Parametric Bandits: The Generalized Linear Case

Sarah Filippi
LTCI
Telecom ParisTech et CNRS
Paris, France
filippi@telecom-paristech.fr

Olivier Cappé
LTCI
Telecom ParisTech et CNRS
Paris, France
cappe@telecom-paristech.fr

Aurélien Garivier
LTCI
Telecom ParisTech et CNRS
Paris, France
garivier@telecom-paristech.fr

Csaba Szepesvári
RLAI Laboratory
University of Alberta
Edmonton, Canada
szepesva@ualberta.ca

Abstract

We consider structured multi-armed bandit problems based on the Generalized Linear Model (GLM) framework of statistics. For these bandits, we propose a new algorithm, called GLM-UCB. We derive finite time, high probability bounds on the regret of the algorithm, extending previous analyses developed for the linear bandits to the non-linear case. The analysis highlights a key difficulty in generalizing linear bandit algorithms to the non-linear case, which is solved in GLM-UCB by focusing on the reward space rather than on the parameter space. Moreover, as the actual effectiveness of current parameterized bandit algorithms is often poor in practice, we provide a tuning method based on asymptotic arguments, which leads to significantly better practical performance. We present two numerical experiments on real-world data that illustrate the potential of the GLM-UCB approach.

Keywords: multi-armed bandit, parametric bandits, generalized linear models, UCB, regret minimization.

1 Introduction

In the classical K-armed bandit problem, an agent selects at each time step one of the K arms and receives a reward that depends on the chosen action. The aim of the agent is to choose the sequence of arms to be played so as to maximize the cumulated reward. There is a fundamental trade-off between gathering experimental data about the reward distribution (exploration) and exploiting the arm which seems to be the most promising.

In the basic multi-armed bandit problem, also called the independent bandits problem, the rewards are assumed to be random and distributed independently according to a probability distribution that is specific to each arm –see [1, 2, 3, 4] and references therein. Recently, *structured bandit problems* in which the distributions of the rewards pertaining to each arm are connected by a common unknown parameter have received much attention [5, 6, 7, 8, 9]. This model is motivated by the many practical applications where the number of arms is large, but the payoffs are interrelated. Up to know, two different models were studied in the literature along these lines. In one model, in each times step, a side-information, or context, is given to the agent first. The payoffs of the arms depend both on this side information and the index of the arm. Thus the optimal arm changes with the context [5, 6, 9]. In the second, simpler model, that we are also interested in here, there is no side-information, but the agent is given a model that describes the possible relations between the arms' payoffs. In particular, in "linear bandits" [10, 8, 11, 12], each

arm $a \in \mathsf{A}$ is associated with some d-dimensional vector $m_a \in \mathbb{R}^d$ known to the agent. The expected payoffs of the arms are given by the inner product of their associated vector and some fixed, but initially unknown parameter vector θ_*. Thus, the expected payoff of arm a is $m_a'\theta_*$, which is linear in θ_*.[1]

In this article, we study a richer *generalized linear model* (GLM) in which the expectation of the reward conditionally to the action a is given by $\mu(m_a'\theta_*)$, where μ is a real-valued, non-linear function called the (inverse) link function. This generalization allows to consider a wider class of problems, and in particular cases where the rewards are counts or binary variables using, respectively, Poisson or logistic regression. Obviously, this situation is very common in the fields of marketing, social networking, web-mining (see example of Section 5.2 below) or clinical studies.

Our first contribution is an "optimistic" algorithm, termed GLM-UCB, inspired by the *Upper Confidence Bound* (UCB) approach [2]. GLM-UCB generalizes the algorithms studied by [10, 8, 12]. Our next contribution are finite-time bounds on the statistical performance of this algorithm. In particular, we show that the performance depends on the dimension of the parameter but not on the number of arms, a result that was previously known in the linear case. Interestingly, the GLM-UCB approach takes advantage of the particular structure of the parameter estimate of generalized linear models and *operates only in the reward space*. In contrast, the parameter-space confidence region approach adopted by [8, 12] appears to be harder to generalize to non-linear regression models. Our second contribution is a tuning method based on asymptotic arguments. This contribution addresses the poor empirical performance of the current algorithms that we have observed for small or moderate sample-sizes when these algorithms are tuned based on finite-sample bounds.

The paper is organized as follows. The generalized linear bandit model is presented in Section 2, together with a brief survey of needed statistical results. Section 3 is devoted to the description of the GLM-UCB algorithm, which is compared to related approaches. Section 4 presents our regret bounds, as well as a discussion, based on asymptotic arguments, on the optimal tuning of the method. Section 5 reports the results of two experiments on real data sets.

2 Generalized Linear Bandits, Generalized Linear Models

We consider a structured bandit model with a finite, but possibly very large, number of arms. At each time t, the agent chooses an arm A_t from the set A (we shall denote the cardinality of A by K). The prior knowledge available to the agent consists of a collection of vectors $\{m_a\}_{a\in\mathsf{A}}$ of features which are specific to each arm and a so-called (inverse) *link function* $\mu : \mathbb{R} \to \mathbb{R}$.

The *generalized linear bandit model* investigated in this work is based on the assumption that the payoff R_t received at time t is conditionally independent of the past payoffs and choices and it satisfies

$$\mathbb{E}\left[R_t \middle| A_t\right] = \mu(m_{A_t}'\theta_*)\,, \tag{1}$$

for some unknown parameter vector $\theta_* \in \mathbb{R}^d$. This framework generalizes the linear bandit model considered by [10, 8, 12]. Just like the linear bandit model builds on linear regression, our model capitalizes on the well-known statistical framework of Generalized Linear Models (GLMs). The advantage of this framework is that it allows to address various, specific reward structures widely found in applications. For example, when rewards are binary-valued, a suitable choice of μ is $\mu(x) = \exp(x)/(1+\exp(x))$, leading to the *logistic regression model*. For integer valued rewards, the choice $\mu(x) = \exp(x)$ leads to the *Poisson regression model*. This can be easily extended to the case of *multinomial (or polytomic) logistic regression*, which is appropriate to model situations in which the rewards are associated with categorical variables.

To keep this article self-contained, we briefly review the main properties of GLMs [13]. A univariate probability distribution is said to belong to a *canonical exponential family* if its density with respect to a reference measure is given by

$$p_\beta(r) = \exp\left(r\beta - b(\beta) + c(r)\right)\,, \tag{2}$$

where β is a real parameter, $c(\cdot)$ is a real function and the function $b(\cdot)$ is assumed to be twice continuously differentiable. This family contains the Gaussian and Gamma distributions when the reference measure is the Lebesgue measure and the Poisson and Bernoulli distributions when the reference measure is the counting measure on the integers. For a random variable R with density

[1] Throughout the paper we use the prime to denote transposition.

defined in (2), $\mathbb{E}(R) = \dot{b}(\beta)$ and $\mathrm{Var}(R) = \ddot{b}(\beta)$, where $\dot{b}$ and $\ddot{b}$ denote, respectively, the first and second derivatives of b. In addition, $\ddot{b}(\beta)$ can also be shown to be equal to the Fisher information matrix for the parameter β. The function b is thus strictly convex.

Now, assume that, in addition to the response variable R, we have at hand a vector of covariates $X \in \mathbb{R}^d$. The canonical GLM associated to (2) postulates that $p_\theta(r|x) = p_{x'\theta}(r)$, where $\theta \in \mathbb{R}^d$ is a vector of parameter. Denote by $\mu = \dot{b}$ the so-called *inverse link function*. From the properties of b, we know that μ is continuously differentiable, strictly increasing, and thus one-to-one. The maximum likelihood estimator $\hat{\theta}_t$, based on observations $(R_1, X_1), \ldots (R_{t-1}, X_{t-1})$, is defined as the maximizer of the function

$$\sum_{k=1}^{t-1} \log p_\theta(R_k|X_k) = \sum_{k=1}^{t-1} R_k X_k'\theta - b(X_k'\theta) + c(R_k)\,,$$

a strictly concave function in θ.[2] Upon differentiating, we obtain that $\hat{\theta}_t$ is the unique solution of the following estimating equation

$$\sum_{k=1}^{t-1} \left(R_k - \mu(X_k'\theta)\right) X_k = 0\,, \tag{3}$$

where we have used the fact that $\mu = \dot{b}$. In practice, the solution of (3) may be found efficiently using, for instance, Newton's algorithm.

A semi-parametric version of the above model is obtained by assuming only that $\mathbb{E}_\theta[R|X] = \mu(X'\theta)$ without (much) further assumptions on the conditional distribution of R given X. In this case, the estimator obtained by solving (3) is referred to as the *maximum quasi-likelihood estimator*. It is a remarkable fact that this estimator is consistent under very general assumptions as long as the design matrix $\sum_{k=1}^{t-1} X_k X_k'$ tends to infinity [14]. As we will see, this matrix also plays a crucial role in the algorithm that we propose for bandit optimization in the generalized linear bandit model.

3 The GLM-UCB Algorithm

According to (1), the agent receives, upon playing arm a, a random reward whose expected value is $\mu(m_a'\theta_*)$, where $\theta_* \in \Theta$ is the unknown parameter. The parameter set Θ is an arbitrary closed subset of $\mathbb{R}^d$. Any arm with largest expected reward is called *optimal*. The aim of the agent is to quickly find an optimal arm in order to maximize the received rewards. The *greedy action* $\mathrm{argmax}_{a \in \mathsf{A}}\, \mu(m_a'\hat{\theta}_t)$ may lead to an unreliable algorithm which does not sufficiently explore to guarantee the selection of an optimal arm. This issue can be addressed by resorting to an "optimistic approach". As described by [8, 12] in the linear case, an optimistic algorithm consists in selecting, at time t, the arm

$$A_t = \operatorname*{argmax}_a \max_\theta \mathbb{E}_\theta\left[R_t \,|\, A_t = a\right] \text{ s.t. } \|\theta - \hat{\theta}_t\|_{M_t} \leq \rho(t)\,, \tag{4}$$

where ρ is an appropriate, "slowly increasing" function,

$$M_t = \sum_{k=1}^{t-1} m_{A_k} m_{A_k}' \tag{5}$$

is the design matrix corresponding to the first $t-1$ timesteps and $\|v\|_M = \sqrt{v'Mv}$ denotes the matrix norm induced by the positive semidefinite matrix M. The region $\|\theta - \hat{\theta}_t\|_{M_t} \leq \rho(t)$ is a confidence ellipsoid around the estimated parameter $\hat{\theta}_t$. Generalizing this approach beyond the case of linear link functions looks challenging. In particular, in GLMs, the relevant confidence regions may have a more complicated geometry in the parameter space than simple ellipsoids. As a consequence, the benefits of this form of optimistic algorithms appears dubious.[3]

An alternative approach consists in directly determining an upper confidence bound for the expected reward of each arm, thus choosing the action a that maximizes

$$\mathbb{E}_{\hat{\theta}_t}\left[R_t \,|\, A_t = a\right] + \rho(t)\|m_a\|_{M_t^{-1}}\,.$$

[2] Here, and in what follows log denotes the natural logarithm.

[3] Note that maximizing $\mu(m_a'\theta)$ over a convex confidence region is equivalent to maximizing $m_a'\theta$ over the same region since μ is strictly increasing. Thus, computationally, this approach is not more difficult than it is for the linear case.

In the linear case the two approaches lead to the same solution [12]. Interestingly, for non-linear bandits, the second approach looks more appropriate.

In the rest of this section, we apply this second approach to the GLM bandit model defined in (1). According to (3), the maximum quasi-likelihood estimator of the parameter in the GLM is the unique solution of the estimating equation

$$\sum_{k=1}^{t-1} \left(R_k - \mu(m'_{A_k} \hat{\theta}_t) \right) m_{A_k} = 0 \,, \tag{6}$$

where $A_1, \ldots, A_{t-1}$ denote the arms played so far and $R_1, \ldots, R_{t-1}$ are the corresponding rewards. Let $g_t(\theta) = \sum_{k=1}^{t-1} \mu(m'_{A_k}\theta) m_{A_k}$ be the invertible function such that the estimated parameter $\hat{\theta}_t$ satisfies $g_t(\hat{\theta}_t) = \sum_{k=1}^{t-1} R_k m_{A_k}$. Since $\hat{\theta}_t$ might be outside of the set of admissible parameters Θ, we "project it" to Θ, to obtain $\tilde{\theta}_t$:

$$\tilde{\theta}_t = \operatorname*{argmin}_{\theta \in \Theta} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{M_t^{-1}} = \operatorname*{argmin}_{\theta \in \Theta} \left\| g_t(\theta) - \sum_{k=1}^{t-1} R_k m_{A_k} \right\|_{M_t^{-1}} . \tag{7}$$

Note that if $\hat{\theta}_t \in \Theta$ (which is easy to check and which happened to hold always in the examples we dealt with) then we can let $\tilde{\theta}_t = \hat{\theta}_t$. This is important since computing $\tilde{\theta}_t$ is non-trivial and we can save this computation by this simple check. The proposed algorithm, GLM-UCB, is as follows:

Algorithm 1 GLM-UCB

Input: $\{m_a\}_{a \in \mathsf{A}}$
Play actions $a_1, \ldots, a_d$, receive $R_1, \ldots, R_d$.
for $t > d$ **do**
  Estimate $\hat{\theta}_t$ according to (6)
  if $\hat{\theta}_t \in \Theta$ **let** $\tilde{\theta}_t = \hat{\theta}_t$ **else** compute $\tilde{\theta}_t$ according to (7)
  Play the action $A_t = \operatorname{argmax}_a \left\{ \mu(m'_a \tilde{\theta}_t) + \rho(t) \|m_a\|_{M_t^{-1}} \right\}$, receive R_t
end for

At time t, for each arm a, an upper bound $\mu(m'_a \tilde{\theta}_t) + \beta_t^a$ is computed, where the "exploration bonus" $\beta_t^a = \rho(t) \|m_a\|_{M_t^{-1}}$ is the product of two terms. The quantity $\rho(t)$ is a slowly increasing function; we prove in Section 4 that $\rho(t)$ can be set to guarantee high-probability bounds on the expected regret (for the actual form used, see (8)). Note that the leading term of β_t^a is $\|m_a\|_{M_t^{-1}}$ which decreases to zero as t increases.

As we are mostly interested in the case when the number of arms K is much larger than the dimension d, the algorithm is simply initialized by playing actions $a_1, \ldots, a_d$ such that the vectors $m_{a_1} \ldots, m_{a_d}$ form a basis of $\mathcal{M} = \text{span}(m_a, a \in \mathsf{A})$. Without loss of generality, here and in what follows we assume that the dimension of $\mathcal{M}$ is equal to d. Then, by playing $a_1, \ldots, a_d$ in the first d steps the agent ensures that M_t is invertible for all t. An alternative strategy would be to initialize $M_0 = \lambda_0 I$, where I is the $d \times d$ identify matrix.

3.1 Discussion

The purpose of this section is to discuss some properties of Algorithm 1, and in particular the interpretation of the role played by $\|m_a\|_{M_t^{-1}}$.

Generalizing UCB The standard UCB algorithm for K arms [2] can be seen as a special case of GLM-UCB where the vectors of covariates associated with the arms form an orthogonal system and $\mu(x) = x$. Indeed, take $d = K$, $\mathsf{A} = \{1, \ldots, K\}$, define the vectors $\{m_a\}_{a \in \mathsf{A}}$ as the canonical basis $\{\mathbf{e}_a\}_{a \in \mathsf{A}}$ of $\mathbb{R}^d$, and take $\theta \in \mathbb{R}^d$ the vector whose component θ_a is the expected reward for arm a.

Then, M_t is a diagonal matrix whose a-th diagonal element is the number $N_t(a)$ of times the a-th arm has been played up to time t. Therefore, the exploration bonus in GLM-UCB is given by $\beta_t^a = \rho(t)/\sqrt{N_t(a)}$. Moreover, the maximum quasi-likelihood estimator $\hat{\theta}_t$ satisfies $\bar{R}_t^a = \hat{\theta}_t(a)$

for all $a \in \mathsf{A}$, where $\bar{R}_t^a = \frac{1}{N_t(a)} \sum_{k=1}^{t-1} \mathbb{I}_{\{A_t=a\}} R_k$ is the empirical mean of the rewards received while playing arm a. Algorithm 1 then reduces to the familiar UCB algorithm. In this case, it is known that the expected cumulated regret can be controlled upon setting the slowly varying function ρ to $\rho(t) = \sqrt{2\log(t)}$, assuming that the range of the rewards is bounded by one [2].

Generalizing linear bandits Obviously, setting $\mu(x) = x$, we obtain a linear bandit model. In this case, assuming that $\Theta = \mathbb{R}^d$, the algorithm will reduce to those described in the papers [8, 12]. In particular, the maximum quasi-likelihood estimator becomes the least-squares estimator and as noted earlier, the algorithm behaves identically to one which chooses the parameter optimistically within the confidence ellipsoid $\{\theta : \|\theta - \hat{\theta}_t\|_{M_t} \leq \rho(t)\}$.

Dependence in the Number of Arms In contrast to an algorithm such as UCB, Algorithm 1 does not need that all arms be played even once.[4] To understand this phenomenon, observe that, as $M_{t+1} = M_t + m_{A_t} m'_{A_t}$, $\|m_a\|^2_{M_{t+1}^{-1}} = \|m_a\|^2_{M_t^{-1}} - \left(m'_a M_t^{-1} m_{A_t}\right)^2 / (1 + \|m_{A_t}\|^2_{M_t^{-1}})$ for any arm a. Thus the exploration bonus β^a_{t+1} decreases for all arms, except those which are exactly orthogonal to m_{A_t} (in the M_t^{-1} metric). The decrease is most significant for arms that are colinear to m_{A_t}. This explains why the regret bounds obtained in Theorems 1 and 2 below depend on d but not on K.

4 Theoretical analysis

In this section we first give our finite sample regret bounds and then show how the algorithm can be tuned based on asymptotic arguments.

4.1 Regret Bounds

To quantify the performance of the GLM-UCB algorithm, we consider the *cumulated (pseudo) regret* defined as the expected difference between the optimal reward obtained by always playing an optimal arm and the reward received following the algorithm:

$$\text{Regret}_T = \sum_{t=1}^{T} \mu(m'_{a^*}\theta_*) - \mu(m'_{A_t}\theta_*)\,.$$

For the sake of the analysis, in this section we shall assume that the following assumptions hold:

Assumption 1. *The link function $\mu : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, Lipschitz with constant k_μ and such that $c_\mu = \inf_{\theta\in\Theta, a\in\mathsf{A}} \dot{\mu}(m'_a\theta) > 0$.*

For the logistic function $k_\mu = 1/4$, while the value of c_μ depends on $\sup_{\theta\in\Theta, a\in\mathsf{A}} |m'_a\theta|$.

Assumption 2. *The norm of covariates in $\{m_a : a \in \mathsf{A}\}$ is bounded: there exists $c_m < \infty$ such that for all $a \in \mathsf{A}$, $\|m_a\|_2 \leq c_m$.*

Finally, we make the following assumption on the rewards:

Assumption 3. *There exists $R_{\max} > 0$ such that for any $t \geq 1$, $0 \leq R_t \leq R_{\max}$ holds a.s. Let $\epsilon_t = R_t - \mu(m'_{A_t}\theta_*)$. For all $t \geq 1$, it holds that $\mathbb{E}\left[\epsilon_t | m_{A_t}, \epsilon_{t-1}, \ldots, m_{A_2}, \epsilon_1, m_{A_1}\right] = 0$ a.s.*

As for the standard UCB algorithm, the regret can be analyzed in terms of the difference between the expected reward received playing an optimal arm and that of the best sub-optimal arm:

$$\Delta(\theta_*) = \min_{a:\mu(m'_a\theta_*)<\mu(m'_{a^*}\theta_*)} \mu(m'_{a^*}\theta_*) - \mu(m'_a\theta_*)\,.$$

Theorem 1 establishes a high probability bound on the regret underlying using GLM-UCB with

$$\rho(t) = \frac{2k_\mu \kappa R_{\max}}{c_\mu} \sqrt{2d\log(t)\log(2\,d\,T/\delta)}\,, \tag{8}$$

where T is the fixed time horizon, $\kappa = \sqrt{3 + 2\log(1 + 2c_m^2/\lambda_0)}$ and λ_0 denotes the smallest eigenvalue of $\sum_{i=1}^d m_{a_i} m'_{a_i}$, which by our previous assumption is positive.

[4] Of course, the linear bandit algorithms also share this property with our algorithm.

Theorem 1 (Problem Dependent Upper Bound)**.** *Let $s = \max(1, c_m^2/\lambda_0)$. Then, under Assumptions 1–3, for all $T \geq 1$, the regret satisfies:*

$$\mathbb{P}\left(\textit{Regret}_T \leq (d+1)R_{\max} + \frac{C\,d^2}{\Delta(\theta_*)}\log^2\left[s\,T\right]\log\left[\frac{2d\,T}{\delta}\right]\right) \geq 1-\delta \quad \textit{with} \quad C = \frac{32\kappa^2 R_{\max}^2 k_\mu^2}{c_\mu^2}\,.$$

Note that the above regret bound depends on the true value of θ_* through $\Delta(\theta_*)$. The following theorem provides an upper-bound of the regret independently of the θ_*.

Theorem 2 (Problem Independent Upper Bound)**.** *Let $s = \max(1, c_m^2/\lambda_0)$. Then, under Assumptions 1–3, for all $T \geq 1$, the regret satisfies*

$$\mathbb{P}\left(\textit{Regret}_T \leq (d+1)R_{\max} + Cd\log\left[s\,T\right]\sqrt{T\log\left[\frac{2dT}{\delta}\right]}\right) \geq 1-\delta \quad \textit{with} \quad C = \frac{8R_{\max}k_\mu\kappa}{c_\mu}\,.$$

The proofs of Theorems 1–2 can be found in the supplementary material. The main idea is to use the explicit form of the estimator given by (6) to show that

$$\left|\mu(m'_{A_t}\theta_*) - \mu(m'_{A_t}\hat{\theta}_t)\right| \leq \frac{k_\mu}{c_\mu}\|m_{A_t}\|_{M_t^{-1}}\left\|\sum_{k=1}^{t-1} m_{A_k}\,\epsilon_k\right\|_{M_t^{-1}}\,.$$

Bounding the last term on the right-hand side is then carried out following the lines of [12].

4.2 Asymptotic Upper Confidence Bound

Preliminary experiments carried out using the value of $\rho(t)$ defined equation (8), including the case where μ is the identity function –i.e., using the algorithm described by [8, 12], revealed poor performance for moderate sample sizes. A look into the proof of the regret bound easily explains this observation as the mathematical involvement of the arguments is such that some approximations seem unavoidable, in particular several applications of the Cauchy-Schwarz inequality, leading to pessimistic confidence bounds. We provide here some asymptotic arguments that suggest to choose significantly smaller exploration bonuses, which will in turn be validated by the numerical experiments presented in Section 5.

Consider the canonical GLM associated with an inverse link function μ and assume that the vectors of covariates X are drawn independently under a fixed distribution. This *random design* model would for instance describe the situation when the arms are drawn randomly from a fixed distribution. Standard statistical arguments show that the Fisher information matrix pertaining to this model is given by $J = \mathbb{E}[\dot{\mu}(X'\theta_*)XX']$ and that the maximum likelihood estimate $\hat{\theta}_t$ is such that $t^{-1/2}(\hat{\theta}_t - \theta_*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J^{-1})$, where $\xrightarrow{\mathcal{D}}$ stands for convergence in distribution. Moreover, $t^{-1}M_t \xrightarrow{\text{a.s.}} \Sigma$ where $\Sigma = \mathbb{E}[XX']$. Hence, using the delta-method and Slutsky's lemma

$$\|m_a\|_{M_t^{-1}}^{-1}(\mu(m'_a\hat{\theta}_t) - \mu(m'_a\theta_*)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \dot{\mu}(m'_a\theta_*)\|m'_a\|_{\Sigma^{-1}}^{-2}\|m'_a\|_{J^{-1}}^2)\,.$$

The right-hand variance is smaller than k_μ/c_μ as $J \succeq c_\mu\Sigma$. Hence, for *any sampling distribution* such that J and Σ are positive definite and sufficiently large t and small δ,

$$\mathbb{P}\left(\|m_a\|_{M_t^{-1}}^{-1}(\mu(m'_a\hat{\theta}_t) - \mu(m'_a\theta_*)) > \sqrt{2k_\mu/c_\mu\log(1/\delta)}\right)$$

is asymptotically bounded by δ. Based on the above asymptotic argument, we postulate that using $\rho(t) = \sqrt{2k_\mu/c_\mu\log(t)}$, i.e., inflating the exploration bonus by a factor of $\sqrt{k_\mu/c_\mu}$ compared to the usual UCB setting, is sufficient. This is the setting used in the simulations below.

5 Experiments

To the best of our knowledge, there is currently no public benchmark available to test bandit methods on real world data. On simulated data, the proposed method unsurprisingly outperforms its competitors when the data is indeed simulated from a well-specified generalized linear model. In order to evaluate the potential of the method in more challenging scenarios, we thus carried out two experiments using real world datasets.

5.1 Forest Cover Type Data

In this first experiment, we test the performance of the proposed method on a toy problem using the "Forest Cover Type dataset" from the UCI repository. The dataset (centered and normalized with constant covariate added, resulting in 11-dimensional vectors, ignoring all categorical variables) has been partitioned into $K = 32$ clusters using unsupervised k-means. The values of the response variable for the data points assigned to each cluster are viewed as the outcomes of an arm while the centroid of the cluster is taken as the 11-dimensional vector of covariates characteristic of the arm. To cast the problem into the logistic regression framework, each response variable is binarized by associating the first class ("Spruce/Fir") to a response $R = 1$ and all other six classes to $R = 0$. The proportions of responses equal to 1 in each cluster (or, in other word, the expected reward associated with each arm) ranges from 0.354 to 0.992, while the proportion on the complete set of 581,012 data points is equal to 0.367. In effect, we try to locate as fast as possible the cluster that contains the maximal proportion of trees from a given species. We are faced with a 32-arm problem in a 11-dimensional space with binary rewards. Obviously, the logistic regression model is not satisfied, although we do expect some regularity with respect to the position of the cluster's centroid as the logistic regression trained on all data reaches a 0.293 misclassification rate.

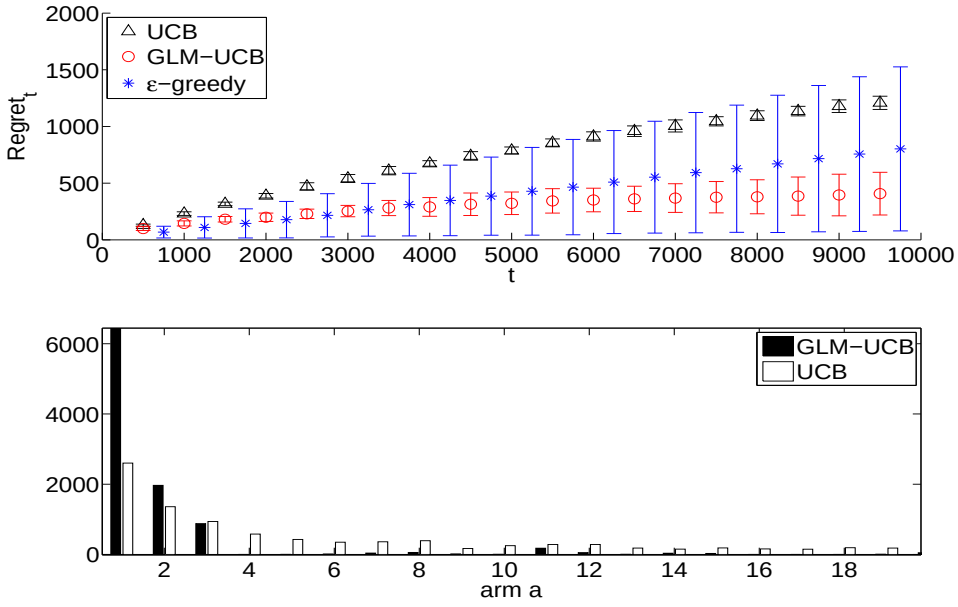

Figure 1: Top: Regret of the UCB, GLM-UCB and the ϵ-greedy algorithms. Bottom: Frequencies of the 20 best arms draws using the UCB and GLM-UCB.

We compare the performance of three algorithms. First, the GLM-UCB algorithm, with parameters tuned as indicated in Section 4.2. Second, the standard UCB algorithm that ignores the covariates. Third, an ϵ-greedy algorithm that performs logistic regression and plays the best estimated action, $A_t = \text{argmax}_a \, \mu(m'_a \hat{\theta}_t)$, with probability $1 - \epsilon$ (with $\epsilon = 0.1$). We observe in the top graph of Figure 1 that the GLM-UCB algorithm achieves the smallest average regret by a large margin. When the parameter is well estimated, the greedy algorithm may find the best arm in little time and then leads to small regrets. However, the exploration/exploitation tradeoff is not correctly handled by the ϵ-greedy approach causing a large variability in the regret. The lower plot of Figure 1 shows the number of times each of the 20 best arms have been played by the UCB and GLM-UCB algorithms. The arms are sorted in decreasing order of expected reward. It can be observed that GML-UCB only plays a small subset of all possible arms, concentrating on the bests. This behavior is made possible by the predictive power of the covariates: by sharing information between arms, it is possible to obtain sufficiently accurate predictions of the expected rewards of all actions, even for those that have never (or rarely) been played.

5.2 Internet Advertisement Data

In this experiment, we used a large record of the activity of internet users provided by a major ISP. The original dataset logs the visits to a set of 1222 pages over a six days period corresponding to about 5.10^8 page visits. The dataset also contains a record of the users clicks on the ads that were presented on these pages. We worked with a subset of 208 ads and 3.10^5 users. The pages (ads) were partitioned in 10 (respectively, 8) categories using Latent Dirichlet Allocation [15] applied to their respective textual content (in the case of ads, the textual content was that of the page pointed to by the ad's link). This second experiment is much more challenging, as the predictive power of the sole textual information turns out to be quite limited (for instance, Poisson regression trained on the entire data does not even correctly identify the best arm).

The action space is composed of the 80 pairs of pages and ads categories: when a pair is chosen, it is presented to a group of 50 users, randomly selected from the database, and the reward is the number of recorded clicks. As the average reward is typically equal to 0.15, we use a logarithmic link function corresponding to Poisson regression. The vector of covariates for each pair is of dimension 19: it is composed of an intercept followed by the concatenation of two vectors of dimension 10 and 8 representing, respectively, the categories of the pages and the ads. In this problem, the covariate vectors do not span the entire space; to address this issue, it is sufficient to consider the pseudo-inverse of M_t instead of the inverse.

On this data, we compared the GLM-UCB algorithm with the two alternatives described in Section 5.1. Figure 2 shows that GLM-UCB once again outperforms its competitors, even though the margin over UCB is now less remarkable. Given the rather limited predictive power of the covariates in this example, this is an encouraging illustration of the potential of techniques which use vectors of covariates in real-life applications.

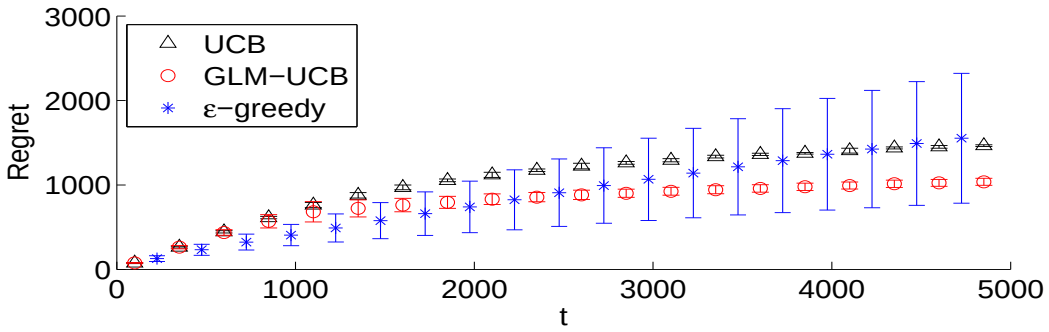

Figure 2: Comparison of the regret of the UCB, GLM-UCB and the ϵ-greedy ($\epsilon = 0.1$) algorithm on the advertisement dataset.

6 Conclusions

We have introduced an approach that generalizes the linear regression model studied by [10, 8, 12]. As in the original UCB algorithm, the proposed GLM-UCB method operates directly in the reward space. We discussed how to tune the parameters of the algorithm to avoid exaggerated optimism, which would slow down learning. In the numerical simulations, the proposed algorithm was shown to be competitive and sufficiently robust to tackle real-world problems. An interesting open problem (already challenging in the linear case) consists in tightening the theoretical results obtained so far in order to bridge the gap between the existing (pessimistic) confidence bounds and those suggested by the asymptotic arguments presented in Section 4.2, which have been shown to perform satisfactorily in practice.

Acknowledgments

This work was supported in part by AICML, AITF, NSERC, PASCAL2 under n°216886, the DARPA GALE project under n°HR0011-08-C-0110 and Orange Labs under contract n°289365.

References

[1] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[3] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge Univ Pr, 2006.

[4] J. Audibert, R. Munos, and Cs. Szepesvári. Tuning bandit algorithms in stochastic environments. *Lecture Notes in Computer Science*, 4754:150, 2007.

[5] C.C. Wang, S.R. Kulkarni, and H.V. Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

[6] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems*, pages 817–824, 2008.

[7] S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. *International Conference on Machine learning*, pages 721–728, 2007.

[8] V. Dani, T.P. Hayes, and S.M. Kakade. Stochastic linear optimization under bandit feedback. *Conference on Learning Theory*, 2008.

[9] S.M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine learning*, pages 440–447. ACM, 2008.

[10] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[11] Y. Abbasi-Yadkori, A. Antos, and Cs. Szepesvári. Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*, 2009.

[12] P. Rusmevichientong and J.N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[13] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[14] K. Chen, I. Hu, and Z. Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Annals of Statistics*, 27(4):1155–1163, 1999.

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 14:601–608, 2002.

[16] V.H. De La Pena, M.J. Klass, and T.L. Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of Probability*, 32(3):1902–1933, 2004.

[17] P. Rusmevichientong and J.N. Tsitsiklis. Linearly parameterized bandits. Arxiv preprint arXiv:0812.3465v2, 2008.

A Supplementary material: Proofs

Before proving Theorems 1 and 2, we provide some preliminary results presented sections A.1 and A.2

A.1 Tail inequalities for vector-valued martingales

We need the following result about vector-valued martingales, extracted from [12].

Lemma 1. *Let $(\mathcal{F}_k; k \geq 0)$ be a filtration, $(m_k; k \geq 0)$ be an $\mathbb{R}^d$-valued stochastic process adapted to $(\mathcal{F}_k)$, $(\eta_k; k \geq 1)$ be a real-valued martingale difference process adapted to $(\mathcal{F}_k)$. Assume that η_k is conditionally sub-Gaussian in the sense that there exists some $R > 0$ such that for any $\gamma \geq 0$, $k \geq 1$,*

$$\mathbb{E}[\exp(\gamma \eta_k) \,|\, \mathcal{F}_{k-1}] \leq \exp\left(\frac{\gamma^2 R^2}{2}\right) \quad \text{a.s.} \tag{9}$$

Consider the martingale $\xi_t = \sum_{k=1}^t m_{k-1}\eta_k$ and the process $M_t = \sum_{k=1}^t m_{k-1}m'_{k-1}$. Assume that with probability one the smallest eigenvalue of M_d is lower bounded by some positive constant λ_0 and that $\|m_k\|_2 \leq c_m$ holds a.s. for any $k \geq 0$.

The following hold true: Let

$$\kappa = \sqrt{3 + 2\log(1 + 2c_m^2/\lambda_0)}. \tag{10}$$

For any $x \in \mathbb{R}^d$, $0 < \delta \leq 1/e$, $t \geq \max(d, 2)$, with probability at least $1 - \delta$,

$$|x'\xi_t| \leq \kappa R\sqrt{2\log t}\,\sqrt{\log(1/\delta)}\;\|x\|_{M_t}. \tag{11}$$

Further, for any $0 < \delta < \min(1, d/e)$, $t \geq \max(d, 2)$, with probability at least $1 - \delta$,

$$\|\xi_t\|_{M_t^{-1}} \leq \kappa R\sqrt{2\,d\log t}\,\sqrt{\log(d/\delta)}. \tag{12}$$

The proof of (11) is based on an exponential inequality of [16] and is adopted from that of Lemma B.4 of [17]. Given (11), inequality (12) follows by some algebra from (11).

Proof. In order to prove (11), we shall use Corollary 2.2 of [16] which states the following: Pick some random variables A and $B \geq 0$ such that

$$\mathbb{E}\left[\exp\left\{\gamma A - \frac{\gamma^2}{2}B^2\right\}\right] \leq 1 \quad \text{for all } \gamma \in \mathbb{R}\,. \tag{13}$$

Then, for all $c \geq \sqrt{2}$, and all $y > 0$,

$$\mathbb{P}\left(|A| \geq c\sqrt{(B^2 + y)\left(1 + \frac{1}{2}\log\left(\frac{B^2}{y} + 1\right)\right)}\right) \leq \exp\left\{-\frac{c^2}{2}\right\}. \tag{14}$$

We apply this inequality to the random variables $A = x'\xi_t/R$ and $B = \|x\|_{M_t}$, where $x \in \mathbb{R}^d$ is some fixed vector. We first check if the so-defined A, B satisfy (13). Pick any $\gamma \in \mathbb{R}$. We first study $\gamma A - (\gamma B)^2/2$. We have

$$\gamma A - (\gamma B)^2/2 = \frac{\gamma\, x'\xi_t}{R} - \frac{\gamma^2\, x'M_t x}{2} = \sum_{k=1}^t D_k\,,$$

where

$$D_k = \tfrac{\gamma}{R}\, x' m_{k-1}\eta_k - \tfrac{\gamma^2}{2} x'\, m_{k-1}m'_{k-1}\, x = \tfrac{\gamma}{R}\, x' m_{k-1}\eta_k - \tfrac{\gamma^2}{2}(x'm_{k-1})^2\,.$$

Now, observe that thanks to (9), $\mathbb{E}\left[\exp(D_k) \,|\, \mathcal{F}_{k-1}\right] \leq 1$. Let $P_k = \exp(D_k)$. Noting that P_k is $\mathcal{F}_k$-adapted,

$$\begin{aligned}
\mathbb{E}\left[\exp(\gamma A - \gamma B^2/2)\right] &= \mathbb{E}\left[P_1 \cdots P_{t-1}P_t\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\,P_1 \cdots P_{t-1}P_t \,|\, \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[P_1 \cdots P_{t-1}\,\mathbb{E}\left[\,P_t \,|\, \mathcal{F}_{t-1}\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\,P_1 \cdots P_{t-1} \,|\, \mathcal{F}_{t-2}\right]\right] = \mathbb{E}\left[P_1 \cdots P_{t-2}\mathbb{E}\left[\,P_{t-1} \,|\, \mathcal{F}_{t-2}\right]\right] \\
&\;\;\vdots \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\,P_1 \,|\, \mathcal{F}_0\right]\right] \leq 1
\end{aligned}$$

which finishes the verification of (13). Now, choose $y = \lambda_0 \|x\|_2^2$ to get from (14) that for all $0 < \delta \le 1/e$, $t \ge 1$, with probability $1 - \delta$,

$$|x'\xi_t| \le R\sqrt{\left(\|x\|_{M_t}^2 + \lambda_0\|x\|_2^2\right)\left(1 + \frac{1}{2}\log\left(1 + \frac{\|x\|_{M_t}^2}{\lambda_0\|x\|_2^2}\right)\right)}\sqrt{2\log\left(\frac{1}{\delta}\right)}\,. \tag{15}$$

Noting that for $t \ge \max(d, 2)$, $\lambda_0\|x\|_2^2 \le \|x\|_{M_t}^2 \le t\|x\|_2^2 c_m^2$, we have $\|x\|_{M_t}^2 + \lambda_0\|x\|_2^2 \le 2\|x\|_{M_t}^2$ and $1 + \frac{1}{2}\log\left(1 + \frac{\|x\|_{M_t}^2}{\lambda_0\|x\|_2^2}\right) \le 1 + \frac{1}{2}\log\left(1 + \frac{tc_m^2}{\lambda_0}\right) \le \kappa^2\log(t)/2$, thanks to the definition of κ. Indeed, it is easy to verify that the slope of function $1 + \frac{1}{2}\log(1 + c_m^2 t/\lambda_0)$ is below that of $\kappa^2\log(t)/2$ for any $t \ge 1$ provided that $\kappa \ge 1$. Hence, the last inequality holds if it holds true for $t = 2$, which, after reordering the terms gives the constraint

$$\kappa \ge \sqrt{\frac{2 + \log(1 + 2c_m^2/\lambda_0)}{\log 2}}.$$

Upper bounding $2/\log 2$ by 3 and $1/\log 2$ by 2, we get the definition of κ, which indeed satisfies $\kappa \ge 1$.

Hence, when (15) holds, it also holds that

$$|x'\xi_t| \le \kappa R\|x\|_{M_t}\sqrt{\log(t)}\sqrt{2\log\left(\frac{1}{\delta}\right)}\,. \tag{16}$$

which is exactly (11).

Now, let us turn to proving (12). Denote by S_t the symmetric, positive definite matrix such that $S_t^2 = M_t$ and, for all $1 \le i \le d$, let $\mathbf{e}_i$ be the i^{th} unit vector (i.e., for all $j \ne i$, $\mathbf{e}_{ij} = 0$ and $\mathbf{e}_{ii} = 1$). Noting that the identity matrix can be written as $I = \sum_{i=1}^d \mathbf{e}_i\mathbf{e}_i'$, we have $\|\xi_t\|_{M_t^{-1}}^2 = \xi_t' M_t^{-1}\xi_t = \xi_t' S_t^{-1} I S_t^{-1}\xi_t = \sum_{i=1}^d \xi_t' S_t^{-1}\mathbf{e}_i\mathbf{e}_i' S_t^{-1}\xi_t$. Therefore, for any constant $\tau > 0$,

$$\begin{aligned}\mathbb{P}\left[\|\xi_t\|_{M_t^{-1}}^2 \ge d\tau^2\right] &= \mathbb{P}\left[\sum_{i=1}^d \xi_t' S_t^{-1}\mathbf{e}_i\mathbf{e}_i' S_t^{-1}\xi_t \ge d\tau^2\right] \le \sum_{i=1}^d \mathbb{P}\left[\xi_t' S_t^{-1}\mathbf{e}_i\mathbf{e}_i' S_t^{-1}\xi_t \ge \tau^2\right] \\ &\le \sum_{i=1}^d \mathbb{P}\left[|\xi_t' S_t^{-1}\mathbf{e}_i| \ge \tau\right]\,.\end{aligned}$$

Applying (11) with $x = S_t^{-1}\mathbf{e}_i$, and $\tau = \kappa R\left\|S_t^{-1}\mathbf{e}_i\right\|_{M_t}\sqrt{\log(t)}\sqrt{2\log\left(\frac{d}{\delta}\right)}$, $0 < \delta < \min(1, d/e)$, $t \ge \max(d, 2)$, and using the fact that $\left\|S_t^{-1}\mathbf{e}_i\right\|_{M_t} = 1$, we have

$$\mathbb{P}\left[\|\xi_t\|_{M_t^{-1}}^2 \ge 2d\kappa^2 R^2\log(t)\,\log\left(\frac{d}{\delta}\right)\right] \le \delta\,,$$

thus, finishing the proof. □

Remark 1. *Note that if $\eta_k \in [\alpha_k - R, \alpha_k + R]$ holds almost surely for some $\mathcal{F}_{k-1}$-measurable random variable α_k then, using Hoeffding's lemma (see, e.g., Lemma A.1 of [3]), we get that for all $\gamma \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp\left\{\gamma\eta_k\right\} | \mathcal{F}_{k-1}\right] \le \exp\left\{\gamma\mathbb{E}\left[\eta_k | \mathcal{F}_{k-1}\right]\right\}\exp\left\{\frac{4R^2\gamma^2}{8}\right\} = \exp\left\{\frac{\gamma^2 R^2}{2}\right\}\,,$$

showing that (η_k) satisfies the sub-Gaussian conditions (9)*. In particular, this holds if $|\eta_k| \le R$ holds almost surely.*

A.2 A bound on the prediction error

In this section we prove some bounds on the error of predicting the mean-rewards.

We start with the following result:

Proposition 1. *Take any δ, t such that $0 < \delta < \min(1, d/e)$, $1 + \max(d, 2) \le t \le T$. Let $\tilde{A}_t$ be any* A*-valued random variable. Let*

$$\beta_t^a(\delta) = \frac{2\, k_\mu \kappa R_{\max}}{c_\mu} \|m_a\|_{M_t^{-1}} \sqrt{2\, d \log t} \sqrt{\log(d/\delta)}\,, \tag{17}$$

where κ is defined by (10)*. Then, with probability at least $1 - \delta$, it holds that*

$$\left|\mu(m'_{\tilde{A}_t}\theta_*) - \mu(m'_{\tilde{A}_t}\tilde{\theta}_t)\right| \le \beta_t^{\tilde{A}_t}(\delta)\,.$$

Proof. Pick a time t such that $d + 1 \le t \le T$ and an action $a \in \mathsf{A}$. We start with bounding $\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right|$. Since μ is Lipschitz, we have $|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)| \le k_\mu |m'_a(\theta_* - \tilde{\theta}_t)|$. By Assumption 1, ∇g_t is continuous,[5] hence, by the Fundamental Theorem of Calculus,

$$g_t(\theta_*) - g_t(\tilde{\theta}_t) = G_t(\theta_* - \tilde{\theta}_t)\,,$$

where

$$G_t = \int_0^1 \nabla g_t\left(s\theta_* + (1-s)\tilde{\theta}_t\right) ds\,.$$

Now, for any $\theta \in \Theta$, $\nabla g_t(\theta) = \sum_{k=1}^{t-1} m_{A_k} m'_{A_k} \dot{\mu}(m'_{A_k}\theta)$. Therefore, thanks to Assumption 1, we have $G_t \succeq c_\mu M_t \succeq c_\mu M_d \succ 0$, where in the last step we used that the first d actions are such that $M_d \succeq \lambda_0 I \succ 0$. Thus, G_t is positive definite and, hence, it is also non-singular. Therefore,

$$\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right| \le k_\mu \left|m'_a G_t^{-1}(g_t(\theta_*) - g_t(\tilde{\theta}_t))\right|\,.$$

Since G_t^{-1} is also positive definite, we get

$$\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right| \le k_\mu \|m_a\|_{G_t^{-1}} \left\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\right\|_{G_t^{-1}}\,. \tag{18}$$

Since $G_t \succeq c_\mu M_t$ implies that $G_t^{-1} \preceq c_\mu^{-1} M_t^{-1}$, $\|x\|_{G_t^{-1}} \le \frac{1}{\sqrt{c_\mu}} \|x\|_{M_t^{-1}}$ holds for arbitrary $x \in \mathbb{R}^d$. Hence,

$$\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right| \le \frac{k_\mu}{c_\mu} \|m_a\|_{M_t^{-1}} \left\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\right\|_{M_t^{-1}}\,.$$

Now,

$$\begin{aligned}
\left\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\right\|_{M_t^{-1}} &\le \left\|g_t(\theta_*) - g_t(\hat{\theta}_t)\right\|_{M_t^{-1}} + \left\|g_t(\hat{\theta}_t) - g_t(\tilde{\theta}_t)\right\|_{M_t^{-1}} \\
&\le 2\left\|g_t(\theta_*) - g_t(\hat{\theta}_t)\right\|_{M_t^{-1}}\,,
\end{aligned}$$

where the first inequality follows from the triangle inequality and second follows since by assumption $\theta_* \in \Theta$ and because of the optimizing property of $\tilde{\theta}_t$ within Θ.

Thanks to the definition of $\hat{\theta}_t$, and using $\epsilon_k = R_k - \mu(m'_{A_k}\theta_*)$, $\xi_t \stackrel{\text{def}}{=} g_t(\hat{\theta}_t) - g_t(\theta_*) = \sum_{k=1}^{t-1} m_{A_k}\epsilon_k$. Therefore,

$$\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right| \le \frac{2\, k_\mu}{c_\mu} \|m_a\|_{M_t^{-1}} \|\xi_t\|_{M_t^{-1}}\,.$$

Since this holds simultaneously for all $a \in \mathsf{A}$, it also holds when a is replaced by *any* A-valued random variable $\tilde{A}_t$:

$$\left|\mu(m'_{\tilde{A}_t}\theta_*) - \mu(m'_{\tilde{A}_t}\tilde{\theta}_t)\right| \le \frac{2\, k_\mu}{c_\mu} \|m_{\tilde{A}_t}\|_{M_t^{-1}} \|\xi_t\|_{M_t^{-1}}\,. \tag{19}$$

Now, let us use Lemma 1 to bound $\|\xi_t\|_{M_t^{-1}}$. Set $m_k = m_{A_{k+1}}$ ($k = 0, 1, \ldots$), $\eta_k = \epsilon_k$ ($k = 1, 2, \ldots$), $\mathcal{F}_k = \sigma(m_s, \eta_s; s \le k)$. Due to Assumption 3, $\mathbb{E}[\eta_k|\mathcal{F}_{k-1}] = \mathbb{E}[\eta_k|m_{k-1}, \eta_{k-1}, \ldots, m_1, \eta_1, m_0] = \mathbb{E}[\epsilon_k|m_{A_k}, \epsilon_{k-1}, \ldots, m_{A_2}, \epsilon_1, m_{A_1}] = 0$. Since by the

[5] For all $x \in \mathbb{R}^d$, $\nabla g_t(x)$ denotes the Jacobian matrix of g_t at point x.

same assumption, $|\epsilon_k| \le R_{\max}$, we may choose $R = R_{\max}$ by Remark 1. Further, by Assumption 2, $\|m_k\|_2 = \|m_{A_{k+1}}\|_2 \le \max_{a\in\mathsf{A}} \|m_a\|_2 \le c_m$, and, by the choice of the first d actions, $\sum_{k=1}^d m_{k-1}m'_{k-1} = \sum_{k=1}^d m_{A_k}m'_{A_k} \succeq \lambda_0 I$. Therefore, all the assumptions of the Lemma are met and we can conclude that for any $0 < \delta < \min(1, d/e)$, $t \ge 1 + \max(d, 2)$, with probability at least $1 - \delta$,

$$\|\xi_t\|_{M_t^{-1}} \le \kappa R_{\max}\sqrt{2\,d\log t}\,\sqrt{\log(d/\delta)}, \tag{20}$$

where κ is defined by (10).

By chaining (19) and (20), we get that on the event when (20) holds, we also have

$$\left|\mu(m'_{\tilde{A}_t}\theta_*) - \mu(m'_{\tilde{A}_t}\tilde{\theta}_t)\right| \le \frac{2\,k_\mu\kappa R_{\max}}{c_\mu}\|m_{\tilde{A}_t}\|_{M_t^{-1}}\sqrt{2\,d\log t}\,\sqrt{\log(d/\delta)}\,,$$

finishing the proof. □

Proposition 1 implies the following bound on the immediate mean regret:

Proposition 2. *For all δ such that $0 < \delta \le \min(1, 2Td/e)$, simultaneously for all $t \in \{1 + \max(d, 2), \dots, T\}$,*

$$\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*) \le 2\,\beta_t^{A_t}\left(\tfrac{\delta}{2T}\right).$$

holds with probability at least $1 - \delta$.

Proof. Fix $t \in \{1 + \max(d, 2), \dots, T\}$ and let δ be as in the statement. Consider the decomposition

$$\begin{aligned}\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*) = &\left(\mu(m'_{a_*}\theta_*) - \mu(m_{a_*}\tilde{\theta}_t)\right)\\ &+ \left(\mu(m_{a_*}\tilde{\theta}_t) - \mu(m_{A_t}\tilde{\theta}_t)\right) + \left(\mu(m_{A_t}\tilde{\theta}_t) - \mu(m'_{A_t}\theta_*)\right).\end{aligned}$$

Now, according to Proposition 1, outside of an event of measure bounded by $\delta/(2T)$,

$$\mu(m'_{a_*}\theta_*) - \mu(m'_{a_*}\tilde{\theta}_t) \le \beta_t^{a_*}(\delta/(2T))\,.$$

Also, outside of an event of measure bounded by $\delta/(2T)$,

$$\mu(m'_{A_t}\theta_*) - \mu(m'_{A_t}\tilde{\theta}_t) \le \beta_t^{A_t}(\delta/(2T))\,.$$

Further, by the definition of A_t,

$$\begin{aligned}\mu(m_{a_*}\tilde{\theta}_t) - \mu(m_{A_t}\tilde{\theta}_t) &= \mu(m_{a_*}\tilde{\theta}_t) + \beta_t^{a_*}(\delta/(2T)) - \mu(m_{A_t}\tilde{\theta}_t) - \beta_t^{a_*}(\delta/(2T))\\ &\le \mu(m_{A_t}\tilde{\theta}_t) + \beta_t^{A_t}(\delta/(2T)) - \mu(m_{A_t}\tilde{\theta}_t) - \beta_t^{a_*}(\delta/(2T))\\ &= \beta_t^{A_t}(\delta/(2T)) - \beta_t^{a_*}(\delta/(2T)).\end{aligned}$$

Chaining the inequalities and using a union bound gives the final result. □

According to the previous proposition, the behavior of the immediate regret at time step t is bounded by $2\beta_t^{A_t}(\delta/2T) = 2\rho(t)\|m_{A_t}\|_{M_t^{-1}} \le 2\rho(T)\|m_{A_t}\|_{M_t^{-1}}$. Therefore, with $t_0 = 1 + \max(d, 2)$, outside of an event of probability at most δ, we can bound the cumulated regret up to time T by

$$\text{Regret}_T \le (t_0 - 1)R_{\max} + \sum_{t=t_0}^{T} \min\left\{\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*), R_{\max}\right\} \tag{21}$$

$$\le (t_0 - 1)R_{\max} + 2\,\rho(T)\sum_{t=t_0}^{T} \min\left\{\|m_{A_t}\|_{M_t^{-1}}, 1\right\}, \tag{22}$$

where the last inequality follows from the fact that $R_{\max} \le 2\rho(T)$ by definition of $\rho(T)$. Note that $\|m_{A_t}\|_{M_t^{-1}}$ is expected to become small as t gets large. This motivates us to bound a sum of $\|m_{A_t}\|^2_{M_t^{-1}}$. For technical reasons that will become clear later, we bound $\sum_{t=d}^{T} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\}$.

Proposition 3. *Let $t_0 \geq d+1$. Then,*

$$\sum_{t=t_0}^{T} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\} \leq 2\,d\,\log\left(\frac{c_m^2 T}{\lambda_0}\right) \quad \text{a.s.}\,.$$

Proof. This proof follows the steps of the proof of Lemma 9 of [8]. By the definition of M_{t+1}, we have

$$\begin{aligned}\det\left(M_{t+1}\right) &= \det\left(M_t + m_{A_t} m_{A_t}'\right) = \det\left(M_t\right) \det\left(I + M_t^{-1/2} m_{A_t} (M_t^{-1/2} m_{A_t})'\right) \\ &= \det\left(M_t\right)\left(1 + \|m_{A_t}\|^2_{M_t^{-1}}\right) = \det\left(M_{t_0}\right) \prod_{k=t_0}^{t}\left(1 + \|m_{A_k}\|^2_{M_k^{-1}}\right),\end{aligned}$$

where the last line follows from the fact that $1 + \|m_{A_t}\|^2_{M_t^{-1}}$ is an eigenvalue of the matrix $I + M_t^{-1/2} m_{A_t} (M_t^{-1/2} m_{A_t})'$ and that all the other eigenvalues are equal to 1. Thus, using the fact that $x \leq 2\log(1+x)$ which holds for any $0 \leq x \leq 1$, we have

$$\begin{aligned}\sum_{t=t_0}^{T} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\} &\leq 2 \sum_{t=t_0}^{T} \log\left(1 + \|m_{A_t}\|^2_{M_t^{-1}}\right) \\ &= 2\log \prod_{t=t_0}^{T}\left(1 + \|m_{A_t}\|^2_{M_t^{-1}}\right) \\ &= 2\log\left(\frac{\det(M_{T+1})}{\det(M_{t_0})}\right).\end{aligned}$$

Note that the trace of M_{t+1} is upper-bounded by $t\,c_m^2$. Then, since the trace of the positive definite matrix M_{t+1} is equal to the sum of its eigenvalues and $\det(M_{t+1})$ is the product of its eigenvalues, we have $\det(M_{t+1}) \leq (t c_m^2)^d$. In addition, $\det(M_{t_0}) \geq \lambda_0^d$ since $t_0 \geq d+1$. Thus,

$$\sum_{t=t_0}^{T} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\} \leq 2\,d \log\left(\frac{c_m^2 T}{\lambda_0}\right).$$

□

A.3 Proof of the Main Theorems

A.3.1 Proof of Theorem 1

Proof. We start from (21), where $t_0 = 1 + \max(d, 2)$. According to the definition of $\Delta(\theta_*)$ whenever A_t is a suboptimal action, $\mu(m_{a_*}'\theta_*) - \mu(m_{A_t}'\theta_*) \geq \Delta(\theta_*)$, while in the other case we have $\mu(m_{a_*}'\theta_*) - \mu(m_{A_t}'\theta_*) = 0$. In both cases, we can write

$$\mu(m_{a_*}'\theta_*) - \mu(m_{A_t}'\theta_*) \leq \frac{(\mu(m_{a_*}'\theta_*) - \mu(m_{A_t}'\theta_*))^2}{\Delta(\theta_*)}\,.$$

According to Proposition 2, with probability $1-\delta$, simultaneously for all $t \in \{t_0, \ldots, T\}$,

$$\mu(m_{a_*}'\theta_*) - \mu(m_{A_t}'\theta_*) \leq 2\beta_t^{A_t}(\delta/(2T)) = 2\rho(t)\,\|m_{A_t}\|_{M_t^{-1}}\,.$$

Therefore, on the event when these inequalities holds, we have

$$\begin{aligned}\sum_{t=t_0}^{T} \min\left\{\mu(m_{a_*}'\theta_*) - \mu(m_{A_t}'\theta_*), R_{\max}\right\} &\leq \sum_{t=t_0}^{T} \min\left\{4\frac{\rho(t)^2}{\Delta(\theta_*)}\|m_{A_t}\|^2_{M_t^{-1}}, R_{\max}\right\} \\ &\leq 4\frac{\rho(T)^2}{\Delta(\theta_*)} \sum_{t=t_0}^{T} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\}\end{aligned}$$

where the last inequality follows from the fact that $\Delta(\theta_*) \leq R_{\max} \leq 4\rho(T)^2/R_{\max}$ and that $\rho(.)$ is an increasing function. Combining this with the bound of Proposition 3, we get

$$\sum_{t=t_0}^{T} \min\left\{\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*), R_{\max}\right\} \leq 8\,d\,\frac{\rho(T)^2}{\Delta(\theta_*)}\,\log\left(\frac{c_m^2 T}{\lambda_0}\right).$$

Plugging in the definition of $\rho(T)$, we get that it holds with probability $1-\delta$ that

$$\begin{aligned}\text{Regret}_T &\leq (t_0-1)R_{\max} + \sum_{t=t_0}^{T} \min\left\{\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*), R_{\max}\right\} \\ &\leq (t_0-1)R_{\max} + \frac{32\,d^2\,\kappa^2 R_{\max}^2 k_\mu^2}{c_\mu^2 \Delta(\theta_*)}\,\log(T)\log(2d\,T/\delta)\log\left(\frac{c_m^2 T}{\lambda_0}\right).\end{aligned}$$

□

A.3.2 Proof of Theorem 2

Proof. Let $t_0 = 1 + \max(d, 2)$. According to Proposition 2, (22) holds with probability $1-\delta$, so it remains to bound

$$\sum_{t=t_0}^{T} \min\left\{\|m_{A_t}\|_{M_t^{-1}}, 1\right\}.$$

Using the Cauchy-Schwarz inequality and Proposition 3, we have

$$\begin{aligned}\sum_{t=t_0}^{T} \min\left\{\|m_{A_t}\|_{M_t^{-1}}, 1\right\} &\leq \sqrt{T}\sqrt{\sum_{t=t_0}^{T} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\}} \\ &\leq \sqrt{T}\,\sqrt{2d\log(c_m^2 T/\lambda_0)}.\end{aligned}$$

Combining with (22) and using the definition of $\rho(\cdot)$ gives

$$\begin{aligned}\text{Regret}_T &\leq (t_0-1)R_{\max} + 2\,\rho(T)\,\sqrt{2\,d\,T\log(c_m^2 T/\lambda_0)} \\ &= (t_0-1)R_{\max} + 8\,d\,\frac{k_\mu \kappa R_{\max}}{c_\mu}\,\sqrt{T\,\log(T)\,\log(c_m^2 T/\lambda_0)\log(2Td/\delta)} \\ &\leq (d+1)R_{\max} + 8\,d\,\frac{k_\mu \kappa R_{\max}}{c_\mu}\,\log(s\,T)\,\sqrt{T\,\log(2Td/\delta)},\end{aligned}$$

where $s = \max\left(\frac{c_m^2}{\lambda_0}, 1\right)$, thus, finishing the proof. □