Sequential Transfer in Multi-armed Bandit with Finite Set of Models

Mohammad Gheshlaghi Azar, Alessandro Lazaric, Emma Brunskill

Sequential Transfer in Multi-armed Bandit with Finite Set of Models

Mohammad Gheshlaghi Azar *
School of Computer Science
CMU

Alessandro Lazaric †
INRIA Lille - Nord Europe
Team SequeL

Emma Brunskill *
School of Computer Science
CMU

Abstract

Learning from prior tasks and transferring that experience to improve future performance is critical for building lifelong learning agents. Although results in supervised and reinforcement learning show that transfer may significantly improve the learning performance, most of the literature on transfer is focused on batch learning tasks. In this paper we study the problem of *sequential transfer in online learning*, notably in the multi–armed bandit framework, where the objective is to minimize the total regret over a sequence of tasks by transferring knowledge from prior tasks. We introduce a novel bandit algorithm based on a method-of-moments approach for estimating the possible tasks and derive regret bounds for it.

1 Introduction

Learning from prior tasks and transferring that experience to improve future performance is a key aspect of intelligence, and is critical for building lifelong learning agents. Recently, multi-task and transfer learning received much attention in the supervised and reinforcement learning (RL) setting with both empirical and theoretical encouraging results (see recent surveys by Pan and Yang, 2010; Lazaric, 2011). Most of these works focused on scenarios where the tasks are batch learning problems, in which a training set is directly provided to the learner. On the other hand, the online learning setting (Cesa-Bianchi and Lugosi, 2006), where the learner is presented with samples in a sequential fashion, has been rarely considered (see Mann and Choe (2012); Taylor (2009) for examples in RL and Sec. E of **?** for a discussion on related settings).

The multi–armed bandit (MAB) (Robbins, 1952) is a simple yet powerful framework formalizing the online learning with partial feedback problem, which encompasses a large number of applications, such as clinical trials, web advertisements and adaptive routing. In this paper we take a step towards understanding and providing formal bounds on transfer in stochastic MABs. We focus on a *sequential transfer* scenario where an (online) learner is acting in a series of tasks drawn from a stationary distribution over a finite set of MABs. The learning problem, within each task, can be seen as a standard MAB problem with a fixed number of steps. Prior to learning, the model parameters of each bandit problem are not known to the learner, nor does it know the distribution probability over the bandit problems. Also, we assume that the learner is not provided with the identity of the tasks throughout the learning. To act efficiently in this setting, it is crucial to define a mechanism for transferring knowledge across tasks. In fact, the learner may encounter the same bandit problem over and over throughout the learning, and an efficient algorithm should be able to leverage the knowledge obtained in previous tasks, when it is presented with the same problem again. To address this problem one can transfer the estimates of all the possible models from prior tasks to the current one. Once these models are accurately estimated, we show that an extension of the *UCB* algorithm (Auer et al., 2002) is able to efficiently exploit this prior knowledge and reduce the regret through tasks (Sec. 3).

* {mazar,ebrun}@cs.cmu.edu
† alessandro.lazaric@inria.fr

The main contributions of this paper are two-fold: **(i)** we introduce the *tUCB* algorithm which transfers the model estimates across the tasks and uses this knowledge to achieve a better performance than *UCB*. We also prove that the new algorithm is guaranteed to perform as well as *UCB* in early episodes, thus avoiding any *negative transfer* effect, and then to approach the performance of the ideal case when the models are all known in advance (Sec. 4.4). **(ii)** To estimate the models we rely on a new variant of method of moments, robust tensor power method (RTP) (Anandkumar et al., 2013, 2012b) and extend it to the multi-task bandit setting[1]:we prove that *RTP* provides a consistent estimate of the means of all arms (for all models) as long as they are pulled at least three times per task and prove sample complexity bounds for it (Sec. 4.2). Finally, we report some preliminary results on synthetic data confirming the theoretical findings (Sec. 5). An extended version of this paper containing proofs and additional comments is available in (**?**).

2 Preliminaries

We consider a stochastic MAB problem defined by a set of arms $\mathcal{A} = \{1, \dots, K\}$, where each arm $i \in \mathcal{A}$ is characterized by a distribution ν_i and the samples (rewards) observed from each arm are independent and identically distributed. We focus on the setting where there exists a set of models $\Theta = \{\theta = (\nu_1, \dots, \nu_K)\}$, $|\Theta| = m$, which contains all the possible bandit problems. We denote the mean of an arm i, the best arm, and the best value of a model $\theta \in \Theta$ respectively by $\mu_i(\theta)$, $i_*(\theta)$, $\mu_*(\theta)$. We define the arm gap of an arm i for a model θ as $\Delta_i(\theta) = \mu_*(\theta) - \mu_i(\theta)$, while the model gap for an arm i between two models θ and θ' is defined as $\Gamma_i(\theta, \theta') = |\mu_i(\theta) - \mu_i(\theta')|$. We also assume that arm rewards are bounded in $[0, 1]$. We consider the sequential transfer setting where at each episode j the learner interacts with a task $\bar{\theta}^j$, drawn from a distribution ρ over Θ, for n steps. The objective is to minimize the (pseudo-)regret $\mathcal{R}_J$ over J episodes measured as the difference between the rewards obtained by pulling $i_*(\bar{\theta}^j)$ and those achieved by the learner:

$$\mathcal{R}_J = \sum_{j=1}^{J} \mathcal{R}_n^j = \sum_{j=1}^{J} \sum_{i \neq i^*} T_{i,n}^j \Delta_i(\bar{\theta}^j), \tag{1}$$

where $T_{i,n}^j$ is the number of pulls to arm i after n steps of episode j. We also introduce some tensor notation. Let $X \in \mathbb{R}^K$ be a random realization of the rewards of all arms from a random model. All the realizations are i.i.d. conditional on a model $\bar{\theta}$ and $\mathbb{E}[X|\theta = \bar{\theta}] = \mu(\theta)$, where the i-th component of $\mu(\theta) \in \mathbb{R}^K$ is $[\mu(\theta)]_i = \mu_i(\theta)$. Given realizations X^1 , X^2 and X^3, we define the second moment matrix $M_2 = \mathbb{E}[X^1 \otimes X^2]$ such that $[M_2]_{i,j} = \mathbb{E}[X_i^1 X_j^2]$ and the third moment tensor $M_3 = \mathbb{E}[X^1 \otimes X^2 \otimes X^3]$ such that $[M_2]_{i,j,l} = \mathbb{E}[X_i^1 X_j^2 X_l^3]$. Since the realizations are conditionally independent, we have that, for every $\theta \in \Theta$, $\mathbb{E}[X^1 \otimes X^2|\theta] = \mathbb{E}[X^1|\theta] \otimes \mathbb{E}[X^2|\theta] = \mu(\theta) \otimes \mu(\theta)$ and this allows us to rewrite the second and third moments as $M_2 = \sum_\theta \rho(\theta)\mu(\theta)^{\otimes 2}$, $M_3 = \sum_\theta \rho(\theta)\mu(\theta)^{\otimes 3}$, where $v^{\otimes p} = v \otimes v \otimes \cdots v$ is the p-th tensor power. Let A be a 3rd order member of the tensor product of the Euclidean space $\mathbb{R}^K$ (as M_3), then we define the multilinear map as follows. For a set of three matrices $\{V_i \in \mathbb{R}^{K \times m}\}_{1 \le i \le 3}$, the (i_1, i_2, i_3) entry in the 3-way array representation of $A(V_1, V_2, V_3) \in \mathbb{R}^{m \times m \times m}$ is $[A(V_1, V_2, V_3)]_{i_1,i_2,i_3} := \sum_{1 \le j_1, j_2, j_3 \le n} A_{j_1,j_2,j_3} [V_1]_{j_1,i_1} [V_2]_{j_2,i_2} [V_3]_{j_3,i_3}$. We also use different norms: the Euclidean norm $\|\cdot\|$; the Frobenius norm $\|\cdot\|_F$; the matrix max-norm $\|A\|_{\max} = \max_{ij} |[A]_{ij}|$.

3 Multi-arm Bandit with Finite Models

Before considering the transfer problem, we show that a simple variation to *UCB* allows us to effectively exploit the knowledge of Θ and obtain a significant reduction in the regret. The *mUCB* (model-UCB) algorithm in Fig. 1 takes as input a set of models Θ including the current (unknown) model $\bar{\theta}$. At each step t, the algorithm computes a subset $\Theta_t \subseteq \Theta$ containing only the models whose means $\mu_i(\theta)$ are *compatible* with the current estimates $\hat{\mu}_{i,t}$ of the means $\mu_i(\bar{\theta})$ of the current model, obtained averaging

> **Require:** Set of models Θ, number of steps n
> **for** $t = 1, \dots, n$ **do**
> Build $\Theta_t = \{\theta : \forall i, |\mu_i(\theta) - \hat{\mu}_{i,t}| \le \varepsilon_{i,t}\}$
> Select $\theta_t = \arg\max_{\theta \in \Theta_t} \mu_*(\theta)$
> Pull arm $I_t = i_*(\theta_t)$
> Observe sample x_{I_t} and update
> **end for**

Figure 1: The *mUCB* algorithm.

[1]Notice that estimating the models involves solving a latent variable model estimation problem, for which RTP is the state-of-the-art.

$T_{i,t}$ pulls, and their uncertainty $\varepsilon_{i,t}$ (see Eq. 2 for an explicit definition of this term). Notice that it is enough that one arm does not satisfy the compatibility condition to discard a model θ. Among all the models in Θ_t, *mUCB* first selects the model with the largest optimal value and then it pulls its corresponding optimal arm. This choice is coherent with the *optimism in the face of uncertainty* principle used in UCB-based algorithms, since *mUCB* always pulls the optimal arm corresponding to the optimistic model compatible with the current estimates $\hat{\mu}_{i,t}$. We show that *mUCB* incurs a regret which is never worse than *UCB* and it is often significantly smaller.

We denote the set of arms which are optimal for at least a model in a set Θ' as $\mathcal{A}_*(\Theta') = \{i \in \mathcal{A} : \exists \theta \in \Theta' : i_*(\theta) = i\}$. The set of models for which the arms in $\mathcal{A}'$ are optimal is $\Theta(\mathcal{A}') = \{\theta \in \Theta : \exists i \in \mathcal{A}' : i_*(\theta) = i\}$. The set of optimistic models for a given model $\bar{\theta}$ is $\Theta_+ = \{\theta \in \Theta : \mu_*(\theta) \geq \mu_*(\bar{\theta})\}$, and their corresponding optimal arms $\mathcal{A}_+ = \mathcal{A}_*(\Theta_+)$. The following theorem bounds the expected regret (similar bounds hold in high probability). The lemmas and proofs (using standard tools from the bandit literature) are available in Sec. B of **?**.

Theorem 1. *If mUCB is run with $\delta = 1/n$, a set of m models Θ such that the $\bar{\theta} \in \Theta$ and*

$$\varepsilon_{i,t} = \sqrt{\log(mn^2/\delta)/(2T_{i,t-1})}, \tag{2}$$

where $T_{i,t-1}$ is the number of pulls to arm i at the beginning of step t, then its expected regret is

$$\mathbb{E}[\mathcal{R}_n] \leq K + \sum\nolimits_{i \in \mathcal{A}_+} \frac{2\Delta_i(\bar{\theta}) \log\big(mn^3\big)}{\min_{\theta \in \Theta_{+,i}} \Gamma_i(\theta, \bar{\theta})^2} \leq K + \sum\nolimits_{i \in \mathcal{A}_+} \frac{2 \log\big(mn^3\big)}{\min_{\theta \in \Theta_{+,i}} \Gamma_i(\theta, \bar{\theta})}, \tag{3}$$

where $\mathcal{A}_+ = \mathcal{A}_*(\Theta_+)$ is the set of arms which are optimal for at least one optimistic model Θ_+ and $\Theta_{+,i} = \{\theta \in \Theta_+ : i_*(\theta) = i\}$ is the set of optimistic models for which i is the optimal arm.

Remark (comparison to *UCB*). The *UCB* algorithm incurs a regret

$$\mathbb{E}[\mathcal{R}_n(\text{UCB})] \leq O\Big(\sum\nolimits_{i \in \mathcal{A}} \frac{\log n}{\Delta_i(\bar{\theta})}\Big) \leq O\Big(K \frac{\log n}{\min_i \Delta_i(\bar{\theta})}\Big).$$

We see that *mUCB* displays two major improvements. The regret in Eq. 3 can be written as

$$\mathbb{E}[\mathcal{R}_n(\text{mUCB})] \leq O\Big(\sum\nolimits_{i \in \mathcal{A}_+} \frac{\log n}{\min_{\theta \in \Theta_{+,i}} \Gamma_i(\theta, \bar{\theta})}\Big) \leq O\Big(|\mathcal{A}_+| \frac{\log n}{\min_i \min_{\theta \in \Theta_{+,i}} \Gamma_i(\theta, \bar{\theta})}\Big).$$

This result suggests that *mUCB* tends to discard all the models in Θ_+ from the most optimistic down to the actual model $\bar{\theta}$ which, with high-probability, is never discarded. As a result, even if other models are still in Θ_t, the optimal arm of $\bar{\theta}$ is pulled until the end. This significantly reduces the set of arms which are actually pulled by *mUCB* and the previous bound only depends on the number of arms in $\mathcal{A}_+$, which is $|\mathcal{A}_+| \leq |\mathcal{A}_*(\Theta)| \leq K$. Furthermore, for all arms i, the minimum gap $\min_{\theta \in \Theta_{+,i}} \Gamma_i(\theta, \bar{\theta})$ is guaranteed to be larger than the arm gap $\Delta_i(\bar{\theta})$ (see Lem. 4 in Sec. B of **?**), thus further improving the performance of *mUCB* w.r.t. *UCB*.

4 Online Transfer with Unknown Models

We now consider the case when the set of models is unknown and the regret is cumulated over multiple tasks drawn from ρ (Eq. 1). We introduce *tUCB* (transfer-UCB) which transfers estimates of Θ, whose accuracy is improved through episodes using a method-of-moments approach.

4.1 The transfer-UCB Bandit Algorithm

Fig. 2 outlines the structure of our online transfer bandit algorithm *tUCB* (transfer-UCB). The algorithm uses two sub-algorithms, the bandit algorithm *umUCB* (*uncertain model-UCB*), whose objective is to minimize the regret at each episode, and *RTP* (*robust tensor power* method) which at each episode j computes an estimate $\{\hat{\mu}_i^j(\theta)\}$ of the arm means of all the models. The bandit algorithm *umUCB* in Fig. 3 is an extension of the *mUCB* algorithm. It first computes a set of models Θ_t^j whose means $\hat{\mu}_i(\theta)$ are compatible with the current estimates $\hat{\mu}_{i,t}$. However, unlike the case where the exact models are available, here the models themselves are estimated and the uncertainty ε^j in their means (provided as input to *umUCB*) is taken into account in the definition of Θ_t^j. Once

Require: number of arms K, number of models m, constant $C(\theta)$.
Initialize estimated models $\Theta^1 = \{\hat{\mu}_i^1(\theta)\}_{i,\theta}$, samples $R \in \mathbb{R}^{J\times K\times n}$
for $j = 1, 2, \dots, J$ **do**
  Run $R^j = \text{umUCB}(\Theta^j, n)$
  Run $\Theta^{j+1} = \text{RTP}(R, m, K, j, \delta)$
end for

Figure 2: The *tUCB* algorithm.

Require: set of models Θ^j, num. steps n
Pull each arm three times
for $t = 3K+1, \dots, n$ **do**
  Build $\Theta_t^j = \{\theta : \forall i, |\hat{\mu}_i^j(\theta) - \hat{\mu}_{i,t}| \le \varepsilon_{i,t} + \varepsilon^j\}$
  Compute $B_t^j(i;\theta) = \min\big\{(\hat{\mu}_i^j(\theta)+\varepsilon^j), (\hat{\mu}_{i,t}+\varepsilon_{i,t})\big\}$
  Compute $\theta_t^j = \arg\max_{\theta\in\Theta_t^j} \max_i B_t^j(i;\theta)$
  Pull arm $I_t = \arg\max_i B_t^j(i;\theta_t^j)$
  Observe sample $R(I_t, T_{i,t}) = x_{I_t}$ and update
end for
return Samples R

Figure 3: The *umUCB* algorithm.

Require: samples $R \in \mathbb{R}^{j\times n}$, number of models m and arms K, episode j
Estimate the second and third moment $\widehat{M}_2$ and $\widehat{M}_3$ using the reward samples from R (Eq. 4)
Compute $\widehat{D} \in \mathbb{R}^{m\times m}$ and $\widehat{U} \in \mathbb{R}^{K\times m}$ (m largest eigenvalues and eigenvectors of $\widehat{M}_2$ resp.)
Compute the whitening mapping $\widehat{W} = \widehat{U}\widehat{D}^{-1/2}$ and the tensor $\widehat{T} = \widehat{M}_3(\widehat{W},\widehat{W},\widehat{W})$
Plug $\widehat{T}$ in Alg. 1 of Anandkumar et al. (2012b) and compute eigen-vectors/values $\{\widehat{v}(\theta)\}$, $\{\widehat{\lambda}(\theta)\}$
Compute $\widehat{\mu}^j(\theta) = \widehat{\lambda}(\theta)(\widehat{W}^\mathsf{T})^+\widehat{v}(\theta)$ for all $\theta \in \Theta$
return $\Theta^{j+1} = \{\widehat{\mu}^j(\theta) : \theta \in \Theta\}$

Figure 4: The robust tensor power (*RTP*) method (Anandkumar et al., 2012b).

the active set is computed, the algorithm computes an upper-confidence bound on the value of each arm i for each model θ and returns the best arm for the most optimistic model. Unlike in *mUCB*, due to the uncertainty over the model estimates, a model θ might have more than one optimal arm, and an upper-confidence bound on the mean of the arms $\hat{\mu}_i(\theta) + \varepsilon^j$ is used together with the upper-confidence bound $\hat{\mu}_{i,t} + \varepsilon_{i,t}$, which is directly derived from the samples observed so far from arm i. This guarantees that the B-values are always consistent with the samples generated from the actual model $\bar{\theta}^j$. Once *umUCB* terminates, *RTP* (Fig. 4) updates the estimates of the model means $\widehat{\mu}^j(\theta) = \{\hat{\mu}_i^j(\theta)\}_i \in \mathbb{R}^K$ using the samples obtained from each arm i. At the beginning of each task *umUCB* pulls all the arms 3 times, since *RTP* needs at least 3 samples from each arm to accurately estimate the 2nd and 3rd moments (Anandkumar et al., 2012b). More precisely, *RTP* uses all the reward samples generated up to episode j to estimate the 2nd and 3rd moments (see Sec. 2) as

$$\widehat{M}_2 = j^{-1}\sum\nolimits_{l=1}^{j} \overline{\mu}_{1l}\otimes\overline{\mu}_{2l}, \qquad \text{and} \qquad \widehat{M}_3 = j^{-1}\sum\nolimits_{l=1}^{j} \overline{\mu}_{1l}\otimes\overline{\mu}_{2l}\otimes\overline{\mu}_{3l}, \tag{4}$$

where the vectors $\overline{\mu}_{1l}, \overline{\mu}_{2l}, \overline{\mu}_{3l} \in \mathbb{R}^K$ are obtained by dividing the $T_{i,n}^l$ samples observed from arm i in episode l in three batches and taking their average (e.g., $[\overline{\mu}_{1l}]_i$ is the average of the first $T_{i,n}^l/3$ samples).[2] Since $\overline{\mu}_{1l}, \overline{\mu}_{2l}, \overline{\mu}_{3l}$ are independent estimates of $\mu(\bar{\theta}^l)$, $\widehat{M}_2$ and $\widehat{M}_3$ are consistent estimates of the second and third moments M_2 and M_3. *RTP* relies on the fact that the model means $\mu(\theta)$ can be recovered from the spectral decomposition of the symmetric tensor $T = M_3(W,W,W)$, where W is a whitening matrix for M_2, i.e., $M_2(W,W) = \mathbf{I}^{m\times m}$ (see Sec. 2 for the definition of the mapping $A(V_1,V_2,V_3)$). Anandkumar et al. (2012b) (Thm. 4.3) have shown that under some mild assumption (see later Assumption 1) the model means $\{\mu(\theta)\}$, can be obtained as $\mu(\theta) = \lambda(\theta)Bv(\theta)$, where $(\lambda(\theta), v(\theta))$ is a pair of eigenvector/eigenvalue for the tensor T and $B := (W^\mathsf{T})^+$.Thus the *RTP* algorithm estimates the eigenvectors $\widehat{v}(\theta)$ and the eigenvalues $\widehat{\lambda}(\theta)$, of the $m\times m\times m$ tensor $\widehat{T} := \widehat{M}_3(\widehat{W},\widehat{W},\widehat{W})$.[3] Once $\widehat{v}(\theta)$ and $\widehat{\lambda}(\theta)$ are computed, the estimated mean vector $\widehat{\mu}^j(\theta)$ is obtained by the inverse transformation $\widehat{\mu}^j(\theta) = \widehat{\lambda}(\theta)\widehat{B}\widehat{v}(\theta)$, where $\widehat{B}$ is the pseudo inverse of $\widehat{W}^\mathsf{T}$(for a detailed description of RTP algorithm see Anandkumar et al., 2012b).

[2]Notice that $1/3([\overline{\mu}_{1l}]_i + [\overline{\mu}_{2l}]_i + [\overline{\mu}_{1l}]_i) = \hat{\mu}_{i,n}^l$, the empirical mean of arm i at the end of episode l.

[3]The matrix $\widehat{W} \in \mathbb{R}^{K\times m}$ is such that $\widehat{M}_2(\widehat{W},\widehat{W}) = \mathbf{I}^{m\times m}$, i.e., $\widehat{W}$ is the whitening matrix of $\widehat{M}_2$. In general $\widehat{W}$ is not unique. Here, we choose $\widehat{W} = \widehat{U}\widehat{D}^{-1/2}$, where $\widehat{D} \in \mathbb{R}^{m\times m}$ is a diagonal matrix consisting of the m largest eigenvalues of $\widehat{M}_2$ and $\widehat{U} \in \mathbb{R}^{K\times m}$ has the corresponding eigenvectors as its columns.

4.2 Sample Complexity of the Robust Tensor Power Method

umUCB requires as input ε^j, i.e., the uncertainty of the model estimates. Therefore we need sample complexity bounds on the accuracy of $\{\hat{\mu}_i(\theta)\}$ computed by *RTP*. The performance of *RTP* is directly affected by the error of the estimates $\widehat{M_2}$ and $\widehat{M_3}$ w.r.t. the true moments. In Thm. 2 we prove that, as the number of tasks j grows, this error rapidly decreases with the rate of $\sqrt{1/j}$. This result provides us with an upper-bound on the error ε^j needed for building the confidence intervals in *umUCB*. The following definition and assumption are required for our result.

Definition 1. *Let $\Sigma_{M_2} = \{\sigma_1, \sigma_2, \dots, \sigma_m\}$ be the set of m largest eigenvalues of the matrix M_2. Define $\sigma_{\min} := \min_{\sigma \in \Sigma_{M_2}} \sigma$, $\sigma_{\max} := \max_{\sigma \in \Sigma_{M_2}} \sigma$ and $\lambda_{\max} := \max_\theta \lambda(\theta)$. Define the minimum gap between the distinct eigenvalues of M_2 as $\Gamma_\sigma := \min_{\sigma_i \neq \sigma_l}(|\sigma_i - \sigma_l|)$.*

Assumption 1. *The mean vectors $\{\mu(\theta)\}_\theta$ are linear independent and $\rho(\theta) > 0$ for all $\theta \in \Theta$.*

We now state our main result which is in the form of a high probability bound on the estimation error of mean reward vector of every model $\theta \in \Theta$.

Theorem 2. *Pick $\delta \in (0,1)$. Let $C(\Theta) := C_3 \lambda_{\max} \sqrt{\sigma_{\max}/\sigma_{\min}^3}\,(\sigma_{\max}/\Gamma_\sigma + 1/\sigma_{\min} + 1/\sigma_{\max})$, where $C_3 > 0$ is a universal constant. Then under Assumption 1 there exist constants $C_4 > 0$ and a permutation π on Θ, such that for all $\theta \in \Theta$, we have w.p. $1 - \delta$*

$$\|\mu(\theta) - \widehat{\mu}^j(\pi(\theta))\| \leq \varepsilon_j \triangleq C(\Theta) K^{2.5} m^2 \sqrt{\frac{\log(K/\delta)}{j}} \quad \textit{after} \quad j \geq \frac{C_4 m^5 K^6 \log(K/\delta)}{\min(\sigma_{\min}, \Gamma_\sigma)^2 \sigma_{\min}^3 \lambda_{\min}^2}. \tag{5}$$

Remark (computation of $C(\Theta)$). As illustrated in Fig. 3, *umUCB* relies on the estimates $\widehat{\mu}^j(\theta)$ and on their accuracy ε^j. Although the bound reported in Thm. 2 provides an upper confidence bound on the error of the estimates, it contains terms which are not computable in general (e.g., $\sigma_{\min}$). In practice, $C(\Theta)$ should be considered as a parameter of the algorithm.This is not dissimilar from the parameter usually introduced in the definition of $\varepsilon_{i,t}$ in front of the square-root term in *UCB*.

4.3 Regret Analysis of *umUCB*

We now analyze the regret of *umUCB* when an estimated set of models Θ^j is provided as input. At episode j, for each model θ we define the set of non-dominated arms (i.e., potentially optimal arms) as $\mathcal{A}_*^j(\theta) = \{i \in \mathcal{A} : \nexists i', \hat{\mu}_i^j(\theta) + \varepsilon^j < \hat{\mu}_{i'}^j(\theta) - \varepsilon^j\}$. Among the non-dominated arms, when the actual model is $\bar{\theta}^j$, the set of optimistic arms is $\mathcal{A}_+^j(\theta; \bar{\theta}^j) = \{i \in \mathcal{A}_*^j(\theta) : \hat{\mu}_i^j(\theta) + \varepsilon^j \geq \mu^*(\bar{\theta}^j)\}$. As a result, the set of optimistic models is $\Theta_+^j(\bar{\theta}^j) = \{\theta \in \Theta : \mathcal{A}_+^j(\theta; \bar{\theta}^j) \neq \emptyset\}$. In some cases, because of the uncertainty in the model estimates, unlike in *mUCB*, not all the models $\theta \neq \bar{\theta}^j$ can be discarded, not even at the end of a very long episode. Among the optimistic models, the set of models that cannot be discarded is defined as $\widetilde{\Theta}_+^j(\bar{\theta}^j) = \{\theta \in \Theta_+^j(\bar{\theta}^j) : \forall i \in \mathcal{A}_+^j(\theta; \bar{\theta}^j), |\hat{\mu}_i^j(\theta) - \mu_i(\bar{\theta}^j)| \leq \varepsilon^j\}$. Finally, when we want to apply the previous definitions to a set of models Θ' instead of single model we have, e.g., $\mathcal{A}_*^j(\Theta'; \bar{\theta}^j) = \bigcup_{\theta \in \Theta'} \mathcal{A}_*^j(\theta; \bar{\theta}^j)$.

The proof of the following results are available in Sec. D of **?**, here we only report the number of pulls, and the corresponding regret bound.

Corollary 1. *If at episode j umUCB is run with $\varepsilon_{i,t}$ as in Eq. 2 and ε^j as in Eq. 5 with a parameter $\delta' = \delta/2K$, then for any arm $i \in \mathcal{A}$, $i \neq i_*(\bar{\theta}^j)$ is pulled $T_{i,n}$ times such that*

$$\begin{cases} T_{i,n} \leq \min\left\{ \dfrac{2\log\left(2mKn^2/\delta\right)}{\Delta_i(\bar{\theta}^j)^2}, \dfrac{\log\left(2mKn^2/\delta\right)}{2\min_{\theta \in \Theta_{i,+}^j(\bar{\theta}^j)} \widehat{\Gamma}_i(\theta; \bar{\theta}^j)^2} \right\} + 1 & \textit{if } i \in \mathcal{A}_1^j \\ T_{i,n} \leq 2\log\left(2mKn^2/\delta\right)/(\Delta_i(\bar{\theta}^j)^2) + 1 & \textit{if } i \in \mathcal{A}_2^j \\ T_{i,n} = 0 & \textit{otherwise} \end{cases}$$

w.p. $1 - \delta$, where $\Theta_{i,+}^j(\bar{\theta}^j) = \{\theta \in \Theta_+^j(\bar{\theta}^j) : i \in \mathcal{A}_+(\theta; \bar{\theta}^j)\}$ is the set of models for which i is among their optimistic non-dominated arms, $\widehat{\Gamma}_i(\theta; \bar{\theta}^j) = \Gamma_i(\theta, \bar{\theta}^j)/2 - \varepsilon^j$, $\mathcal{A}_1^j = \mathcal{A}_+^j(\Theta_+^j(\bar{\theta}^j); \bar{\theta}^j) - \mathcal{A}_+^j(\widetilde{\Theta}_+^j(\bar{\theta}^j); \bar{\theta}^j)$ (i.e., set of arms only proposed by models that can be discarded), and $\mathcal{A}_2^j = \mathcal{A}_+^j(\widetilde{\Theta}_+^j(\bar{\theta}^j); \bar{\theta}^j)$ (i.e., set of arms only proposed by models that cannot be discarded).

The previous corollary states that arms which cannot be optimal for any optimistic model (i.e., the optimistic non-dominated arms) are never pulled by *umUCB*, which focuses only on arms in $i \in \mathcal{A}_+^j(\Theta_+^j(\bar\theta^j);\bar\theta^j)$. Among these arms, those that may help to remove a model from the active set (i.e., $i \in \mathcal{A}_1^j$) are potentially pulled less than *UCB*, while the remaining arms, which are optimal for the models that cannot be discarded (i.e., $i \in \mathcal{A}_2^j$), are simply pulled according to a *UCB* strategy. Similar to *mUCB*, *umUCB* first pulls the arms that are more *optimistic* until either the active set Θ_t^j changes or they are no longer optimistic (because of the evidence from the actual samples). We are now ready to derive the per-episode regret of *umUCB*.

Theorem 3. *If umUCB is run for n steps on the set of models Θ^j estimated by RTP after j episodes with $\delta = 1/n$, and the actual model is $\bar\theta^j$, then its expected regret (w.r.t. the random realization in episode j and conditional on $\bar\theta^j$) is*

$$\mathbb{E}[\mathcal{R}_n^j] \leq K + \sum\nolimits_{i \in \mathcal{A}_1^j} \log\big(2mKn^3\big) \min\Big\{2/\Delta_i(\bar\theta^j)^2, 1/\big(2\min_{\theta\in\Theta^j_{i,+}(\bar\theta^j)} \widehat{\Gamma}_i(\theta;\bar\theta^j)^2\big)\Big\}\Delta_i(\bar\theta^j)$$
$$+ \sum\nolimits_{i \in \mathcal{A}_2^j} 2\log\big(2mKn^3\big)/\Delta_i(\bar\theta^j).$$

Remark (negative transfer). The transfer of knowledge introduces a bias in the learning process which is often beneficial. Nonetheless, in many cases transfer may result in a bias towards wrong solutions and a worse learning performance, a phenomenon often referred to as *negative transfer*. The first interesting aspect of the previous theorem is that *umUCB* is guaranteed to never perform worse than *UCB* itself. This implies that *tUCB* never suffers from negative transfer, even when the set Θ^j contains highly uncertain models and might bias *umUCB* to pull suboptimal arms.

Remark (improvement over *UCB*). In Sec. 3 we showed that *mUCB* exploits the knowledge of Θ to focus on a restricted set of arms which are pulled less than *UCB*. In *umUCB* this improvement is not as clear, since the models in Θ are not known but are estimated online through episodes. Yet, similar to *mUCB*, *umUCB* has the two main sources of potential improvement w.r.t. to *UCB*. As illustrated by the regret bound in Thm. 3, *umUCB* focuses on arms in $\mathcal{A}_1^j \cup \mathcal{A}_2^j$ which is potentially a smaller set than $\mathcal{A}$. Furthermore, the number of pulls to arms in $\mathcal{A}_1^j$ is smaller than for *UCB* whenever the estimated model gap $\widehat{\Gamma}_i(\theta;\bar\theta^j)$ is bigger than $\Delta_i(\bar\theta^j)$. Eventually, *umUCB* reaches the same performance (and improvement over *UCB*) as *mUCB* when j is big enough. In fact, the set of optimistic models reduces to the one used in *mUCB* (i.e., $\Theta_+^j(\bar\theta^j) \equiv \Theta_+(\bar\theta^j)$) and all the optimistic models have only optimal arms (i.e., for any $\theta \in \Theta_+$ the set of non-dominated optimistic arms is $\mathcal{A}_+(\theta;\bar\theta^j) = \{i_*(\theta)\}$), which corresponds to $\mathcal{A}_1^j \equiv \mathcal{A}_*(\Theta_+(\bar\theta^j))$ and $\mathcal{A}_2^j \equiv \{i_*(\bar\theta^j)\}$, which matches the condition of *mUCB*. For instance, for any model θ, in order to have $\mathcal{A}_*(\theta) = \{i_*(\theta)\}$, for any arm $i \neq i_*(\theta)$ we need that $\hat\mu_i^j(\theta) + \varepsilon^j \leq \hat\mu^j_{i_*(\theta)}(\theta) - \varepsilon^j$. Thus after

$$j \geq \frac{2C(\Theta)}{\min\limits_{\bar\theta\in\Theta} \min\limits_{\theta\in\Theta_+(\bar\theta)} \min_i \Delta_i(\theta)^2} + 1.$$

episodes, all the optimistic models have only one optimal arm independently from the actual identity of the model $\bar\theta^j$. Although this condition may seem restrictive, in practice *umUCB* starts improving over *UCB* much earlier, as illustrated in the numerical simulation in Sec. 5.

4.4 Regret Analysis of *tUCB*

Given the previous results, we derive the bound on the cumulative regret over J episodes (Eq. 1).

Theorem 4. *If tUCB is run over J episodes of n steps in which the tasks $\bar\theta^j$ are drawn from a fixed distribution ρ over a set of models Θ, then its cumulative regret is*

$$\mathcal{R}_J \leq JK + \sum\nolimits_{j=1}^J \sum\nolimits_{i\in\mathcal{A}_1^j} \min\Big\{\frac{2\log\big(2mKn^2/\delta\big)}{\Delta_i(\bar\theta^j)^2}, \frac{\log\big(2mKn^2/\delta\big)}{2\min_{\theta\in\Theta^j_{i,+}(\bar\theta^j)} \widehat{\Gamma}_i^j(\theta;\bar\theta^j)^2}\Big\}\Delta_i(\bar\theta^j)$$
$$+ \sum\nolimits_{j=1}^J \sum\nolimits_{i\in\mathcal{A}_2^j} \frac{2\log\big(2mKn^2/\delta\big)}{\Delta_i(\bar\theta^j)},$$

w.p. $1-\delta$ w.r.t. the randomization over tasks and the realizations of the arms in each episode.

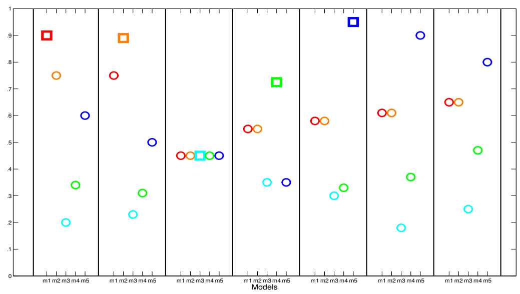

Figure 5: Set of models Θ.

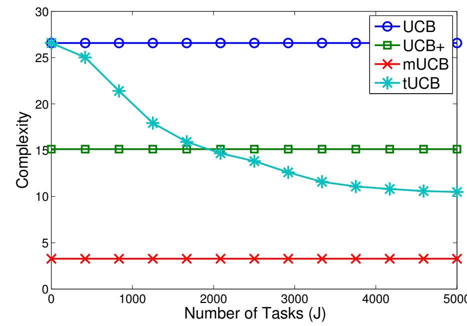

Figure 6: Complexity over tasks.

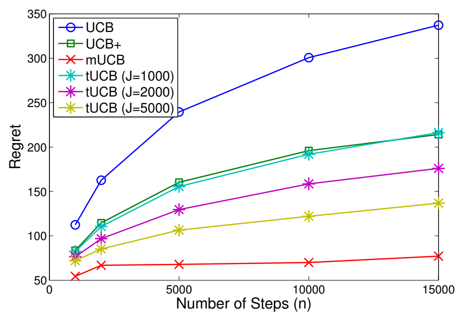

Figure 7: Regret of *UCB*, *UCB+*, *mUCB*, and *tUCB* (avg. over episodes) vs episode length.

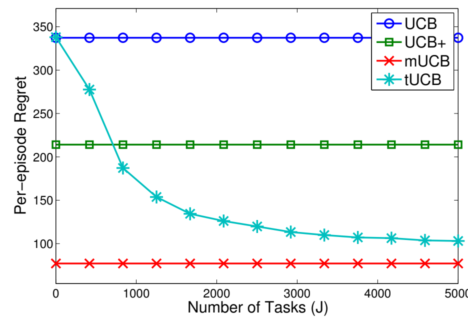

Figure 8: Per-episode regret of *tUCB*.

This result immediately follows from Thm. 3 and it shows a linear dependency on the number of episodes J. This dependency is the price to pay for not knowing the identity of the current task $\bar{\theta}^j$. If the task was revealed at the beginning of the task, a bandit algorithm could simply cluster all the samples coming from the same task and incur a much smaller cumulative regret with a logarithmic dependency on episodes and steps, i.e., $\log(nJ)$. Nonetheless, as discussed in the previous section, the cumulative regret of *tUCB* is never worse than for *UCB* and as the number of tasks increases it approaches the performance of *mUCB*, which fully exploits the prior knowledge of Θ.

5 Numerical Simulations

In this section we report preliminary results of *tUCB* on synthetic data. The objective is to illustrate and support the previous theoretical findings. We define a set Θ of $m = 5$ MAB problems with $K = 7$ arms each, whose means $\{\mu_i(\theta)\}_{i,\theta}$ are reported in Fig. 5 (see Sect. F in **?** for the actual values), where each model has a different color and squares correspond to optimal arms (e.g., arm 2 is optimal for model θ_2). This set of models is chosen to be challenging and illustrate some interesting cases useful to understand the functioning of the algorithm.[4] Models θ_1 and θ_2 only differ in their optimal arms and this makes it difficult to distinguish them. For arm 3 (which is optimal for model θ_3 and thus potentially selected by *mUCB*), all the models share exactly the same mean value. This implies that no model can be discarded by pulling it. Although this might suggest that *mUCB* gets stuck in pulling arm 3, we showed in Thm. 1 that this is not the case. Models θ_1 and θ_5 are challenging for *UCB* since they have small minimum gap. Only 5 out of the 7 arms are actually optimal for a model in Θ. Thus, we also report the performance of *UCB+* which, under the assumption that Θ is known, immediately discards all the arms which are not optimal ($i \notin \mathcal{A}^*$) and performs *UCB* on the remaining arms. The model distribution is uniform, i.e., $\rho(\theta) = 1/m$.

Before discussing the transfer results, we compare *UCB*, *UCB+*, and *mUCB*, to illustrate the advantage of the prior knowledge of Θ w.r.t. *UCB*. Fig. 7 reports the per-episode regret of the three

[4]Notice that although Θ satisfies Assumption 1, the smallest singular value $\sigma_{\min} = 0.0039$ and $\Gamma_\sigma = 0.0038$, thus making the estimation of the models difficult.

algorithms for episodes of different length n (the performance of *tUCB* is discussed later). The results are averaged over all the models in Θ and over 200 runs each. All the algorithms use the same confidence bound $\varepsilon_{i,t}$. The performance of *mUCB* is significantly better than both *UCB*, and *UCB+*, thus showing that *mUCB* makes an efficient use of the prior of knowledge of Θ. Furthermore, in Fig. 6 the horizontal lines correspond to the value of the regret bounds up to the n dependent terms and constants[5] for the different models in Θ averaged w.r.t. ρ for the three algorithms (the actual values for the different models are in the supplementary material). These values show that the improvement observed in practice is accurately predicated by the upper-bounds derived in Thm. 1.

We now move to analyze the performance of *tUCB*. In Fig. 8 we show how the per-episode regret changes through episodes for a transfer problem with $J = 5000$ tasks of length $n = 5000$. In *tUCB* we used ε^j as in Eq. 5 with $C(\Theta) = 2$. As discussed in Thm. 3, *UCB* and *mUCB* define the boundaries of the performance of *tUCB*. In fact, at the beginning *tUCB* selects arms according to a *UCB* strategy, since no prior information about the models Θ is available. On the other hand, as more tasks are observed, *tUCB* is able to transfer the knowledge acquired through episodes and build an increasingly accurate estimate of the models, thus approaching the behavior of *mUCB*. This is also confirmed by Fig. 6 where we show how the complexity of *tUCB* changes through episodes. In both cases (regret and complexity) we see that *tUCB* does not reach the same performance of *mUCB*. This is due to the fact that some models have relatively small gaps and thus the number of episodes to have an accurate enough estimate of the models to reach the performance of *mUCB* is much larger than 5000 (see also the Remarks of Thm. 3). Since the final objective is to achieve a small global regret (Eq. 1), in Fig. 7 we report the cumulative regret averaged over the total number of tasks (J) for different values of J and n. Again, this graph shows that *tUCB* outperforms *UCB* and that it tends to approach the performance of *mUCB* as J increases, for any value of n.

6 Conclusions and Open Questions

In this paper we introduce the transfer problem in the multi–armed bandit framework when a tasks are drawn from a finite set of bandit problems. We first introduced the bandit algorithm *mUCB* and we showed that it is able to leverage the prior knowledge on the set of bandit problems Θ and reduce the regret w.r.t. *UCB*. When the set of models is unknown we define a method-of-moments variant (*RTP*) which consistently estimates the means of the models in Θ from the samples collected through episodes. This knowledge is then transferred to *umUCB* which performs no worse than *UCB* and tends to approach the performance of *mUCB*. For these algorithms we derive regret bounds, and we show preliminary numerical simulations. To the best of our knowledge, this is the first work studying the problem of transfer in multi-armed bandit. It opens a series of interesting directions, including whether explicit model identification can improve our transfer regret.

Optimality of tUCB. At each episode, *tUCB* transfers the knowledge about Θ acquired from previous tasks to achieve a small per-episode regret using *umUCB*. Although this strategy guarantees that the per-episode regret of *tUCB* is never worse than UCB, it may not be the optimal strategy in terms of the cumulative regret through episodes. In fact, if J is large, it could be preferable to run a *model identification* algorithm instead of *umUCB* in earlier episodes so as to improve the quality of the estimates $\hat{\mu}_i(\theta)$. Although such an algorithm would incur a much larger regret in earlier tasks (up to linear), it could approach the performance of *mUCB* in later episodes much faster than done by *tUCB*. This trade-off between *identification* of the models and *transfer* of knowledge may suggest that different algorithms than *tUCB* are possible.

Unknown model-set size. In some problems the size of model set m is not known to the learner and needs to be estimated. This problem can be addressed by estimating the rank of matrix M_2 which equals to m (Kleibergen and Paap, 2006). We also note that one can relax the assumption that $\rho(\theta)$ needs to be positive (see Assumption 1) by using the estimated model size as opposed to m, since M_2 depends not on the means of models with $\rho(\theta) = 0$.

Acknowledgments. This research was supported by the National Science Foundation (NSF award #SBE-0836012). We would like to thank Sham Kakade and Animashree Anandkumar for valuable discussions. A. Lazaric would like to acknowledge the support of the Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER through the' Contrat de Projets Etat Region (CPER) 2007-2013', and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 231495 (project CompLACS).

[5]For instance, for *UCB* we compute $\sum_i 1/\Delta_i$.

References

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. E. (2012). Contextual bandit learning with predictable rewards. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS'12)*.

Anandkumar, A., Foster, D. P., Hsu, D., Kakade, S., and Liu, Y.-K. (2012a). A spectral algorithm for latent dirichlet allocation. In *Proceedings of Advances in Neural Information Processing Systems 25 (NIPS'12)*, pages 926–934.

Anandkumar, A., Ge, R., Hsu, D., and Kakade, S. M. (2013). A tensor spectral approach to learning mixed membership community models. *Journal of Machine Learning Research*, 1:65.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2012b). Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559.

Anandkumar, A., Hsu, D., and Kakade, S. M. (2012c). A method of moments for mixture models and hidden markov models. In *Proceeding of the 25th Annual Conference on Learning Theory (COLT'12)*, volume 23, pages 33.1–33.34.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47:235–256.

Azar, M. G., Lazaric, A., and Brunskill, E. (2013). Sequential transfer in multi-armed bandit with finite set of models. *CoRR*, abs/1307.6887.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. (2010). Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11:2901–2934.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Dekel, O., Long, P. M., and Singer, Y. (2006). Online multitask learning. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT'06)*, pages 453–467.

Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd international conference on Algorithmic learning theory*, ALT'11, pages 174–188, Berlin, Heidelberg. Springer-Verlag.

Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1):97–126.

Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. In *Proceedings of Advances in Neural Information Processing Systems 20 (NIPS'07)*.

Lazaric, A. (2011). Transfer in reinforcement learning: a framework and a survey. In Wiering, M. and van Otterlo, M., editors, *Reinforcement Learning: State of the Art*. Springer.

Lugosi, G., Papaspiliopoulos, O., and Stoltz, G. (2009). Online multi-task learning with hard constraints. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09)*.

Mann, T. A. and Choe, Y. (2012). Directed exploration in reinforcement learning with transferred knowledge. In *Proceedings of the Tenth European Workshop on Reinforcement Learning (EWRL'12)*.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

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

Saha, A., Rai, P., Daumé III, H., and Venkatasubramanian, S. (2011). Online learning of multiple tasks and their relationships. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS'11)*, Ft. Lauderdale, Florida.

Stewart, G. W. and Sun, J.-g. (1990). *Matrix perturbation theory*. Academic press.

Taylor, M. E. (2009). *Transfer in Reinforcement Learning Domains*. Springer-Verlag.

Wedin, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Appendices

A Table of Notation

Symbol	*Explanation*
$\mathcal{A}$	Set of arms
Θ	Set of models
K	Number of arms
m	Number of models
J	Number of episodes
n	Number of steps per episode
t	Time step
$\bar{\theta}$	Current model
Θ_t	Active set of models at time t
ν_i	Distribution of arm i
$\mu_i(\theta)$	Mean of arm i for model θ
$\mu(\theta)$	Vector of means of model θ
$\hat{\mu}_{i,t}$	Estimate of $\mu_i(\bar{\theta})$ at time t
$\hat{\mu}_i^j(\theta)$	Estimate of $\mu_i(\theta)$ by RTP for model θ and arm i at episode j
$\widehat{\mu}^j(\theta)$	Estimate of $\mu(\theta)$ by RTP for model θ at episode j
Θ^j	Estimated model of RTP after j episode
ε^j	Uncertainty of the estimated model by RTP at episode j
$\varepsilon_{i,t}$	Model uncertainty at time t
δ	Probability of failure
$i_*(\theta)$	Best arm of model θ
$\mu_*(\theta)$	Optimal value of model θ
$\Delta_i(\theta)$	Arm gap of an arm i for a model θ
$\Gamma_i(\theta,\theta')$	Model gap for an arm i between two models θ and θ'
M_2	2nd-order moment
M_3	3rd-order moment
$\widehat{M}_2$	Empirical 2nd-order moment
$\widehat{M}_3$	Empirical 3rd-order moment
$\|\cdot\|$	Euclidean norm
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|_{\max}$	Matrix max-norm
$\mathcal{R}_J$	Pseudo-regret
$T_{i,n}^j$	The number of pulls to arm i after n steps of episode j
$\mathcal{A}_*(\Theta')$	Set of arms which are optimal for at least a model in a set Θ'
$\Theta(\mathcal{A}')$	Set of models for which the arms in $\mathcal{A}'$ are optimal
Θ_+	Set of optimistic models for a given model $\bar{\theta}$
$\mathcal{A}_+$	Set of optimal arms corresponds to Θ_+
W	Whitening matrix of M_2
$\widehat{W}$	Empirical whitening matrix
T	M_2 under the linear transformation W
$\widehat{T}$	$\widehat{M}_2$ under the linear transformation $\widehat{W}$
D	Diagonal matrix consisting of the m largest eigenvalues of M_2
$\widehat{D}$	Diagonal matrix consisting of the m largest eigenvalues of $\widehat{M}_2$
U	$K \times m$ matrix with the corresponding eigenvectors of D as its columns
$\widehat{U}$	$K \times m$ matrix with the corresponding eigenvectors of $\widehat{D}$ as its columns
$\lambda(\theta)$	Eigenvalue of T associated with θ
$v(\theta)$	Eigenvector of T associated with θ
$\widehat{\lambda}(\theta)$	Eigenvalue of $\widehat{T}$ associated with θ
$\widehat{v}(\theta)$	Eigenvector of $\widehat{T}$ associated with θ
Σ_{M_2}	Set of m largest eigenvalues of the matrix M_2
$\sigma_{\min}$	Minimum eigenvalue of M_2 among the m-largest
$\sigma_{\max}$	Maximum eigenvalue of M_2
$\lambda_{\max}$	Maximum eigenvalue of T
Γ_σ	Minimum gap between the eigenvalues of M_2
$C(\Theta)$	$O\left(\lambda_{\max}\sqrt{\frac{\sigma_{\max}}{\sigma_{\min}^3}}\left(\frac{\sigma_{\max}}{\Gamma_\sigma}+\frac{1}{\sigma_{\min}}+\frac{1}{\sigma_{\max}}\right)\right)$
$\pi(\theta)$	Permutation on θ
$\mathcal{A}_*^j(\theta)$	Set of non-dominated arms for model θ at episode j
$\widetilde{\Theta}_+^j$	Set of models that cannot be discarded at episode j
$\Theta_{i,+}^j$	Set of models for which i is among the optimistic non-dominated arms at episode j

B Proofs of Section 3

Lemma 1. *mUCB never pulls arms which are not optimal for at least one model, that is* $\forall i \notin \mathcal{A}_*(\Theta)$, $T_{i,n} = 0$ *with probability 1. Notice also that* $|\mathcal{A}_*(\Theta)| \leq |\Theta|$.

Lemma 2. *The actual model* $\bar{\theta}$ *is never discarded with high-probability. Formally, the event* $\mathcal{E} = \{\forall t = 1, \ldots, n, \bar{\theta} \in \Theta_t\}$ *holds with probability* $\mathbb{P}[\mathcal{E}] \geq 1 - \delta$ *if*

$$\varepsilon_{i,t} = \sqrt{\frac{1}{2T_{i,t-1}} \log\left(\frac{mn^2}{\delta}\right)},$$

where $T_{i,t-1}$ *is the number of pulls to arm* i *at the beginning of step* t *and* $m = |\Theta|$.

In the previous lemma we implicitly assumed that $|\Theta| = m \leq K$. In general, the best choice in the definition of $\varepsilon_{i,t}$ has a logarithmic factor with $\min\{|\Theta|, K\}$.

Lemma 3. *On event* $\mathcal{E}$, *all the arms* $i \notin \mathcal{A}_*(\Theta_+)$, *i.e., arms which are not optimal for any of the optimistic models, are never pulled, i.e.,* $T_{i,n} = 0$ *with probability* $1 - \delta$.

The previous lemma suggests that *mUCB* tends to discard all the models in Θ_+ from the most optimistic down to the actual model $\bar{\theta}$ which, on event $\mathcal{E}$, is never discarded. As a result, even if other models are still in Θ_t, the optimal arm of $\bar{\theta}$ is pulled until the end. Finally, we show that the model gaps of interest (see Thm. 1) are always bigger than the arm gaps.

Lemma 4. *For any model* $\theta \in \Theta_+$, $\Gamma_{i_*(\theta)}(\theta, \bar{\theta}) \geq \Delta_{i_*(\theta)}(\bar{\theta})$.

Proof of Lem. 1. From the definition of the algorithm we notice that I_t can only correspond to the optimal arm i_* of one model in the set Θ_t. Since Θ_t can at most contain all the models in Θ, all the arms which are not optimal are never pulled. □

Proof of Lem. 2. We compute the probability of the complementary event $\mathcal{E}^C$, that is that event on which there exist at least one step $t = 1, \ldots, n$ where the true model $\bar{\theta}$ is not in Θ_t. By definition of Θ_t, we have that

$$\mathcal{E} = \{\forall t, \bar{\theta} \in \Theta_t\} = \{\forall t, \forall i \in \mathcal{A}, |\mu_i - \hat{\mu}_{i,t}| \leq \varepsilon_{i,t}\},$$

then

$$\mathbb{P}[\mathcal{E}^C] = \mathbb{P}[\exists t, i, |\mu_i - \hat{\mu}_{i,t}| \geq \varepsilon_{i,t}] \leq \sum_{t=1}^{n} \sum_{i \in \mathcal{A}} \mathbb{P}[|\mu_i - \hat{\mu}_{i,t}| \geq \varepsilon_{i,t}] = \sum_{t=1}^{n} \sum_{i \in \mathcal{A}^*(\Theta)} \mathbb{P}[|\mu_i - \hat{\mu}_{i,t}| \geq \varepsilon_{i,t}]$$

where the upper-bounding is a simple union bound and the last passage comes from the fact that the probability for the arms which are never pulled is always 0 according to Lem. 1. At time t, $\hat{\mu}_{i,t}$ is the empirical average of the $T_{i,t-1}$ samples observed from arm i up to the beginning of round t. We define the confidence $\varepsilon_{i,t}$ as

$$\varepsilon_{i,t} = \sqrt{\frac{1}{2T_{i,t-1}} \log\left(\frac{|\Theta| n^\alpha}{\delta}\right)},$$

where $\delta \in (0, 1)$ and α is a constant chosen later. Since $T_{i,t-1}$ is a random variable, we need to take an additional union bound over $T_{i,t-1} = 1, \ldots, t-1$ thus obtaining

$$\begin{aligned}
\mathbb{P}[\mathcal{E}^C] &\leq \sum_{t=1}^{n} \sum_{i \in \mathcal{A}^*(\Theta)} \sum_{T_{i,t-1}=1}^{t-1} \mathbb{P}[|\mu_i - \hat{\mu}_{i,t}| \geq \varepsilon_{i,t}] \\
&\leq \sum_{t=1}^{n} \sum_{i \in \mathcal{A}^*(\Theta)} \sum_{T_{i,t-1}=1}^{t-1} 2 \exp\big(-2T_{i,t-1}\varepsilon_{i,t}^2\big) \leq n(n-1) \frac{|\mathcal{A}^*(\Theta)|\delta}{|\Theta| n^\alpha}.
\end{aligned}$$

Since $|\mathcal{A}^*(\Theta)| < |\Theta|$ (see Lem. 1) and by taking $\alpha = 2$ we finally have $\mathbb{P}[\mathcal{E}^C] \leq \delta$. □

Proof of Lem. 3. On event $\mathcal{E}$, Θ_t always contains the true model $\bar{\theta}$, thus only models with larger optimal value could be selected as the optimistic model $\theta_t = \arg\max_{\theta \in \Theta_t} \mu_*(\theta)$, thus restricting the focus of the algorithm only to the models in Θ_+ and their respective optimal arms. □

Proof of Lem. 4. By definition of Θ_+ we have $\mu_{i_*(\theta)}(\theta) = \mu_*(\theta) > \mu_*(\bar{\theta})$ and by definition of optimal arm we have $\mu_*(\bar{\theta}) > \mu_{i_*(\theta)}(\bar{\theta})$, hence $\mu_*(\theta) > \mu_{i_*(\theta)}(\bar{\theta})$. Recalling the definition of model gap, we have $\Gamma_{i_*(\theta)}(\theta) = |\mu_{i_*(\theta)}(\theta) - \mu_{i_*(\theta)}(\bar{\theta})| = \mu_*(\theta) - \mu_{i_*(\theta)}(\bar{\theta})$, where we used the definition of $\mu_*(\theta)$ and the previous inequality. Using the definition of arm gap Δ_i, we obtain

$$\Gamma_{i_*(\theta)}(\theta, \bar{\theta}) = \mu_*(\theta) - \mu_{i_*(\theta)}(\bar{\theta}) \geq \mu_*(\bar{\theta}) - \mu_{i_*(\theta)}(\bar{\theta}) = \Delta_{i_*(\theta)}(\bar{\theta}),$$

which proves the statement. □

Proof of Thm. 1. We decompose the expected regret as

$$\mathbb{E}[\mathcal{R}_n] = \sum_{i \in \mathcal{A}} \Delta_i \mathbb{E}[T_{i,n}] = \sum_{i \in \mathcal{A}_*(\Theta)} \Delta_i \mathbb{E}[T_{i,n}] \leq n\mathbb{P}\{\mathcal{E}^C\} + \sum_{i \in \mathcal{A}_+} \Delta_i \mathbb{E}[T_{i,n}|\mathcal{E}],$$

where the refinement on the sum over arms follows from Lem. 1 and 3 and the high probability event $\mathcal{E}$. In the following we drop the dependency on $\bar{\theta}$ and we write $\mu_i(\bar{\theta}) = \mu_i$.

We now bound the regret when the correct model is always included in Θ_t. On event $\mathcal{E}$, only the restricted set of *optimistic* models $\Theta_+ = \{\theta \in \Theta : \mu_*(\theta) \geq \mu_*\}$ is actually used by the algorithm. Thus we need to compute the number of pulls to the suboptimal arms before all the models in Θ_+ are discarded from Θ_t. We first compute the number of pulls to an arm i needed to discard a model θ on event $\mathcal{E}$. We notice that

$$\theta \in \Theta_t \Leftrightarrow \{\forall i \in \mathcal{A}, |\mu_i(\theta) - \hat{\mu}_{i,t}| \leq \varepsilon_{i,t}\},$$

which means that a model θ is included only when all its means are *compatible* with the current estimates. Since we consider event $\mathcal{E}$, $|\mu_i - \hat{\mu}_{i,t}| \leq \varepsilon_{i,t}$, thus $\theta \in \Theta_t$ only if for all $i \in \mathcal{A}$

$$2\varepsilon_{i,t} \geq \Gamma_i(\theta, \bar{\theta}),$$

which corresponds to

$$T_{i,t-1} \leq \frac{2}{\Gamma_i(\theta, \bar{\theta})^2} \log\left(\frac{|\Theta| n^2}{\delta}\right), \tag{6}$$

which implies that if there exists at least one arm i for which at time t the number of pulls $T_{i,t}$ exceeds the previous quantity, then $\forall s > t$ we have $\theta \notin \Theta_t$ (with probability $\mathbb{P}(\mathcal{E})$). To obtain the final bound on the regret, we recall that the algorithm first selects an optimistic model θ_t and then it pulls the corresponding optimal arm until the optimistic model is not discarded. Thus we need to compute the number of times the optimal arm of the optimistic model is pulled before the model is discarded. More formally, since we know that on event $\mathcal{E}$ we have that $T_{i,n} = 0$ for all $i \notin \mathcal{A}_+$, the constraints of type (6) could only be applied to the arms $i \in \mathcal{A}_+$. Let t be the last time arm i is pulled, which coincides, by definition of the algorithm, with the last time any of the models in $\Theta_{+,i} = \{\theta \in \Theta_+ : i_*(\theta) = i\}$ (i.e., the optimistic models recommending i as the optimal arm) is included in Θ_t. Then we have that $T_{i,t-1} = T_{i,n} - 1$ and the fact that i is pulled corresponds to the fact the a model $\theta_i \in \Theta_{+,i}$ is such that

$$\theta_i \in \Theta_t \land \forall \theta' \in \Theta_t, \mu_*(\theta_i) > \mu_*(\theta'),$$

which implies that (see Eq. 6)

$$T_{i,n} \leq \frac{2}{\min_{\theta \in \Theta_{+,i}} \Gamma_i(\theta, \bar{\theta})^2} \log\left(\frac{|\Theta| n^2}{\delta}\right) + 1. \tag{7}$$

where the minimum over $\Theta_{+,i}$ guarantees that all the optimistic models with optimal arm i are actually discarded.

Grouping all the conditions, we obtain the expected regret

$$\mathbb{E}[\mathcal{R}_n] \leq K + \sum_{i \in \mathcal{A}_+} \frac{2\Delta_i(\bar{\theta})}{\min_{\theta \in \Theta_{+,i}} \Gamma_i(\theta, \bar{\theta})^2} \log\left(|\Theta| n^3\right)$$

with $\delta = 1/n$. Finally we can apply Lem. 4 which guarantees that for any $\theta \in \Theta_{+,i}$ the gaps $\Gamma_i(\theta, \bar{\theta}) \geq \Delta_i(\bar{\theta})$ and obtain the final statement. □

Remark (proof). The proof of the theorem considers a worst case. In fact, while pulling the optimal arm of the optimistic model $i_*(\theta_t)$ we do not consider that the algorithm might actually discard other models, thus reducing Θ_t before the optimistic model is actually discarded. More formally, we assume that for any $\theta \in \Theta_t$ not in $\Theta_{+,i}$ the number of steps needed to be discarded by pulling $i_*(\theta_t)$ is larger than the number of pulls needed to discard θ_t itself, which corresponds to

$$\min_{\theta \in \Theta_{+,i}} \Gamma_i^2(\theta, \bar{\theta}) \geq \max_{\substack{\theta \in \Theta^+ \\ \theta \notin \Theta_{+,i}}} \Gamma_i^2(\theta, \bar{\theta}).$$

Whenever this condition is not satisfied, the analysis is suboptimal since it does not fully exploit the structure of the problem and *mUCB* is expected to perform better than predicted by the bound.

Remark (comparison to *UCB* with hypothesis testing). An alternative strategy is to pair *UCB* with hypothesis testing of fixed confidence δ. Let $\Gamma_{\min}(\bar{\theta}) = \min_i \min_\theta \Gamma_i(\theta, \bar{\theta})$, if at time t there exists an arm i such that $T_{i,t} > 2\log(2/\delta)\Gamma_{\min}^2$, then all the models $\theta \neq \bar{\theta}$ can be discarded with probability $1-\delta$. Since from the point of view of the hypothesis testing the exploration strategy is unknown, we can only assume that after τ steps we have $T_{i,\tau} \geq \tau/K$ for at least one arm i. Thus after $\tau > 2K\log(2/\delta)/\Gamma_{\min}^2$ steps, the hypothesis testing returns a model $\hat{\theta}$ which coincides with $\bar{\theta}$ with probability $1-\delta$. If $\tau \leq n$, from time τ on, the algorithm always pulls $I_t = i_*(\hat{\theta})$ and incurs a zero regret with high probability. If we assume $\tau \leq n$, the expected regret is

$$\mathbb{E}[\mathcal{R}_n(\text{UCB+Hyp})] \leq O\bigg(\sum_{i \in \mathcal{A}} \frac{\log n\tau}{\Delta_i}\bigg) \leq O\bigg(K\frac{\log n\tau}{\Delta}\bigg).$$

We notice that this algorithm only has a mild improvement w.r.t. standard *UCB*. In fact, in *UCB* the big-O notation hides the constants corresponding to the exponent of n in the logarithmic term. This suggests that whenever τ is much smaller than n, then there might be a significant improvement. On the other hand, since τ has an inverse dependency w.r.t. $\Gamma_{\min}$, it is very easy to build model sets Θ where $\Gamma_{\min} = 0$ and obtain an algorithm with exactly the same performance as *UCB*.

C Sample Complexity Analysis of *RTP*

In this section we provide the full sample complexity analysis of the *RTP* algorithm. In our analysis we rely on some results of Anandkumar et al. (2012b). Anandkumar et al. (2012b) have provided perturbation bounds on the error of the orthonormal eigenvectors $\widehat{v}(\theta)$ and the corresponding eigenvalues $\widehat{\lambda}(\theta)$ in terms of the perturbation error of the transformed tensor $\epsilon = \|T - \widehat{T}\|$ (see Anandkumar et al., 2012b, Thm 5.1). However, this result does not provide us with the sample complexity bound on the estimation error of model means. Here we complete their analysis by proving a sample complexity bound on the ℓ_2-norm of the estimation error of the means $\|\mu(\theta) - \widehat{\mu}(\theta)\|$.

We follow the following steps in our proof: **(i)** we bound the error ϵ in terms of the estimation errors $\epsilon_2 := \|\widehat{M}_2 - M_2\|$ and $\epsilon_3 := \|\widehat{M}_3 - M_3\|$ (Lem. 6). **(ii)** we prove high probability bounds on the error ϵ_2 and ϵ_3 using some standard concentration inequality results (Lem. 7). The bounds on the errors of the estimates $\widehat{v}(\theta)$ and $\widehat{\lambda}(\theta)$ immediately follow from combining the results of Lem. 6, Lem. 7 and Thm. 5. **(iii)** Based on these bounds we then prove our main result by bounding the estimation error associated with the inverse transformation $\widehat{\mu}(\theta) = \widehat{\lambda}(\theta)\widehat{B}\widehat{v}(\theta)$ in high probability.

We begin by recalling the perturbation bound of Anandkumar et al. (2012b):

Theorem 5 (Anandkumar et al., 2012b). *Pick $\eta \in (0,1)$. Define $W := UD^{-1/2}$, where $D \in \mathbb{R}^{m \times m}$ is the diagonal matrix of the m largest eigenvalues of M_2 and $U \in \mathbb{R}^{K \times m}$ is the matrix with the eigenvectors associated with the m largest eigenvalues of M_2 as its columns. Then W is a linear mapping which satisfies $W^\mathsf{T} M_2 W = \mathbf{I}$. Let $\widehat{T} = T + E \in \mathbb{R}^{m \times m \times m}$, where the 3rd order moment tensor $T = M_3(W, W, W)$ is symmetric and orthogonally decomposable in the form of $\sum_{\theta \in \Theta} \lambda(\theta) v(\theta)^{\otimes 3}$, where each $\lambda(\theta) > 0$ and $\{v(\theta)\}_\theta$ is an orthonormal basis. Define $\epsilon := \|E\|$ and $\lambda_{\max} = \max_\theta \lambda(\theta)$. Then there exist some constants $C_1, C_2 > 0$, some polynomial function $f(\cdot)$, and a permutation π on Θ such that the following holds w.p. $1-\eta$*

$$\|v(\theta) - \widehat{v}(\pi(\theta))\| \leq 8\epsilon/\lambda(\theta),$$

$$|\lambda(\theta) - \widehat{\lambda}(\pi(\theta))| \leq 5\epsilon,$$

for $\epsilon \le C_1 \frac{\lambda_{\min}}{m}$, $L > \log(1/\eta) f(k)$ and $N \ge C_2(\log(k) + \log\log(\lambda_{\max}/\epsilon))$, where N and L are the internal parameters of RTP algorithm.

For ease of exposition we consider the *RTP* algorithm in asymptotic case, i.e., $N, L \to \infty$ and $\eta \approx 1$. We now prove bounds on the perturbation error ϵ in terms of the estimation error ϵ_2 and ϵ_3. This requires bounding the error between $W = UD^{-1/2}$ and $\widehat{W} = \widehat{U}\widehat{D}^{-1/2}$ using the following perturbation bounds on $\|U - \widehat{U}\|$, $\|\widehat{D}^{-1/2} - D^{-1/2}\|$ and $\|\widehat{D}^{1/2} - D^{1/2}\|$.

Lemma 5. *Assume that $\epsilon_2 \le 1/2 \min(\Gamma_\sigma, \sigma_{\min})$, then we have*

$$\|\widehat{D}^{-1/2} - D^{-1/2}\| \le \frac{2\epsilon_2}{(\sigma_{\min})^{3/2}}, \quad \textit{and} \quad \|\widehat{D}^{1/2} - D^{1/2}\| \le \frac{\epsilon_2}{\sigma_{\max}}, \quad \textit{and} \quad \|\widehat{U} - U\| \le \frac{2\sqrt{m}\epsilon_2}{\Gamma_\sigma}.$$

Proof. Here we just prove bounds on $\|\widehat{D}^{-1/2} - D^{-1/2}\|$ and $\|\widehat{U} - U\|$. The bound on $\|\widehat{D}^{-1/2} - D^{-1/2}\|$ can be proven using a similar argument to that used for bounding $\|\widehat{D}^{1/2} - D^{1/2}\|$. Let $\widehat{\Sigma}_m = \{\widehat{\sigma}_1, \widehat{\sigma}_2, \dots, \widehat{\sigma}_m\}$ be the set of m largest eigenvalues of the matrix $\widehat{M}_2$. We have

$$\begin{aligned}\|\widehat{D}^{-1/2} - D^{-1/2}\| &\overset{(1)}{=} \max_{1\le i\le m} \left| \sqrt{\frac{1}{\sigma_i}} - \sqrt{\frac{1}{\widehat{\sigma}_i}} \right| = \max_{1\le i\le m} \left(\frac{\left| \frac{1}{\sigma_i} - \frac{1}{\widehat{\sigma}_i} \right|}{\sqrt{\frac{1}{\sigma_i}} + \sqrt{\frac{1}{\widehat{\sigma}_i}}} \right) \\ &\le \max_{1\le i\le m} \left(\sqrt{\sigma_i} \left| \frac{1}{\sigma_i} - \frac{1}{\widehat{\sigma}_i} \right| \right) \le \max_{1\le i\le m} \left| \frac{\sigma_i - \widehat{\sigma}_i}{\sqrt{\sigma_i}\widehat{\sigma}_i} \right| \overset{(2)}{\le} \frac{\epsilon_2}{\sqrt{\sigma_{\min}}(\sigma_{\min} - \epsilon_2)} \overset{(3)}{\le} \frac{2\epsilon_2}{(\sigma_{\min})^{3/2}},\end{aligned}$$

where in (1) we use the fact that the spectral norm of matrix is its largest singular value, which in case of a diagonal matrix coincides with its biggest element, in (2) we rely on the result of Weyl (see Stewart and Sun, 1990, Thm. 4.11, p. 204) for bounding the difference between σ_i and $\widehat{\sigma}_i$, and in (3) we make use of the assumption that $\epsilon_2 \le 1/2\sigma_{\min}$.

In the case of $\|U - \widehat{U}\|$ we rely on the perturbation bound of Wedin (1972). This result guarantees that for any positive definite matrix A the difference between the eigenvectors of A and the perturbed $\widehat{A}$ (also positive definite) is small whenever there is a minimum gap between the eigenvalues of $\widehat{A}$ and A. More precisely, for any positive definite matrix A and $\widehat{A}$ such that $||A - \widehat{A}|| \le \epsilon_A$, let the minimum eigengap be $\Gamma_{A\leftrightarrow\widehat{A}} := \min_{j\ne i} |\sigma_i - \widehat{\sigma}_j|$, then we have

$$\|u_i - \widehat{u}_i\| \le \frac{\epsilon_A}{\Gamma_{A\leftrightarrow\widehat{A}}}, \tag{8}$$

where (u_i, σ_i) is an eigenvalue/vector pair for the matrix A. Based on this result we now bound the error $\|U - \widehat{U}\|$

$$\|U - \widehat{U}\| \le \|U - \widehat{U}\|_F \le \sqrt{\sum_i \|u_i - \widehat{u}_i\|^2} \overset{(1)}{\le} \frac{\sqrt{m}\epsilon_2}{\Gamma_{M_2\leftrightarrow\widehat{M}_2}} \overset{(2)}{\le} \frac{\sqrt{m}\epsilon_2}{\Gamma_\sigma - \epsilon_2} \overset{(3)}{\le} \frac{2\sqrt{m}\epsilon_2}{\Gamma_\sigma},$$

where in (1) we rely on Eq. 8 and in (2) we rely on the definition of the gap as well as Weyl's inequality. Finally, in (3) We rely on the fact that $\epsilon_2 \le 1/2\Gamma_\sigma$ for bounding denominator from below.

Our result also holds for those cases where the multiplicity of some of the eigenvalues are greater than 1. Note that for any eigenvalue λ with multiplicity l the linear combination of the corresponding eigenvectors $\{v_1, v_2, \dots, v_l\}$ is also an eigenvector of the matrix. Therefore, in this case it suffices to bound the difference between the eigenspaces of two matrix. The result of Wedin (1972) again applies to this case and bounds the difference between the eigenspaces in terms of the perturbation ϵ_2 and Γ_σ. $\square$

We now bound ϵ in terms of ϵ_2 and ϵ_3.

Lemma 6. *Let $\mu_{\max} := \max_\theta \|\mu(\theta)\|$, if $\epsilon_2 \le 1/2\min(\Gamma_\sigma, \sigma_{\min})$, then the estimation error ϵ is bounded as*

$$\epsilon \le \left(\frac{m}{\sigma_{\min}}\right)^{3/2} \left(10\epsilon_2 \left(\frac{1}{\Gamma_\sigma} + \frac{1}{\sigma_{\min}}\right)(\epsilon_3 + \mu_{\max}^3) + \epsilon_3\right).$$

Proof. Based on the definitions of T and $\widehat{T}$ we have

$$\begin{aligned}\epsilon = \|T-\widehat{T}\| &= \|M_3(W,W,W) - \widehat{M}_3(\widehat{W},\widehat{W},\widehat{W})\| \\ &\le \|M_3(W,W,W) - \widehat{M}_3(W,W,W)\| + \|\widehat{M}_3(W,W,W) - \widehat{M}_3(W,W,\widehat{W})\| \\ &\quad + \|\widehat{M}_3(W,W,\widehat{W}) - \widehat{M}_3(W,\widehat{W},\widehat{W})\| + \|\widehat{M}_3(W,\widehat{W},\widehat{W}) - \widehat{M}_3(\widehat{W},\widehat{W},\widehat{W})\| \\ &= \|E_{M_3}(W,W,W)\| + \|\widehat{M}_3(W,W,W-\widehat{W})\| + \|\widehat{M}_3(W,W-\widehat{W},\widehat{W})\| \\ &\quad + \|\widehat{M}_3(W-\widehat{W},\widehat{W},\widehat{W})\|,\end{aligned} \tag{9}$$

where $E_{M_3} = M_3 - \widehat{M}_3$. We now bound the terms in the r.h.s. of Eq. 9 in terms of ϵ_3 and ϵ_2. We begin by bounding $\|E_{M_3}(W,W,W)\|$:

$$\begin{aligned}\|E_{M_3}(W,W,W)\| &\le \|E_{M_3}\|\|W\|^3 \le \|E_{M_3}\|\|U\|^3\|D^{-1}\|^{3/2} \le \|E_{M_3}\|\|U\|_F^3\|D^{-1}\|^{3/2} \\ &\overset{(1)}{=} \left(\frac{m}{\sigma_{\min}}\right)^{3/2}\|E_{M_3}\| \le \left(\frac{m}{\sigma_{\min}}\right)^{3/2}\epsilon_3,\end{aligned} \tag{10}$$

where in (1) we use the fact that U is an orthonormal matrix and D is diagonal. In the case of $\|\widehat{M}_3(W,W,W-\widehat{W})\|$ we have

$$\begin{aligned}&\|\widehat{M}_3(W,W,W-\widehat{W})\| \le \|W\|^2\|W-\widehat{W}\|\|\widehat{M}_3\| \le \|W\|^2\|W-\widehat{W}\|(\|\widehat{M}_3 - M_3\| + \|M_3\|) \\ &\overset{(1)}{\le} \|W\|^2\|W-\widehat{W}\|(\epsilon_3 + \mu_{\max}^3) \le \|W\|^2\|UD^{-1/2} - \widehat{U}\widehat{D}^{-1/2}\|(\epsilon_3+\mu_{\max}^3) \\ &\le \|W\|^2(\|(U-\widehat{U})D^{-1/2}\| + \|\widehat{U}(\widehat{D}^{-1/2} - D^{-1/2})\|)\left(\epsilon_3 + \mu_{\max}^3\right) \\ &\le \|W\|^2\left(\frac{\|U-\widehat{U}\|}{\sqrt{\sigma_{\min}}} + \sqrt{m}\|\widehat{D}^{-1/2} - D^{-1/2}\|\right)\left(\epsilon_3+\mu_{\max}^3\right).\end{aligned}$$

where in (1) we use the definition of M_3 as a linear combination of the tensor product of the means $\mu(\theta)$. This result combined with the result of Lem. 5 and the fact that $\|W\| \le \sqrt{m/\sigma_{\min}}$ (see Eq. 10) implies that

$$\begin{aligned}\|\widehat{M}_3(W,W,W-\widehat{W})\| &\le \frac{m}{\sigma_{\min}}\left(\frac{2\sqrt{m}\epsilon_2}{\Gamma_\sigma\sqrt{\sigma_{\min}}} + \frac{2\sqrt{m}\epsilon_2}{(\sigma_{\min})^{3/2}}\right)\left(\epsilon_3+\mu_{\max}^3\right) \\ &\le 2\epsilon_2\left(\frac{m}{\sigma_{\min}}\right)^{3/2}\left(\frac{1}{\Gamma_\sigma} + \frac{1}{\sigma_{\min}}\right)\left(\epsilon_3+\mu_{\max}^3\right).\end{aligned} \tag{11}$$

Likewise one can prove the following perturbation bounds for $\widehat{M}_3(W,W-\widehat{W},\widehat{W})$ and $\widehat{M}_3(W,W-\widehat{W},\widehat{W})$:

$$\begin{aligned}\|\widehat{M}_3(W,W-\widehat{W},\widehat{W})\| &\le 2\sqrt{2}\epsilon_2\left(\frac{m}{\sigma_{\min}}\right)^{3/2}\left(\frac{1}{\Gamma_\sigma} + \frac{1}{\sigma_{\min}}\right)\left(\epsilon_3+\mu_{\max}^3\right) \\ \|\widehat{M}_3(W-\widehat{W},\widehat{W},\widehat{W})\| &\le 4\epsilon_2\left(\frac{m}{\sigma_{\min}}\right)^{3/2}\left(\frac{1}{\Gamma_\sigma} + \frac{1}{\sigma_{\min}}\right)\left(\epsilon_3+\mu_{\max}^3\right).\end{aligned} \tag{12}$$

The result then follows by plugging the bounds of Eq. 10, Eq. 11 and Eq. 12 into Eq. 9. □

We now prove high-probability bounds on ϵ_3 and ϵ_2 when M_2 and M_3 are estimated by sampling.

Lemma 7. *For any $\delta \in (0,1)$, if $\widehat{M}_2$ and $\widehat{M}_3$ are computed with samples from j episodes, then we that with probability $1-\delta$:*

$$\epsilon_3 \le K^{1.5}\sqrt{\frac{6\log(2K/\delta)}{j}} \qquad \textit{and} \qquad \epsilon_2 \le 2K\sqrt{\frac{\log(2K/\delta)}{j}}.$$

Proof. Using some norm inequalities for the tensors we obtain

$$\epsilon_3 = \|M_3 - \widehat{M}_3\| \le K^{1.5}\|M_3 - \widehat{M}_3\|_{\max} = K^{1.5}\max_{i,j,x}|[M_3]_{i,j,x} - [\widehat{M}_3]_{i,j,x}|.$$

A similar argument leads to the bound of $K \max_{i,j} |[M_2]_{i,j} - [\widehat{M}_2]_{i,j}|$ on ϵ_2. One can easily show that, for every $1 \leq i, j, x \leq K$, the term $[M_3]_{i,j,x} - [\widehat{M}_3]_{i,j,x}$ and $[M_3]_{i,j,x} - [\widehat{M}_3]_{i,j,x}$ can be expressed as a sum of martingale differences with the maximum value $1/j$. The result then follows by applying the Azuma's inequality (e.g., see Cesa-Bianchi and Lugosi, 2006, appendix, pg. 361) and taking the union bound. □

We now draw our attention to the proof of our main result.

Proof of Thm. 2. We begin by deriving the condition of Eq. 5. The assumption on ϵ_2 in Lem. 6 and the result of Lem. 7 hold at the same time, w.p. $1-\delta$, if the following inequality holds

$$2K\sqrt{\frac{\log(2K/\delta)}{j}} \leq 1/2 \min(\Gamma_\sigma, \sigma_{\min}).$$

By solving the bound w.r.t. j we obtain

$$j \geq \frac{16K^2 \log(2K/\delta)}{\min(\Gamma_\sigma, \sigma_{\min})^2}. \tag{13}$$

A similar argument applies in the case of the assumption on ϵ in Thm. 5. The results of Thm. 5 and Lem. 6 hold at the same time if we have

$$\varepsilon \leq \left(\frac{mK}{\sigma_{\min}}\right)^{3/2} \left(20\epsilon_2 \left(\frac{1}{\Gamma_\sigma} + \frac{1}{\sigma_{\min}}\right) + \epsilon_3\right) \leq C_1 \frac{\lambda_{\min}}{m},$$

where in the first inequality we used that $\varepsilon_3 \leq K^{3/2}$ and $\mu_{\max}^3 \leq K^{3/2}$ by their respective definitions. This combined with high probability bounds of Lem. 7 on ϵ_1 and ϵ_2 implies

$$\left(\frac{m}{\sigma_{\min}}\right)^{1.5} \left(20K^{2.5}\sqrt{\frac{\log(4K/\delta)}{j}} \left(\frac{1}{\Gamma_\sigma} + \frac{1}{\sigma_{\min}}\right) + K^{1.5}\sqrt{\frac{6\log(4K/\delta)}{j}}\right) \leq C_1 \frac{\lambda_{\min}}{m}.$$

By solving this bound w.r.t. j (and some simplifications) we obtain w.p. $1-\delta$

$$j \geq \frac{43^2 m^5 K^6 \log(4K/\delta)}{C_1 \sigma_{\min}^3 \lambda_{\min}^2} \left(\frac{1}{\Gamma_\sigma} + \frac{1}{\sigma_{\min}}\right)^2.$$

Combining this result with that of Eq.13 and taking the union bound leads to the bound of Eq. 5 on the minimum number of samples.

We now draw our attention to the main result of the theorem. We begin by bounding $\|\mu(\theta) - \widehat{\mu}(\pi(\theta))\|$ in terms of estimation error term ϵ_3 and ϵ_2:

$$\begin{aligned}
&\|\mu(\theta) - \widehat{\mu}(\pi(\theta))\| = \|\lambda(\theta)Bv(\theta) - \widehat{\lambda}(\pi(\theta))\widehat{B}\widehat{v}(\pi(\theta))\| \\
\leq& \|(\lambda(\pi(\theta)) - \widehat{\lambda}(\theta))Bv(\pi(\theta))\| + \|\widehat{\lambda}(\theta)(B - \widehat{B})v(\pi(\theta))\| + \|\widehat{\lambda}(\theta)\widehat{B}(v(\pi(\theta)) - \widehat{v}(\theta))\| \\
\leq& |\lambda(\theta) - \widehat{\lambda}(\pi(\theta))|\|B\| + \widehat{\lambda}(\pi(\theta))\|B - \widehat{B}\| + \widehat{\lambda}(\pi(\theta))\|\widehat{B}\|\|v(\theta) - \widehat{v}(\pi(\theta))\|,
\end{aligned} \tag{14}$$

where in the last line we rely on the fact that both $v(\theta)$ and $\widehat{v}(\pi(\theta))$ are normalized vectors. We first bound the term $\|B - \widehat{B}\|$:

$$\begin{aligned}
\|B - \widehat{B}\| &= \|UD^{1/2} - \widehat{U}\widehat{D}^{1/2}\| \leq \|(U - \widehat{U})D^{1/2}\| + \|\widehat{U}(D^{1/2} - \widehat{D}^{1/2})\| \\
&\overset{(1)}{\leq} \frac{2\sqrt{m}\epsilon_2\sigma_{\max}}{\Gamma_\sigma} + \frac{\sqrt{m}\epsilon_2}{\sigma_{\max}} \leq \sqrt{m}\epsilon_2\left(\frac{2\sigma_{\max}}{\Gamma_\sigma} + \frac{1}{\sigma_{\max}}\right),
\end{aligned}$$

where in (1) we make use of the result of Lem. 5. Furthermore, we have

$$\|\widehat{B}\| = \|\widehat{U}\widehat{D}^{1/2}\| \leq \sqrt{m\widehat{\sigma}_{\max}} \leq \sqrt{m}(\sigma_{\max}^{1/2} + \epsilon_2^{1/2}) \leq \sqrt{m}(\sigma_{\max}^{1/2} + \sigma_{\min}^{1/2}) \leq \sqrt{2m\sigma_{\max}},$$

where we used the condition on ϵ_2. This combined with Eq.14 and the result of Thm 5 and Lem. 6 implies

$$
\begin{aligned}
&\|\mu(\pi(\theta)) - \widehat{\mu}(\theta)\| \\
&\overset{(1)}{\leq} 5\sqrt{m\sigma_{\max}}\epsilon + \sqrt{m}\epsilon_2\left(\lambda(\theta)+\epsilon\right)\left(\frac{2\sigma_{\max}}{\Gamma_\sigma} + \frac{1}{\sigma_{\max}}\right) + \frac{8\epsilon}{\lambda(\theta)}\sqrt{2m\sigma_{\max}}\left(\lambda(\theta)+\epsilon\right) \\
&\overset{(2)}{\leq} 5\sqrt{m\sigma_{\max}}\epsilon + \sqrt{m}\epsilon_2\left(\lambda(\theta)+5C_1\frac{\sigma_{\min}}{m}\right)\left(\frac{2\sigma_{\max}}{\Gamma_\sigma} + \frac{1}{\sigma_{\max}}\right) + 8\sqrt{2m\sigma_{\max}}\left(1+5C_1\frac{\sigma_{\min}}{m}\right)\epsilon \\
&\leq 5\sqrt{m\sigma_{\max}}\left(\frac{m}{\sigma_{\min}}\right)^{3/2}\left(10\epsilon_2\left(\frac{1}{\Gamma_\sigma}+\frac{1}{\sigma_{\min}}\right)\left(\epsilon_3+\mu_{\max}^3\right)+\epsilon_3\right) \\
&\quad + \sqrt{m}\epsilon_2\left(\lambda(\theta)+5C_1\frac{\sigma_{\min}}{m}\right)\left(\frac{2\sigma_{\max}}{\Gamma_\sigma}+\frac{1}{\sigma_{\max}}\right) \\
&\quad + 8\sqrt{2m\sigma_{\max}}\left(1+\frac{5C_1}{m}\right)\left(\frac{m}{\sigma_{\min}}\right)^{3/2}\left(10\epsilon_2\left(\frac{1}{\Gamma_\sigma}+\frac{1}{\sigma_{\min}}\right)\left(\epsilon_3+\mu_{\max}^3\right)+\epsilon_3\right).
\end{aligned}
$$

where in (1) we used $||B|| \leq \sqrt{m\sigma_{\max}}$, the bound on $\widehat{\lambda}(\pi(\theta)) \leq \lambda(\theta) + 5\epsilon$, $\|v(\theta) - \widehat{v}(\pi(\theta))\| \leq 8\epsilon/\lambda(\theta)$, in (2) we used $\lambda(\theta) = 1/\sqrt{\rho(\theta)} \geq 1$ and the condition that $\varepsilon \leq 5C_1\sigma_{\min}/m$. The result then follows by combining this bound with the high probability bound of Lem. 7 and taking union bound as well as collecting the terms. □

D Proofs of Section 4.3

Lemma 8. *At episode j, the arms $i \notin \mathcal{A}_*^j(\Theta; \bar{\theta}^j)$ are never pulled, i.e., $T_{i,n} = 0$.*

Lemma 9. *If umUCB is run with*

$$
\varepsilon_{i,t} = \sqrt{\frac{1}{2T_{i,t-1}}\log\left(\frac{2mKn^2}{\delta}\right)}, \qquad \varepsilon^j = C(\Theta)\sqrt{\frac{1}{j}\log\left(\frac{2mKJ}{\delta}\right)}, \tag{15}
$$

where $C(\Theta)$ is defined in Thm. 2, then the event $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$ is such that $\mathbb{P}[\mathcal{E}] \geq 1-\delta$ where $\mathcal{E}_1 = \{\forall \theta, t, i, |\hat{\mu}_{i,t} - \mu_i(\theta)| \leq \varepsilon_{i,t}\}$ and $\mathcal{E}_2 = \{\forall j, \theta, i, |\hat{\mu}_i^j(\theta) - \mu_i(\theta)| \leq \varepsilon^j\}$.

Notice that the event $\mathcal{E}$ implies that for any episode j and step t, the actual model is always in the active set, i.e., $\bar{\theta}^j \in \Theta_t^j$.

Lemma 10. *At episode j, all the arms $i \notin \mathcal{A}_+^j(\Theta_+^j(\bar{\theta}^j); \bar{\theta}^j)$ are never pulled on event $\mathcal{E}$, i.e., $T_{i,n} = 0$ with probability $1-\delta$.*

Lemma 11. *At episode j, the arms $i \in \mathcal{A}_+^j(\Theta_+^j(\bar{\theta}^j); \bar{\theta}^j)$ are never pulled more than with a UCB strategy, i.e.,*

$$
T_{i,n}^j \leq \frac{2}{\Delta_i(\bar{\theta}^j)^2}\log\left(\frac{2mKn^2}{\delta}\right) + 1, \tag{16}
$$

with probability $1-\delta$.

Notice that for *UCB* the logarithmic term in the previous statement would be $\log(Kn^2/\delta)$ which would represent a negligible constant fraction improvement w.r.t. *umUCB* whenever the number of models is of the same order of the number of arms.

Lemma 12. *At episode j, for any model $\theta \in (\Theta_+^j(\bar{\theta}^j) - \widetilde{\Theta}^j(\bar{\theta}^j))$ (i.e., an optimistic model that can be discarded), the number of pulls to any arm $i \in \mathcal{A}_+^j(\theta; \bar{\theta}^j)$ needed before discarding θ is*

$$
T_{i,n}^j \leq \frac{1}{2\big(\Gamma_i(\theta, \bar{\theta}^j)/2 - \varepsilon^j\big)^2}\log\left(\frac{2mKn^2}{\delta}\right) + 1, \tag{17}
$$

with probability $1-\delta$.

Proof of Lem. 8. We first notice that the algorithm only pulls arms recommended by a model $\theta \in \Theta_t^j$. Let $\hat{i}_*(\theta) = \arg\max_i B_t^j(i;\theta)$ with $\theta \in \Theta_t^j$, and $i \in \mathcal{A}_*^j(\theta;\bar{\theta}^j)$. According to the selection process, we have

$$B_t^j(i;\theta) < B_t^j(\hat{i}_*;\theta).$$

Since $\theta \in \Theta_t^j$ we have that for any i, $|\hat{\mu}_{i,t} - \hat{\mu}_i^j(\theta)| \le \varepsilon_{i,t} + \varepsilon^j$ which leads to $\hat{\mu}_i^j(\theta) - \varepsilon^j \le \hat{\mu}_{i,t} + \varepsilon_{i,t}$. Since $\hat{\mu}_i^j(\theta) - \varepsilon^j \le \hat{\mu}_i^j(\theta) + \varepsilon^j$, then we have that

$$\hat{\mu}_i^j(\theta) - \varepsilon^j \le \min\{\hat{\mu}_{i,t} + \varepsilon_{i,t}, \hat{\mu}_i^j(\theta) + \varepsilon^j\} = B_t^j(i;\theta).$$

Furthermore from the definition of the B-values we deduce that

$$B_t^j(\hat{i}_*;\theta) \le \hat{\mu}_{\hat{i}_*}^j(\theta) + \varepsilon^j.$$

Bringing together the previous inequalities, we obtain

$$\hat{\mu}_i^j(\theta) - \varepsilon^j \le \hat{\mu}_{\hat{i}_*}^j(\theta) + \varepsilon^j.$$

which is a contradiction with the definition of non-dominated arms $\mathcal{A}_*^j(\Theta;\bar{\theta}^j)$. □

Proof of Lem. 9. The probability of $\mathcal{E}_1$ is computed in Lem. 2 with the difference that now we need an extra union bound over all the models and that the union bound over the arms cannot be restricted to the number of models. The probability of $\mathcal{E}_2$ follows from Thm. 2. □

Proof of Lem. 10. We first recall that on event $\mathcal{E}$, at any episode j, the actual model $\bar{\theta}^j$ is always in the active set Θ_t^j. If an arm i is pulled, then according to the selection strategy, there exists a model $\theta \in \Theta_t$ such that

$$B_t^j(i;\theta) \ge B_t^j(\hat{i}_*(\bar{\theta}^j);\bar{\theta}^j).$$

Since $\hat{i}_*(\bar{\theta}^j) = \arg\max_i B_t^j(i;\bar{\theta}^j)$, then $B_t^j(\hat{i}_*(\bar{\theta}^j);\bar{\theta}^j) \ge B_t^j(i_*(\bar{\theta}^j);\bar{\theta}^j)$ where $i_*(\bar{\theta}^j)$ is the true optimal arm of $\bar{\theta}^j$. By definition of $B(i;\theta)$, on event $\mathcal{E}$ we have that $B_t^j(i_*(\bar{\theta}^j);\bar{\theta}^j) \ge \mu_*(\bar{\theta}^j)$ and that $B_t^j(i;\theta) \le \hat{\mu}_i^j(\theta) + \varepsilon^j$. Grouping these inequalities we obtain

$$\hat{\mu}_i^j(\theta) + \varepsilon^j \ge \mu_*(\bar{\theta}^j),$$

which, together with Lem. 8, implies that $i \in \mathcal{A}_+^j(\theta;\bar{\theta}^j)$ and that this set is not empty, which corresponds to $\theta \in \Theta_+^j(\bar{\theta}^j)$. □

Proof of Lem. 11. Let t be the last time arm i is pulled ($T_{i,t-1} = T_{i,n} + 1$), then according to the selection strategy we have

$$B_t^j(i;\theta_t^j) \ge B_t^j(\hat{i}_*(\bar{\theta}^j);\bar{\theta}^j) \ge B_t^j(i_*;\bar{\theta}^j),$$

where $i_* = i_*(\bar{\theta}^j)$. Using the definition of B, we have that on event $\mathcal{E}$

$$B_t^j(i_*(\bar{\theta}^j);\bar{\theta}^j) = \min\big\{(\hat{\mu}_{i_*}^j(\bar{\theta}^j) + \varepsilon^j);(\hat{\mu}_{i_*,t} + \varepsilon_{i_*,t})\big\} \ge \mu_*(\bar{\theta}^j)$$

and

$$B_t^j(i;\theta_t^j) \le \hat{\mu}_{i,t} + \varepsilon_{i,t} \le \mu_i(\bar{\theta}^j) + 2\varepsilon_{i,t}.$$

Bringing the two conditions together we have

$$\mu_i(\bar{\theta}^j) + 2\varepsilon_{i,t} \ge \mu_*(\bar{\theta}^j) \Rightarrow 2\varepsilon_{i,t} \ge \Delta_i(\bar{\theta}^j),$$

which coincides with the (high-probability) bound on the number of pulls for i using a *UCB* algorithm and leads to the statement by definition of $\varepsilon_{i,t}$. □

Proof of Lem. 12. According to Lem. 10, a model θ can only propose arms in $\mathcal{A}_+^j(\theta;\bar{\theta}^j)$. Similar to the analysis of *mUCB*, θ is discarded from Θ_t^j with high probability after t steps and j episodes if

$$2(\varepsilon_{i,t} + \varepsilon^j) \le \Gamma_i(\theta, \bar{\theta}^j).$$

At round j, if $\varepsilon^j \ge \Gamma_i(\theta,\bar{\theta}^j)/2$ then the algorithm will never be able to pull i enough to discard θ (i.e., the uncertainty on θ is too large), but since $i \in \mathcal{A}_*^j(\theta;\bar{\theta}^j)$, this corresponds to the case when $\theta \in \widetilde{\Theta}^j(\bar{\theta}^j)$. Thus, the condition on the number of pulls to i is derived from the inequality

$$\varepsilon_{i,t} \le \Gamma_i(\theta,\bar{\theta}^j)/2 - \varepsilon^j.$$

□

E Related Work

As discussed in the introduction, transfer in online learning has been rarely studied. In this section we review possible alternatives and a series of settings which are related to the problem we consider in this paper.

Models estimation. Although in *tUCB* we use *RTP* for the estimation of the model means, a wide number of other algorithms could be used, in particular those based on the method of moments (MoM). Recently a great deal of progress has been made regarding the problem of parameter estimation in LVM based on the method of moments approach (MoM) (Anandkumar et al., 2012c,a,b). The main idea of *MoM* is to match the empirical moments of the data with the model parameters that give rise to nearly the same corresponding population quantities. In general, matching the model parameters to the observed moments may require solving systems of high-order polynomial equations which is often computationally prohibitive. However, for a rich class of LVMs, it is possible to efficiently estimate the parameters only based on the low-order moments (up to the third order) (Anandkumar et al., 2012c). Prior to *RTP* various scenarios for *MoM* are considered in the literature for different classes of LVMs using different linear algebra techniques to deal with the empirical moments Anandkumar et al. (2012c,a). The variant introduced in (Anandkumar et al., 2012c, Algorithm B) recovers the matrix of the means $\{\mu(\theta)\}$ up to a permutation in columns without any knowledge of ρ. Also, theoretical guarantees in the form of sample complexity bounds with polynomial dependency on the parameters of interest have been provided for this algorithm. The excess correlation analysis (ECA) (Alg. 5 in Anandkumar et al. (2012a)) generalizes the idea of the *MoM* to the case that ρ is not fixed anymore but sampled from some Dirichlet distribution. The parameters of this Dirichlet distribution is not to be known by the learner.[6] In this case again we can apply a variant of *MoM* to recover the models.

Online Multi-task. In the online multi-task learning the task change at each step ($n = 1$) but at the end of each step both the true label (in the case of online binary classification) and the identity of the task are revealed. A number of works (Dekel et al., 2006; Saha et al., 2011; Cavallanti et al., 2010; Lugosi et al., 2009) focused on this setting and showed how the samples coming from different tasks can be used to perform multi-task learning and improve the worst-case performance of an online learning algorithm compared to using all the samples separately.

Contextual Bandit. In contextual bandit (e.g., see Agarwal et al., 2012; Langford and Zhang, 2007), at each step the learner observes a context x_t and has to choose the arm which is best for the context. The contexts belong to an arbitrary (finite or continuous) space and are drawn from a stationary distribution. This scenario resembles our setting where tasks arrive in a sequence and are drawn from a ρ. The main difference is that in our setting the learner does not observe explicitly the context and it repeatedly interact with that context for n steps. Furthermore, in general in contextual bandits some similarity between contexts is used, while here the models are completely independent.

Non-stationary Bandit. When the learning algorithm does not know when the actual change in the task happens, then the problem reduces to learning in a piece-wise stationary environment. Garivier and Moulines (2011) introduces a modified version of *UCB* using either a sliding window or discounting to *track* the changing distributions and they show, when optimally tuned w.r.t. the number of switches R, it achieves a (worst-case) expected regret of order $O(\sqrt{TR})$ over a total number of steps T and R switches. Notice that this could be also considered as a partial transfer algorithm. Even in the case when the switch is directly observed, if T is too short to learn from scratch and to identify similarity with other previous tasks, one option is just to transfer the averages computed before the switch. This clearly introduces a transfer bias that could be smaller than the regret cumulated in the attempt of learning from scratch. This is not surprising since transfer is usually employed whenever the number of samples that can be collected from the task at hand is relatively small. If we applied this algorithm to our setting $T = nJ$ and $R = J$, the corresponding performance would be $O(J\sqrt{n})$, which matches the worst-case performance of *UCB* (and *tUCB* as well) on J tasks. This result is not surprising since the advantage of knowing the switching points (every n steps) could always be removed by carefully choosing the worst possible tasks. Nonetheless, whenever we are not facing a worst case, the non-stationary *UCB* would have a much worse performance than *tUCB*.

[6] We only need to know sum of the parameters of the Dirichlet distribution α_0.

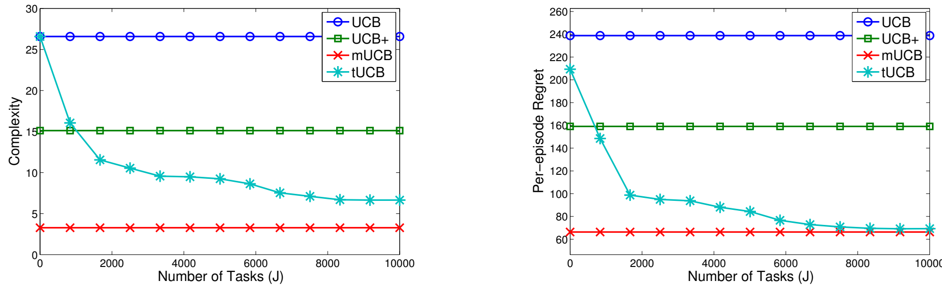

Figure 9: Complexity and per-episode regret of *tUCB* over tasks.

F Numerical Simulations

	Arm1	Arm2	Arm3	Arm4	Arm5	Arm6	Arm7
θ_1	0.9	0.75	0.45	0.55	0.58	0.61	0.65
θ_2	0.75	0.89	0.45	0.55	0.58	0.61	0.65
θ_3	0.2	0.23	0.45	0.35	0.3	0.18	0.25
θ_4	0.34	0.31	0.45	0.725	0.33	0.37	0.47
θ_5	0.6	0.5	0.45	0.35	0.95	0.9	0.8

Table 1: Models.

	UCB	*UCB+*	*mUCB*
θ_1	22.31	14.87	2.33
θ_2	23.32	15.58	8.48
θ_3	33.91	25.21	2.08
θ_4	17.91	11.17	3.48
θ_5	35.41	8.76	0
avg	26.57	15.11	3.27

Table 2: Complexity of *UCB*, *UCB+*, and *mUCB*.

In Table 1 we report the actual values of the means of the arms of the models in Θ, while in Table 2 we compare the complexity of *UCB*, *UCB+*, and *mUCB*, for all the different models and on average. Finally, the graphs in Fig. 9 are an extension up to $J = 10000$ of the performance of *tUCB* for $n = 5000$ reported in the main text.