Global Multi-armed Bandits with Hölder Continuity

Onur Atan
UCLA

Cem Tekin
UCLA

Mihaela van der Schaar
UCLA

Abstract

Standard Multi-Armed Bandit (MAB) problems assume that the arms are independent. However, in many application scenarios, the information obtained by playing an arm provides information about the remainder of the arms. Hence, in such applications, this informativeness can and should be exploited to enable faster convergence to the optimal solution. In this paper, formalize a new class of multi-armed bandit methods, *Global Multi-armed Bandit* (GMAB), in which arms are *globally* informative through a *global parameter*, i.e., choosing an arm reveals information about *all* the arms. We propose a greedy policy for the GMAB which always selects the arm with the highest estimated expected reward, and prove that it achieves *bounded parameter-dependent regret*. Hence, this policy selects suboptimal arms only finitely many times, and after a finite number of initial time steps, the *optimal arm* is selected in *all* of the remaining time steps with probability one. In addition, we also study how the *informativeness* of the arms about each other's rewards affects the speed of learning. Specifically, we prove that the parameter-free (worst-case) regret is sublinear in time, and decreases with the informativeness of the arms. We also prove a sublinear in time *Bayesian risk* bound for the GMAB which reduces to the well-known Bayesian risk bound for linearly parameterized bandits when the arms are *fully informative*. GMABs have applications ranging from drug dosage control to dynamic pricing.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

1 Introduction

In this paper we study a new class of multi-armed bandit (MAB) problems which we name the Global MAB (GMAB). In the GMAB problem, a learner sequentially selects one of the available K arms with the goal of maximizing its total expected reward. We assume that expected reward of arm k is $\mu_k(\theta_*)$, where $\theta_* \in \Theta$ is an unknown global parameter. For the given global parameter θ_*, the reward of each arm follows an i.i.d. process. The learner knows the expected reward function $\mu_k(\cdot)$ of all the arms k. In this setting an arm k is informative about another arm k' because the learner can estimate the expected reward of arm k' by using the estimated reward of arm k and the expected reward functions $\mu_k(\cdot)$ and $\mu_{k'}(\cdot)$. Under mild assumptions on the expected reward functions, we prove that a greedy policy which always selects the arm with the highest estimated expected reward achieves *bounded regret*, which is independent of time. In other words, suboptimal arms are selected only finitely many times before converging to the optimal arm. This is the case because the expected arm rewards in standard MAB problems [18] are independent of each other, in contrast to the GMAB problem in which the expected rewards of the arms are related through θ_*.

One prominent example of GMAB is provided in [20] in which the expected rewards of the arms are known linear functions of a global parameter. Under this assumption, [20] proves that the greedy policy achieves bounded regret. Proving finite regret bounds under this linearity assumption becomes possible since all the arms are fully informative about each other, i.e., rewards obtained from an arm can be used to estimate the expected reward of the other arms using a linear transformation on the obtained rewards.

In this paper we consider a more general model in which the expected reward functions are *Hölder continuous*, which requires to use a non-linear estimator to exploit the *weak informativeness*. Thus, our model includes the case when the expected reward functions are linear functions as a special case. However, while our regret results are a generalization of the results in [20], our analysis of the regret is more complicated since the arms are not fully informative as in the linear case. Thus, deriving regret bounds in our setting requires us to develop new proof techniques. However, we also show that our learning algorithm and the regret bounds reduce to the ones in [20] when the arms have linear reward functions. In addition to the bounded regret bound (which depends on the value of the global parameter θ_*), we also provide a parameter-free regret bound and a bound on the Bayesian risk given a distribution $f(\cdot)$ over the parameter space Θ, which matches known upper bound $\Omega(\log T)$ for the linear reward functions [20]. Both of these bounds are sublinear in time and depend on the informativeness of the arms with respect to the other arms, subsequently referred to shortly as informativeness.

Many applications can be formalized as a GMAB, where the reward functions are *Hölder continuous* in the global parameter. An import example is clinical trials involving drug dosage control [17]. Another example is dynamic pricing [9]. In dynamic pricing, an agent sequentially selects a price from a set of prices $\mathcal{P}$ with the objective of maximizing its revenue over a finite time horizon. At time t, the agent first selects a price $p \in \mathcal{P}$, and then observes the amount of sales, which is given as $S_{p,t}(\Lambda) = \bar{F}_p(\Lambda) + \epsilon_t$, where $\bar{F}(.)$ is modulating function and ϵ_t is the noise term with zero mean. The modulating function is the purchase probability of an item of price p given the market size Λ. Here, the market size is the global parameter, which is unknown and needs to be learned by setting any price and observing the sales related to that price. Commonly used modulating functions include the exponential and logistic functions.

1.1 Summary of contributions

- We formalize a new class of structured MAB problems, which we refer to as Global MABs. This class of problems represents a generalization of the linearly parametrized bandits in [20].

- For GMABs, we propose a greedy policy that always selects the arm with the highest estimated expected reward. We prove that the greedy policy achieves bounded regret independent of time horizon T, depending on θ_*.

- In addition to proving that the regret is bounded (which is related to the asymptotic behavior), we also show how the regret increases over time by identifying and characterizing three regimes of growth: first, the regret increases at most sublinearly over time until a first threshold (that depends on the informativeness) after which it increases at most logarithmically over time until a second threshold, before converging to a finite number asymptotically. These thresholds have the property that they are decreasing in the informativeness.

- We prove a sublinear in time worst-case (parameter-free) regret bound. The rate of increase in time decreases with the informativeness of the arms, meaning that the regret will increase slower when the informativeness is high.

- Given a distribution over the set of global parameter values, we prove a Bayesian risk bound that depends on the informativeness. When the arms are *fully informative*, such as in the case of linearly parametrized bandits [20], our Bayesian risk bound and our proposed greedy policy reduce to the well known Bayesian risk bound and the greedy policy in [20], respectively.

1.2 Related Work

Numerous types of MAB problems have been defined and investigated in the past decade - these include stochastic bandits [1, 4, 5, 13, 18], Bayesian bandits [2, 7, 15, 16, 23], contextual bandits [3, 19, 22], combinatorial bandits [12], and many other variants. Instead of comparing our method against all these MAB variants, we group the existing literature based on the main theme of this paper: exploiting the informativeness of an arm to learn about the rewards of other arms. We call a MAB problem *non-informative* if the reward observations of any arm do not reveal any information about the expected rewards of any other arms. Examples of non-informative MAB are the stochastic bandits [5,18] and the bandits with local parameters [2,15]. In these problems the regret grows at least logarithmically in time, since each arm should be selected at least logarithmically many times to identify the optimal arm. We call a MAB problem *group-informative* if the reward observations from an arm provide information about the rewards of a known group of other arms but not all the arms. Examples of group-informative MAB problems are combinatorial bandits [12], contextual bandits [3, 19, 22] and structured bandits [11, 21]. In these problems the regret grows at least logarithmically over time since at least one suboptimal arm should be selected at least logarithmically many times to identify groups of arms that are suboptimal. We call a MAB problem *globally-informative* if the reward observations from an arm provides information about the rewards of *all* the arms. The proposed GMABs include the linearly-parametrized MABs in [20] as a subclass. Therefore, we prove a bounded regret for a larger class of problems.

Another related work is [14], in which the optimal arm selection strategy is derived for the infinite time horizon learning problem, when the arm rewards are parametrized with known priors, and the future rewards are discounted. However, in the Gittins' formulation of the MAB problem, the parameters of the arms are different from each other, and the discounting allows the learner to efficiently solve the optimization problem related to arm selection by decoupling the joint optimization problem into K individual optimization problems - one for each arm. In contrast, we do not assume known priors, and the learner in our case does not solve an optimization problem but rather learns the global parameter through its reward observations.

Another seemingly related learning scenario is the experts setting [8], where after an arm is chosen, the rewards of all arms are observed and their estimated rewards is updated. Hence, there is no tradeoff between exploration and exploitation and finite regret bounds can be achieved in an expert system with finite number of arms and stochastic arm rewards. However, unlike in the expert setting, the GMABs achieve finite regret bounds while observing *only* the reward of the selected arm. Hence, the arm reward estimation procedure in GMABs requires forming reward estimates by collectively considering the observed rewards from all the arms, which is completely different than in the expert systems, in which the expected reward of an arm is estimated only by using the past reward observations from that arm.

2 Global Multi-Armed Bandits

2.1 Problem Formulation

The set of all arms is denoted by $\mathcal{K}$ and the number of arms is $K = |\mathcal{K}|$, where $|\cdot|$ is the cardinality operator. The reward obtained by playing an arm $k \in \mathcal{K}$ at time t is given by a random variable $X_{k,t}$. We assume that for $t \geq 1$ and $k \in \mathcal{K}$, $X_{k,t}$ is drawn independently from an unknown distribution $\nu_k(\theta_*)$ with support $[0,1]$.[1] The learner knows that the expected reward of an arm $k \in \mathcal{K}$ is a (Hölder continuous, invertible) function of the global parameter θ_*, which is given by $E_{X_{k,t} \sim \nu_k(\theta_*)}(X_{k,t}) = \mu_k(\theta_*)$, where $\mu_k : \Theta \to [0,1]$ and $E[\cdot]$ denotes the expectation. Hence, the true expected reward of arm k is equal to $\mu_k(\theta_*)$.

Assumption 1. *(i) For each $k \in \mathcal{K}$, the reward function μ_k is invertible on $[0,1]$.*
(ii) For each $k \in \mathcal{K}$ and $y, y' \in \mathcal{Y}_k$, there exists $D_1 > 0$ and $0 < \gamma_1 \leq 1$ such that $|\mu_k^{-1}(y) - \mu_k^{-1}(y')| \leq D_1 |y - y'|^{\gamma_1}$, where μ_k^{-1} is the inverse reward function for arm k.
(iii) For each $k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$ there exists $D_2 > 0$ and $0 < \gamma_2 \leq 1$, such that $|\mu_k(\theta) - \mu_k(\theta')| \leq D_2 |\theta - \theta'|^{\gamma_2}$.
where $\mathcal{Y}_k \subseteq [0,1]$ is the range of $\mu_k(\cdot)$ on Θ.

Assumption 1 ensures that the reward obtained from an arm can be used to update the estimated expected rewards of the other arms. The last two conditions are Hölder conditions on the reward and inverse reward functions, which enable us to define the informativeness. It turns out that the invertibility of the reward functions is a crucial assumption that is required to achieve bounded regret. We illustrate this by a counter example when we discuss parameter dependent regret bounds.

There are many reward functions that satisfy Assumption 1. Examples include: (i) exponential functions such as $\mu_k(\theta) = a \exp(b\theta)$ for some $a > 0$, (ii) linear and piecewise linear functions, and (iii) sub-linear and super-linear functions in θ which are invertible in Θ such as $\mu_k(\theta) = a\theta^\gamma$ with $\gamma > 0$.

The goal of the learner is to choose a sequence of arms (one at each time) $\boldsymbol{I} := (I_1, \ldots, I_T)$ up to to time T to maximize its expected total reward. This corresponds to minimizing the regret which is the expected total loss due to not always selecting the optimal arm, i.e., the arm with the highest expected reward. Let $k^*(\theta_*) := \arg\max_{k \in \mathcal{K}} \mu_k(\theta_*)$ be the set of optimal arms and $\mu^*(\theta_*) := \max_{k \in \mathcal{K}} \mu_k(\theta_*)$ be the expected reward of the optimal arm for true value of global parameter θ_*. The cumulative regret of learning algorithm which selects arm I_t until time horizon T is defined as

$$\mathrm{Reg}(\theta_*, T) := \sum_{t=1}^{T} r_t(\theta_*),$$

where $r_t(\theta_*)$ is the one step regret given by $r_t(\theta_*) := \mu^*(\theta_*) - \mu_{I_t}(\theta_*)$ for global parameter θ_*. In the following sections we will derive regret bounds both as a function of θ_* (parameter-dependent regret) and independent from θ_* (worst-case or parameter-free regret).

2.2 Greedy Policy

In this section, we propose a *Greedy Policy* (GP) for the GMAB problem, which selects the arm with the highest estimated expected reward at each time t. Different from previous works in MABs [5, 18] in which the expected reward estimate of an arm only depends on the reward observations from that arm, the proposed GP uses a *global parameter estimate* $\hat{\theta}_t$ for the global parameter, which is given by $\hat{\theta}_t := \sum_{k=1}^{K} w_k(t) \hat{\theta}_{k,t}$, where $w_k(t)$ is the weight of arm k at time t and $\hat{\theta}_{k,t}$ is the estimate of the global parameter based only on the reward observations from arm k until time t. Let $\mathcal{X}_{k,t}$ denote the set of rewards obtained from the selections of arm k by time t, i.e., $\mathcal{X}_{k,t} = (X_{k,t})_{\tau < t \,|\, I_\tau = k}$ and $\hat{X}_{k,t}$ be the sample mean estimate of the rewards obtained from arm k by time t, i.e., $\hat{X}_{k,t} := \left(\sum_{x \in \mathcal{X}_{k,t}} x\right) / |\mathcal{X}_{k,t}|$. The proposed GP

[1] The set $[0,1]$ is just a convenient normalization. In general, we only need that distribution has a bounded support.

Input : μ_k for each $k \in \mathcal{K}$.
Initialization: $w_k(0) = 0, \hat{\theta}_{k,1} = 0, N_k(0) = 0$ for all $k \in \mathcal{K}$.
while $t \geq 1$ **do**
  if $t = 1$ **then**
    Randomly select arm I_t from the set $\mathcal{K}$
  else
    Select the arm $I_t \in \arg\max_{k\in\mathcal{K}} \mu_k\left(\hat{\theta}_{t-1}\right)$
  end if
  Observe the reward $X_{I_t,t}$
  $\hat{X}_{k,t} = \hat{X}_{k,t-1}$ for all $k \in \mathcal{K} \setminus I_t$
  $\hat{X}_{I_t,t} = \frac{\hat{X}_{I_t,t} + X_{I_t,t} N_k(t)}{N_k(t)+1}$
  if $X_{I_t,t} \in \mathcal{Y}_k$ **then**
    Update individual estimates for global parameter as $\hat{\theta}_{k,t} = \mu_k^{-1}\left(\hat{X}_{k,t}\right)$ for all $k \in \mathcal{K}$
  else
    Update individual estimates as $\hat{\theta}_{k,t} = \arg\min_{\theta\in\Theta} |\mu_k(\theta) - \hat{X}_{k,t}|$
  end if
  Update counters $N_{I_t}(t) = N_{I_t}(t-1) + 1$
  Update the rest $N_k(t) = N_k(t-1)$ for all $k \in \mathcal{K} \setminus I_t$
  Update weights $w_k(t) = \frac{N_k(t)}{t}$ for all $k \in \mathcal{K}$
  $\hat{\theta}_t = \sum_{k=1}^{K} w_k(t) \hat{\theta}_{k,t}$
end while

Figure 1: Pseudocode of the greedy policy.

operates as follows for any time $t \geq 2$: (i) the arm with highest expected reward by parameter estimate $\hat{\theta}_{t-1}$ is selected, i.e., $I_t \in \arg\max_{k\in\mathcal{K}} \mu_k\left(\hat{\theta}_{t-1}\right)$, (ii) reward $X_{I_t,t}$ is obtained and individual reward estimates $\hat{X}_{k,t}$ are updated for $k \in \mathcal{K}$, (iii) the individual estimate of the global parameter for each arm k is updated in the following way. If there is no global parameter $\theta \in \Theta$ such that $\mu_k\left(\hat{\theta}_{k,t}\right) = \hat{X}_{k,t}$ for arm k, then individual estimate for arm k is updated such a way that the gap between $\mu_k\left(\hat{\theta}_{k,t}\right)$ and $\hat{X}_{k,t}$ is minimized, i.e $\hat{\theta}_{k,t} = \arg\min_{\theta\in\Theta} |\mu_k(\theta) - \hat{X}_{k,t}|$. Otherwise, individual estimate of arm k is updated as $\hat{\theta}_{k,t} = \mu_k^{-1}\left(\hat{X}_{k,t}\right)$, (iv) the weight of each arm k is updated as $w_k(t) = N_k(t)/(t)$, where $N_k(t)$ is the number of times the arm k is played until time t. For $t = 1$, the GP selects randomly among the set of arms. The pseudocode of the GP is given in Fig. 1.

3 Regret Analysis for the GP

3.1 Definitions and Preliminary Results

In this subsection we define the tools that will be used in deriving the regret bounds. Consider any arm $k \in \mathcal{K}$. Its *optimality region* is defined as $\Theta_k := \{\theta \in \Theta \mid k \in k^*(\theta)\}$. Clearly, we have $\bigcup_{k\in\mathcal{K}} \Theta_k = \Theta$. If $\Theta_k = \emptyset$ for an arm k, this implies that there exists no global parameter values for which arm k is optimal. Since there exists an arm k' such that $\mu_{k'}(\theta) > \mu_k(\theta)$ for any $\theta \in \Theta$ for an arm with $\Theta_k = \emptyset$, the GP will discard arm k after $t = 1$. Therefore, without loss of generality we assume that $\Theta_k \neq \emptyset$ for all $k \in \mathcal{K}$. For global parameter $\theta_* \in \Theta$, we define the *suboptimality gap* of an arm $k \in \mathcal{K} \setminus k^*(\theta_*)$ as $\delta_k(\theta_*) := \mu^*(\theta_*) - \mu_k(\theta_*)$. For parameter θ_*, the minimum suboptimality gap is defined as $\delta_{\min}(\theta_*) := \min_{k\in\mathcal{K}\setminus k^*(\theta_*)} \delta_k(\theta_*)$.

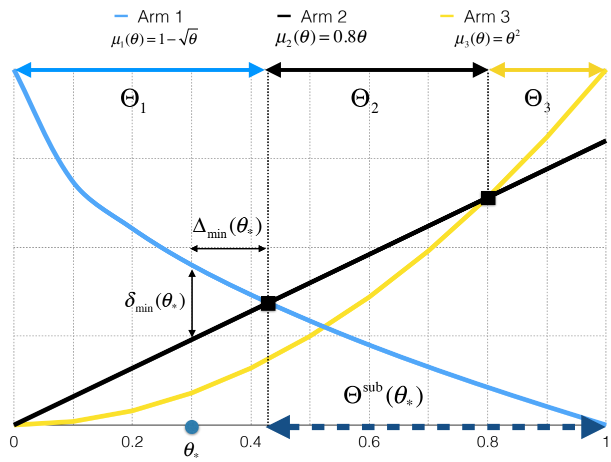

Figure 2: Illustration of minimum suboptimality gap and suboptimality distance

Recall that the expected reward estimate for arm k is equal to its expected reward corresponding to the global parameter estimate. We will show that as more arms are selected, the global parameter estimate will converge to the true value of the global parameter. However, if θ_* lies close to the boundary of the optimality region of $k^*(\theta_*)$, the global parameter estimate may fall outside of the optimality region of $k^*(\theta_*)$ for a large number of time steps, thereby resulting in a large regret. Let $\Theta^{\text{sub}}(\theta_*)$ be the suboptimality region for given global parameter θ_*, which is defined as the subset of parameter space in which an arm in the set $\mathcal{K} \setminus k^*(\theta_*)$ is optimal, i.e $\Theta^{\text{sub}}(\theta_*) = \cup_{k'\in\mathcal{K}\setminus k^*(\theta_*)} \Theta_{k'}$. In order to bound the expected number of such deviations from the optimality region, for any arm k we define a metric called the *suboptimality distance,* which is equal to the smallest distance between the value of the global parameter and suboptimality region.

Definition 1. *For a given global parameter θ_*, the suboptimality distance is defined as*

$$\Delta_{min}(\theta_*) := \begin{cases} \inf_{\theta'\in\Theta^{sub}(\theta_*)} |\theta_* - \theta'| & \text{if } \Theta^{sub}(\theta_*) \neq \emptyset \\ 1 & \text{if } \Theta^{sub}(\theta_*) = \emptyset \end{cases}$$

From the definition of the suboptimality distance it is evident that the GP always selects an optimal arm in $k^*(\theta_*)$ when $\hat{\theta}_t$ is within $\Delta_{\min}(\theta_*)$ of the global parameter θ_*. An illustration of suboptimality gap and suboptimality distance is given in Fig. 2 for a GMAB problem instance with 3 arms and reward functions $\mu_1(\theta) = 1 - \sqrt{\theta}$, $\mu_2(\theta) = 0.8\theta$ and $\mu_3(\theta) = \theta^2$.

In the following lemma, we show that minimum suboptimality distance is nonzero for any global parameter

θ_*. This result ensures that we can identify the optimal arm within finite amount of time.

Lemma 1. *Given any $\theta_* \in \Theta$, there exists a constant $\epsilon_{\theta_*} = \delta_{\min}(\theta_*)^{1/\gamma_2} / (2D_2)^{1/\gamma_2}$, where D_2 and γ_2 are the constants given in Assumption 1 such that $\Delta_{\min}(\theta_*) \geq \epsilon_{\theta_*}$. In other words, the minimum suboptimality distance is always positive.*

For notational brevity, we denote in the remainder of the paper $\Delta_{\min}(\theta_*)$ and $\delta_{\min}(\theta_*)$ as Δ_* and δ_*, respectively.

Lemma 2. *Consider a run of the GP until time t. Then, the following relation between $\hat{\theta}_t$ and θ_* holds with probability 1: $|\hat{\theta}_t - \theta_*| \leq \sum_{k=1}^{K} w_k(t) D_1 |\hat{X}_{k,t} - \mu_k(\theta_*)|^{\gamma_1}$*

Lemma 2 shows that the gap between the global parameter estimate and the true value of the global parameter is bounded by a weighted sum of the gaps between the estimated expected rewards and the true expected rewards of the arms.

Lemma 3. *For given global parameter θ_*, the one step regret of the GP is bounded by $r_t(\theta_*) = \mu^*(\theta_*) - \mu_{I_t}(\theta_*) \leq 2D_2|\theta_* - \hat{\theta}_t|^{\gamma_2}$ with probability 1, where I_t is the arm selected by the GP at time $t \geq 2$.*

Lemma 3 ensures that the one step loss decreases as $\hat{\theta}_t$ approaches to θ_*. Since the regret at time T is the sum of the one step losses up to time T, we will bound the regret by bounding the expected distance between $\hat{\theta}_t$ and θ_*.

Given a parameter value θ_*, let $\mathcal{G}^x_{\theta_*, \hat{\theta}_t} := \{|\theta_* - \hat{\theta}_t| > x\}$ be the event that the distance between the global parameter estimate and its true value exceeds x. Similarly, let $\mathcal{F}^k_{\theta_*, \hat{\theta}_t}(x) := \{|\hat{X}_{k,t} - \mu_k(\theta_*)| > x\}$ be the event that the distance between the sample mean reward estimate of arm k and the true expected reward of arm k exceeds x. The following lemma relates these events.

Lemma 4. *For any $t \geq 2$ and given global parameter θ_*, we have $\mathcal{G}^x_{\theta_*, \hat{\theta}_t} \subseteq \cup_{k=1}^{K} \mathcal{F}^k_{\theta_*, \hat{\theta}_t}\left(\left(\frac{x}{D_1}\right)^{\frac{1}{\gamma_1}}\right)$ with probability 1.*

This lemma follows from the decomposition given in Lemma 2. This lemma will be used to bound the probability of event $\mathcal{G}^x_{\theta_*, \hat{\theta}_t}$ in terms of probabilities of the events $\mathcal{F}^k_{\theta_*, \hat{\theta}_t}\left(\left(\frac{x}{D_1}\right)^{\frac{1}{\gamma_1}}\right)$.

3.2 Parameter-Free Regret Analysis

The following theorem bounds the expected regret of the GP in one step.

Theorem 1. *Under Assumption 1, for a given global parameter θ_*, the expected one-step regret of the GP is bounded by $E[r_t(\theta_*)] = O\left(t^{-\frac{\gamma_1\gamma_2}{2}}\right)$.*

Theorem 1 does not only prove that the expected loss incurred in one step by the GP goes to zero with time but also bounds the expected loss that will be incurred at any time step t.[2] This is a *worst-case* bound in the sense that it does not depend on θ_*. Using this result, we derive the parameter-free regret bound in the next theorem.

Theorem 2. *Under Assumption 1, for given global parameter θ, the parameter-free regret of the GP is bounded by $E[Reg(\theta_*, T)] = O\left(K^{\frac{\gamma_1\gamma_2}{2}} T^{1-\frac{\gamma_1\gamma_2}{2}}\right)$.*

Note that the parameter-free regret bound is sublinear both in terms of the time horizon T and the number of arms K. Moreover, it depends on the form of the reward functions given in Assumption 1. The Hölder exponent γ_1 on the inverse reward functions characterizes the informativeness of an arm about the other arms. The informativeness of an arm k can be viewed as the information obtained about the expected rewards of the other arms from the rewards observed from arm k. The informativeness is maximized for the case when the inverse reward functions are linear or piecewise linear, i.e., $\gamma_1 = 1$. It is increasing γ_1, which results in the regret decreasing with the informativeness. On the other hand, the Hölder exponent γ_2 is related to the loss due to suboptimal arm selections, which decreases with γ_2. Both of these observations follow from Lemma 2 and 3. As a consequence, the parameter-free regret is decreasing in both γ_1 and γ_2.

When the reward functions are linear or piecewise linear, we have $\gamma_1 = \gamma_2 = 1$; hence, the parameter-free regret is $O\left(\sqrt{T}\right)$, which matches with the worst-case regret bound of standard MAB algorithms in which a linear estimator is used [6] and bounds given for linearly parametrized bandits [20].

3.3 Parameter-Dependent Regret Analysis

Although the regret bound derived in the previous section holds for any global parameter value, it is easy to see that the performance of the GP depends on the true value of the global parameter. For example, it is easier to identify the optimal arm in GMAB problems with large suboptimality distance than GMAB problems which have small suboptimality distance. In this section, we prove a regret bound that depends on the suboptimality distance. Moreover, our regret bound is characterized by three regimes of growth: sublinear

[2] The asymptotic notation is only used for a succinct representation, to hide the constants and highlight the time dependence. This bound holds not just asymptotically but for any finite t.

growth followed by logarithmic growth followed by a constant bound.

The boundaries of these regimes are defined by parameter-dependent (problem-specific) constants.

Definition 2. *Let $C_1(\Delta_*)$ be the least integer τ which satisfied $\tau \geq \frac{D_1^{\frac{2}{\gamma_1}} K}{2\Delta_*^{\frac{2}{\gamma_1}}} \log(\tau)$ and let $C_2(\Delta_*)$ be the least integer τ which satisfied $\tau \geq \frac{D_1^{\frac{2}{\gamma_1}} K}{\Delta_*^{\frac{2}{\gamma_1}}} \log(\tau)$.*

The constants $C_1(\Delta_*)$ and $C_2(\Delta_*)$ depend on the informativeness (Hölder exponent γ_1) and global parameter θ_*. We define the expected regret between time T_1 and T_2 for global parameter θ_* as

$$R_{\theta_*}(T_1, T_2) := \sum_{T_1}^{T_2} E\left[r_t(\theta_*)\right].$$

The following theorem gives a three regime parameter-dependent regret bound.

Theorem 3. *Under Assumptions 1, the regret of the GP is bounded as follows:*
(i) If $1 \leq T \leq C_1(\Delta_*)$, the regret is sublinear in time, i.e.,

$$R_{\theta_*}(0, T) = O\left(T^{1-\frac{\gamma_1\gamma_2}{2}}\right),$$

(ii) If $C_1(\Delta_*) \leq T \leq C_2(\Delta_*)$, the regret is logarithmic in time, i.e.,

$$R_{\theta_*}(C_1(\Delta_*, T)) \leq 1 + 2K \log\left(\frac{T}{C_1(\Delta_*)}\right),$$

(iii) If $T \geq C_2(\Delta_*)$, the regret is bounded, i.e.,

$$R_{\theta_*}(C_2(\Delta_*), T) \leq K\frac{\pi^2}{3}.$$

Corollary 1. *The regret of the GP is bounded, i.e., $\lim_{T\to\infty} Reg(T, \theta_*) < \infty$.*

These results are obtained when Assumption 1 holds, which implies that the reward functions are invertible. We provide a counter example for a non-invertible reward function to show that bounded regret is not possible for general non-invertible reward functions. In each time $t \leq T$ in each regime in Theorem 3, the probability of selecting a suboptimal arm is bounded by different functions of t, which leads to different growth rates of the regret bound depending on the value of T. For instance, when $C_1(\Delta_*) \leq t \leq C_2(\Delta_*)$, the probability of selecting a suboptimal arm is in the order of t^{-1}; hence, the GP achieves the logarithmic regret, when $t \geq C_2(\Delta_*)$, the probability of selecting a suboptimal arm is in the order of t^{-2}, which makes the probability of selecting a suboptimal arm infinitely often zero. In conclusion, the GP achieves bounded regret. Note that a bounded regret is the striking difference between the standard MAB algorithms [5, 18] and the proposed policy.

Theorem 4. *The sequence of arms selected by the GP converges to the optimal arm almost surely, i.e., $\lim_{t\to\infty} I_t = k^*(\theta_*)$ with probability 1.*

Theorem 4 implies that a suboptimal arm is selected by GP only finitely many times. In other words, there exists a finite number such that selection of GP is the optimal arm after that number with probability 1. This is the biggest difference between MAB algorithms [5, 18] in which suboptimal arms are selected infinitely many times and the proposed GP.

Although the parameter dependent regret bound is finite, since $\lim_{\Delta_*\to 0} C_1(\Delta_*) = \infty$, in the worst-case, this bound reduces to the parameter-free regret bound given in Theorem 2.

4 Bayesian Risk Analysis of the GP

In this section, assuming that global parameter is drawn from an unknown distribution $f(\theta_*)$ on Θ, we provide an analysis of the Bayesian risk, which is defined as follows:

$$\mathrm{Risk}(T) = E_{\theta_*\sim f(\theta_*)}\left[E_{\boldsymbol{X}_t\sim\boldsymbol{\nu}}\left[\sum_{t=1}^{T} r_t(\theta)\,|\theta_* = \theta\right]\right],$$

$\boldsymbol{\nu} = \times_{k=1}^{K} \nu_k(\theta_*)$ is the joint distribution of the rewards given the parameter value is θ_*. The Bayesian risk is equal to the expected regret with respect to the distribution of the global parameter $f(\theta_*)$. Since suboptimality distance is a function of global parameter θ_*, there is a prior distribution on the minimum sub optimality distance, which we denote as $g(\Delta_*)$. A simple upper bound on the Bayesian risk can be obtained by taking the expectation of the regret bound given in Theorem 2 with respect to θ_*, which gives the bound $\mathrm{Risk}(T) = O\left(T^{1-\frac{\gamma_1\gamma_2}{2}}\right)$. Next, we will show that a tighter regret bound on the Bayesian risk can be derived if the following assumption holds.

Assumption 2. *The prior distribution on the global parameter is such that minimum suboptimality distance Δ_* has a bounded density function, i.e., $g(\Delta_*) \leq B$. One example of this is the case when $f(\theta_*)$ is bounded.*

Assumption 2 is satisfied for many instances of the GMAB problem. An example is a GMAB problem with two arms, $f(\theta_*) \sim \mathrm{Uniform}([0, 1])$, $\mu_1(\theta_*) = \theta_*$ and $\mu_2(\theta_*) = 1 - \theta_*$. For this example we have $g(\Delta_*) \leq 2$ for $\Delta_* \in [0, 0.5]$.

Theorem 5. *Under Assumptions 1 and 2, the Bayesian risk of the greedy policy is bounded by*
(i) $Risk(T) = O(\log T)$, for $\gamma_1\gamma_2 = 1$.
(ii) $Risk(T) = O\left(T^{1-\gamma_1\gamma_2}\right)$, for $\gamma_1\gamma_2 < 1$.

Our Bayesian risk bound for the GP coincides with the Bayesian risk bound for the linearly-parametrized

MAB problem given in [20] when the arms are fully informative, i.e., $\gamma_1\gamma_2 = 1$. For this case, the optimality of the Bayesian risk bound is established in [20], in which a lower bound of $\Omega(\log T)$ is proven. Similar to the parameter-free regret bound given in Theorem 2, the Bayesian risk is also decreasing with the informativeness, and minimized for the case when the arms are fully informative.

5 Concluding Remarks and Future Work

In this paper we introduce a new class of MAB problems called global multi-armed bandits. This general class of GMAB problems encompasses the previously introduced linearly-parametrized bandits as a special case. We proved that the regret for the GMABs has three regimes, which we characterized for the regret bound, and showed that the parameter-dependent regret is bounded, i.e., it is asymptotically finite. In addition to this, we also proved a parameter-free regret bound and a Bayesian risk bound, both of which grow sublinearly over time, where the rate of growth depends on the informativeness of the arms.

Future work includes extension of global informativeness to group informativeness. Our global informativeness assumption can be relaxed to *group informativeness*. When the arms are group informative, reward observations from an arm only provides information about the rewards of the arms that are within the same group with the original arm. Another extension of our work is the relaxation of invertibility assumption of reward functions. We can extend it to a setting where the reward functions are known to lie in an epsilon neighborhood of the exact reward function. Such a neighborhood can be constructed based on prior knowledge and/or data. For instance, this prior knowledge can come from other learners that used our algorithm in environments with different theta values for a given time horizon. Then, we can show that if the difference between the estimated reward functions and the exact reward functions is small enough across the (global) parameter space, the GP can still achieve a bounded regret.

6 Proofs

In this section, we provide the proofs of theorems. The proofs of lemmas are given in the supplementary material. Let $\boldsymbol{w}(t) := (w_1(t), \dots, w_K(t))$ be the vector of weights and $\boldsymbol{N}(t) := (N_1(t), \dots, N_k(t))$ be the vector of counters at time t. We have $\boldsymbol{w}(t) = \frac{1}{t}\boldsymbol{N}(t)$. Since $\boldsymbol{N}(t)$ depends on the history, they are both random variables depending on the obtained rewards.

6.1 Proof of Theorem 1

By lemma 3 and Jensen's inequality, we have

$$E[r_t(\theta_*)] \leq 2D_2 E\left[|\theta_* - \hat{\theta}_t|\right]^{\gamma_2}. \quad (1)$$

By using Lemma 2 and Jensen's inequality, we have

$$E\left[|\theta_* - \hat{\theta}_t|\right] \leq$$
$$D_1 E\left[\sum_{k=1}^{K} w_k(t)\, E\left[|\hat{X}_{k,t} - \mu_k(\theta_*)| \,|\boldsymbol{w}(t)\right]^{\gamma_1}\right], \quad (2)$$

where $E[\cdot|\cdot]$ denotes the conditional expectation. Note that $\hat{X}_{k,t} = \frac{\sum_{x\in\mathcal{X}_{k,t}} x}{N_k(t)}$ and $E_{x\sim\nu_k(\theta_*)}[x] = \mu_k(\theta_*)$. Therefore, we can bound $E\left[|\hat{X}_{k,t} - \mu_k(\theta_*)| \,|\boldsymbol{w}(t)\right]$ for each $k \in \mathcal{K}$ using Chernoff- Hoeffding inequality. For each $k \in \mathcal{K}$, we have

$$\begin{aligned} &E\left[|\hat{X}_{k,t} - \mu_k(\theta_*)| \,|\boldsymbol{w}(t)\right] \\ &= \int_{x=0}^{1} \Pr\left(|\hat{X}_{k,t} - \mu_k(\theta_*)| > x|\boldsymbol{w}(t)\right) \mathrm{d}x \\ &\leq \int_{x=0}^{\infty} 2\exp\left(-2x^2 N_k(t)\right) \mathrm{d}x \leq \sqrt{\frac{\pi}{2N_k(t)}}, \quad (3) \end{aligned}$$

where $N_k(t) = t w_k(t)$ is a random variable. The first inequality is a result of the Chernoff-Hoeffding bound. Combining (2) and (3), we get

$$E\left[|\theta_* - \hat{\theta}_t|\right] \leq D_1 \left(\frac{\pi}{2t}\right)^{\frac{\gamma_1}{2}} E\left[\sum_{k=1}^{K} w_k(t)^{1-\frac{\gamma_1}{2}}\right]. \quad (4)$$

Since $w_k(t) \leq 1$ for all $k \in \mathcal{K}$, and $\sum_{k=1}^{K} w_k(t) = 1$ for any possible $\boldsymbol{w}(t)$, we have $E\left[\sum_{k=1}^{K} w_k(t)^{1-\frac{\gamma_1}{2}}\right] \leq K^{\frac{\gamma_1}{2}}$. Then, combining (1) and (4), we have

$$E[r_t(\theta_*)] \leq D_1^{\gamma_2} D_2 \frac{\pi}{2}^{\frac{\gamma_1\gamma_2}{2}} K^{\frac{\gamma_1\gamma_2}{2}} \frac{1}{t^{\frac{\gamma_1\gamma_2}{2}}}.$$

6.2 Proof of Theorem 2

The bound is consequence of Theorem 1 and inequality given in [10], i.e.,

$$E[\mathrm{Reg}(\theta_*, T)] \leq 1 + \frac{D_1^{\gamma_2} D_2 \frac{\pi}{2}^{\frac{\gamma_1\gamma_2}{2}} K^{\frac{\gamma_1\gamma_2}{2}}}{1 - \frac{\gamma_1\gamma_2}{2}} \left(1 + T^{1-\frac{\gamma_1\gamma_2}{2}}\right).$$

6.3 Proof of Theorem 3

We need to bound the probability of the event that $I_t \neq k^*(\theta_*)$. Since at time t, the arm with the highest $\mu_k\left(\hat{\theta}_t\right)$ is selected by the GP, $\hat{\theta}_t$ should lie in $\Theta \setminus \Theta_{k^*(\theta_*)}$ for GP to select a suboptimal arm. Therefore, we can write,

$$\{I_t \neq k^*(\theta_*)\} = \{\hat{\theta}_t \in \Theta \setminus \Theta_{k^*(\theta_*)}\} \subseteq \mathcal{G}^{\Delta_*}_{\theta_*, \hat{\theta}_t}. \quad (5)$$

By Lemma 4 and (5), we have

$$\begin{aligned}
&\Pr\left(I_t \neq k^*\left(\theta_*\right)\right) \\
&\leq \sum_{k=1}^{K} E\left[E\left[I\left(\mathcal{F}^k_{\theta_*,\hat{\theta}_t}\left(\left(\frac{x}{D_1}\right)^{\frac{1}{\gamma_1}}\right)\right)|\boldsymbol{N}(t)\right]\right] \\
&= \sum_{k=1}^{K} E\left[\Pr\left(\mathcal{F}^k_{\theta_*,\hat{\theta}_t}\left(\left(\frac{x}{D_1}\right)^{\frac{1}{\gamma_1}}\right)|\boldsymbol{N}(t)\right)\right] \\
&\leq \sum_{k=1}^{K} 2E\left[\exp\left(-2\left(\frac{\Delta_*}{D_1}\right)^{\frac{2}{\gamma_1}} N_k(t)\right)\right] \\
&\leq 2K\exp\left(-2\left(\frac{\Delta_*}{D_1}\right)^{\frac{2}{\gamma_1}}\frac{t}{K}\right), \qquad (6)
\end{aligned}$$

where the first inequality is followed by union bound and second inequality is obtained by using the Chernoff-Hoeffding bound. The last inequality is obtained by using the worst-case selection processes $N_k(t) = \frac{t}{K}$. We have $\Pr\left(I_t \neq k^*\left(\theta_*\right)\right) \leq \frac{1}{t}$ for $t > C_1\left(\Delta_*\right)$ and $\Pr\left(I_t \neq k^*\left(\theta_*\right)\right) \leq \frac{1}{t^2}$ for $t > C_2\left(\Delta_*\right)$. The bound in the first regime is the result of Theorem 2. The bound in the second and third regimes is obtained by summing the probability given in (6) from $C_1\left(\Delta_*\right)$ to T and $C_2\left(\Delta_*\right)$ to T, respectively.

6.4 Proof of Theorem 4

Let $(\Omega, \mathcal{F}, P)$ denote probability space, where Ω is the sample set and $\mathcal{F}$ is the σ-algebra that the probability measure Pr is defined on. Let $\omega \in \Omega$ denote a sample path. We will prove that there exists event $N \in \mathcal{F}$ such that $\Pr(N) = 0$ and if $\omega \in N^c$, then $\lim_{t\to\infty} I_t(\omega) = k^*(\theta_*)$. Define the event $\mathcal{E}_t := \{I_t \neq k^*(\theta_*)\}$. We show in the proof of Theorem 3 that $\sum_{t=1}^{T}\Pr(\mathcal{E}_t) < \infty$. By Borel-Cantelli lemma, we have

$$\Pr\left(\mathcal{E}_t \text{ infintely often}\right) = \Pr\left(\limsup_{t\to\infty}\mathcal{E}_t\right) = 0.$$

Define $N := \limsup_{t\to\infty}\mathcal{E}_t$, where $\Pr(N) = 0$. We have,

$$N^{\mathrm{c}} = \liminf_{t\to\infty}\mathcal{E}^{\mathrm{c}}_t,$$

, where $\Pr(N^{\mathrm{c}}) = 1 - \Pr(N) = 1$, which means that $I_t = k^*(\theta_*)$ for all t except for a finite number.

6.5 Proof of Theorem 5

Proof. The one step loss due to suboptimal arm selection with global parameter estimate $\hat{\theta}_t$ is given in Lemma 3. Recall that we have

$$\{I_t \neq k^*(\theta_*)\} \subseteq \{|\theta_* - \hat{\theta}_t| > \Delta_*\}.$$

Let $Y_{\theta_*,\hat{\theta}_t} := |\theta_* - \hat{\theta}_t|$. Then, we have

$$\begin{aligned}
&\mathrm{Risk}(T) \\
&\leq 2D_2\sum_{t=1}^{T} E_{\theta_*\sim f(\theta)}\left[E_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[Y^{\gamma_2}_{\theta_*,\hat{\theta}_t} I\left(Y_{\theta_*,\hat{\theta}_t} > \Delta_*\right)\right]\right] \\
&\leq 2D_2\sum_{t=1}^{T} E_{\theta_*\sim f(\theta)}\left[E_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[Y_{\theta_*,\hat{\theta}_t} I\left(Y_{\theta_*,\hat{\theta}_t} > \Delta_*\right)\right]\right]^{\gamma_2},
\end{aligned}$$

, where $I(.)$ is the indicator function which is 1 if the statement is true and zero otherwise. The first inequality followed by Lemma 2. The second inequality is by Jensen's inequality and the fact that $I(.) = I^{\gamma}(.)$ for any $\gamma > 0$. We now focus on the expectation expression for some arbitrary t. Let $f(\theta)$ denote the density function of global parameter.

$$\begin{aligned}
&E_{\theta_*\sim f(\theta)}\left[E_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[Y_{\theta_*,\hat{\theta}_t} I\left(Y_{\theta_*,\hat{\theta}_t} > \Delta_*\right)\right]\right] \\
&= \int_{\theta_*=0}^{1} f(\theta_*)\int_{x=0}^{\infty}\Pr\left(Y_{\theta_*,\hat{\theta}_t} I\left(Y_{\theta_*,\hat{\theta}_t} > \Delta_*\right) \geq x\right)\mathrm{d}x\,\mathrm{d}\theta \\
&= \int_{\theta_*=0}^{1} f(\theta_*)\int_{x=\Delta_*}^{\infty}\Pr\left(Y_{\theta_*,\hat{\theta}_t} \geq x\right)\mathrm{d}x\,\mathrm{d}\theta \\
&= \int_{\Delta=0}^{1} g(\Delta)\int_{x=\Delta}^{\infty}\Pr\left(Y_{\theta_*,\hat{\theta}_t} \geq x\right)\mathrm{d}x\,\mathrm{d}\Delta,
\end{aligned}$$

where the last equation is followed by change of variables in integral. Note that we have by Theorem 3

$$\Pr\left(Y_{\theta_*,\hat{\theta}_t} \geq x\right) \leq 2K\exp\left(-2x^{\frac{2}{\gamma_1}} D_1^{-\frac{2}{\gamma_1}}\frac{t}{K}\right).$$

Then, we have

$$\begin{aligned}
&E_{\theta\sim f(\theta)}\left[E_{\boldsymbol{X}\sim\boldsymbol{\nu}}\left[Y_{\theta_*,\hat{\theta}_t} I\left(Y_{\theta_*,\hat{\theta}_t} > \Delta_*\right)\right]\right] \\
&\leq 2KB\int_{\Delta=0}^{1}\exp\left(-2\Delta^{\frac{2}{\gamma_1}} D_1^{-\frac{2}{\gamma_1}}\frac{t}{K}\right)\mathrm{d}\Delta \\
&\int_{y=0}^{\infty}\exp\left(-2y^{\frac{2}{\gamma_1}} D_1^{-\frac{2}{\gamma_1}}\frac{t}{K}\right)\mathrm{d}y \\
&= 2KB\left(\frac{\gamma_1}{2}2^{-\frac{\gamma_1}{2}} D_1 K^{\frac{\gamma_1}{2}}\Gamma\left(\frac{\gamma_1}{2}\right)\right)^2 t^{-\gamma_1},
\end{aligned}$$

where the inequality follows from the change of variable $y = x - \Delta$ and then the fact that $(y+\Delta)^{\frac{2}{\gamma_1}} \geq y^{\frac{2}{\gamma_1}} + \Delta^{\frac{2}{\gamma_1}}$ since $\frac{2}{\gamma_1} \geq 1$. By summing these from 1 to T, we get

$$\mathrm{Risk}(T) \leq \begin{cases} 1 + A\left(1 + 2\log T\right) & \text{if } \gamma_1\gamma_2 = 1 \\ 1 + A\left(1 + \frac{1}{1-\gamma_1\gamma_2}T^{1-\gamma_1\gamma_2}\right) & \text{if } \gamma_1\gamma_2 < 1 \end{cases}$$

where $A = 2D_2\left(\frac{B\gamma_1^2 D_1^2 K^{1+\gamma_1}}{2^{1+\gamma_1}}\Gamma^2\left(\frac{\gamma_1}{2}\right)\right)$. $\square$

Acknowlegment

This work is partially supported by a DDDAS Airforce grant.

References

[1] R. Agrawal, D. Teneketzis, and V. Anantharam. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: finite parameter space. *IEEE Transactions on Automatic Control*, pages 258–267, 1989.

[2] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory (COLT)*, 2012.

[3] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *In Proceedings of Thirtieth International Conference on Machine Learning (ICML)*, 2013.

[4] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Rsearch*, pages 397–422, 2002.

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

[6] S. Bubeck and N. Cesa Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 2012.

[7] S. Bubeck and C. Y. Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, pages 638–646. 2013.

[8] Cesa-Bianchi, N., Y. Freund, D. P. Helmhold, D. Haussler, R. E. Schapire, and M. K. Warmuth. How to use expert advice. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on the Theory of Computing*, pages 382–391, 1993.

[9] Y. Chen and V. Farias. Simple policies for dynamic pricing with imperfect forecasts. *Operations Research*, 2013.

[10] E. Chlebus. An approximate formula for a partial sum of divergent p-series. *Applied Mathematics Letters*, 22(5):732–737, 2009.

[11] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvari. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems (NIPS)*, pages 586–594, 2011.

[12] Y. Gai, B. Krishnamachari, and R. Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, 20(5):1466–1478, 2012.

[13] A. Garivier and O. Cappe. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory (COLT)*, 2011.

[14] J. C. Gittins. Multi-armed bandit allocation indices. *Wiley Interscience Series in Systems and Optimization. John Wiley and Son*, 1989.

[15] E. Kaufmann, C. O., and A. Garivier. On bayesian upper confidence bounds for bandit problems. In *Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[16] N. Korda, E. Kaufmann, and R. E., Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Neural Information Processing Systems (NIPS)*, 2013.

[17] T. Lai and H. Robbins. Adaptive design in regression and control. In *Proceedings of the National Academy of Sciences*, 1978.

[18] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math*, 6, 1985.

[19] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1096–1023, 2008.

[20] A. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. *IEEE Trans. Automatic Control*, 54:2787–2802, 2009.

[21] P. Rusmevichientong and J. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 5:395–411, 2010.

[22] A. Slivkins. Contextual bandits with similarity information. In *Journal of Machine Learning Research*, 2014.

[23] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.