{Z}_1, \dots, \eta_{t-1}, \mathbf{Z}_{t-1}$. Now we further upper bound r_t^{TS} while bounding the probability of event $\widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}}$ occurring in (4.24). This is done in the following lemma, whose proof (omitted here) is identical to the proof of [Abeille and Lazaric, 2017, Lemma 3], where ellipsoids are replaced by their sketched counterparts.

Lemma 8. *Assume that* $\mathcal{D}^{\text{TS}}$ *is a TS-sampling distribution with anti-concentration parameter* p*. Then, for* $\mathbf{Z} \sim \mathcal{D}^{\text{TS}}$ *we have that*

$$\mathbb{P}\left(\widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}} \,\middle|\, \widetilde{E}_{t-1}, \mathcal{F}_{t-1}\right) \ge \frac{p}{2} \qquad t = 1, \dots, T\,.$$

We now proceed with the main argument of the proof. Using $g(\widetilde{\mathbf{w}}^{\text{TS}}) = \|\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}})\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}$,

$$\begin{aligned}
\mathbb{E}\left[g(\widetilde{\mathbf{w}}^{\text{TS}}) \,\middle|\, \widetilde{E}_{t-1}, \mathcal{F}_{t-1}\right] &\ge \mathbb{E}\left[g(\widetilde{\mathbf{w}}^{\text{TS}})\mathbb{I}\{\widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}}\} \,\middle|\, \widetilde{E}_{t-1}, \mathcal{F}_{t-1}\right]\\
&= \mathbb{E}\left[g(\widetilde{\mathbf{w}}^{\text{TS}}) \,\middle|\, \widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}}, \widetilde{E}_{t-1}, \mathcal{F}_{t-1}\right]\mathbb{P}\left(\widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}} \,\middle|\, \widetilde{E}_{t-1}, \mathcal{F}_{t-1}\right)\\
&\ge \mathbb{E}\left[g(\widetilde{\mathbf{w}}^{\text{TS}}) \,\middle|\, \widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}}, \widetilde{E}_{t-1}, \mathcal{F}_{t-1}\right]\frac{p}{2} \qquad \text{(by Lemma 8.)}
\end{aligned}$$

The above combined with (4.24) implies that

$$\begin{aligned}
r_t^{\text{TS}} &\le 2\widetilde{\gamma}_{t-1}(\delta')\,\mathbb{E}\left[g(\widetilde{\mathbf{w}}^{\text{TS}})\mathbb{I}\left\{\widetilde{E}_{t-1}\right\}\,\middle|\,\widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}}, \mathcal{F}_{t-1}\right] \\
&= 2\widetilde{\gamma}_{t-1}(\delta')\,\mathbb{E}\left[g(\widetilde{\mathbf{w}}^{\text{TS}})\,\middle|\,\widetilde{\mathbf{w}}^{\text{TS}} \in W_{t-1}^{\text{OPT-TS}}, \widetilde{E}_{t-1}, \mathcal{F}_{t-1}\right]\mathbb{P}\big(\widetilde{E}_{t-1}\big) \\
&\le \frac{4}{p}\widetilde{\gamma}_{t-1}(\delta')\,\mathbb{E}\left[g(\widetilde{\mathbf{w}}^{\text{TS}})\,\middle|\,\widetilde{E}_{t-1}, \mathcal{F}_{t-1}\right] \ . \qquad (4.25)
\end{aligned}$$

Finally, summing (4.25) over time we get

$$\sum_{t=1}^{T} r_t^{\text{TS}} \le \frac{4}{p}\left(\max_{t=0,\dots,T}\{\widetilde{\gamma}_t(\delta')\}\right)\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}})\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\,\middle|\,\mathcal{F}_{t-1}\right] \ .$$

Note that we can already bound $\widetilde{\gamma}_t$ using (4.9). However, we cannot bound the expectation right away, so we rewrite the above as follows

$$\sum_{t=1}^{T} r_t^{\text{TS}} \le \frac{4}{p}\left(\max_{t=0,\dots,T}\{\widetilde{\gamma}_t(\delta')\}\right)\left(\sum_{t=1}^{T}\|X_t\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}} + M_T\right) \qquad (4.26)$$

where we introduce the martingale

$$M_T = \sum_{t=1}^{T}\left(\mathbb{E}\left[\|\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}})\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\,\middle|\,\mathcal{F}_{t-1}\right] - \|X_t\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right) \ .$$

Next, we use the Azuma-Hoeffding (see Proposition 5 in the Appendix) inequality to upper-bound M_T. Now verify that for any $t = 1, \dots, T$,

$$M_t - M_{t-1} = \mathbb{E}\left[\|\mathbf{x}^\star(\widetilde{\mathbf{w}}^{\text{TS}})\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\,\middle|\,\mathcal{F}_{t-1}\right] - \|X_t\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}} \le \frac{2L}{\sqrt{\lambda}} \ .$$

Thus, by the Azuma-Hoeffding inequality, with probability at least $1 - \delta/2$ we have

$$M_T \le \sqrt{\frac{4LT}{\lambda}\ln\left(\frac{4}{\delta}\right)} \ . \qquad (4.27)$$

Now we focus our attention on the remaining term:

$$\begin{aligned}
\sum_{t=1}^{T}\|X_t\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}} &\le \sum_{t=1}^{T}\min\left\{\frac{L}{\sqrt{\lambda}}, \|X_t\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} \\
&\le \max\left\{1, \frac{L}{\sqrt{\lambda}}\right\}\sum_{t=1}^{T}\min\left\{1, \|X_t\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\} \\
&\le \max\left\{1, \frac{L}{\sqrt{\lambda}}\right\}\sqrt{T\sum_{t=1}^{T}\min\left\{1, \|X_t\|^2_{\widetilde{\mathbf{V}}_{t-1}^{-1}}\right\}} \qquad \text{(by Cauchy-Schwartz)} \\
&\overset{\widetilde{\mathcal{O}}}{=} \max\left\{1, \frac{1}{\sqrt{\lambda}}\right\}\sqrt{(1+\varepsilon_m)\left(d\ln(1+\varepsilon_m) + m\right)T} \qquad (4.28)
\end{aligned}$$

where the last step is due to Lemma 4.
For brevity denote $\widetilde{m} = m + d\ln(1+\varepsilon_m)$. Now, we substitute into (4.26) the bound (4.27) on M_T, the bound (4.28), and the bound (4.9) on $\widetilde{\gamma}_t$. This gives

$$\begin{aligned}
\sum_{t=1}^{T} r_t^{\text{TS}} &\overset{\widetilde{\mathcal{O}}}{=} \sqrt{d}\left(R\sqrt{\widetilde{m}\,(1+\varepsilon_m)} + S\sqrt{\lambda}\cdot(1+\varepsilon_m)\right)\left(\max\left\{1, \frac{1}{\sqrt{\lambda}}\right\}\sqrt{(1+\varepsilon_m)\widetilde{m}T} + \sqrt{\frac{T}{\lambda}}\right) \\
&\overset{\widetilde{\mathcal{O}}}{=} \max\left\{1, \frac{1}{\sqrt{\lambda}}\right\}\widetilde{m}\,(1+\varepsilon_m)^{\frac{3}{2}}\left(R + S\sqrt{\lambda}\right)\sqrt{dT} \qquad (4.29)
\end{aligned}$$

which holds with high probability (due to Azuma-Hoeffding inequality).
Now we bound the remaining RLS term of the regret. In particular,

$$\begin{aligned}
\sum_{t=1}^{T} r_t^{\text{RLS}} &= \sum_{t=1}^{T} \mathbb{I}\{E_{t-1}\}\left(X_t^\top \widetilde{\mathbf{w}}_{t-1}^{\text{TS}} - X_t^\top \mathbf{w}^\star\right) \\
&= \sum_{t=1}^{T} \mathbb{I}\{E_{t-1}\}\left(X_t^\top \widetilde{\mathbf{w}}_{t-1}^{\text{TS}} - X_t^\top \widetilde{\mathbf{w}}_{t-1}\right) + \sum_{t=1}^{T} \mathbb{I}\{E_{t-1}\}\left(X_t^\top \widetilde{\mathbf{w}}_{t-1} - X_t^\top \mathbf{w}^\star\right) \\
&\le \sum_{t=1}^{T} \mathbb{I}\{E_{t-1}\}\,\|X_t\|_{\widetilde{\mathbf{V}}_{t-1}}\|\widetilde{\mathbf{w}}_{t-1}^{\text{TS}} - \widetilde{\mathbf{w}}_{t-1}\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}} \\
&\quad + \sum_{t=1}^{T} \mathbb{I}\{E_{t-1}\}\,\|X_t\|_{\widetilde{\mathbf{V}}_{t-1}}\|\widetilde{\mathbf{w}}_{t-1} - \mathbf{w}^\star\|_{\widetilde{\mathbf{V}}_{t-1}^{-1}} && \text{(by Cauchy-Schwartz)} \\
&\le \sum_{t=1}^{T} \|X_t\|_{\widetilde{\mathbf{V}}_{t-1}}\widetilde{\gamma}_{t-1}(\delta') && \text{(by definition of event } \widetilde{E}_{t-1}^{\text{TS}}) \\
&\quad + \sum_{t=1}^{T} \|X_t\|_{\widetilde{\mathbf{V}}_{t-1}}\widetilde{\beta}_{t-1}(\delta') && \text{(by definition of event } \widetilde{E}_{t-1}) \\
&\overset{\widetilde{\mathcal{O}}}{=} \max\left\{1, \frac{1}{\sqrt{\lambda}}\right\}\sqrt{\widetilde{m}(1+\varepsilon_m)T} && \text{(using (4.28))} \\
&\quad \cdot d\left(R\sqrt{\widetilde{m}\,(1+\varepsilon_m)} + S\sqrt{\lambda}\cdot(1+\varepsilon_m)\right) \\
& && \text{(using Theorem 9 to bound } \widetilde{\beta} \text{ and (4.9) to bound } \widetilde{\gamma}) \\
&\overset{\widetilde{\mathcal{O}}}{=} \max\left\{1, \frac{1}{\sqrt{\lambda}}\right\}\left(R\widetilde{m}\,(1+\varepsilon_m) + S\sqrt{\lambda}\sqrt{\widetilde{m}}\,(1+\varepsilon_m)^{\frac{3}{2}}\right)\sqrt{dT} \\
&\overset{\widetilde{\mathcal{O}}}{=} \max\left\{1, \frac{1}{\sqrt{\lambda}}\right\}\widetilde{m}\,(1+\varepsilon_m)^{\frac{3}{2}}\left(R + S\sqrt{\lambda}\right)\sqrt{dT}\,. && (4.30)
\end{aligned}$$

Hence, combining (4.22), (4.29), and (4.30) gives, with high probability,

$$\mathbb{I}\{E_T\}\,R_T = \sum_{t=1}^{T} r_t^{\text{TS}} + \sum_{t=1}^{T} r_t^{\text{RLS}} \overset{\widetilde{\mathcal{O}}}{=} \max\left\{1, \frac{1}{\sqrt{\lambda}}\right\}\widetilde{m}\,(1+\varepsilon_m)^{\frac{3}{2}}\left(R + S\sqrt{\lambda}\right)\sqrt{dT}$$

The proof is concluded by observing that Lemma 6 proves that E_T also holds with high probability.

4.6 Experiments

In this section we present experiments on six publicly available classification datasets.

Setup. The idea of our experimental setup is similar to the one described by Cesa-Bianchi et al. [2013b]. Namely, we convert a K-class classification problem into a contextual bandit problem as follows: given a dataset of labeled instances $(\mathbf{x}, y) \in \mathbb{R}^d \times \{1, \dots, K\}$, we partition it into K subsets according to the class labels. Then we create K sequences by drawing a random permutation of each subset. At each step t the decision set D_t is obtained by picking the t-th instance from each one of these K sequences. Finally, rewards are determined by choosing a class $y \in \{1, \dots, K\}$ and then consistently assigning reward 1 to all instances labeled with y and reward 0 to all remaining instances.

Datasets. We perform experiments on six publicly available datasets for multiclass classification from the `openml` repository Vanschoren et al. [2013] —dataset IDs 1461, 23, 32, 182, 22, and 44, see the table below here for details.

Dataset	Examples	Features	Classes
Bank	45k	17	2
SatImage	6k	37	6
Spam	4k	58	2
Pendigits	11k	17	10
MFeat	2k	48	10
CMC	1.4k	10	3

Baselines. The hyperparameters β (confidence ellipsoid radius) and λ (RLS regularization parameter) are selected on a validation set of size 100 via grid search on $(\beta, \lambda) \in \{1, 10^2, 10^3, 10^4\} \times \{10^{-2}, 10^{-1}, 1\}$ for OFUL, and $\{1, 10^2, 10^3\} \times \{10^{-2}, 10^{-1}, 1, 10^2\}$ for linear TS.

Results We observe that on three datasets, Figure 4.1, sketched algorithms indeed do not suffer a substantial drop in performance when compared to the non-sketched ones, even when the sketch size amounts to 60% of the context space dimension. This demonstrates that sketching successfully captures relevant subspace information relatively to the goal of maximizing reward.

Because the FD-sketching procedure considered in this paper is essentially performing online PCA, it is natural to ask how our sketched algorithms would compare to their non-sketched version run on the best m-dimensional subspace (computed by running PCA on the entire dataset). In Figure 4.2, we compare SOFUL and sketched linear TS to their non-sketched versions. In particular, we keep $60\%, 40\%$, and 20% of the top principal components, and notice that, like in Figure 4.1, there are cases with little or no loss in performance.

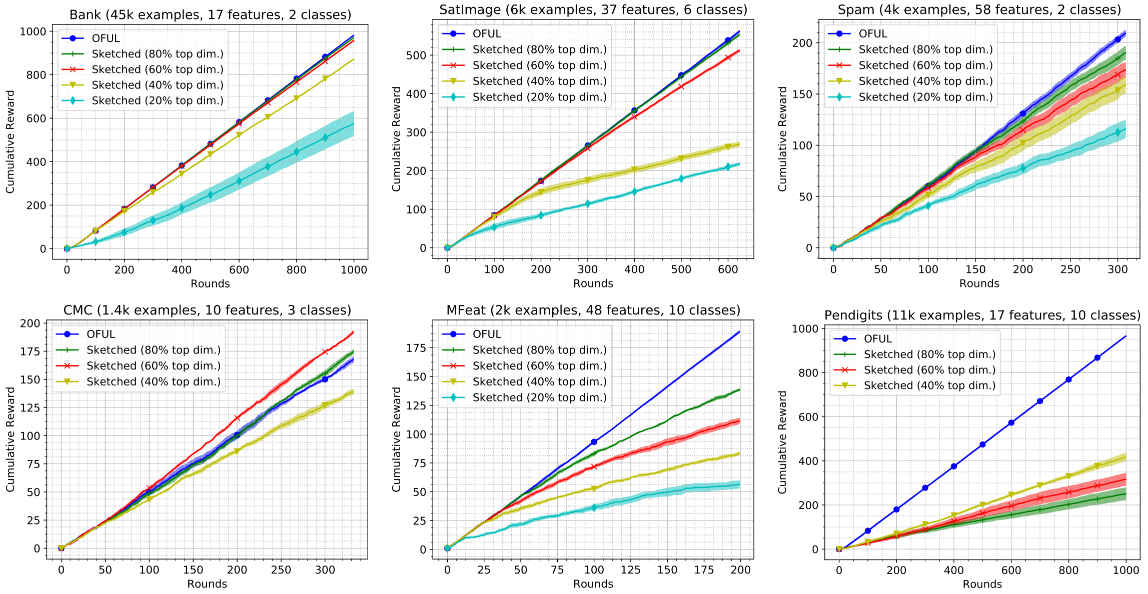

Figure 4.1: Comparison of SOFUL to OFUL on six real-world datasets and for different sketch sizes. Note that, in some cases, a sketch size equal to 80% and even 60% of the context space dimension does not significantly affect the perfomance.

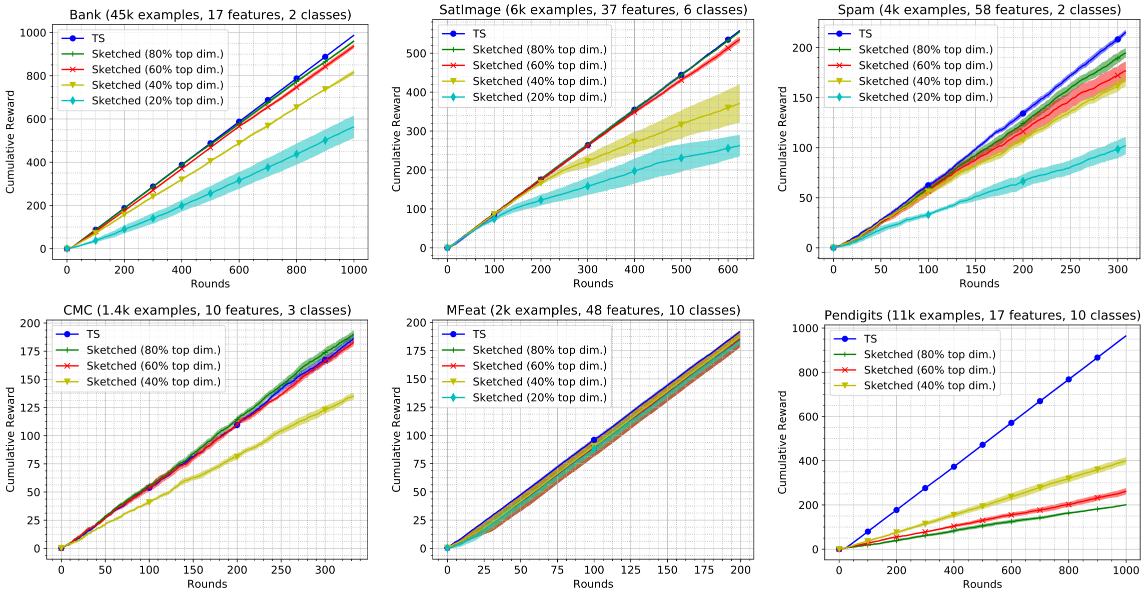

Figure 4.1: Comparison of sketched linear TS to linear TS on six real-world datasets and for different sketch sizes. Note that, in some cases, a sketch size equal to 80% and even 60% of the context space dimension does not significantly affect the perfomance.

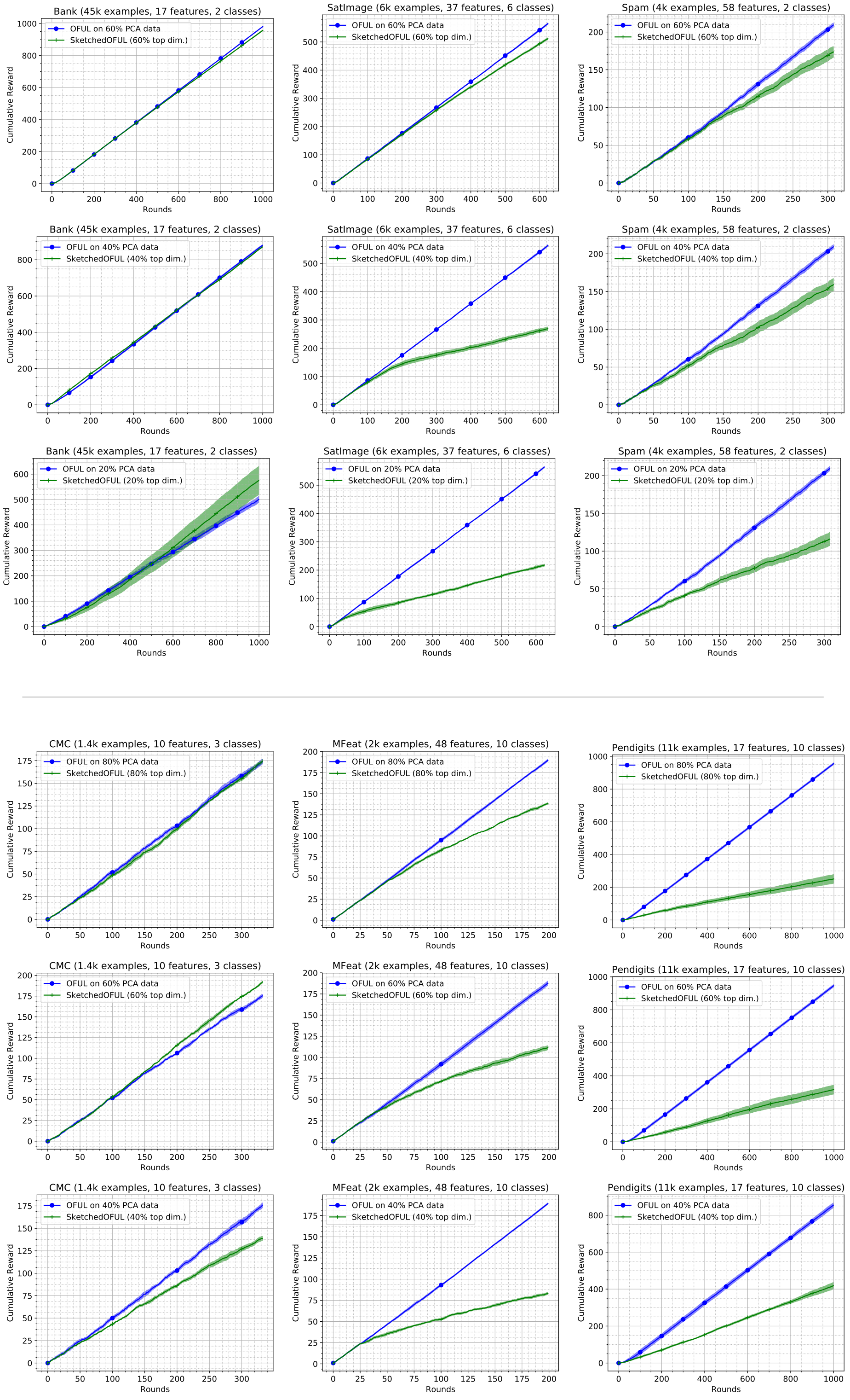

Figure 4.2: Comparison of OFUL run on the best m-dimensional subspace against SOFUL run with sketch size m. Rows show m as a fraction of the context space dimension: $60\%, 40\%, 20\%$ (for the first three datasets), while columns correspond to different datasets. Note that, in some cases (with sketch size m of size at least 60%), SOFUL performs as well as if the best m-dimensional subspace had been known in hindsight.

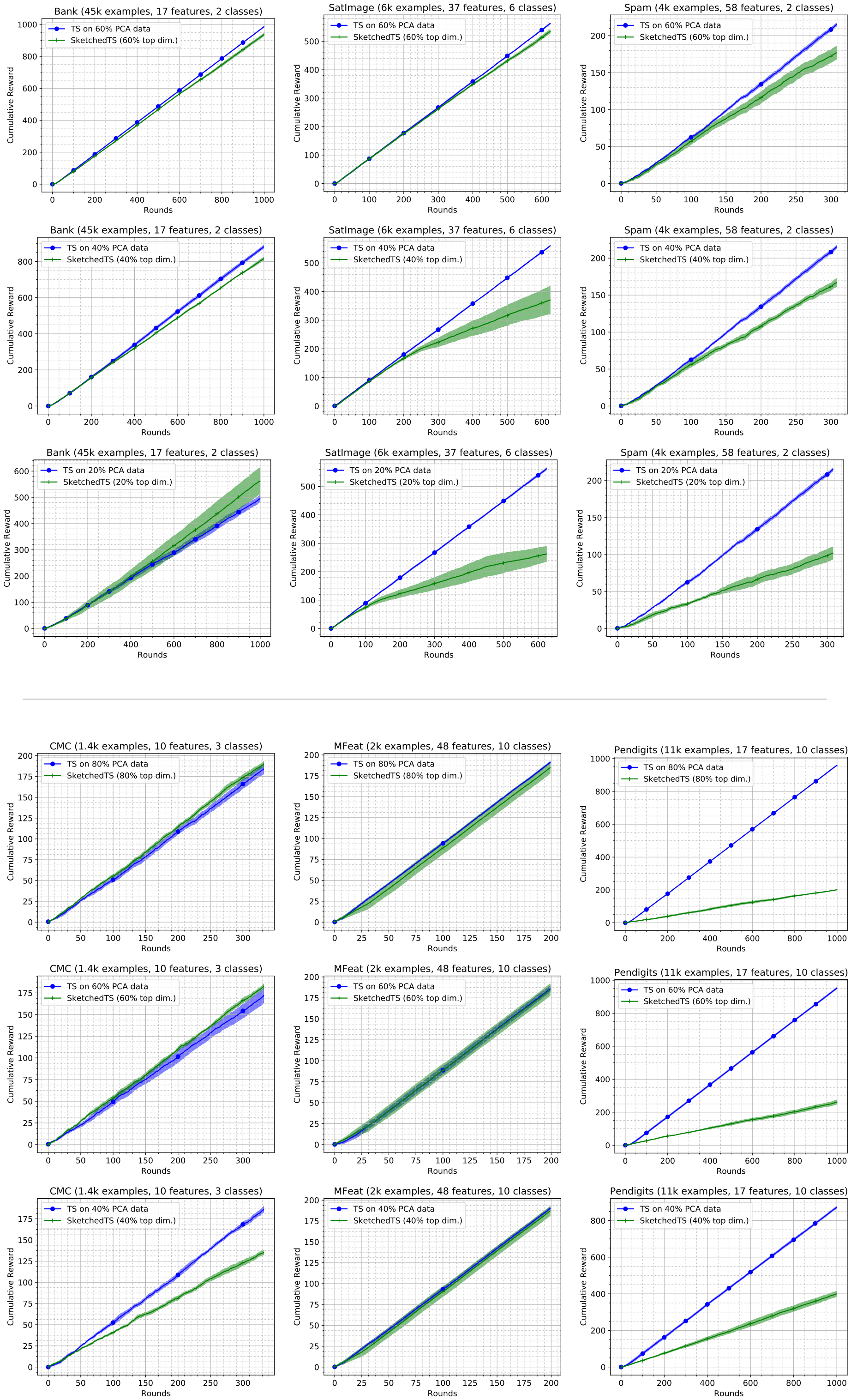

Figure 4.3: Comparison of linear TS run on the best m-dimensional subspace against sketched linear TS run with sketch size m. Rows show m as a fraction of the context space dimension: $60\%, 40\%, 20\%$ (for the first three datasets), while rows correspond to different datasets. Note that, in some cases (with sketch size m of size at least 60%), sketched linear TS performs as well as if the best m-dimensional subspace had been known in hindsight.

4.7 Conclusions

We tackle the efficiency of two well-known stochastic linear bandit algorithms: OFUL and Thompson Sampling. Adopting Frequent Directions, a deterministic online sketching technique, we show that a sketch of size m allows a $O(md)$ update time for both algorithms, as opposed to $\Omega(d^2)$ required by their non-sketched versions in general (where d is the dimension of context vectors). A more general analysis to not predefined sketch sizes have been lately proposed in Calandriello et al. [2019], Chen et al. [2020].

Chapter 5

Meta-learning with Stochastic Linear Bandits

Abstract

In Chapter 4 we have focused on obtaining efficient linear bandit agents. Here, we investigate meta-learning procedures aiming to speedup the learning phase by taking advantage of tasks similarity. 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 OFUL algorithm introduced in Section 2.2. The introduced regularization is defined as 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.

5.1 Introduction

As we were in Chapter 4, we are concerned with linear bandits (Abbasi-Yadkori et al. [2011], Chu et al. [2011], Auer [2003]). Our study builds upon the OFUL algorithm introduced in Section 2.2, 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 and Zhang [2005], Maurer and 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 and 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 and 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 empirically and theoretically (Azar et al. [2013], Calandriello et al. [2014], Zhang and 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. [2013b], 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 successfully considered within the supervised setting setting, to the setting of linear stochastic bandits.

Contributions. Our contribution is threefold. First, we introduce in Section 5.3 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 5.4 and 5.5 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 5.7 we compare experimentally the proposed methods with respect to the standard OFUL algorithm on both synthetic and real data.

5.2 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 $\mathbf{w}_1, \ldots, \mathbf{w}_N, \ldots$ 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.

5.3 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 5.7 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

$$\widehat{\mathbf{w}}_t^{\mathbf{h}} = \left(\mathbf{V}_t^\lambda\right)^{-1} \mathbf{X}_t^\top (\mathbf{Y}_t - \mathbf{X}_t \mathbf{h}) + \mathbf{h}. \tag{5.1}$$

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 Theorem 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 13. *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*

$$\mathcal{C}_t^{\mathbf{h}}(\delta) = \left\{\mathbf{w} \in \mathbb{R}^d : \left\|\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}\right\|_{\mathbf{V}_t^\lambda} \leq \lambda^{\frac{1}{2}} \left\|\mathbf{h} - \mathbf{w}^*\right\|_2 + \sqrt{2 \log\left(\frac{\det\left(\mathbf{V}_t^\lambda\right)^{1/2}}{\det(\lambda I)^{1/2}\, \delta}\right)} = \beta_t^{\mathbf{h}}(\delta)\right\}. \tag{5.2}$$

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

5.3.1 Regret Analysis with Fixed Bias

Given the confidence set defined in Theorem 13 and the *optimism principle* translated into selecting the next arm according to Equation 2.12, we can analyze the expected pseudo-regret (see Eq. (2.9) for its formal definition) depending on the value of $\mathbf{h}$.

Lemma 9. *(REG-OFUL Expected Regret) Under the same assumptions of Theorem 13, 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:*

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

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 Algorithm [Abbasi-Yadkori et al., 2011].

Corollary 2. *Under the conditions of Lemma 9, 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*

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

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*

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

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 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.

5.3.2 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,

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

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

Lemma 10. *(Transfer Regret Bound) Under the same conditions in Theorem 13 and Lemma 9, the expected transfer regret of BIAS-OFUL can be upper bounded as:*

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

Proof. The first statement is the expectation with respect to the *task*-distribution ρ applied to Lemma 9, while the second follows by applying Jensen's inequality. □

We can now replicate what we have done in Corollary 2 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 3. *Under the same assumptions in Theorem 13 and Lemma 9, and setting $\lambda = \frac{1}{T\mathrm{Var}_{\mathbf{h}}}$, 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*

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

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

$$\mathcal{R}(T,\rho) \le \left[1+\sqrt{Td\log\left(T+\frac{T^3LVar_{\overline{\mathbf{w}}}}{d}\right)}\right]C\sqrt{d\log\left(1+\frac{T^2LVar_{\overline{\mathbf{w}}}}{d}\right)}.$$

Proof. These results directly follow from Lemma 10. We have picked $\lambda = \frac{1}{T\mathrm{Var}_{\mathbf{h}}}$ in order to highlight the multiplicative term $\log(1+\mathrm{Var}_{\mathbf{h}})$ which tends to zero according to the variance $\mathrm{Var}_{\mathbf{h}}$ of the task-distribution ρ. □

Algorithm 10 Within Task Algorithm: BIAS-OFUL

```
Input: $\lambda > 0, \widehat{\mathbf{h}}_0 \in \mathbb{R}^d$
$\widehat{\mathbf{w}}_0^{\mathbf{h}} = \widehat{\mathbf{h}}_0, \mathbf{V}_0^{-1} = \frac{1}{\lambda}\mathbf{I}$.
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 $\widehat{\mathbf{h}}_{j,t}^{\lambda}$ according to the meta-algorithm
    UPDATE $\widehat{\mathbf{w}}_t^{\mathbf{h}}$ using Equation 5.1
end for
```

Algorithm 11 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 10 with parameter $\widehat{\mathbf{h}}_{j,0}^{\lambda}$
end for
```

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)*

$$\mathit{Var}_{\overline{\mathbf{w}}} = \mathbb{E}_{\mathbf{w}\sim\rho} \left\|\mathbf{w} - \overline{\mathbf{w}}\right\|_2^2 \ll \mathbb{E}_{\mathbf{w}\sim\rho} \left\|\mathbf{w}\right\|_2^2 = \mathit{Var}_{\mathbf{0}}. \tag{5.3}$$

Notice also that the choice $\lambda = 1/(T\mathrm{Var}_{\mathbf{h}})$, implies that, as $\mathrm{Var}_{\overline{\mathbf{w}}}$ tends to 0, the regret upper bound of the oracle case tends to zero too reflecting the result of Corollary 2. 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 $\overline{\mathbf{w}}$.

5.4 A High Variance Solution

In this section, we present our first meta-learning method. We begin by introducing some additional notation. We let $\mathbf{x}_{j,t}^{\mathbf{h}}$ be the arm pulled by the BIAS-OFUL algorithm (Algorithm 10) at round t-th of the j-th task. We denote by $\mathbf{V}_{j,T} = \sum_{s=1}^{T} \mathbf{x}_{j,s}^{\mathbf{h}}\mathbf{x}_{j,s}^{\mathbf{h}\top}$ 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*

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

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.

5.4.1 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 2.10 on the dataset Z_j without considering any bias.

$$\widehat{\mathbf{h}}^{\lambda}_{N,t} = \frac{1}{NT+t}\left(\sum_{j=1}^{N} T\widehat{\mathbf{w}}^{\lambda}_{j,T} + t\widehat{\mathbf{w}}^{\lambda}_{N+1,t}\right). \tag{5.4}$$

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

Theorem 14. *(Transfer Regret Bound). Let the assumptions of Lemma 10 hold and let $\widehat{\mathbf{h}}^{\lambda}_{N,t}$ be defined as in Equation (5.4). Then, it hold that*

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

where the mean estimation error can be bound as

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

Here, $\beta^{\lambda}_{j}\left(\frac{1}{T}\right)$ refers to the confidence interval computed with OFUL (see Theorem 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 3, this time setting $\mathbf{h} = \widehat{\mathbf{h}}^{\lambda}_{N,T}$, and then observe that

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

□

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 Smale and Zhou [see, e.g., 2007, Lemma 2].

Lemma 11. *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 $\forall j \in [N] : \|\mathbf{w}_j\| \le S$, then for any $0 < \delta < 1$, it holds, with probability at least $1 - \delta$*

$$H(N, \overline{\mathbf{w}}) \le \frac{2\ \log(2/\delta)\ S}{N} + \sqrt{\frac{2\ \log(2/\delta)\ \mathrm{Var}_{\mathbf{0}}}{N}}.$$

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 $\max_{1\le j\le N} \beta_j^\lambda(1/T)\lambda_{\min}^{-1/2}(\mathbf{V}_{j,T}^\lambda)$, associated with the worst past task. By relying on linear regression results Lai and Wei [1982] we have that $\lambda_{\min}(\mathbf{V}_j) \ge \log T$. Moreover, as $\lambda_{\min}(\mathbf{V}_j^\lambda) \ge \lambda + \lambda_{\min}(\mathbf{V}_j)$, 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.

5.5 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 $\widehat{\mathbf{w}}_1, \dots, \widehat{\mathbf{w}}_N$ in Equation 5.4. 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 $\mathbf{A}$.

5.5.1 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

$$\widehat{\mathbf{h}}_{N,t}^\lambda = \left(\widetilde{\mathbf{V}}_{N,t-1}^\lambda\right)^{-1} \widetilde{\mathbf{b}}_{N,t-1}. \tag{5.5}$$

Our next result provides a bound on the transfer regret of this proposed strategy. The proof is presented in Section 5.6.4 of the appendix.

Theorem 15. *(Transfer Regret Bound). Let the assumptions of Lemma 10 hold and let $\widehat{\mathbf{h}}^{\lambda}_{N,t}$ be defined as in Equation (5.5). Then, the following upper bound holds*

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

where the mean estimation error can be bound as

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

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

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

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 $\sigma_{\max}\big(\mathbf{V}_{j,T}\widetilde{\mathbf{V}}^{-1}_{N,T}\big)$.

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: $2(N+1)\max_{j=1,\ldots,N+1}\widetilde{H}(N+1,\mathbf{w}_j)$ which increases with the tasks misalignment $\sigma_{\max}\big(\mathbf{V}_{j,T}\widetilde{\mathbf{V}}^{-1}_{N,T}\big)$.

5.5.2 Tasks Misalignment

We now analyze the tasks misalignment factors appearing in Theorem 15, namely, the quanitities $\sigma_{\max}\big(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\big)$ 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 $\sigma_{\max}\big(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}_{N,t}^{-1}\big)$. Indeed, since we can have at most d misaligned arms, we have the following upper bound $\frac{d}{N}$ to the term $\sigma_{\max}\big(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}_{N,t}^{-1}\big)$. 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 $\sigma_{\max}\big(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}_{N,t}^{-1}\big)$ 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 10) share the same basis:*

$$\mathbf{V}_i = \mathbf{P}\Sigma_i\mathbf{P}^*, \quad \forall i \in [N]. \tag{5.6}$$

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: $\sigma_{\max}\big(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}_{N,t}^{-1}\big) \leq 1$. 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 $\sigma_{\max}\big(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}_{N,t}^{-1}\big) \leq 1/(N - M)$.

5.5.3 Smallest Global Eigenvalue

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 and 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 $\min_{j\in[N]}\lambda_{\max}(V_j) = \mathcal{O}(T)$, this condition implies that $\min_{j\in[N]}\lambda_{\min}(V_j) \geq \log T$.

5.6 Proofs

5.6.1 Biased Confidence Set Definition (Theorem 13)

Proof. Starting from the biased-regularized estimation of Equation 5.1,

$$\begin{aligned}\widehat{\mathbf{w}}_t^{\mathbf{h}} &= \left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{X}_t^{\top}(\mathbf{Y}_t - \mathbf{X}_t\mathbf{h}) + \mathbf{h} = \left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{X}_t^{\top}(\mathbf{X}_t\mathbf{w}^* + \boldsymbol{\eta}_t - \mathbf{X}_t\mathbf{h}) + \mathbf{h}\\ &= \left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{X}_t^{\top}\mathbf{X}_t\mathbf{w}^* + \left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{X}_t^{\top}\boldsymbol{\eta}_t - \left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{X}_t^{\top}\mathbf{X}_t\mathbf{h} + \mathbf{h}\end{aligned}$$

Given this construction we can obtain the following equalities:

$$\begin{aligned}\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}^* &= \left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{X}_t\boldsymbol{\eta}_t + \mathbf{h} - \left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{X}_t^{\top}\mathbf{X}_t\mathbf{h} - \lambda\left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{w}^*\\ &= \left(\mathbf{V}_t^{\lambda}\right)^{-1}\mathbf{X}_t\boldsymbol{\eta}_t + \left(\lambda\left(\mathbf{V}_t^{\lambda}\right)^{-1}\right)(\mathbf{h} - \mathbf{w}^*)\end{aligned}$$

Then, for any $\mathbf{x} \in \mathbb{R}^d$ the following holds:

$$\begin{aligned}\mathbf{x}^{\top}\left(\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}^*\right) &= \langle \mathbf{x}, \mathbf{X}_t\boldsymbol{\eta}_t\rangle_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}} + \lambda\langle \mathbf{x}, \mathbf{h}\rangle_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}} - \lambda\langle \mathbf{x}, \mathbf{w}^*\rangle_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}}\\ &\leq \|\mathbf{x}\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}}\left(\|\mathbf{X}_t\boldsymbol{\eta}_t\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}} + \lambda\|\mathbf{h} - \mathbf{w}^*\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}}\right)\end{aligned}$$

where in the last step we have applied Cauchy-Schwarz inequality. Plugging in $\mathbf{x} = \mathbf{V}_t^{\lambda}(\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}^*)$ we obtain:

$$\|\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}^*\|_{\mathbf{V}_t^{\lambda}}^2 \leq \|\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}^*\|_{\mathbf{V}_t^{\lambda}}\left(\|\mathbf{X}_t\boldsymbol{\eta}_t\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}} + \lambda\|\mathbf{h} - \mathbf{w}^*\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}}\right)$$

finally by dividing both sides by $\left\|\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}^*\right\|_{\mathbf{V}_t^{\lambda}}$ we obtain:

$$\left\|\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}^*\right\|_{\mathbf{V}_t^{\lambda}} \leq \|\mathbf{X}_t\boldsymbol{\eta}_t\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}} + \lambda\left\|\mathbf{h} - \mathbf{w}^*\right\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}}.$$

Finally we bound the noisy term $\|\mathbf{X}_t\boldsymbol{\eta}_t\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}}$ by leveraging on Theorem 1 Abbasi-Yadkori et al. [2011], obtaining:

$$\left\|\widehat{\mathbf{w}}_t^{\mathbf{h}} - \mathbf{w}^*\right\|_{\mathbf{V}_t^{\lambda}} \leq R\sqrt{2\log\left(\frac{\det\left(\mathbf{V}_t^{\lambda}\right)^{1/2}}{\det(\lambda I)^{1/2}\delta}\right)} + \lambda^{\frac{1}{2}}\left\|\mathbf{h} - \mathbf{w}^*\right\|_2 = \beta_t^{\mathbf{h}}(\delta) \qquad (5.7)$$

where we have used the fact that: $\|\mathbf{h} - \mathbf{w}^*\|_{\left(\mathbf{V}_t^{\lambda}\right)^{-1}}^2 \leq \frac{1}{\lambda_{\min}\left(\mathbf{V}_t^{\lambda}\right)}\left\|\mathbf{h} - \mathbf{w}^*\right\|_2^2 \leq \frac{1}{\lambda}\left\|\mathbf{h} - \mathbf{w}^*\right\|_2^2$. □

5.6.2 Regret Analysis with Fixed Bias (Lemma 9)

Proof. We start by analysing the instantaneous regret as follows:

$$\begin{aligned}r_t &= \langle \mathbf{w}^*, \mathbf{x}_t^*\rangle - \langle \mathbf{w}^*, \mathbf{x}_t^{\mathbf{h}}\rangle = \langle \mathbf{w}^*, \mathbf{x}_t^*\rangle - \langle \widetilde{\mathbf{w}}_t^{\mathbf{h}}, \mathbf{x}_t^{\mathbf{h}}\rangle + \langle \widetilde{\mathbf{w}}_t^{\mathbf{h}}, \mathbf{x}_t^{\mathbf{h}}\rangle - \langle \mathbf{w}^*, \mathbf{x}_t^{\mathbf{h}}\rangle\\ &\leq \langle \widetilde{\mathbf{w}}_t^{\mathbf{h}}, \mathbf{x}_t\rangle - \langle \mathbf{w}^*, \mathbf{x}_t^{\mathbf{h}}\rangle = \langle \widehat{\mathbf{w}}_{t-1}^{\mathbf{h}} - \mathbf{w}^*, \mathbf{x}_t^{\mathbf{h}}\rangle + \langle \widetilde{\mathbf{w}}_t^{\mathbf{h}} - \widehat{\mathbf{w}}_{t-1}^{\mathbf{h}}, \mathbf{x}_t^{\mathbf{h}}\rangle\\ &\leq \left\|\widehat{\mathbf{w}}_{t-1}^{\mathbf{h}} - \mathbf{w}^*\right\|_{\mathbf{V}_{t-1}^{\lambda}}\left\|\mathbf{x}_t^{\mathbf{h}}\right\|_{\mathbf{V}_{t-1}^{\lambda}} + \left\|\widetilde{\mathbf{w}}_t^{\mathbf{h}} - \widehat{\mathbf{w}}_{t-1}^{\mathbf{h}}\right\|_{\mathbf{V}_{t-1}^{\lambda}}\left\|\mathbf{x}_t^{\mathbf{h}}\right\|_{\mathbf{V}_{t-1}^{\lambda}} \leq 2\beta_{t-1}^{\mathbf{h}}(\delta)\left\|\mathbf{x}_t^{\mathbf{h}}\right\|_{\mathbf{V}_{t-1}^{\lambda}}\end{aligned}$$

where in the first inequality we have leveraged on the fact that $\left(\widetilde{\mathbf{w}}_t^{\mathbf{h}}, \mathbf{x}_t\right)$ is optimistic and in the last the ellipsoid bound specified in Equation 5.7. The bound of the cumulative regret follows from the bound Abbasi-Yadkori et al. [2011], hence with probability at least $1-\delta$, for all $T \geq 0$:

$$
\begin{aligned}
R(T, \mathbf{w}^*) &\leq \sqrt{T \sum_{t=1}^{T} {r_t}^2} \leq 4\sqrt{T \log\left(\det\left(\mathbf{V}_t^{\lambda}\right)\right) - \log\left(\det(\lambda \mathbf{I})\right)}\beta_T^{\mathbf{h}}(\delta) \\
&\leq 4\sqrt{Td \log\left(1 + \frac{TL}{\lambda d}\right)}\left(\lambda^{\frac{1}{2}} \left\|\mathbf{w}^* - \mathbf{h}\right\|_2 + R\sqrt{2\log(1/\delta) + d\log\left(1 + TL/(\lambda d)\right)}\right)
\end{aligned}
$$

where the last two steps follow from Lemma 11 of Abbasi-Yadkori et al. [2011] and the definition of $\beta^{\mathbf{h}}(\delta)$ (Equation 5.7). The stated result is derived analogously to Corollary 19.3 of Lattimore and Szepesvári [2018] considering $\delta = \frac{1}{T}$. □

5.6.3 Right Bias Value (Corollary 2)

Proof. We start by considering the **oracle** scenario which is given by $\mathbf{h} = \mathbf{w}^*$.

$$
\begin{aligned}
&\lim_{\lambda \to \infty}\left[C\sqrt{Td\log\left(1 + \frac{TL}{\lambda d}\right)}\left(R\sqrt{d\log(T + T^2 L/(\lambda d))}\right)\right] \\
&= C\sqrt{Td\log(1)}\left(R\sqrt{d\log(T + T^2 L/(\lambda d))}\right) = 0
\end{aligned}
$$

As far as the **independent task learning** scenario concerns, the following holds:

$$
\begin{aligned}
&\lim_{\lambda \to \infty} C\sqrt{Td\log\left(1 + \frac{TL}{\lambda d}\right)}\left(\lambda^{\frac{1}{2}} S + R\sqrt{d\log(T + T^2 L/(\lambda d))}\right) \\
&= \lim_{\epsilon \to 0} C\sqrt{Td\log\left(1 + \epsilon\right)}\left(S\sqrt{\frac{TL}{\epsilon d}} + R\sqrt{d\log(T + T^2 L/(\lambda d))}\right) \\
&= \lim_{\epsilon \to 0} C\left[ST\sqrt{\frac{Ld}{d}\frac{\log\left(1+\epsilon\right)}{\epsilon}} + Rd\sqrt{T\log\left(1+\epsilon\right)\log(T + T^2 L/(\lambda d))}\right] \\
&= \lim_{\epsilon \to 0} C\left[ST\sqrt{L} + Rd\sqrt{T\log\left(1+\epsilon\right)\log(T + T^2 L/(\lambda d))}\right)\Bigg] = CTS\sqrt{L}
\end{aligned}
$$

where we have used the substitution $\epsilon = \frac{TL}{\lambda d}$ and the fact that $\lim_{\epsilon \to 0} \frac{\log(1+\epsilon)}{\epsilon} \to 1$. □

5.6.4 Global Ridge Regression Transfer Regret Bound (Theorem 15)

We start by presenting two Lemmas which are necessary to obtain the final bound. Firstly, we need to introduce an additional variable:

$$\overline{\mathbf{h}}'_{N,t+1} = \left(\widetilde{\mathbf{V}}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\mathbf{V}_{j,T}\mathbf{w}_j + \mathbf{V}_{N+1,t}\mathbf{w}_{N+1}\right)$$

We will then split the analysis by studying separately the *estimation error* $\widehat{\mathbf{h}}^{\lambda}_{N,t+1} - \overline{\mathbf{h}}'_{N,t+1}$ (Lemma 12) and the *estimation bias* $\overline{\mathbf{h}}'_{N,t+1} - \overline{\mathbf{h}}_{N,t+1}$ (Lemma 13).

Lemma 12. *The following rewriting holds:*

$$\widehat{\mathbf{h}}^{\lambda}_{N,t+1} - \overline{\mathbf{h}}'_{N,t+1} = \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\eta_{j,s} + \sum_{s=1}^{t}\mathbf{x}_{N+1,s}\eta_{N+1,s}\right) - \lambda\left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\overline{\mathbf{h}}'_{N,t+1}$$

Proof.

$$\begin{aligned}
\widehat{\mathbf{h}}^{\lambda}_{N,t+1} &= \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\widetilde{\mathbf{b}}_{N,t} = \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}y_{j,s} + \sum_{s=1}^{t}\mathbf{x}_{N+1,s}y_{N+1,s}\right) \\
&= \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\left(\mathbf{x}_{j,s}^{\top}\mathbf{w}_j + \eta_{j,s}\right) + \sum_{s=1}^{t}\mathbf{x}_{N+1,s}\left(\mathbf{x}_{N+1,s}^{\top}\mathbf{w}_{N+1} + \eta_{N+1,s}\right)\right) \\
&= \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\eta_{j,s} + \right. \\
&\quad \left. + \sum_{s=1}^{t}\mathbf{x}_{N+1,s}\eta_{N+1,s}\right) + \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\mathbf{x}_{j,s}^{\top}\mathbf{w}_j + \sum_{s=1}^{t}\mathbf{x}_{s}\mathbf{x}_{N+1,s}^{\top}\mathbf{w}_{N+1}\right) \\
&= \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\eta_{j,s} + \sum_{s=1}^{t}\mathbf{x}_{N+1,s}\eta_{N+1,s}\right) + \\
&\quad + \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\widetilde{\mathbf{V}}_{N,t}\left(\widetilde{\mathbf{V}}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\mathbf{V}_{j,T}\mathbf{w}_j + \mathbf{V}_{N+1,t}\mathbf{w}_{N+1}\right) \\
&= \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\eta_{j,s} + \sum_{s=1}^{t}\mathbf{x}_{N+1,s}\eta_{N+1,s}\right) + \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\widetilde{\mathbf{V}}_{N,t}\overline{\mathbf{h}}'_{N,t+1} + \\
&\quad + \lambda\left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left[\overline{\mathbf{h}}'_{N,t+1} - \overline{\mathbf{h}}'_{N,t+1}\right] \\
&= \left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\eta_{j,s} + \sum_{s=1}^{t}\mathbf{x}_{N+1,s}\eta_{N+1,s}\right) + \overline{\mathbf{h}}'_{N,t+1} - \lambda\left(\widetilde{\mathbf{V}}^{\lambda}_{N,t}\right)^{-1}\overline{\mathbf{h}}'_{N,t+1}
\end{aligned}$$

which gives the claimed result. □

Lemma 13. *According to what we have done in Section 5.4, we use:*

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

Differently from $\overline{\mathbf{h}}'_{N,t}$ this definition is a weighted average of the vectors of the N completed tasks. Hence, we have:

$$\begin{aligned}\left\|\overline{\mathbf{w}} - \overline{\mathbf{h}}'_{N,t}\right\| &\leq \frac{1}{NT+t}\sum_{j=1}^{N}\Big[\left\|\overline{\mathbf{w}} - \overline{\mathbf{h}}_{N,t}\right\| + (NT+t)\left\|\overline{\mathbf{h}}_{N,t} - \overline{\mathbf{h}}_{N,t}\right\|'\Big] \\ &= H_\rho(N+1, \overline{\mathbf{w}}) + \left\|\overline{\mathbf{h}}_{N,t} - \overline{\mathbf{h}}'_{N,t}\right\|\end{aligned}$$

where we have denoted with $H_\rho(N+1,\overline{\mathbf{w}})$ according to what we have done in Section 5.4. We can now focus on the term $\left\|\overline{\mathbf{h}}'_{N,t} - \overline{\mathbf{h}}_{N,t}\right\|$ which can be equivalently rewritten as follows:

$$\begin{aligned}\left\|\overline{\mathbf{h}}'_{N,t+1} - \overline{\mathbf{h}}_{N,t+1}\right\| &= \left\|\left(\widetilde{\mathbf{V}}_{N,t}\right)^{-1}\sum_{j-1}^{N}(\mathbf{V}_{j,T}\mathbf{w}_j + \mathbf{V}_{N+1,t}\mathbf{w}_{N+1}) - \overline{\mathbf{h}}_{N,t}\right\| \\ &\leq \sum_{j=1}^{N}\left|\widetilde{\mathbf{V}}^{-1}_{N,t}\mathbf{V}_{j,T}\right|\left\|\mathbf{w}_j - \overline{\mathbf{h}}_{N,t}\right\| + \left|\widetilde{\mathbf{V}}^{-1}_{N,t}\mathbf{V}_{N+1,t}\right|\left\|\mathbf{w}_{N+1} - \overline{\mathbf{h}}_{N,t}\right\| \\ &\leq \sum_{j=1}^{N} H_\rho(N+1,\mathbf{w}_j)\left|\widetilde{\mathbf{V}}^{-1}_{N,t}\mathbf{V}_{j,T}\right| + H_\rho(N+1,\mathbf{w})\left|\widetilde{\mathbf{V}}^{-1}_{N,t}\mathbf{V}_t\right| \\ &= \sum_{j=1}^{N} H_\rho(N+1,\mathbf{w}_j)\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right) + H_\rho(N+1,\mathbf{w}_{N+1})\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right) \\ &\leq (N+1)\max_{j=1,\dots,N+1}\left(H_\rho(N+1,\mathbf{w}_j)\sigma_{\max}\left(\mathbf{V}_{j,t}\widetilde{\mathbf{V}}^{-1}_{N,t}\right)\right) \\ &= (N+1)\max_{j=1,\dots,N+1}\widetilde{H}(N+1,\mathbf{w}_j)\end{aligned}$$

We have used the fact that the matrix norm of a given matrix A induced by the Euclidean norm corresponds to the spectral norm, which is the largest singular value of the matrix $\sigma_{\max}(A)$.

We can now bound the transfer regret bound incurred by the second approach.

Proof. We start the analysis from the result of Lemma 10:

$$\mathcal{R}(T,\rho) \leq d\sqrt{T\log\left(1 + \frac{T^2L\left(\mathbb{E}_{\mathbf{w}\sim\rho}\left[\|\mathbf{w}-\mathbf{h}\|_2^2\right]\right)}{d}\right)}$$

we can then set the hyperparameter $\mathbf{h} = \widehat{\mathbf{h}}^{\lambda}_{N,T}$ and focusing on the first term in brackets we obtain:

$$\sqrt{\mathbb{E}_{\mathbf{w}\sim\rho}\left[\left\|\mathbf{w} - \widehat{\mathbf{h}}^{\lambda}_{N,T}\right\|_2^2\right]} \leq \sqrt{\mathrm{Var}_{\overline{\mathbf{w}}}} + \sqrt{\epsilon_{N,t}(\rho)}$$

According to Lemma 13 the following rewriting holds:

$$\sqrt{\epsilon_{N,t}(\rho)} \leq H_\rho(N+1, \overline{\mathbf{w}}) + (N+1)\max_{j=1,\dots,N+1} \widetilde{H}(N+1, j) + \left\|\overline{\mathbf{h}}'_{N,T} - \widehat{\mathbf{h}}^{\lambda}_{N,T}\right\|_2$$

It remains only to apply Lemma 12 which gives:

$$\begin{aligned}
\left\|\overline{\mathbf{h}}'_{N,T} - \widehat{\mathbf{h}}^{\lambda}_{N,T}\right\|_2 &= \left\|\left(\widetilde{\mathbf{V}}^{\lambda}_{N,T}\right)^{-1}\left(\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\eta_{j,s} + \sum_{s=1}^{T}\mathbf{x}_s\eta_s\right)\right\|_2 + \left\|\lambda\left(\widetilde{\mathbf{V}}^{\lambda}_{N,T}\right)^{-1}\overline{\mathbf{h}}'_{N,T}\right\|_2 \\
&\leq \left\|\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\eta_{j,s} + \sum_{s=1}^{T}\mathbf{x}_s\eta_s\right\|_{\left(\widetilde{\mathbf{V}}^{\lambda}_{N,T}\right)^{-2}} + \lambda\left\|\overline{\mathbf{h}}'_{N,T}\right\|_{\left(\widetilde{\mathbf{V}}^{\lambda}_{N,T}\right)^{-2}} \\
&\leq \frac{1}{\lambda_{\min}^{\frac{1}{2}}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})}\left\|\sum_{j=1}^{N}\sum_{s=1}^{T}\mathbf{x}_{j,s}\eta_{j,s} + \sum_{s=1}^{T}\mathbf{x}_s\eta_s\right\|_{\left(\widetilde{\mathbf{V}}^{\lambda}_{N,T}\right)^{-1}} + \frac{1}{\lambda_{\min}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})}\left\|\overline{\mathbf{h}}'_{N,T}\right\|_2 \\
&\leq \frac{1}{\lambda_{\min}^{\frac{1}{2}}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})}R\sqrt{2\log\left(T + \frac{(NT+T)TL^2}{\lambda d}\right)} + \frac{\left\|\overline{\mathbf{h}}_{N,T}\right\|_2}{\lambda_{\min}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})} + \left\|\overline{\mathbf{h}}'_{N,T} - \overline{\mathbf{h}}_{N,T}\right\|_2 \\
&\leq \frac{1}{\lambda_{\min}^{\frac{1}{2}}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})}R\sqrt{2\log\left(T + \frac{(NT+T)TL^2}{\lambda d}\right)} + \frac{S}{\lambda_{\min}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})} + \left\|\overline{\mathbf{h}}'_{N,T} - \overline{\mathbf{h}}_{N,T}\right\|_2 \\
&\leq \frac{1}{\lambda_{\min}^{\frac{1}{2}}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})}R\sqrt{2\log\left(T + \frac{(NT+T)TL^2}{\lambda d}\right)} + \frac{S}{\lambda_{\min}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})} \\
&\quad + (N+1)\max_{j=1,\dots,N+1}\widetilde{H}(N+1, j)
\end{aligned}$$

where in the last inequality we have applied once more Lemma 13. We can now introduce $\nu_{\min} = \lambda_{\min}\big(\widetilde{\mathbf{V}}_{N,T}\big)$ as the minimum eigenvalue of the global covariance matrix without regularization which gives the following bound:

$$\frac{1}{\lambda_{\min}(\widetilde{\mathbf{V}}^{\lambda}_{N,T})} \leq \frac{1}{\lambda + \nu_{\min}}$$

putting everything together gives the claimed result:

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

□

5.7 Experiments

In this section we test the real effectiveness of the proposed approaches. The theoretical results stated that the method presented in Section 5.4 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 5.5 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 5.3, 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 Algorithm 3, 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 5.1. This can be noticed by observing that the gap $\left\|\widehat{\mathbf{h}}^\lambda - \mathbf{w}^*\right\|$ equally affects all the available arms.

5.7.1 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 10 and 11 and estimate $\widehat{\mathbf{h}}$ as per Equations 5.4 and Equation 5.5, respectively. The Oracle policy knows the mean task parameter $\overline{\mathbf{w}}$ and uses it as the bias $\mathbf{h}$ in BIAS-OFUL (Corollary 3 (ii)). Analogously, the ITL policy consists of BIAS-OFUL with bias set equal to $\mathbf{0}$, see Corollary 3 (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 $\mathcal{D}$.

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

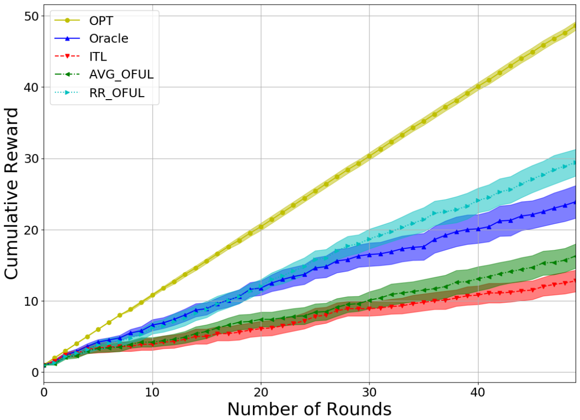

Figure 5.1: Cumulative reward measured after $N = 10$ tasks and averaged over 10 test tasks, with $\lambda = 1$.

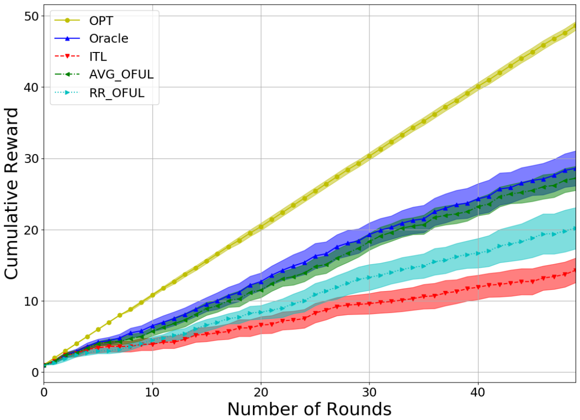

Figure 5.2: Cumulative reward measured after $N = 10$ tasks and averaged over 10 test tasks, with $\lambda = 100$.

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 $\mathcal{N}\big(0.5, 1\big)$, to compute the rewards. Finally, we assigned reward 1 to the arm having the maximum final reward, 0 to the others. In Figures 5.1 and 5.2, 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

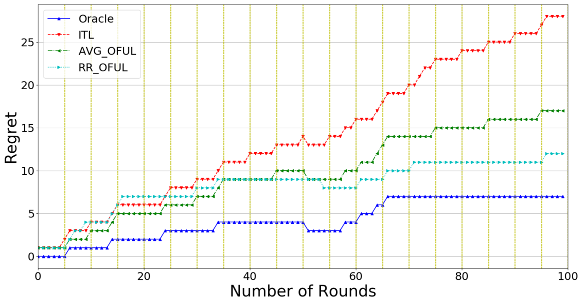

Figure 5.3: Empirical Transfer regret associated with Lastfm.

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 5.3 (and Figure 5.4) 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 3 and Assumption 1.

Movielens Here we consider the Movielens data Harper and 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 users we observe better results by using the KMeans clustering algorithm over the user vectors. The results displayed in Figure 5.4 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 5.4 considers $\lambda = 1$.

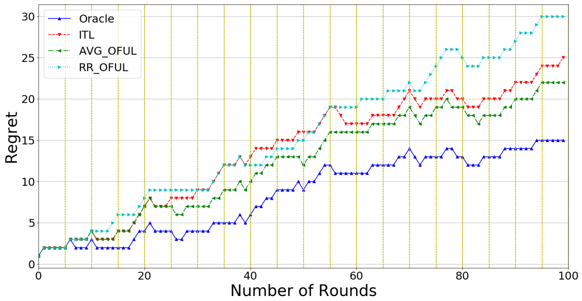

Figure 5.4: Empirical Transfer regret associated with Movielens.

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 15).

5.8 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.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

In this thesis we have investigated three aspects concerning the application of stochastic bandits in the recommender system domain.

- In Chapter 3 we formalized a novel non-stationary stochastic setting. Here, the expected payoff of each arm is parametrized by the delay since the time the arm was last played. We shown that finding the optimal policy is NP-hard even when all the parameters are known. Then, we introduced a class of ranking policies approximating the reward of the optimal one up to a constant factor. We proposed a simple algorithm to learn the best ranking policy. Finally, we studied the empirical performance of the introduced solution on different synthetic problem instances.

- Then, in Chapter 4 we enhance the performance of two well known linear bandit algorithms: OFUL and Thompson Sampling. We made them more efficient by using Frequent Directions, a deterministic online sketching technique. We analyze the impact of this approximation strategy in terms of regret bounds and computational costs. The results pointed out the importance of choosing the right sketch size. Indeed, if from one side we would like to pick a small size to reduce the time complexity, from the other, keeping few features means incurring a big approximation error in the regret bounds.

- Lastly, in Chapter 5, we formalized the learning-to-learn problem with linear bandit tasks. Differently from the existing works, we considered the case where tasks arrived sequentially and they are sampled from an unknown distribution. We design two transfer mechanisms to take advantage of past tasks with the objective of reducing the exploration phase. While the first solution may suffer a potentially high variance, the second might incur a strong bias resulting in a possible negative transfer. We were able to corroborate all the theoretical results with synthetic and real experiments.

6.2 Future Works

Each of these problems opens the way for interesting and closely-connected investigations. The most immediate extension regarding the adoption of sketching techniques to linear bandits involves the investigation of adaptive algorithms that automatically learn the sketch size based on data. This extension was actually investigated in Calandriello et al. [2019]. There, authors replaced Frequent Directions with a randomized matrix sketching technique based on leverage score sampling which gives an accurate low-rank approximation of the covariance matrix. Our formalization of the learning-to-learn problem with bandit tasks leaves an important number of open problems. For instance, we are interested in analyzing the more challenging scenario where the bandit tasks have arms whose expected reward is a linear function of the arms images in a reproducing kernel Hilbert space. As for the non-stationary setting, immediate extensions include the case where arms are represented by vectors of features and their payoff is a linear regression of the mentioned vectors. Indeed, in this case, adopting arm elimination procedures is not straightforward anymore.

Chapter 7

Appendix

7.1 Concentration Inequalities

Lemma 14 (Markov's inequality)**.** *Let us consider a nonnegative random variable X which admits $\mathbb{E}[X]$. Then for any $a > 0$, the following result holds*

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Since $X > 0$,

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty xp(x)dx = \int_0^a xp(x)dx + \int_a^\infty xp(x)dx \\ &\geq \int_a^\infty xp(x)dx \geq a \int_a^\infty p(x)dx = a\mathbb{P}(X > a).\end{aligned}$$

□

Proposition 4 (Chernoff-Hoeffding inequality)**.** *Let $X_1, \ldots, X_n$ be random variables with common range $[0, 1]$ and such that $\mathbb{E}[X_t|x_1, \ldots, X_{t-1}] = \mu$. Let $S_n = X_1 + \cdots + X_n$. Then for all $\epsilon \geq 0$,*

$$\mathbb{P}\left(S_n \geq n\mu + \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{n}\right)$$

and

$$\mathbb{P}\left(S_n \leq n\mu - \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{n}\right)$$

Proof. By applying Lemma 14, for any $a > 0$,

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > s\right] \leq \frac{\mathbb{E}\left[\exp\left(a\sum_{s=1}^n (X_s - \mathbb{E}[X_s])\right)\right]}{\exp(at)}$$

Finally, bounding the numerator using [Cesa-Bianchi and Lugosi, 2006, lemma A.1] and minimizing the obtained bound in a we obtain the first inequality. The second one is obtained by symmetry. □

Definition 2 (Martingale)**.** *A sequence of random variables $Y_1, \dots, Y_n$ is named martingale with respect to the sequence of random variables $X_1, \dots, X_n$ if, for every integer $i > 0$,*

$$\begin{aligned} &\mathbb{E}[|Y_i|] < \infty \\ &\mathbb{E}[Y_i | X_1, \dots, X_{i-1}] = Y_{i-1}. \end{aligned}$$

When the expectation in second condition is upper bounded by Y_{i-1}, the sequence $Y_1, Y_2, \dots$ is named supermartingale. Conversely, when we have $\mathbb{E}[Y_i|X_1, \dots, X_{i-1}] \geq Y_{i-1}$, the sequence $Y_1, Y_2, \dots$ is called submartingale.

Proposition 5 (Azuma-Hoeffding inequality)**.** *If a supermartingale Y_t corresponding to a filtration $\mathcal{F}_t$ satisfies $|Y_t - Y_{t-1}| \leq c_t$ for some constant c_t for $t = 1, 2, \dots, T$ and $Y_0 = 0$. Then for any $t > 0$,*

$$\mathbb{P}\left(Y_T - Y_0 \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{s=1}^{T} c_s^2}\right) .$$

Proof. We consider the function $f(x) = \exp(\lambda x)$ which is convex in x for any $\lambda \in \mathbb{R}$. Relying on the convexity and considering $|\frac{x}{c_i}| \leq 1$ we have

$$\begin{aligned} \exp(\lambda x) &= f\Big(\frac{1}{2}(\frac{x}{c_i} + 1)c_i + \frac{1}{2}(1 - \frac{x}{c_i})(-c_i)\Big) \\ &\leq \frac{1}{2}\Big(\frac{x}{c_i} + 1\Big)f(c_i) + \frac{1}{2}\Big(1 - \frac{x}{c_i}\Big)f(-c_i) \\ &= \frac{f(c_i) + f(-c_i)}{2} + \frac{f(c_i) - f(-c_i)}{2}x \end{aligned}$$

Furthermore, for all α

$$\begin{aligned} \frac{\exp(\alpha) + \exp(-\alpha)}{2} &= \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} + \sum_{k=0}^{\infty} \frac{(-1)^k \alpha^k}{k!} = \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{(2k)!} \\ &\leq \sum_{k=0}^{\infty} \frac{\left(\frac{\alpha^2}{2}\right)^k}{k!} = \exp\left(\frac{\alpha^2}{2}\right) \end{aligned}$$

We can conclude that for every x such that $|x/c_i| \leq 1$, we have,

$$\exp\Big(\lambda x\Big) \leq \exp\left(\frac{c_i^2}{2}\right) + \frac{\exp(\lambda c_i) - \exp(-\lambda c_i)}{2}x \tag{7.1}$$

Getting back to the martingale sequence $Y_1, Y_2, \dots, Y_T$. For every $t \geq 0$ and every $\lambda > 0$ we have

$$\begin{aligned} \mathbb{P}(Y_T - Y_0 \geq t) &= \mathbb{P}\Big(\exp\big(\lambda(Y_n - Y_0)\big) \geq \exp(\lambda t)\Big) \\ &\leq \exp(-\lambda t)\,\mathbb{E}\left[\exp(\lambda Y_T)\right] \\ &= \exp(-\lambda t)\,\mathbb{E}\Big[\exp\big(\lambda \sum_{i=1}^{T}(Y_i - Y_{i-1})\big)\Big]. \end{aligned}$$

Using the tower property of conditional expectation and considering the filtration $\mathcal{F}_{T-1}$ we have

$$\mathbb{E}\left[\exp\big(\lambda(Y_T - Y_{T-1})\big)\exp\big(\lambda\sum_{i=1}^{T-1}(Y_i - Y_{i-1})\big)|\mathcal{F}_{T-1}\right]$$

$$= \exp\big(\lambda\sum_{i=1}^{T-1}(Y_i - Y_{i-1})\big)\,\mathbb{E}\left[\exp\big(\lambda(Y_T - Y_{T-1})\big)|\mathcal{F}_{T-1}\right]$$

$$\leq \exp\left(\lambda\sum_{i=1}^{T-1}(Y_i - Y_{i-1})\right)\left(\exp\left(\frac{\lambda^2 c_T^2}{2}\right) + \frac{\exp(\lambda c_i) - \exp(-\lambda c_i)}{2}\,\mathbb{E}\left[Y_T - T_{T-1}|\mathcal{F}_{T-1}\right]\right)$$

Being a martingale implies $\mathbb{E}[Y_T - Y_{T-1}|\mathcal{F}_{T-1}] = 0$, which gives

$$\mathbb{E}\left[\exp\big(\lambda\sum_{i=1}^{T}(Y_i - Y_{i-1})\big)\right] \leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{T-1}(Y_i - Y_{i-1})\right)\right]\exp\left(\frac{\lambda^2 c_T^2}{2}\right)$$

We can now obtain the following upper bound on $\mathbb{P}(Y_T - Y_0 \geq t)$,

$$\exp(-\lambda t)\exp\left(\frac{\sum_{i=1}^{N}\lambda^2 c_i^2}{2}\right).$$

Finally, optimizing with respect to λ we obtain

$$\mathbb{P}(Y_T - Y_0 \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{T} c_i^2}\right)$$

□

7.2 Tools from the analysis of linear bandits

Lemma 15 (Generalized Woodbury matrix identity)**.** *[Higham, 2008, Theorem 1.35] Let* $\mathbf{A} \in \mathbb{C}^{d\times m}$ *and* $\mathbf{B} \in \mathbb{C}^{m\times d}$*, with* $d \geq m$*, and assume that* $\mathbf{BA}$ *is nonsingular. Let* f *be defined on the spectrum of* $\alpha\mathbf{I}_{d\times d} + \mathbf{AB}$*, and if* $d = m$ *let* f *be defined at* α*. Then*

$$f(\alpha\mathbf{I}_{d\times d} + \mathbf{AB}) = f(\alpha\mathbf{I}_{d\times d}) + \mathbf{A}(\mathbf{BA})^{-1}\left(f(\alpha\mathbf{I}_{m\times m} + \mathbf{BA}) - f(\alpha\mathbf{I}_{m\times m})\right)\mathbf{B}.$$

Proof. By [Higham, 2008, Theorem 1.32] the given assumption on f implies that f is defined on the spectrum of $\alpha\mathbf{I}_{m\times m}+\mathbf{BA}$ and at α. Let now $g(t) = f[\alpha+t,\alpha] = t^{-1}\big(f(\alpha+t)-f(\alpha)\big)$, then $f(\alpha + t) = f(\alpha) + tg(t)$. Finally, using [Higham, 2008, Corollary 1.34],

$$\begin{aligned} f(\alpha + \mathbf{AB}) &= f(\alpha)\mathbf{I}_{d\times d} + \mathbf{AB}g(\mathbf{AB}) \\ &= f(\alpha)\mathbf{I}_{d\times d} + \mathbf{A}g(\mathbf{BA})\mathbf{B} \\ &= f(\alpha)\mathbf{I}_{d\times d} + \mathbf{A}(\mathbf{BA})^{-1}(f(\alpha\mathbf{I}_{m\times m} + \mathbf{BA}) - f(\alpha)\mathbf{I}_{m\times m})\mathbf{B}, \end{aligned}$$

as required. □

Lemma 16 (AM-GM Inequality)**.** *For any set of non-negative real numbers, the arithmetic mean of the set is greater than or equal to the geometric mean of the set. Algebraically, this is expressed as follows. For a set of non-negative real numbers $a_1, a_2, \ldots, a_n$, the following always holds:*

$$\frac{a_1 + a_2 + \ldots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \cdots a_n}$$

Using the shorthand notation for summations and products:

$$\sum_{i=1}^{n} \frac{a_i}{n} \geq \prod_{i=1}^{n} a_i^{\frac{1}{n}}.$$

Proof. We note that the function $x \mapsto \ln x$ is strictly concave. Then by Jensen's Inequality,

$$\ln \sum_i \lambda_i a_i \geq \sum_i \lambda_i \ln a_i = \ln \prod_i a_i^{\lambda_i},$$

with equality if and only if all the a_i are equal. Since $x \mapsto \ln x$ is a strictly increasing function, it then follows that

$$\sum_i \lambda_i a_i \geq \prod_i a_i^{\lambda_i},$$

with equality if and only if all the a_i are equal, as desired. □

Lemma 17 (Determinant-trace inequality)**.** *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$ $\|\mathbf{x}_s\|_2 \leq L$. Let $\mathbf{V}_t^\lambda = \lambda \mathbf{I} + \sum_{s=1}^t \mathbf{x}_s \mathbf{x}_s^\top$ for some $\lambda > 0$. As proved in [Abbasi-Yadkori et al., 2011, Lemma 10], the following holds*

$$\ln \det \left(\mathbf{V}_t^\lambda\right) \leq d \ln \left(\lambda + \frac{tL^2}{d}\right).$$

Proof. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of the covariance matrix $\mathbf{V}_t^\lambda$. Since by construction $\mathbf{V}_t^\lambda$ is positive definite, its eigenvalues are positive. Furthermore, by the definition of determinant we have $\det\left(\mathbf{V}_t^\lambda\right) = \prod_{s=1}^d \lambda_s$ and $\mathrm{tr}\left(\mathbf{V}_t^\lambda\right) = \sum_{s=1}^d \lambda_i$. By using the AM-GM inequality (Lemma 16) we have $\det\left(\mathbf{V}_t^\lambda\right) \leq \left(\mathrm{tr}\left(\mathbf{V}_t^\lambda\right)/d\right)^d$. It remains only to upper bound the trace:

$$\mathrm{tr}\left(\mathbf{V}_t^\lambda\right) = \mathrm{tr}\left(\lambda \mathbf{I}\right) + \sum s = 1^t \mathrm{tr}\left(\mathbf{x}_s \mathbf{x}_s^\top\right) = d\lambda + \sum_{s=1}^{t} \|\mathbf{x}_s\|_2^2 \leq d\lambda + tL^2.$$

□

Lemma 18 (Ridge leverage scores)**.** *Coherently to the notation used in Lemma 17, as proved in [Abbasi-Yadkori et al., 2011, lemma 11], we have*

$$\sum_{t=1}^{T} \min\left\{1, \|\mathbf{x}_t\|^2_{\left(\mathbf{V}_{t-1}^\lambda\right)^{-1}}\right\} \leq 2 \ln \left(\frac{\det\left(\mathbf{V}_T^\lambda\right)}{\lambda \mathbf{I}}\right). \tag{7.2}$$

For $\lambda \geq \max\{1, L^2\}$, we also have that

$$\sum_{t=1}^{T} \|\mathbf{x}_t\|^2_{\left(\mathbf{V}_{t-1}^\lambda\right)^{-1}} \leq 2d \ln \left(1 + \frac{TL^2}{\lambda d}\right). \tag{7.3}$$

Proof. The determinant of the regularized covariance matrix $\mathbf{V}_T^\lambda$ ca be rewritten as

$$\det\left(\mathbf{V}_T^\lambda\right) = \det\left(\mathbf{V}_{T-1}^\lambda + \mathbf{x}_T\mathbf{x}_T^\top\right) = \det\left(\mathbf{V}_{T-1}^\lambda\right)\det\left(\mathbf{I} + \left(\mathbf{V}_{T-1}^\lambda\right)^{-1/2}\mathbf{x}_T\mathbf{x}_T^\top\left(\mathbf{V}_{T-1}^\lambda\right)^{-1/2}\right)$$
$$= \det\left(\lambda\mathbf{I}\right)\prod_{t=1}^{T}\left(1 + \|\mathbf{x}_t\|^2_{\left(\mathbf{V}_{t-1}^\lambda\right)^{-1}}\right) \tag{7.4}$$

Using $\log(1+t) \leq t$ we have the following bound

$$\log\left(\det\left(\mathbf{V}_T^\lambda\right)\right) \leq \log\det\left(\lambda\mathbf{I}\right) + \sum_{t=1}^{T}\|\mathbf{x}_t\|^2_{\left(\mathbf{V}_{t-1}^\lambda\right)^{-1}}.$$

Furthermore, combining $x \leq 2\log(1+x)$ which holds for $x \in [0,1]$ with 7.4 we obtain

$$\sum_{t=1}^{T}\min\left\{1, \|\mathbf{x}_t\|^2_{\left(\mathbf{V}_{t-1}^\lambda\right)^{-1}}\right\} \leq 2\sum_{t=1}^{T}\log\left(1 + \|\mathbf{x}_t\|^2_{\left(\mathbf{V}_{t-1}^\lambda\right)^{-1}}\right) = 2\log\left(\frac{\det(\mathbf{V}_T^\lambda)}{\det(\lambda\mathbf{I})}\right)$$

As soon as we pick a large enough value $\lambda > 0$, we can upper bound the sum $\sum_{t=1}^{T}\|\mathbf{x}_t\|^2_{\left(\mathbf{V}_{t-1}^\lambda\right)^{-1}}$ as a function of $\log\left(\det\left(\mathbf{V}_T^\lambda\right)\right)$. Furthermore, for any $t > 0$ we have

$$\|\mathbf{x}_t\|^2_{\left(\mathbf{V}_{t-1}^\lambda\right)^{-1}} \leq \lambda_{\min}^{-1}\left(\mathbf{V}_T^\lambda\right)\|\mathbf{x}_t\|_2^2 \leq \frac{L^2}{\lambda}.$$

Hence, as soon as $\lambda > \max(1, L^2)$ we have that:

$$\log\left(\frac{\det(\mathbf{V}_T^\lambda)}{\det(\lambda\mathbf{I})}\right) \leq \sum_{t=1}^{T}\|\mathbf{x}_t\|^2_{\left(\mathbf{V}_t^\lambda\right)^{-1}} \leq 2\log\left(\frac{\det(\mathbf{V}_T^\lambda)}{\det(\lambda\mathbf{I})}\right) \tag{7.5}$$

□

Lemma 19 (Self-normalized bound for vector-valued martingales)**.** *Let*

$$S_t = \sum_{s=1}^{t}\eta_s\mathbf{x}_s \qquad t \geq 1$$

where $\eta_1, \eta_2, \ldots$ is a conditionally R-subgaussian real-valued stochastic process and $\mathbf{x}_1, \mathbf{x}_2, \ldots$ is any $\mathbb{R}^d$-valued stochastic process such that $\mathbf{x}_t$ is measurable with respect to the σ-algebra generated by $\eta_1, \ldots, \eta_{t-1}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, $\|S_t\|^2_{\left(\mathbf{V}_t^\lambda\right)^{-1}} \leq B_t(\delta)$ for all $t \geq 0$, where

$$B_t(\delta) = 2R^2\ln\left(\frac{1}{\delta}\det\left(\mathbf{V}_t^\lambda\right)^{\frac{1}{2}}\det\left(\lambda\mathbf{I}\right)^{-\frac{1}{2}}\right). \tag{7.6}$$

Proof. The proof relies on a stopping time construction which goes back to Freedman [1975]. Defining the bad event:

$$\xi_t(\delta) = \left\{ w \in \Omega : \|S_t\|^2_{\left(\mathbf{V}_t^\lambda\right)^{-1}} > 2R^2 \ln\left(\frac{1}{\delta} \det\left(\mathbf{V}_t^\lambda\right)^{\frac{1}{2}} \det\left(\lambda \mathbf{I}\right)^{-\frac{1}{2}}\right) \right\}$$

We want to bound the probability of occurrence of $\bigcup_{t \geq 0} \xi_t(\delta)$. We then define the stopping time $\tau(w) = \min\{t \geq 0 : w \in w \in \xi_t(\delta)\}$ with the convention that $\min \emptyset = \infty$. Further,

$$\bigcup_{t \geq 0} \xi_t(\delta) = \{w : \tau(w) < \infty\}$$

Finally, by [Abbasi-Yadkori et al., 2011, Lemma 9] the following holds

$$\begin{aligned}
\mathbb{P}\left[\bigcup_{t \geq 0} \xi_t(\delta)\right] &= \mathbb{P}[\tau < \infty] \\
&= \mathbb{P}\left[\|S_t\|^2_{\left(\mathbf{V}_t^\lambda\right)^{-1}} > 2R^2 \ln\left(\frac{1}{\delta} \det\left(\mathbf{V}_t^\lambda\right)^{\frac{1}{2}} \det\left(\lambda \mathbf{I}\right)^{-\frac{1}{2}}\right), \tau < \infty\right] \\
&\leq \mathbb{P}\left[\|S_t\|^2_{\left(\mathbf{V}_t^\lambda\right)^{-1}} > 2R^2 \ln\left(\frac{1}{\delta} \det\left(\mathbf{V}_t^\lambda\right)^{\frac{1}{2}} \det\left(\lambda \mathbf{I}\right)^{-\frac{1}{2}}\right)\right] \leq \delta.
\end{aligned}$$

□

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*, NIPS'11, pages 2312–2320, USA, 2011. Curran Associates Inc. ISBN 978-1-61839-599-3. URL http://dl.acm.org/citation.cfm?id=2986459.2986717.

M. Abeille and A. Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learing (ICML)*, pages 127–135, 2013.

Pierre Alquier, The Tien Mai, and Massimiliano Pontil. Regret bounds for lifelong learning. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on rtificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 261–269, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL http://proceedings.mlr.press/v54/alquier17a.html.

Rie Kubota Ando and Tong Zhang. 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.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1747–1754. Omnipress, 2012.

Peter Auer. 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.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 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.

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 - Volume 2*, NIPS'13, pages 2220–2228, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999860.

Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433, 2019.

Amotz Bar-Noy, Randeep Bhatia, Joseph (Seffi) Naor, and Baruch Schieber. Minimizing service and operation costs of periodic scheduling. *Mathematics of Operations Research*, 27:518–544, 2002.

Soumya Basu, Rajat Sen, Sujay Sanghavi, and Sanjay Shakkottai. Blocking bandits. In *Advances in Neural Information Processing Systems*, pages 4785–4794, 2019.

Jonathan Baxter. 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.

Djallel Bouneffouf and Raphael Féraud. Multi-armed bandit problem with known trend. *Neurocomputing*, 205:16–21, 2016.

Guy Bresler, George H Chen, and Devavrat Shah. A latent source model for online collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 3347–3355, 2014.

Guy Bresler, Devavrat Shah, and Luis Filipe Voloch. Collaborative filtering with low regret. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, SIGMETRICS '16, page 207–220, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342667. doi: 10.1145/2896377.2901469. URL https://doi.org/10.1145/2896377.2901469.

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

D. Calandriello, A. Lazaric, and M. Valko. Efficient second-order online kernel learning with adaptive embedding. In *Conference on Neural Information Processing Systems (NIPS)*, pages 6140–6150, 2017.

Daniele Calandriello, Alessandro Lazaric, and Marcello Restelli. Sparse multi-task reinforcement learning. In *NIPS - Advances in Neural Information Processing Systems 26*, 2014. URL https://hal.inria.fr/hal-01073513.

Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. *arXiv preprint arXiv:1903.05594*, 2019.

Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 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.

Giovanni Cavallanti, Nicolò Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *J. Mach. Learn. Res.*, 11:2901–2934, December 2010. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1953026.

Leonardo Cella and Nicolò Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs. In *International Conference on Artificial Intelligence and Statistics*, pages 1168–1177. PMLR, 2020.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. *arXiv preprint arXiv:2005.08531*, 2020.

Nicolò Cesa-Bianchi. *Multi-armed Bandit Problem*. Springer New York, New York, NY, 2016. ISBN 978-1-4939-2864-4.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, pages 1160–1168, 2013a.

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

Cheng Chen, Luo Luo, Weinan Zhang, Yong Yu, and Yijiang Lian. Efficient and robust high-dimensional linear contextual bandits. IJCAI, 2020.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL http://proceedings.mlr.press/v15/chu11a.html.

Corinna Cortes, Giulia DeSalvo, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Discrepancy-based algorithms for non-stationary rested bandits. *arXiv preprint arXiv:1710.10657*, 2017.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10169–10179. Curran Associates, Inc., 2018a.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Incremental learning-to-learn with statistical guarantees. *arXiv preprint arXiv:1803.08089*, 2018b.

Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine*

Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 1566–1575, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Aniket Anand Deshmukh, Urun Dogan, and Clayton Scott. Multi-task learning for contextual bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4851–4859, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, 7:1079–1105, December 2006.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–757–II–765. JMLR.org, 2014.

Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, and Evans Etrue. On context-dependent clustering of bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1253–1262, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/gentile17a.html.

M. Ghashami, E. Liberty, J. M. Phillips, and D. P. Woodruff. Frequent directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5):1762–1792, 2016.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

A. Gonen, F. Orabona, and S. Shalev-Shwartz. Solving ridge regression using sketched preconditioned SVRG. In *International Conference on Machine Learing (ICML)*, pages 1397–1405, 2016.

Sudipto Guha, Kamesh Munagala, and Peng Shi. Approximation algorithms for restless bandit problems. *Journal of ACM*, 58(1):3:1–3:50, December 2010. ISSN 0004-5411. doi: 10.1145/1870103.1870106. URL http://doi.acm.org/10.1145/1870103.1870106.

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

Hoda Heidari, Michael Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI '16)*, pages 1562–1570. AAAI Press, 2016.

N. J. Higham. *Functions of matrices: theory and computation*, volume 104. Siam, 2008.

K.-S. Jun, A. Bhargava, R. Nowak, and R. Willett. Scalable generalized linear bandits: Online computation and hashing. In *Conference on Neural Information Processing Systems (NIPS)*, pages 99–109, 2017.

Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. DCM bandits: Learning to rank with multiple clicks. In *International Conference on Machine Learning*, pages 1215–1224, 2016.

Robert Kleinberg and Nicole Immorlica. Recharging bandits. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 309–319. IEEE, 2018.

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

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776, 2015a.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 535–543, 2015b.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10:154–166, 1982.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2018.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Advances in Neural Information Processing Systems*, pages 3074–3083, 2017.

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

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

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

H. Luo, A. Agarwal, N. Cesa-Bianchi, and J. Langford. Efficient second order online learning by sketching. In *Conference on Neural Information Processing Systems (NIPS)*, pages 902–910, 2016.

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

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

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. 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.

Anastasia Pentina and Ruth Urner. Lifelong learning with weighted majority votes. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3612–3620. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6095-lifelong-learning-with-weighted-majority-votes.pdf.

Ciara Pike-Burke and Steffen Grunewalder. Recovering bandits. In *The 14th European Workshop on Reinforcement Learning (EWRL 2018)*, 2018.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, pages 784–791. ACM, 2008.

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

Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 16–18 Apr 2019.

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

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

Marta Soare, Ouais Alsharif, Alessandro Lazaric, and Joelle Pineau. Multi-task linear bandits. In *NIPS'14 Workshop on Transfer and Multi-task Learning*, 2014.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

Sofía S Villar, Jack Bowden, and James Wason. 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.

Romain Warlop, Alessandro Lazaric, and Jérémie Mary. Fighting boredom in recommender systems with linear reinforcement learning. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 1764–1773, USA, 2018. Curran Associates Inc.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

D. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Xiaotian Yu, Michael R. Lyu, and Irwin King. Cbrap: Contextual bandits with random projection. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 2859–2866. AAAI Press, 2017.

Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandits: A causal approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 1340–1346. AAAI Press, 2017. ISBN 978-0-9992411-0-3. URL http://dl.acm.org/citation.cfm?id=3171642.3171832.