Global Bandits

Onur Atan*, Cem Tekin†, Mihaela van der Schaar*
*Department of Electrical Engineering, University of California, Los Angeles, oatan@ucla.edu, mihaela@ee.ucla.edu
† Department of Electrical and Electronics Engineering, Bilkent University, cemtekin@ee.bilkent.edu.tr

***Abstract*—Multi-armed bandits (MAB) model sequential decision making problems, in which a learner sequentially chooses arms with unknown reward distributions in order to maximize its cumulative reward. Most of the prior work on MAB assumes that the reward distributions of each arm are independent. But in a wide variety of decision problems – from drug dosage to dynamic pricing – the expected rewards of different arms are correlated, so that selecting one arm provides information about the expected rewards of other arms as well. We propose and analyze a class of models of such decision problems, which we call *global bandits*. In the case in which rewards of all arms are deterministic functions of a single unknown parameter, we construct a greedy policy that achieves *bounded regret*, with a bound that depends on the single true parameter of the problem. Hence, this policy selects suboptimal arms only finitely many times with probability one. For this case we also obtain a bound on regret that is *independent of the true parameter*; this bound is sub-linear, with an exponent that depends on the informativeness of the arms (which measures the strength of correlation between expected arm rewards). We also prove matching lower bounds for the worst-case regret of the greedy policy.**

***Index Terms*—Online learning, multi-armed bandits, regret analysis, bounded regret, Bayesian risk, informative arms.**

I. Introduction

Multi-armed bandits (MAB) provide powerful models and algorithms for sequential decision making problems in which the expected reward of each arm (action) is unknown. The goal in MAB problems is to design online learning algorithms which minimize the regret, where the regret is defined as the difference between the total expected reward obtained by an oracle that perfectly knows the expected arm rewards and the total expected reward obtained by the learning algorithm. Standard finite-armed MAB [2] does not impose any dependence between the expected arm rewards. But in a wide variety of decision problems – from drug dosage to dynamic pricing – the expected rewards of different arms are correlated, so that selecting one arm provides information about the expected rewards of other arms as well. In this paper we propose and analyze such a MAB model, which we call *Global Bandits* (GB).

In GB, the expected reward of each arm is a function of a single global parameter. It is assumed that the learner knows these functions but does not know the true value of the parameter. We propose a greedy policy, which constructs an estimate of the global parameter by taking a weighted average of parameter estimates computed separately from the reward observations of each arm. Then, we show that this policy achieves *bounded regret*, where the bound depends on the value of the parameter. This implies that the greedy policy learns the optimal arm, i.e., the arm with the highest expected reward, in finite time. We also obtain a worst-case (parameter independent) bound on the regret of the greedy policy. We show that this bound is sub-linear in time, and its time exponent depends on the *informativeness of the arms*, which is a measure of the strength of correlation between expected arm rewards.

GBs encompass the model studied in [3], in which it is assumed that the expected reward of each arm is a *linear function* of a single global parameter. This is a special case of the more general model we consider in this paper, in which the expected reward of each arm is a Hölder continuous, possibly non-linear function of a single global parameter. On the technical side, non-linear expected reward functions significantly complicates the learning problem. When the expected reward functions are linear, then the information one can infer about the expected reward of arm X by an additional single sample of the reward from arm Y is independent of the history of previous samples from arm Y.[1] However, if reward functions are non-linear, then the additional information that can be inferred about the expected reward of arm X by a single sample of the reward from arm Y is biased. Therefore, the previous samples from arm X and arm Y needs to be incorporated in a special way to ensure that this bias asymptotically converges to 0.

Many applications can be formalized as GBs. Examples include: (i) clinical trials involving similar drugs (e.g., drugs with a similar chemical composition) or treatments which may have similar effects on the patients; (ii) dynamic pricing with the objective of maximizing its revenue over a finite time horizon.

Example 1: Let x_i be the dosage level of the drug for patient i and y_i be the response of patient i. The relationship between the drug dosage and patient response is modeled in [4] as $y_i = M(x_i;\theta_*) - c(x_i) + \epsilon_i$, where $M(\cdot)$ is the response function, θ_* is the slope if the function is linear or the elasticity if the function is exponential or logistic, and $c(x_i)$ is the cost of the dosage level of drug and ϵ_i is i.i.d. zero mean noise. For this model, θ_* becomes the global parameter and the set of drug dosage levels becomes the set of arms.

A preliminary version of this paper is published in AISTATS 2015 [1].

[1]The additional information about the expected reward of arm X that can be inferred from obtaining sample reward r from arm Y is the same as the additional information about about the expected reward of arm X that could be inferred from obtaining the sample reward $L(r)$ from arm X itself, where L is a linear function that depends only on the reward functions themselves.

Example 2: In dynamic pricing, an agent sequentially selects a price from a finite set of prices $\mathcal{P}$ with the objective of maximizing its revenue over a finite time horizon [5]. When the agent first selects a price $p \in \mathcal{P}$, and then observes the amount of sales at time t, we assume that sales $S_{p,t}(\theta)$ is given by

$$S_{p,t}(\theta) = \bar{F}_p(\theta) + \epsilon_t,$$

where $\bar{F}(.)$ is the modulating function, θ is the market size 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 θ. The revenue is then given by $R_{p,t} = p\bar{F}_p(\theta) + p\epsilon_t$. In this example, the market size is the global parameter; this is unknown and needs to be learned by setting prices and observing the revenue related to the set price. Commonly used modulating functions are included in [6]. In Section 5, we illustrate the proposed method on this pricing example.

In addition to the above examples, GBs can also be applied in any setting in which the non-linear parameters of a system need to be estimated in order to take optimal arms. At this point, it is important to note that our work differs from the existing works on non-linear parameter estimation [7]–[9] because its focus is on using these estimates of the environment (i.e. the non-linear parameter) in order to determine what arms to take/decisions to make in an online manner.

In summary, the main contributions of our paper are:

- We propose a non-linear parametric model for MABs, which we refer to as GBs, and a greedy policy, referred to as *Weighted Arm Greedy Policy* (WAGP), which achieves bounded regret.
- We provide different regimes of growth for the finite time analysis of the regret. We define two thresholds and show that the regret increases at most sub-linearly over time until the first threshold and at most logarithmically over time starting from the first threshold until the second threshold.
- We define the concept of *informativeness* of the arms and prove a sub-linear in time worst-case regret bound for WAGP that depends on informativeness. We show that the regret increases slowly when the informativeness is high. Moreover, we also provide a matching lower bound for the worst-case regret and show that the worst-case regret of policies that rely on global parameter estimates can be worse than the worst-case regret of upper confidence bound (UCB) policies [10].
- To alleviate the above problem, we propose another learning algorithm called the *Best of UCB and WAGP (BUW)*, which fuses the decisions of the UCB1 [10] and WAGP in order to achieve both $\mathcal{O}(\sqrt{T})$ worst-case regret and $\mathcal{O}(1)$ parameter dependent regret.
- We also study a non-stationary version of GB, where the value of the global parameter slowly changes over time. For this case, we prove a bound on the time-averaged regret, which depends on the speed of change of the global parameter.
- We simulate our algorithms on a synthetic dynamic pricing dataset and show that they beat other state-of-art MAB algorithms.

This paper is extended version of [1], adding the following contributions. First, it provides two new theoretical results on the main setting, stating mean-squared convergence of the global parameter and a lower bound on the regret. Second, it provides two new algorithms: (i) Best of WAGP and UCB switches between the UCB and WAGP in order to achieve optimal regret bounds, (ii) non-stationary WAGP, that tracks the parameter over time and makes optimal decisions. Third, it provides numerical results on a simulated data inspired by the dynamic pricing example. In addition to these contributions, this manuscript has extended introduction, related work and all proofs of theoretical results.

The remainder of the paper is organized as follows. Related work is discussed in Section II. Problem formulation and the description of the greedy policy is given in Section III.Parameter dependent and worst-case regret bounds for this policy are also proved in Section III. In Section IV, we provide two interesting extensions to the WAGP: the BUW which achieves both $\mathcal{O}(\sqrt{T})$ worst-case regret and $\mathcal{O}(1)$ parameter dependent regret, and the modified WAGP with a time-varying global parameter. Numerical results are given in Section V, followed by the concluding remarks given in Section VI.

II. Related Work

There is a wide strand of literature on MABs including finite armed stochastic MAB [2], [10]–[12], Bayesian MAB [13]–[17], contextual MAB [18]–[20] and distributed MAB [21]–[23]. Depending on the extent of informativeness of the arms, MABs can be categorized into three: non-informative, group informative and globally informative MABs.

A. Non-informative MAB

We call a MAB *non-informative* if the reward observations of any arm do not reveal any information about the rewards of the other arms. Example of non-informative MABs include finite armed stochastic [2], [10] and non-stochastic [24] MABs. Lower bounds derived for these settings point out to the impossibility of bounded regret.

B. Group-informative MAB

We call a MAB *group-informative* if the reward observations from an arm provides information about a group of other arms. Examples include linear contextual bandits [25], [26], multi-dimensional linear bandits [27]–[31] and combinatorial bandits [32], [33]. In these works, the regret is sublinear in time and in the number of arms. For example, [27] assumes a reward structure that is linear in an unknown parameter and shows a regret bound that scales linearly with the dimension of the parameter. It is not possible to achieve bounded regret in any of the above settings since multiple arms are required to be selected logarithmically many times in order to learn the unknown parameters.

Another related work [34] studies a setting that interpolates between the bandit and experts settings. In this setting, the decision-maker obtains not only the reward of the selected arm but also an unbiased estimate of the rewards of a subset

of the other arms where the feedback structure is modeled by a directed graph. However, in our setting, it is not possible to construct an unbiased estimate of the other arms because of the non-linear reward structure.

C. Globally-informative MAB

We call a MAB problem *globally-informative* if the reward observations from an arm provide information about the rewards of *all* the arms [3], [35]. GB belongs to the class of globally-informative MAB and includes the linearly-parametrized MAB [3] as a subclass. Hence, our results reduce to the results of [3] for the special case when expected arm rewards are linear in the parameter.

A related work that falls into this setting is [36], in which the authors prove regret bounds that depend on the learner's uncertainty about the optimal arm. This uncertainty depends on the learner's prior knowledge and prior observations, and affect the constant factors that contribute to the $\mathcal{O}(\sqrt{T})$ regret bound. Whereas, in our problem formulation, we show that the strong dependence of the arms result in a bounded parameter dependent and a sub-linear worst-case regret, whose time order depends on the informativeness of the arms.

Another related learning scenario is the experts setting [37], where after an arm is chosen, the rewards of all arms are observed. Since there is no tradeoff between exploration and exploitation in this setting, finite regret bounds can be achieved when the number of arms is finite and the arm rewards are stochastic. Unlike the experts setting, in GB, finite regret is achievable by only observing the reward of the selected arms.

III. Problem Formulation and Greedy Policy

A. Problem Formulation

There are K arms indexed by the set $\mathcal{K} := \{1, \dots, K\}$. The unknown single-dimensional global parameter is denoted by θ_*, which belongs to the parameter set Θ that is taken to be the unit interval for simplicity of exposition. The random variable $X_{k,t}$ denotes the reward of arm k at time t. $X_{k,t}$ is drawn independently from the other arms from an unknown distribution $\nu_k(\theta_*)$ with support $[0,1]$ for all $t \geq 1$ and $k \in \mathcal{K}$. The expected reward of an arm $k \in \mathcal{K}$ is a Hölder continuous, invertible function of θ_*, which is given by $\mu_k(\theta_*) := \mathrm{E}_{\nu_k(\theta_*)}[X_{k,t}]$, where $\mathrm{E}_\nu[\cdot]$ denotes the expectation taken with respect to distribution ν.

Assumption 1. *(i) For each $k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$ there exists $D_{1,k} > 0$ and $1 < \gamma_{1,k}$, such that*

$$|\mu_k(\theta) - \mu_k(\theta')| \geq D_{1,k}|\theta - \theta'|^{\gamma_{1,k}}.$$

(ii) For each $k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$ there exists $D_{2,k} > 0$ and $0 < \gamma_{2,k} \leq 1$, such that

$$|\mu_k(\theta) - \mu_k(\theta')| \leq D_{2,k}|\theta - \theta'|^{\gamma_{2,k}}.$$

The first assumption is known as Hölder continuity, which ensures that the reward functions are smooth and second assumption ensures that the reward functions are monotonic. The next proposition shows that these assumptions naturally imply that the reward functions are invertible and inverse reward functions are Hölder continuous.

Proposition 1. *Under Assumption 1, the following are true: (i) For each $k \in \mathcal{K}$, the reward functions $\mu_k(\cdot)$ are invertible. (ii) For each $k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$ and $y, y' \in [0,1]$,*

$$|\mu_k^{-1}(y) - \mu_k^{-1}(y')| \leq \bar{D}_{1,k}|y - y'|^{\bar{\gamma}_{1,k}}$$

where $\bar{\gamma}_{1,k} = \frac{1}{\gamma_{1,k}}$ and $\bar{D}_{1,k} = \left(\frac{1}{D_{1,k}}\right)^{\frac{1}{\gamma_{1,k}}}$.

Proposition 1 ensures that the reward functions are invertible, and hence the reward obtained from an arm can be used to update the estimated expected rewards of the other arms. Assumption 1 and Proposition 2 are Hölder continuity conditions on the reward and inverse reward functions, which are used to define the informativeness. Let $\bar{\gamma}_1$ and D_2 be the maximum of the constants $\bar{\gamma}_{1,k}$ and $D_{2,k}$ and $\bar{D}_1$ and γ_2 be the minimum of exponents $\bar{D}_{1,k}$ and $\gamma_{2,k}$, respectively. Assumption 1 is satisfied by the following reward functions: (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 $\gamma > 0$ for $\Theta = [0,1]$.

The learner knows the expected reward of an arm as a function of the global parameter, i.e., $\mu_k(\cdot)$ for each $k \in \mathcal{K}$. At each time step, the learner only observes the random reward of the selected arm. The learner's goal is to maximize its cumulative reward up to any time T. If θ_* was known by the learner, it would always select the optimal arm given by $k^*(\theta_*) := \arg\max_{k \in \mathcal{K}} \mu_k(\theta_*)$. We refer to the policy that selects the arm $k^*(\theta_*)$ as the *oracle* policy and denote the expected reward of the optimal arm by $\mu^*(\theta_*) := \max_{k \in \mathcal{K}} \mu_k(\theta_*)$. We define the one-step regret at time step t as the difference between the expected reward of the oracle policy and the learner's policy that selects arm I_t, which can be written as $r_t(\theta_*) := \mu^*(\theta_*) - \mu_{I_t}(\theta_*)$. Based on this, the cumulative regret of the learner by time T is given by

$$\mathrm{Reg}(\theta_*, T) := \mathbb{E}_\nu \left[\sum_{t=1}^{T} r_t(\theta_*) \right]$$

where the expectation is taken with respect to the distribution of rewards $\nu = \times_{k \in \mathcal{K}} \nu_k(\theta_*)$.

In the next section, we propose a greedy policy which achieves bounded regret (independent of time horizon T). In the following sections we will derive regret bounds both as a function of θ_* (*parameter dependent* regret) and independent of θ_* (*worst-case* regret).

B. Weighted-Arm Greedy Policy (WAGP)

In this section, we propose a greedy policy called the Weighted-Arm Greedy Policy (WAGP). The pseudocode of WAGP is given in Algorithm 1. The WAGP consists of two phases: arm selection phase and parameter update phase.

Let $N_k(t)$ denote the number of times arm k is selected until time step t and $\hat{X}_{k,t}$ denote the reward estimate of arm k at time step t. Initially, all the counters and estimates are set to zero. In the arm selection phase at time step $t > 1$, the

Algorithm 1 The WAGP

Initialization

Inputs: $\mu_k(\cdot)$, $w_k(0) = 0, \hat{\theta}_{k,0} = 0, N_k(0) = 0$ for all $k \in \mathcal{K}$
while $t > 0$ **do**
  if t = 1 **then**
    Select arm I_1 uniformly at random from $\mathcal{K}$
  else
    Select arm $I_t \in \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_{t-1})$ (break ties randomly)
  end if
  $\hat{X}_{k,t} = \hat{X}_{k,t-1}$ for all $k \in \mathcal{K} \setminus I_t$.
  $\hat{X}_{I_t,t} = \frac{N_{I_t}(t-1)\hat{X}_{I_t,t-1}+X_{I_t,t}}{N_{I_t}(t-1)+1}$.
  $\hat{\theta}_{k,t} = \arg\min_{\theta\in\Theta} |\mu_k(\theta) - \hat{X}_{k,t}|$.
  $N_{I_t}(t) = N_{I_t}(t-1) + 1$
  $N_k(t) = N_k(t-1)$ for all $k \in \mathcal{K} \setminus I_t$
  $w_k(t) = N_k(t)/t$ for all $k \in \mathcal{K}$
  $\hat{\theta}_t = \sum_{k=1}^{K} w_k(t)\hat{\theta}_{k,t}$.
end while

WAGP selects the arm I_t with the highest estimated expected reward:

$$I_t \in \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_{t-1})$$

where $\hat{\theta}_{t-1}$ is the global parameter estimate of time step $t-1$ and the ties are broken randomly.[2] In the parameter update phase the WAGP updates: (i) the estimated reward of selected arm I_t, denoted by $\hat{X}_{I_t,t}$, (ii) the global parameter estimate of the selected arm I_t, denoted by $\hat{\theta}_{I_t,t}$, (iii) the global parameter estimate $\hat{\theta}_t$, and (iv) the counters $N_k(t)$. The reward of estimate of arm I_t is updated as:

$$\hat{X}_{I_t,t} = \frac{N_k(t-1)\hat{X}_{I_t,t-1} + X_{I_t,t}}{N_k(t-1)+1}.$$

The reward estimate of the rest of the arms are not updated. The WAGP constructs estimates of global parameter from the rewards of all the arms and combines their estimates using a weighted sum. Let $\hat{\theta}_{k,t}$ denote the global parameter estimate and $w_k(t)$ denote the weight of arm k at time step t. The WAGP updates $\hat{\theta}_{I_t,t}$ of arm I_t in a way that minimizes the distance between $\hat{X}_{I_t,t}$ and $\mu_{I_t}(\theta)$, i.e.,

$$\hat{\theta}_{I_t,t} = \arg\min_{\theta\in\Theta} |\mu_{I_t}(\theta) - \hat{X}_{I_t,t}|.$$

The, the WAGP sets the global parameter estimate as

$$\hat{\theta}_t = \sum_{k=1}^{K} w_k(t)\hat{\theta}_{k,t}$$

where $w_k(t) := N_k(t)/t$. Hence, the WAGP gives more weights to the arms with more reward observations since the confidence on their estimates are higher.

[2]For $t = 1$, the WAGP selects a random arm since there is no prior reward observation that can be used to estimate θ_*.

C. Preliminaries for the Regret Analysis

In this subsection we define the tools that will be used in deriving the regret bounds for the WAGP. Consider any arm $k \in \mathcal{K}$. Its *optimality region* is defined as

$$\Theta_k := \{\theta \in \Theta : k \in k^*(\theta)\}.$$

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 value 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 greedy policy 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 the 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 the parameter θ_*, the minimum suboptimality gap is defined as $\delta_{\min}(\theta_*) := \min_{k\in\mathcal{K}\setminus k^*(\theta_*)} \delta_k(\theta_*)$.

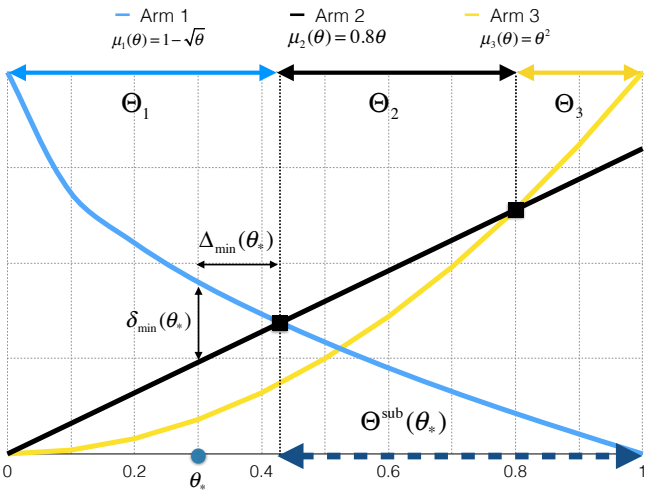


Fig. 1: Illustration of the minimum suboptimality gap and the suboptimality distance.

Recall that the estimated expected reward of 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.,

$$\Theta^{\text{sub}}(\theta_*) := \bigcup_{k'\in\mathcal{K}\setminus k^*(\theta_*)} \Theta_{k'}.$$

In order to bound the expected number of such deviations from the optimality region, for any arm k we define the *suboptimality distance*, which is equal to the smallest distance between the value of the global parameter and the 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 proposed policy always selects an optimal arm in $k^*(\theta_*)$ when $\hat{\theta}_t$ is within $\Delta_{\min}(\theta_*)$ of θ_*.[3] An illustration of the suboptimality gap and the suboptimality distance is given in Fig. 1 for the case with 3 arms and reward functions $\mu_1(\theta) = 1 - \sqrt{\theta}$, $\mu_2(\theta) = 0.8\theta$ and $\mu_3(\theta) = \theta^2$, $\theta \in [0, 1]$.

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 $\epsilon_{\theta_*} = (\delta_{\min}(\theta_*)/2D_2)^{1/\gamma_2}$, 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 is bounded above by a positive number.*

The next lemma 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 2. *For the WAGP the following relation between $\hat{\theta}_t$ and θ_* holds with probability one: $|\hat{\theta}_t - \theta_*| \leq \sum_{k=1}^{K} w_k(t) \bar{D}_1 |\hat{X}_{k,t} - \mu_k(\theta_*)|^{\bar{\gamma}_1}$.*

The following lemma ensures that the one-step regret of the WAGP decreases as $\hat{\theta}_t$ approaches to θ_*.

Lemma 3. *The one-step regret of the WAGP is bounded by $r_t(\theta_*) = \mu^*(\theta_*) - \mu_{I_t}(\theta_*) \leq 2D_2 |\theta_* - \hat{\theta}_t|^{\gamma_2}$ with probability one, for $t \geq 2$.*

Since the regret at time T is the sum of the one-step regrets up to time T, we bound the regret by bounding the expected distance between $\hat{\theta}_t$ and θ_*.

Given a parameter value θ_*, let $\mathcal{G}_{\theta_*, \hat{\theta}_t}(x) := \{|\theta_* - \hat{\theta}_t| > x\}$ 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 relation between these two events are given in the following lemma.

Lemma 4. *For WAGP we have*

$$\mathcal{G}_{\theta_*, \hat{\theta}_t}(x) \subseteq \bigcup_{k=1}^{K} \mathcal{F}^k_{\theta_*, \hat{\theta}_t} \left(\left(\frac{x}{D_1 w_k(t) K} \right)^{\frac{1}{\bar{\gamma}_1}} \right)$$

with probability one, for $t \geq 2$.

Lemma 4 follows from the decomposition given in Lemma 2. This lemma will be used to bound the probability of event $\mathcal{G}_{\theta_*, \hat{\theta}_t}(x)$ in terms of probabilities of the events $\mathcal{F}^k_{\theta_*, \hat{\theta}_t} \left(\left(\frac{x}{D_1 w_k(t) K} \right)^{\frac{1}{\bar{\gamma}_1}} \right)$.

[3]For notational brevity, we also use $\Delta_* := \Delta_{\min}(\theta_*)$ and $\delta_* := \delta_{\min}(\theta_*)$.

D. Worst-case Regret Bounds for the WAGP

First, we show that parameter estimate of the WAGP converges in the mean-squared sense.

Theorem 1. *Under Assumption 1, the global parameter estimate of the WAGP converges to true value of global parameter in mean-squared sense, i.e.,*

$$\lim_{t \to \infty} \mathbb{E}\left[|\hat{\theta}_t - \theta_*|^2 \right] = 0.$$

The following theorem bounds the expected one-step regret of the WAGP.

Theorem 2. *Under Assumption 1, we have for WAGP $\mathbb{E}[r_t(\theta_*)] = \mathcal{O}(t^{-\frac{\bar{\gamma}_1 \gamma_2}{2}})$ for any θ_*.*

Theorem 2 proves that the expected one-step regret of the WAGP converges to zero.[4] This is a *worst-case* bound in the sense that it holds for any θ_*. Using this result, we derive the worst-case regret bound for the WAGP in the next theorem.

Theorem 3. *Under Assumption 1, the worst-case regret of WAGP is*

$$\sup_{\theta_* \in \Theta} Reg(\theta_*, T) \leq \mathcal{O}(K^{\frac{\bar{\gamma}_1 \gamma_2}{2}} T^{1 - \frac{\bar{\gamma}_1 \gamma_2}{2}}).$$

Note that the worst-case regret bound is sublinear both in 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 $\bar{\gamma}_1$ 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., $\bar{\gamma}_1 = 1$. It is increasing with $\bar{\gamma}_1$, 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 Lemmas 2 and 3. As a consequence, the problem independent regret is decreasing in both $\bar{\gamma}_1$ and γ_2.[5]

When the reward functions are linear or piecewise linear, we have $\bar{\gamma}_1 = \gamma_2 = 1$, which is an extreme case of our model; hence, the worst-case regret is $\mathcal{O}(\sqrt{T})$, which matches with (i) the worst-case regret bound of the standard MAB algorithms in which a linear estimator is used [38], and (ii) the bounds obtained for the linearly parametrized bandits [3].

E. Parameter Dependent Regret Bounds for the WAGP

In this section we are concerned with deriving parameter dependent, finite-time regret bounds for WAGP. Specifically, we prove a regret bound that depends on the suboptimality distance. For instance, it is easier to identify the optimal arm in a GB with a large suboptimality distance than a GB

[4]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.

[5]Informativeness can also be regarded as a measure of the strength of correlation between the expected arm rewards.

with a small suboptimality distance. We derive three regimes of growth for our regret bound: Initially the regret grows sublinearly until a threshold is reached. After this threshold, it grows logarithmically until a second threshold, after which its growth is bounded. The thresholds that define the boundaries of these regimes are given below.

Definition 2. *$C_1(\Delta_*)$ is the smallest integer τ such that $\tau \geq \left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \frac{\log(\tau)}{2}$ and $C_2(\Delta_*)$ is the smallest integer τ such that $\tau \geq \left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \log(\tau)$.*

In order to define these constants in a closed form, let us define *glog* function.

Definition 3. *$y = \mathrm{glog}(x)$ if and only if $x = \frac{\exp(y)}{y}$.*

The constants are given as

$$C_1(\Delta_*) = \frac{1}{2}\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \mathrm{glog}\left(\frac{1}{2}\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}}\right),$$
$$C_2(\Delta_*) = \left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \mathrm{glog}\left(\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}}\right) \quad (1)$$

These constants depend on the informativeness (Hölder exponent $\bar{\gamma}_1$) θ_*. We define the expected regret incurred between time steps T_1 and T_2 given θ_* as

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

The parameter dependent regret bound for the WAGP is given in the following theorem.

Theorem 4. *Under Assumption 1, the regret of the WAGP is bounded as follows:*
(i) For $1 \leq T < C_1(\Delta_*)$, the regret grows sublinearly in time, i.e.,

$$R_{\theta_*}(0, T) \leq S_1 + S_2 T^{1-\frac{\bar{\gamma}_1\gamma_2}{2}}$$

where S_1 and S_2 are constants that are independent of the global parameter θ_*, whose exact forms are given in Appendix VII-F. (ii) For $C_1(\Delta_*) \leq T < C_2(\Delta_*)$, the regret grows logarithmically in time, i.e.,

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

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

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

Since $\lim_{\Delta_* \to 0} C_1(\Delta_*) = \infty$, in the worst-case, the bound given in Theorem 4 reduces to the one given in Theorem 3. We can also calculate a Bayesian risk bound for the WAGP by assuming a prior over the global parameter space. This risk bound is given to be $\mathcal{O}(\log T)$ when $\bar{\gamma}_1\gamma_2 = 1$ and $\mathcal{O}(T^{1-\bar{\gamma}_1\gamma_2})$ when $\bar{\gamma}_1\gamma_2 < 1$ [1]. The following corollary characterizes the asymptotic behavior of the regret of the WAGP.

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

The different growth rates for the regret given in Theorem 4 are found by bounding the probability that the WAGP selects a suboptimal arm as a function of t. For instance, when $C_1(\Delta_*) \leq t < C_2(\Delta_*)$, the probability of selecting a suboptimal arm is in the order of t^{-1}; hence, the WAGP achieves 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 regret of the WAGP is bounded asymptotically.

Theorem 5. *The sequence of arms selected by the WAGP converges to the optimal arm almost surely, i.e.,* $\lim_{t\to\infty} I_t = k^*(\theta_*)$ *with probability 1.*

Theorem 5 implies that a suboptimal arm is selected by the WAGP only finitely many times. This is the biggest difference between the achievable performance in the GB and the standard MAB [2], [10], [36] in which every arm needs to be selected infinitely many times asymptotically.

F. Lower Bound on the Worst-case Regret

Theorem 3 shows that the worst-case regret bound is $\mathcal{O}(T^{1-\frac{\bar{\gamma}_1\gamma_2}{2}})$, which implies that the regret is decreasing with Hölder exponents $\bar{\gamma}_1$ and γ_2. In this section, we show that this is the best attainable regret order for the family of policies that use a global estimator $\hat{\theta}_t$. Intuitively, these policies map the observation tuple $(\boldsymbol{M}_t, \boldsymbol{N}_t, t)$ to the set of arms where $\boldsymbol{M}_t = \left[\mu_1^{-1}(\hat{X}_{1,t}), \mu_2^{-1}(\hat{X}_{2,t}), \ldots, \mu_K^{-1}(\hat{X}_{K,t})\right]$ and $\boldsymbol{N}_t = [N_1(t), N_2(t), \ldots, N_K(t)]$.

Theorem 6. *For the family of policies that use a global estimator $\hat{\theta}_t$ and $T \geq 8$, the worst-case regret is lower bounded by* $\sup_{\theta_* \in \Theta} Reg(\theta_*, T) = \Omega(T^{1-\frac{\bar{\gamma}_1\gamma_2}{2}})$.

Theorem 6 is proven by showing that by choosing unfavorably small values for Δ_* (for instance, by letting $\Delta_* = T^{-\bar{\gamma}_1\gamma_2/2}$), the regret can be made to grow as $t^{-\frac{\bar{\gamma}_1\gamma_2}{2}}$ before stabilizing to the finite value. Therefore, there exists a problem instance for which the worst-case regret bound of WAGP matches the lower bound in terms of the time order.

While Theorem 4 shows that the WAGP-type policies result in a great performance improvement compared to the standard MAB algorithms that treat each arm independently for a given problem instance, Theorem 6 points to a deficiency of these type of policies in terms of the worst-case regret. The reason behind this deficiency is that the global policies select arms based on a non-linear estimation of the global parameter. This raises a natural question that we resolve in the next section: Can we achieve both $\mathcal{O}(\sqrt{T})$ worst-case regret (like the UCB-type MAB algorithms) and bounded parameter dependent regret by using a combination of UCB and WAGP-type policies?

IV. Extensions

In this section, we provide two extensions of the WAGP. In the first part, we propose the Best of the UCB and the WAGP

Algorithm 2 The BUW

Inputs: T, $\mu_k(\cdot)$ for each arm k.
Initialization: $\hat{\theta}_{k,t} = 0$, $\hat{\theta}_t = 0$, $N_k(t) = 0$, $\hat{\mu}_k = 0$, $\hat{\Delta}_t = 0$.

while $t \geq 1$ **do**
  if $\hat{\Delta}_t \leq \bar{D}_1 K \left(\frac{\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}}$ **then**
    $I_t \in \arg\max_{k\in\mathcal{K}} \hat{X}_{k,t} + \sqrt{\frac{2\log t}{N_k(t)}}$
  else if $t < C_2(\tilde{\Delta}_t)$ **then**
    $I_t \in \arg\max_{k\in\mathcal{K}} \hat{X}_{k,t} + \sqrt{\frac{2\log t}{N_k(t)}}$
  else
    $I_t \in \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_t)$
  end if
  Update $\hat{X}_{I_t,t}$, $N_k(t)$, $w_k(t)$, $\hat{\theta}_{k,t}$, $\hat{\theta}_t$
  Solve

$$\hat{\Delta}_t = \begin{cases} \inf_{\theta'\in\Theta^{\text{sub}}(\hat{\theta}_t)} |\hat{\theta}_t - \theta'| & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) \neq \emptyset \\ \infty & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) = \emptyset \end{cases}$$

  $\tilde{\Delta}_t = \hat{\Delta}_t - \bar{D}_1 \left(\frac{2K\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}}$.
end while

(BUW), which combines UCB1 and the WAGP to achieve bounded parameter dependent and $O(\sqrt{T})$ worst-case regrets. In the second part, we provide a modified version of the WAGP that works under a time-varying global parameter.

A. The Best of the UCB and the WAGP (BUW)

In the worst-case, the WAGP achieves $\mathcal{O}(T^{1-\frac{\bar{\gamma}_1\gamma_2}{2}})$ regret, which is weaker than $O(\sqrt{T})$ worst-case regret of UCB1. On the other hand, the WAGP achieves bounded parameter dependent regret whereas the UCB1 achieves a logarithmic parameter dependent regret. In this section, we provide an algorithm which combines these two algorithms and achieves both $O(\sqrt{T})$ worst-case regret and bounded parameter dependent regret. The main idea for such an algorithm follows from Theorem 4. Recall that Theorem 4 shows that the WAGP achieves a regret of $O(T^{1-\frac{\bar{\gamma}_1\gamma_2}{2}})$ when $1 < T < C_1(\Delta_*)$. If the BUW follows the recommendations of the UCB when $1 < T < C_1(\Delta_*)$ and the recommendations of the WAGP when $T > C_1(\Delta_*)$, then the algorithm will achieve a worst-case regret bound of $O(\sqrt{T})$ and bounded parameter-dependent regret bound. The only problem in this approach is that suboptimality distance Δ_* is unknown a priori. We can solve this problem by using a data-dependent estimate $\tilde{\Delta}_t$ where $\Delta_* > \tilde{\Delta}_t$ holds with high probability. The following $\tilde{\Delta}_t$ satisfies $\Delta_* > \tilde{\Delta}_t$ with high probability :

$$\tilde{\Delta}_t = \hat{\Delta}_t - \bar{D}_1 \left(\frac{K\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}}$$

where

$$\hat{\Delta}_t = \Delta_{\min}(\hat{\theta}_t) = \begin{cases} \inf_{\theta'\in\Theta^{\text{sub}}(\hat{\theta}_t)} |\hat{\theta}_t - \theta'| & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) \neq \emptyset \\ \infty & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) = \emptyset \end{cases}$$

The pseudo-code for the BUW is given in Fig. 2. Define $C_3(\Delta_*)$ as

$$C_3(\Delta_*) = 2\left(\frac{4D_1K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \text{glog}\left(2\left(\frac{4D_1K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}}\right)$$

The regret bounds for the BUW are given in the following theorem. In theorem, we drop the dependence on the θ_* in order to write the regret bounds more elegantly.

Theorem 7. *Under Assumption 1,the worst-case regret of the BUW is bounded as follows:*

$$\sup_{\theta_*\in\Theta} Reg(\theta_*, T) \leq \mathcal{O}(\sqrt{KT}).$$

Under Assumption 1, the parameter dependent regret of the BUW is bounded as follows:

(i) For $1 \leq T < C_3(\Delta_*)$, the regret grows logarithmically in time, i.e.,

$$R_{\theta_*}(0,T) \leq \left[8\sum_{k:\mu_k<\mu^*} \frac{\log T}{\delta_k}\right] + K\left(1+\frac{\pi^2}{3}\right) + KT^{-2}.$$

(ii) For $T \geq C_3(\Delta_*)$, the growth of the regret is bounded, i.e.,

$$R_{\theta_*}(C_3(\Delta_*/2), T) \leq K\frac{\pi^2}{3} + KT^{-2}.$$

Since the BUW is not in the class of policies that only relies on the global estimator, it is able to achieve the best worst-case regret bound.

B. Learning under Time-varying Global Parameter

In this section, we consider the case when the global parameter changes over time. We denote the global parameter at time t as θ_*^t. The reward of arm k at time t, i.e., $X_{k,t}$, is drawn independently from the distribution $\nu_k(\theta_*^t)$ where $\mathrm{E}[X_{k,t}] = \mu_k(\theta_*^t)$. In order to bound the regret, we impose a restriction on the *speed* of change of the global parameter which is formalized in the following assumption.

Assumption 2. *For any t and t', we have*

$$|\theta_*^t - \theta_*^{t'}| \leq \left|\frac{t}{\tau} - \frac{t'}{\tau}\right|$$

where $\tau > 0$ controls the speed of the change.

The WAGP needs to be modified to handle a non-stationary global parameter since the optimal arm $k^*(\theta_*^t)$ may be changing over time. To do this, the modified WAGP uses only a recent past window of reward observations from the arms when estimating the global parameter. By choosing the window length appropriately, we can balance the regret due to the variation of the global parameter over time given in Assumption 2 and the sample size within the window.

The modified algorithm groups the time steps into rounds $\rho = 1, 2, \ldots$, each having a fixed length of $2\tau_h$, where τ_h is called *half window length*. The key point in the modified algorithm is to keep separate counters for each round and estimate the global parameter in a round based only on observations that are made within the particular window of

each round. Each round ρ is further divided into two sub-rounds. The first sub-round is called passive sub round, while the second one is called the active sub-round. The first round, $\rho = 1$, is an exception where it is both active and passive sub-round.

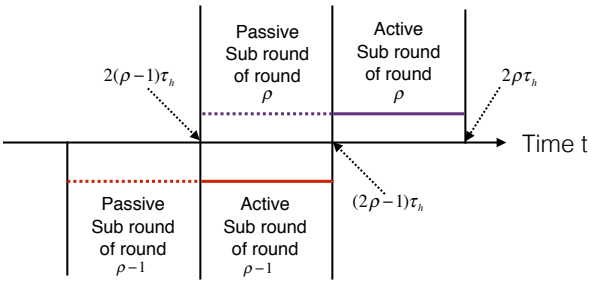


Fig. 2: Operation of the time windowed WAGP.

Let WAGP$_\rho$ be the running instance of the modified WAGP at round ρ. The arm selected at time t is based on WAGP$_\rho$ if time t is in the active sub round of round ρ. Let $N_{k,\rho}(t)$ and $\hat{X}_{k,\rho,t}$ be the number of times arm k is chosen and the estimate of the arm k at round ρ at time t, respectively. At the beginning of each round ρ, the estimates and counters of that round are equal to zero, i.e $N_{k,\rho}(2\tau_h\rho+1) = 0$ and $\hat{X}_{k,\rho,2\tau_h\rho+1} = 0$. However, due to the subround structure, the learner can use the observations from the passive subround of a round when choosing actions in the active subround of a round.

Similar to static parameter case, the WAGP selects the arm with the highest estimated reward at round ρ. Let $\hat{\theta}_{k,\rho,t}$ denote the parameter estimate from arm k at round ρ at time t, which is given by $\hat{\theta}_{k,\rho,t} = \mu_k^{-1}(\hat{X}_{k,\rho,t})$. The global parameter estimate at round ρ is then given by

$$\hat{\theta}_{\rho,t} = \sum_{k=1}^{K} w_{k,\rho}(t)\hat{\theta}_{k,\rho,t},$$

where $w_{k,\rho}(t) = N_{k,\rho}(t)/(t - 2\tau_h(\rho-1))$. The arm with the highest reward estimate at round ρ is selected, i.e.,

$$I_{\rho,t} = \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_{\rho,t})$$

In the static global parameter model, we were able to bound the problem specific regret with a finite constant number (independent of time horizon T) and the parameter-independent regret with a sub-linear in time (T^γ for $\gamma > 0$). However,when global parameter is changing, it is not possible to give a sub-linear or finite regret bounds. Therefore, we focus on the average regret, which is given as

$$\text{Reg}^{\text{ave}}(T) := \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\mu^*(\theta_*^t) - \sum_{t=1}^{T}\mu_{I_t}(\theta_*^t)\right].$$

Next theorem quantifies the average regret bound with respect to the stability and exponent of the drift.

Theorem 8. *Under Assumptions 1 and 2, when the half window length of the time windowed WAGP is set to* $\tau_h = \tau^{\frac{\gamma_2}{(\gamma_2+0.5)}}$*, the average regret is*

$$Reg^{ave}(T) = \mathcal{O}\left(\tau^{\frac{-\bar{\gamma}_1\gamma_2^2}{(2\gamma_2+1)}}\right).$$

Theorem 8 shows that the average regret is bounded by a decreasing function of τ and informativeness. This is expected since the greedy policy is able to track the changes in the parameter when the drift is slow. Since the informativeness of the arms is directly related to the learning rate of the global parameter, the tracking performance of the modified algorithm is increasing with the informativeness.

V. Illustrative Results : A Dynamic Pricing Example

This setting is inspired by the dynamic pricing example formulated in Section 1. We assume that the expected sales $S_{p,t}$ at time t under price p are of the form $\mathbb{E}[S_{p,t}] = (1-\theta p)^2$, where θ characterizes the market size, and is set to 0.4. Note that this is the linear-power demand model used by [6], [39]. The expected revenue is $\mathbb{E}[R_{p,t}] = p(1-\theta p)^2$. Note that reward function is $\mu_p = \mu_p(\theta) = p(1-\theta p)^2$ for this problem instance. We generate random rewards of each price p at each time t by drawing randomly from a Beta distribution with parameters 1 and $(1-\mu_p)/\mu_p$, i.e., $R_{p,t} \sim \text{Beta}(1,(1-\mu_p)/\mu_p)$, and hence $\mathbb{E}[R_{p,t}] = \mu_p$. We set $K = 12$ with $0.4, 0.45, 0.5, \ldots, 0.95$.

Experiment 1 (Comparison): We compare our algorithm with two different benchmarks: Upper Confidence Bound (UCB) of [10] and Uncertainty Ellipsoid (UE) [28]. The UCB algorithm treats each arm independently and learn their expected rewards by exploration. The UE algorithm is proposed for linearly parametrized reward structure with high-dimensional parameter space. In our setting, the UE can be used by setting an arm vector $u_p = \left[p, p^2 p^3\right]$ in order to fit a polynomial with order 3 for the expected rewards. We generate rewards according to setting described above and average the results over 100 iterations. Figure 4 shows that WAGP significantly outperforms UCB by exploiting the correlations between the arms. The significant performance advantage obtained by WAGP as compared to UCB is due to the fact that WAGP is able to focus on good arms early on while UCB learns each arm separately. WAGP selects arm 10 (the best arm) 81.7% of time, arm 9 (the second best arm) 16.4% of time and the rest of the arms 1.9% of time. The UE outperforms the UCB by using (some) of the correlations between the arms, however, fails to achieve the performance of the WAGP. The reason is that the WAGP learns about the parameter by selecting any of the arms, however, the UE needs to select 3 linearly independent arms in order to learn about the parameter.

Experiment 2 (The effect of the sub optimality distance): Table 1 shows the expected regret of WAGP for different θ and hence different $\Delta_{\min}(\theta)$. From this it can be seen that the regret of WAGP is indeed increasing with the sub optimality distance as predicted by Theorem 4.

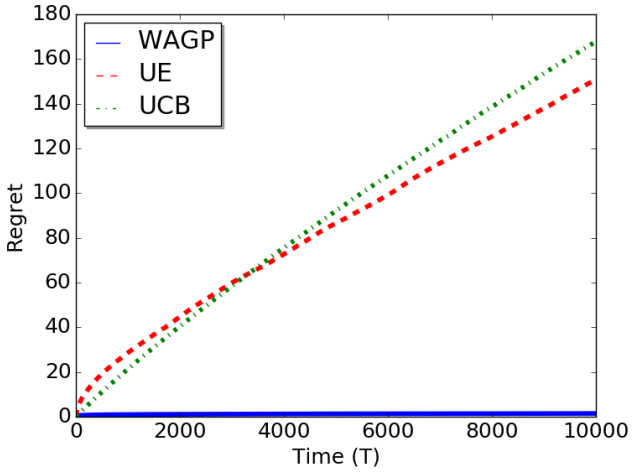


Fig. 3: Comparison of UCB and WAGP for dynamic pricing example

θ	0.2	0.1	0.3	0.8	0.5
$\Delta_{\min}(\theta)$	0.17	0.1	0.07	0.02	0.01
$\mathrm{Reg}(\theta, 1000)$	0.3	0.65	0.72	2.02	2.47

TABLE I: Regret of WAGP for different market size

Experiment 3 (Non-stationary Parameter): We show the performance of the proposed methods for a non-stationary setting. The expected revenue for price p at time t is given by $\mathbb{E}[R_{p,t}] = p(1-\theta_*^t p)^2$. We assume that $\theta_1^* = 0.5$ and $\theta_*^t = \theta_*^{t-1} + Y_t/\tau$ where Y_t is a random variable with $\Pr(Y_t = 1) = 0.6$ and $\Pr(Y_t = -1) = 0.4$ and τ controls the speed of the drift. Hence,

$$|\theta_*^t - \theta_*^{t'}| \leq \left|\frac{t}{\tau} - \frac{t'}{\tau}\right|$$

with probability 1 for all $t, t' \geq 1$.

Figure 4 illustrates modified WAGP under non-stationary dynamic pricing example. We use $\tau = 1000$ to illustrate the tracking performance of the modified WAGP in Figure 4a. Note that $\tau_h = 100$ for this example. The reward observations used to estimate parameter changes for $t = 200, 300 \ldots, 900$. This results in some jumps in the estimate at these times as seen from Figure 4a. From these figures it can be seen that our modified WAGP is able to track the non-stationary global parameter and the slope of the regret is decreasing function of τ as predicted by Theorem 8.

VI. Conclusion

In this paper we introduce a new class of MAB problems called Global Bandits (GB). This general class of GB problems encompasses the previously introduced linearly-parametrized bandits as a special case. We proved that the regret for the GBs 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, 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, and a foresighted MAB, where the arm selection is based on a foresighted policy that explores the arms according to their level of informativeness rather than the greedy policy.

VII. Appendices

A. Preliminaries

In all the proofs given below let $\boldsymbol{w}(t) := (w_1(t), \ldots, w_K(t))$ be the vector of weights and $\boldsymbol{N}(t) := (N_1(t), \ldots, N_k(t))$ be the vector of counters at time t. We have $\boldsymbol{w}(t) = \boldsymbol{N}(t)/t$. Since $\boldsymbol{N}(t)$ depends on the history, they are both random variables that depend on the sequence of obtained rewards.

B. Proof of Proposition 1

(i) Let k and $\theta \neq \theta'$ be arbitrary. Then, by Assumption 1,

$$|\mu_k(\theta) - \mu_k(\theta')| \geq D_{1,k}|\theta - \theta'|^{\gamma_{1,k}} > 0$$

and hence $\mu_k(\theta) \neq \mu_k(\theta')$.

(ii) Suppose $y = \mu_k(\theta)$ and $y' = \mu_k(\theta')$ for some arbitrary θ and θ'. Then, by Assumption 1,

$$|y - y'| \geq D_{1,k}|\mu_k^{-1}(y) - \mu_k^{-1}(y')|^{\gamma_{1,k}}.$$

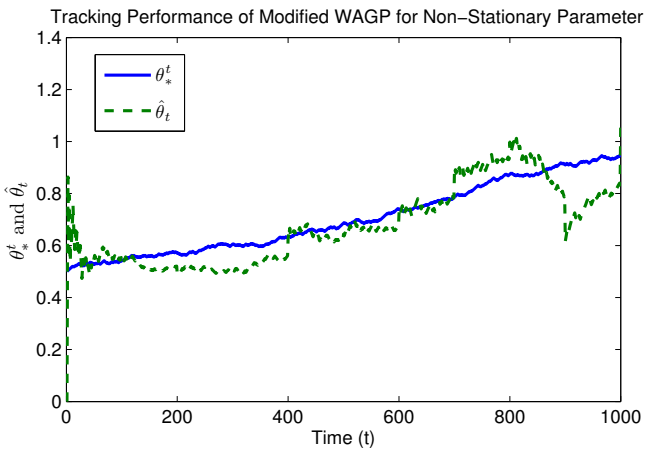
C. Proof of Lemma 1

Consider a parameter value $\theta \in \Theta$. For any suboptimal arm $k \in \mathcal{K} - k^*(\theta)$, we have $\mu_{k^*(\theta)}(\theta) - \mu_k(\theta) \geq \delta_{\min}(\theta) > 0$. We also know that $\mu_k(\theta') \geq \mu_{k^*(\theta)}(\theta')$ for all $\theta' \in \Theta_k$. Hence for any $\theta' \in \Theta_k$ at least one of the following must hold: (i) $\mu_k(\theta') \geq \mu_k(\theta) - \delta_{\min}(\theta)/2$, (ii) $\mu_{k^*(\theta)}(\theta') \leq \mu_{k^*(\theta)}(\theta) + \delta_{\min}(\theta)/2$. If both of the above does not hold, then we must have $\mu_k(\theta') < \mu_{k^*(\theta)}(\theta')$, which is false. This implies that we either have $\mu_k(\theta) - \mu_k(\theta') \leq \delta_{\min}(\theta)/2$ or $\mu_{k^*(\theta)}(\theta) - \mu_{k^*(\theta)}(\theta') \geq -\delta_{\min}(\theta)/2$, or both. Recall that from Assumption 1 we have $|\theta - \theta'| \geq |\mu_k(\theta) - \mu_k(\theta')|^{1/\gamma_2}/D_2^{1/\gamma_2}$. This implies that $|\theta - \theta'| \geq \epsilon_\theta$ for all $\theta' \in \Theta_k$.

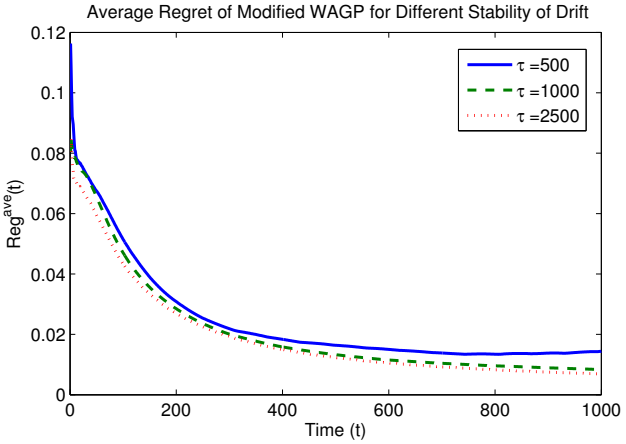
D. Proof of Lemma 2

Assumption 2 ensures that the reward functions are either monotonically increasing or decreasing. We generate imaginary functions that are $\mu_k(\theta) = \tilde{\mu}_k(\theta)$ for $\theta \in \Theta$ and for $y, y' \in [0, 1]$,

$$|\tilde{\mu}_k^{-1}(y) - \tilde{\mu}_k^{-1}(y')| \leq \bar{D}_1|y - y'|^{\bar{\gamma}_1} \tag{2}$$



(a) 1a : Tracking Performance of Modified WAGP



(b) 1b : Expected Regret of Modified WAGP

Fig. 4: Performance of Modified WAGP for Non-Stationary Global Parameter

We have also $\tilde{\mu}_k^{-1}(y) > 1$ when $y > \max_{\theta\in\Theta} \mu_k(\theta)$ and $\tilde{\mu}_k^{-1}(y) < 0$ when $y < \min_{\theta\in\Theta} \mu_k(\theta)$.Then,

$$\begin{aligned}
|\theta_* - \hat{\theta}_t| &= |\sum_{k=1}^{K} w_k(t)\hat{\theta}_{k,t} - \theta_*| \\
&= \sum_{k=1}^{K} w_k(t) \left|\theta_* - \hat{\theta}_{k,t}\right| \\
&\leq \sum_{k=1}^{K} w_k(t)|\tilde{\mu}_k^{-1}(\hat{X}_{k,t}) - \tilde{\mu}_k^{-1}(\tilde{\mu}_k(\theta_*))| \\
&\leq \sum_{k=1}^{K} w_k(t)\bar{D}_1|\hat{X}_{k,t} - \mu_k(\theta_*)|^{\bar{\gamma}_1}, \qquad (3)
\end{aligned}$$

where we need to look at following two cases for the first inequality. The first case is $\hat{X}_{k,t} \in \mathcal{Y}_k$ where the statement immediately follows. The second case is $\hat{X}_{k,t} \notin \mathcal{Y}_k$, where the global parameter estimator $\hat{\theta}_{k,t}$ is either 0 or 1.

E. Proof of Lemma 3

Note that $I_t \in \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_t)$. Therefore, we have

$$\mu_{I_t}(\hat{\theta}_t) - \mu_{k^*(\theta_*)}(\hat{\theta}_t) \geq 0. \qquad (4)$$

Since $\mu^*(\theta_*) = \mu_{k^*(\theta_*)}(\theta_*)$, we have

$$\begin{aligned}
&\mu^*(\theta_*) - \mu_{I_t}(\theta_*) \\
&= \mu_{k^*(\theta_*)}(\theta_*) - \mu_{I_t}(\theta_*) \\
&\leq \mu_{k^*(\theta_*)}(\theta_*) - \mu_{I_t}(\theta_*) + \mu_{I_t}(\hat{\theta}_t) - \mu_{k^*(\theta_*)}(\hat{\theta}_t) \\
&= \mu_{k^*(\theta_*)}(\theta_*) - \mu_{k^*(\theta_*)}(\hat{\theta}_t) + \mu_{I_t}(\hat{\theta}_t) - \mu_{I_t}(\theta_*) \\
&\leq 2D_2|\theta_* - \hat{\theta}_t|^{\gamma_2},
\end{aligned}$$

where the first inequality follows from (4) and the second inequality follows from Assumption 1.

F. Proof of Lemma 4

Observe that

$$\begin{aligned}
&\{|\theta_* - \hat{\theta}_t| \leq x\} \\
&\supseteq \left\{\sum_{k=1}^{K} w_k(t)\bar{D}_1|\hat{X}_{k,t} - \mu_k(\theta_*)|^{\bar{\gamma}_1} \leq x\right\} \\
&\supseteq \bigcup_{k=1}^{K} \left\{|\hat{X}_{k,t} - \mu_k(\theta_*)| \leq \left(\frac{x}{w_k(t)\bar{D}_1 K}\right)^{1/\bar{\gamma}_1}\right\},
\end{aligned}$$

where the first inequality follows from Lemma 2. Then,

$$\begin{aligned}
&\{|\theta_* - \hat{\theta}_t| > x\} \subseteq \\
&\qquad \bigcup_{k=1}^{K} \left\{|\hat{X}_{k,t} - \mu_k(\theta_*)| > \left(\frac{x}{w_k(t)\bar{D}_1 K}\right)^{1/\bar{\gamma}_1}\right\}.
\end{aligned}$$

G. Proof of Theorem 1

Using Lemma 2, the mean-squared error can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[|\theta_* - \hat{\theta}_t|^2\right] \\
&\leq \mathbb{E}\left[\left(\sum_{k=1}^{K} \bar{D}_1 w_k(t)|\hat{X}_{k,t} - \mu_k(\theta_*)|\right)^2\right] \qquad (5) \\
&\leq K\bar{D}_1^2 \sum_{k=1}^{K} \mathbb{E}\left[w_k^2(t)|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1}\right],
\end{aligned}$$

where the inequality follows from the fact that $\sum_{k=1}^{K} a_k \leq K\sum_{k=1}^{K} a_k^2$ for any $a_k > 0$. Then,

$$\begin{aligned}
&\mathbb{E}\left[|\theta_* - \hat{\theta}_t|^2\right] \\
&\leq K\bar{D}_1^2 \mathbb{E}\left[\sum_{k=1}^{K} w_k^2(t)\mathrm{E}\left[|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1}|\boldsymbol{w}(t)\right]\right] \qquad (6) \\
&\leq K\bar{D}_1^2 \mathbb{E}\left[\sum_{k=1}^{K} w_k^2(t)\int_{x=0}^{\infty} \Pr(|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1} \geq x)dx\right],
\end{aligned}$$

where the second inequality follows from the fundamental theorem of expectation. Then, we can bound inner expectation as

$$\begin{aligned}&\int_{x=0}^{\infty} \Pr(|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1} \geq x)dx \\ &\leq \int_{x=0}^{\infty} 2\exp(-x^{\frac{1}{\bar{\gamma}_1}} N_k(t))\, dx. \\ &= 2\bar{\gamma}_1 \Gamma(\bar{\gamma}_1) N_k^{-\bar{\gamma}_1}(t),\end{aligned}$$

where $N_k(t)$ is a random variable and $\Gamma(\cdot)$ is Gamma function. Then, we have

$$\begin{aligned}\mathbb{E}[|\theta_* - \hat{\theta}_t|^2] &\leq 2K\gamma \bar{D}_1^2 \Gamma(\bar{\gamma}_1) \mathbb{E}\left[\sum_{k=1}^{K} \frac{N_k^{2-\bar{\gamma}_1}(t)}{t^2}\right] \\ &\leq 2K\gamma \bar{D}_1^2 \Gamma(\bar{\gamma}_1) t^{-\bar{\gamma}_1},\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}[\sum_{k=1}^{K} N_k^{2-\bar{\gamma}_1}(t)/t^2] \leq t^{-\bar{\gamma}_1}$ for any $N_k(t)$ since $\sum_{k=1}^{K} N_k(t) = t$ and $\bar{\gamma}_1 \leq 1$.

H. Proof of Theorem 2

By Lemma 3 and Jensen's inequality, we have

$$\mathbb{E}[r_t(\theta_*)] \leq 2D_2 \mathrm{E}\left[|\theta_* - \hat{\theta}_t|\right]^{\gamma_2}. \tag{7}$$

Also by Lemma 2 and Jensen's inequality, we have

$$\begin{aligned}&\mathbb{E}\left[|\theta_* - \hat{\theta}_t|\right] \\ &\leq \bar{D}_1 \mathbb{E}\left[\sum_{k=1}^{K} w_k(t) \mathbb{E}\left[|\hat{X}_{k,t} - \mu_k(\theta_*)|\, |\boldsymbol{w}(t)\right]^{\bar{\gamma}_1}\right],\end{aligned} \tag{8}$$

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

$$\begin{aligned}&\mathbb{E}\left[|\hat{X}_{k,t} - \mu_k(\theta_*)|\, |\boldsymbol{w}(t)\right] \\ &= \int_{x=0}^{1} \Pr\left(|\hat{X}_{k,t} - \mu_k(\theta_*)| > x | \boldsymbol{w}(t)\right) \mathrm{d}x \\ &\leq \int_{x=0}^{\infty} 2\exp(-2x^2 N_k(t))\, \mathrm{d}x \\ &\leq \sqrt{\frac{\pi}{2N_k(t)}},\end{aligned} \tag{9}$$

where $N_k(t) = t w_k(t)$ is a random variable and the first inequality is a result of the Chernoff-Hoeffding bound. Combining (8) and (9), we get

$$\mathbb{E}[|\theta_* - \hat{\theta}_t|] \leq 2\bar{D}_1 (\frac{\pi}{2})^{\frac{\bar{\gamma}_1}{2}} \frac{1}{t^{\frac{\bar{\gamma}_1}{2}}} \mathbb{E}\left[\sum_{k=1}^{K} w_k(t)^{1-\frac{\bar{\gamma}_1}{2}}\right]. \tag{10}$$

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 $\mathbb{E}[\sum_{k=1}^{K} w_k(t)^{1-\frac{\bar{\gamma}_1}{2}}] \leq K^{\frac{\bar{\gamma}_1}{2}}$. Then, combining (7) and (10), we have

$$\mathbb{E}[r_t(\theta_*)] \leq 2\bar{D}_1^{\gamma_2} D_2 \frac{\pi}{2}^{\frac{\bar{\gamma}_1\gamma_2}{2}} K^{\frac{\bar{\gamma}_1\gamma_2}{2}} \frac{1}{t^{\frac{\bar{\gamma}_1\gamma_2}{2}}}.$$

I. Proof of Theorem 3

This bound is consequence of Theorem 2 and the inequality given in [40], i.e.,

$$\mathrm{Reg}(\theta_*, T) \leq 1 + \frac{2\bar{D}_1^{\gamma_2} D_2 \frac{\pi}{2}^{\frac{\gamma_1\gamma_2}{2}} K^{\frac{\bar{\gamma}_1\gamma_2}{2}}}{1 - \frac{\bar{\gamma}_1\gamma_2}{2}} (1 + T^{1-\frac{\bar{\gamma}_1\gamma_2}{2}}).$$

J. Proof of Theorem 4

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

$$\begin{aligned}&\{I_t \neq k^*(\theta_*)\} \\ &= \{\hat{\theta}_t \in \Theta \setminus \Theta_{k^*(\theta_*)}\} \subseteq \mathcal{G}^{\Delta_*}_{\theta_*, \hat{\theta}_t}.\end{aligned} \tag{11}$$

By Lemma 4 and (11), we have

$$\begin{aligned}&\Pr(I_t \neq k^*(\theta_*)) \\ &\leq \sum_{k=1}^{K} \mathbb{E}\left[\mathbb{E}\left[I\left(\mathcal{F}^k_{\theta_*, \hat{\theta}_t}\left(\left(\frac{\Delta_*}{w_k(t)\bar{D}_1 K}\right)^{\frac{1}{\bar{\gamma}_1}}\right)\right) | \boldsymbol{N}(t)\right]\right] \\ &\leq \sum_{k=1}^{K} 2\mathbb{E}\left[\exp\left(-2\left(\frac{\Delta_*}{w_k(t)\bar{D}_1 K}\right)^{\frac{2}{\bar{\gamma}_1}} w_k(t) t\right)\right] \\ &\leq 2K\exp\left(-2\left(\frac{\Delta_*}{\bar{D}_1 K}\right)^{\frac{2}{\bar{\gamma}_1}} t\right),\end{aligned} \tag{12}$$

where the first inequality follows from a union bound and the second inequality is obtained by using the Chernoff-Hoeffding bound. The last inequality is obtained by using Lemma 5. We have $\Pr(I_t \neq k^*(\theta_*)) \leq 1/t$ for $t > C_1(\Delta_*)$ and $\Pr(I_t \neq k^*(\theta_*)) \leq 1/t^2$ for $t > C_2(\Delta_*)$. The bound in the first regime is the result of Theorem 3. The bounds in the second and third regimes are obtained by summing the probability given in (12) from $C_1(\Delta_*)$ to T and $C_2(\Delta_*)$ to T, respectively.

K. Proof of Theorem 5

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 P is defined on. Let $\omega \in \Omega$ denote a sample path. We will prove that there exists event $N \in \mathcal{F}$ such that $P(N) = 0$ and if $\omega \in N^{\mathrm{c}}$, 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 4 that $\sum_{t=1}^{T} P(\mathcal{E}_t) < \infty$. By Borel-Cantelli lemma, we have

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

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

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

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

L. Proof of Theorem 6

Consider a problem instance with two arms with reward functions $\mu_1(\theta) = \theta^\gamma$ and $\mu_2(\theta) = 1 - \theta^\gamma$, where γ is an odd integer valued number and rewards are Bernoulli distributed with $X_{1,t} \sim \text{Ber}(\mu_1(\theta))$ and $X_{2,t} \sim \text{Ber}(\mu_2(\theta))$. Then, optimality regions are $\Theta_1 = [2^{-\frac{1}{\gamma}}, 1]$ and $\Theta_2 = [0, 2^{-\frac{1}{\gamma}}]$. Note that $\gamma_2 = 1$ and $\gamma_1 = 1/\gamma$ for this case. Let $\theta^* = 2^{-\frac{1}{\gamma}}$. Consider following two cases with $\theta_1^* = \theta_* + \Delta$ and $\theta_2^* = \theta_* - \Delta$. The optimal arm is 1 in the first case and 2 in the second case. In the first case, one step loss due to choosing arm 2 is bounded by

$$\begin{aligned}
&(\theta^* + \Delta)^\gamma - (1 - (\theta^* + \Delta)^\gamma) \\
&= 2(\theta^* + \Delta)^\gamma - 1 \\
&= 2((\theta^*)^\gamma + \binom{\gamma}{1}(\theta^*)^{\gamma-1}\Delta + O(\Delta)) - 1 \\
&= 2\gamma 2^{\frac{1-\gamma}{\gamma}}\Delta + o(\Delta).
\end{aligned}$$

where $a_1 = 2\gamma 2^{\frac{1-\gamma}{\gamma}}$. Similarly, in the second case, it can be again lower bounded by $a_1\Delta - o(\Delta)$.

Then, we can lower bound the regret as

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{r}_t(\theta^* + \Delta)\right] \\
&\geq a_1\Delta \Pr(\hat{\theta}_t <= \theta^*) \qquad (13)\\
&= \frac{a_1\Delta}{2}(\Pr(\hat{\theta}_t - \theta_* \leq -\Delta) + \Pr(\hat{\theta}_t - \theta_* \leq -\Delta)) \\
&= \frac{a_1\Delta}{2}(\Pr((\hat{\theta}_t - \theta_*)^\gamma \leq -\Delta^\gamma) + \Pr((\hat{\theta}_t - \theta_*)^\gamma \leq -\Delta^\gamma)) \\
&\geq \frac{a_1\Delta}{2}(\Pr((\hat{\theta}_t)^\gamma - (\theta_*)^\gamma \leq -\Delta^\gamma) \qquad (14)\\
&+ \Pr((\hat{\theta}_t)^\gamma - (\theta_*)^\gamma \leq -\Delta^\gamma)), \qquad (15)
\end{aligned}$$

where the last inequality follows from the fact that $(a-b)^\gamma \leq a^\gamma - b^\gamma$ for $\gamma >= 1$. Note that informativeness of both arms are the same and the best estimate can be found when we observe the rewards from the same arm. Therefore, the best estimator of (14) is $\mu_1^{-1}(\hat{X}_{1,t}) = (\hat{X}_{1,t})^{\frac{1}{\gamma}}$ and best estimator of (15) is $\hat{\theta}_t = \mu_2^{-1}(\hat{X}_{2,t}) = (1 - \hat{X}_{2,t})^{\frac{1}{\gamma}}$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{r}_t(\theta^* + \Delta)\right] \\
&\geq \frac{a_1\Delta}{2}\Big(\Pr(\hat{X}_{1,t} - \theta_*^\gamma \leq -\Delta^\gamma) \\
&+ \Pr(\hat{X}_{2,t} - (1 - \theta_*^\gamma) \geq \Delta^\gamma)\Big). \qquad (16)
\end{aligned}$$

Define two processes $\nu_1 = \text{Ber}(\theta_*^\gamma) \otimes \text{Ber}(\theta_*^\gamma - \Delta^\gamma)$ and $\nu_2 = \text{Ber}(\theta_*^\gamma + \Delta^\gamma) \otimes \text{Ber}(\theta_*^\gamma)$ where $\nu_1 \otimes \nu_2$ denotes the product distribution. Let $\Pr_\nu$ denotes probability associated with distribution ν. Then, (16) is equivalent to

$$\text{Reg}(\theta^* + \Delta, T) \geq \frac{a_1\Delta}{2}\sum_{t=1}^{T} \Pr_{\nu_1^{\otimes t}}(I_t = 2) + \Pr_{\nu_2^{\otimes t}}(I_t = 1),$$

where $\nu^{\otimes t}$ is the t times product distribution of ν. Using well-known lower bounding techniques for the minimax risk of hypothesis testing [41], we have

$$\text{Reg}(\theta^* + \Delta, T) \geq \frac{a_1\Delta}{4}\sum_{t=1}^{T} \exp(-\text{KL}(\nu_1^{\otimes t}, \nu_2^{\otimes t})), \qquad (17)$$

where

$$\begin{aligned}
&\text{KL}(\nu_1^{\otimes t}, \nu_2^{\otimes t}) = t\Big(\text{KL}(\text{Ber}(\theta_*^\gamma), \text{Ber}(\theta_*^\gamma + \Delta^\gamma)) \\
&+ \text{KL}(\text{Ber}(\theta_*^\gamma - \Delta^\gamma), \text{Ber}(\theta_*^\gamma)\Big). \qquad (18)
\end{aligned}$$

By using the fact $\text{KL}(p, q) \leq \frac{(p-q)^2}{q(1-q)}$ [42], we can further bound (17) by

$$\begin{aligned}
\text{Reg}(\theta^* + \Delta, T) &\geq \frac{a_1\Delta}{4}\sum_{t=1}^{T} \exp\left(-\frac{2t\Delta^{2\gamma}}{(\theta_*^\gamma + \Delta^\gamma)(1 - \theta_*^\gamma - \Delta^\gamma)}\right) \\
&\geq \frac{S_3}{\Delta^{2\gamma-1}},
\end{aligned}$$

where $S_3 = \frac{1}{64\gamma}(1 - \exp(-16))$ for $T \geq 8$. Then, by setting $\Delta = T^{-\frac{1}{2\gamma}}$, we have

$$\text{Reg}(\theta^* + \Delta, T) \geq \frac{1}{64\gamma}(1 - \exp(-16))T^{1-\frac{1}{2\gamma}}.$$

M. Proof of Theorem 7

First, we show that $|\hat{\theta}_t - \theta_*| = \epsilon$ implies $|\hat{\Delta}_t - \Delta_*| \leq \epsilon$. Four possible cases for $\hat{\Delta}_t$):

- θ_* and $\hat{\theta}_t$ lie in the same optimality interval and Δ_* and $\hat{\Delta}_t$ are computed with respect to the same boundary of that interval.
- θ_* and $\hat{\theta}_t$ lie in the same optimality interval and Δ_* and $\hat{\Delta}_t$ are computed with respect to the different boundaries of that interval.
- θ_* and $\hat{\theta}_t$ lie in adjacent optimality intervals.
- θ_* and $\hat{\theta}_t$ lie in non-adjacent optimality intervals.

In the first case, $|\hat{\theta}_t - \theta_*| = |\hat{\Delta}_t - \Delta_*| = \epsilon$. In the second case, $\hat{\Delta}_t$ can not be larger than $\Delta_* + \epsilon$ since in that case $\hat{\theta}_t$ would be computed with respect to the same boundary of that interval. Similarly, $\hat{\Delta}_t$ can not be smaller than $\Delta_* - \epsilon$ since in that case θ_* would be computed with respect to the same boundary of that interval. In the third and fourth cases, since $|\hat{\theta}_t - \theta_*| = \epsilon$, $\hat{\Delta}_t \leq \epsilon - \Delta_*$, and hence the difference between $\hat{\Delta}_t$ and Δ_* is smaller than ϵ.

Second, we show that $|\hat{\Delta}_t - \Delta_*| < \bar{D}_1\left(\frac{2K\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}}$ holds with high probability.

$$\begin{aligned}
&\Pr\left(|\hat{\Delta}_t - \Delta_*| \geq \bar{D}_1\left(\frac{K\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}}\right) \\
&\leq \Pr\left(|\hat{\theta}_t - \theta_*| \geq \bar{D}_1\left(\frac{K\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}}\right) \\
&\leq \sum_{k=1}^{K} 2\mathbb{E}\left[\exp\left(-2\left(\frac{\bar{D}_1 K\left(\frac{2\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}}}{\bar{D}_1 K w_k(t)}\right)^{\frac{2}{\bar{\gamma}_1}} N_k(t)\right)\Bigg| N_k(t)\right] \\
&\leq \sum_{k=1}^{K} 2\mathbb{E}\left[\exp\left(-4w_k(t)^{1-\frac{2}{\bar{\gamma}_1}}\log T\right)\Big| w_k(t)\right] \\
&\leq 2KT^{-4}, \qquad (19)
\end{aligned}$$

where the second inequality follows from Lemma 4 and Chernoff-Hoeffding inequality and third inequality by Lemma

5. Then, with probability at least $1-2KT^{-3}$, the following inequalities hold for all $1 \le t \le T$:

$$\Delta_* - 2\bar{D}_1 K \left(\frac{2\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}} \le \hat{\Delta}_t - \bar{D}_1 K \left(\frac{2\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}} \le \Delta_*.$$

For $2\bar{D}_1 K \left(\frac{2\log T}{T}\right)^{\frac{\bar{\gamma}_1}{2}} \le \left(\frac{\Delta_*}{2}\right)$, which is when

$$T \ge C_3(\Delta_*) = 2\left(\frac{4D_1K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \operatorname{glog}\left(2\left(\frac{4D_1K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}}\right),$$

with probability $1-2KT^{-3}$, it holds that

$$\frac{\Delta_*}{2} \le \hat{\Delta}_t - \bar{D}_1 \left(\frac{2\log T}{t}\right)^{\frac{\bar{\gamma}_1}{2}} \le \Delta_*.$$

The BUW follows the recommendations of the UCB algorithm when $T < \max\left(C_3(\Delta_*), C_2(\Delta_*/2)\right) = C_3(\Delta_*)$, and follows the recommendations of the WAGP when $T \ge C_3(\Delta_*)$ with probability at least $1-2KT^{-3}$. On the other hand, with probability at most $2KT^{-3}$, regret of the BUW is at most linear in T. Let define $R^g_{\theta_*}(T_1,T_2)$ denote the regret incurred by algorithm $g \in \{BUW, WAGP, UCB\}$. Then, when $T < C_3(\Delta_*)$, the regret of the WAGP can be written as

$$R^{BUW}_{\theta^*}(0,T) \le R^{UCB}_{\theta^*}(0,T) + KT^{-2}$$

and when $T \ge C_3(\Delta_*)$,

$$R^{BUW}_{\theta^*}(C_3(\Delta_*),T) \le R^{WAGP}_{\theta^*}(C_3(\Delta_*),T) + KT^{-2}$$

This concludes the parameter-dependent regret bound of the BUW.

For the worst-case regret bound, observe that for any Δ_*, $R^{WAGP}_{\theta^*}(C_3(\Delta_*),T) \le \mathcal{O}(1)$ by Theorem 4 and since $C_3(\Delta_*) \ge C_2(\Delta_*)$. Hence, the worst-case is when $T < C_2(\Delta_*)$. We can conclude that the worst-case regret of the BUW is the worst-case regret of the UCB1 algorithm, which is given by $\mathcal{O}(\sqrt{KT\log T})$ [10].

N. Proof of Theorem 8

By Lemma 3 and Jensen's inequality, we have

$$\mathbb{E}\left[r_t(\theta^t_*)\right] \le 2D_2\mathbb{E}\left[|\theta^t_* - \hat{\theta}_t|\right]^{\gamma_2}, \tag{20}$$

where

$$\hat{\theta}_t = \frac{\sum_{k=1}^K N_{k,\rho}(t)\mu_k^{-1}(\hat{X}_{k,\rho,t})}{\tau_\rho(t)},$$

where $\sum_{k=1}^K N_{k,\rho}(t) = \tau_\rho(t)$. Then, by using Lemma 2, we have

$$\mathbb{E}\left[\left|\hat{\theta}_t - \theta^t_*\right|\right] \le \frac{\sum_{k=1}^K \bar{D}_1\mathbb{E}\left[N_{k,\rho}(t)\mathbb{E}\left[\left|\hat{X}_{k,\rho,t} - \mu_k(\theta^t_*)\right| |N_{k,\rho}(t)\right]^{\bar{\gamma}_1}\right]}{\tau_\rho(t)}.$$

Let $\mathcal{S}^{\tau_h}_{k,\rho,t}$ be the set of times times that arm k is chosen in round ρ before time t, i.e.,

$$\mathcal{S}^{\tau_h}_{k,\rho,t} = \{t' \le t : I_{t'} = k, 2(\rho-1)\tau_h \le t' \le 2\rho\tau_h\}.$$

Clearly, $|\mathcal{S}^{\tau_h}_{k,\rho,t}| = N_{k,\rho}(t)$. We have,

$$\hat{X}_{k,\rho,t} = \frac{\sum_{t'\in\mathcal{S}^{\tau_h}_{k,\rho,t}} X_{k,t'}}{N_{k,\rho}(t)},$$

where $\mathbb{E}[X_{k,t'}] = \mu_k(\theta^{t'}_*)$ for all $t' \in \mathcal{S}^{\tau_h}_{k,\rho,t}$. Define a random variable $\tilde{X}_{k,t'} = X_{k,t'} - \mu_k(\theta^{t'}_*)$ for all $t' \in \mathcal{S}^{\tau_h}_{k,\rho,t}$, $k \in \mathcal{K}$ and ρ. Observe that $\{\tilde{X}_{k,t'}\}_{t'\in\mathcal{S}^{\tau_h}_{k,\rho,t}}$ is a random sequence with $\mathbb{E}[\tilde{X}_{k,t'}] = 0$ and $\tilde{X}_{k,t'} \in [-1,1]$ almost surely for all $k \in \mathcal{K}$ and ρ. Then,

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{X}_{k,\rho,t} - \mu_k(\theta^t_*)\right| |N_{k,\rho}(t)\right] \\
&\le \mathbb{E}\left[\left|\frac{\sum_{t'\in\mathcal{S}^{\tau_h}_{k,\rho,t}}(X_{k,t'} - \mu_k(\theta^{t'}_*))}{N_{k,\rho}(t)}\right|\right] \\
&\qquad + \frac{\sum_{t'\in\mathcal{S}^{\tau_h}_{k,\rho,t}} |\mu_k(\theta^{t'}_*) - \mu_k(\theta^t_*)|}{N_{k,\rho}(t)} \\
&\le \mathbb{E}\left[\left|\frac{\sum_{t'\in\mathcal{S}^{\tau_h}_{k,\rho,t}} \tilde{X}_{k,t'}}{N_{k,\rho}(t)}\right|\right] + \frac{\sum_{t'\in\mathcal{S}^{\tau_h}_{k,\rho,t}} 2D_2|\theta^{t'}_* - \theta^t_*|^{\gamma_2}}{N_{k,\rho}(t)},
\end{aligned}$$

where for any $t' \in \mathcal{S}_{k,\rho,t}$, $k \in \mathcal{K}$ and ρ,

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{\sum_{t'\in\mathcal{S}^{\tau_h}_{k,\rho,t}} \tilde{X}_{k,t'}}{N_{k,\rho}(t)}\right|\right] \\
&= \int_{x=0}^{\infty} \Pr\left(\left|\frac{\sum_{t'\in\mathcal{S}^{\tau_h}_{k,\rho,t}} \tilde{X}_{k,t'}}{N_{k,\rho}(t)}\right| > x\right) \mathrm{d}x \\
&\le \int_{x=0}^{\infty} 2\exp(-x^2 N_{k,\rho}(t))\,\mathrm{d}x = \sqrt{\frac{\pi}{N_{k,\rho}(t)}},
\end{aligned} \tag{21}$$

where the inequality follows from the Chernoff-Hoeffding bound and

$$|\theta^t_* - \theta^{t'}_*| \le (2\tau_h/\tau), \tag{22}$$

since for all $t, t' \in \mathcal{S}^{\tau_h}_{k,\rho,t}$ we have $|t-t'| \le 2\tau_h$. Then, using (21) and (22), the expected gap between θ^t_* and $\hat{\theta}_t$ can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[|\theta^t_* - \hat{\theta}_t|\right] \\
&\le \frac{\sum_{k=1}^K \bar{D}_1\mathbb{E}\left[N_{k,\rho}(t)(|\sqrt{\frac{\pi}{N_{k,\rho}(t)}} + 2D_2(\frac{2\tau_h}{\tau})^{\gamma_2}|)^{\bar{\gamma}_1}\right]}{\tau_\rho(t)} \\
&\le \frac{\sum_{k=1}^K \bar{D}_1\mathbb{E}\left[N_{k,\rho}(t)(\frac{\pi}{N_{k,\rho}(t)})^{\frac{\bar{\gamma}_1}{2}}\right]}{\tau_\rho(t)} \\
&\qquad + \frac{2D_2^{\bar{\gamma}_1}(2\tau_h/\tau)^{\alpha\gamma_2\bar{\gamma}_1}N_{k,\rho}(t)}{\tau_\rho(t)} \\
&\le \bar{D}_1((\pi K)^{\frac{\bar{\gamma}_1}{2}}\tau_\rho(t)^{-\frac{\bar{\gamma}_1}{2}} + 2D_2^{\bar{\gamma}_1}(2\tau_h/\tau)^{\bar{\gamma}_1\gamma_2}) \\
&\le \bar{D}_1((\pi K)^{\frac{\bar{\gamma}_1}{2}}\tau_h^{-\frac{\bar{\gamma}_1}{2}} + 2D_2^{\bar{\gamma}_1}(2\tau_h/\tau)^{\bar{\gamma}_1\gamma_2}),
\end{aligned}$$

where the second inequality follows from the fact that $(a+b)^\gamma \le a^\gamma + b^\gamma$ for $a, b > 0$ and $0 < \gamma \le 1$, the third inequality is due to the worst case selection process, i.e., $N_{k,\rho}(t) = \tau_\rho(t)/K$ for all $k \in \mathcal{K}$, and the fourth inequality follows from

the fact that $\tau_\rho(t) \geq \tau_h$. By choosing $\tau_h = \tau^b$, we get the optimal $b = \frac{\gamma_2}{0.5+\gamma_2}$. Then, cumulative regret at time T can be bounded as

$$\sum_{t=1}^{T} \mathbb{E}[r_t(\theta_*^t)] \leq \tau^{-\frac{\gamma_2}{0.5+\gamma_2}} + \left(2D_2\bar{D}_1^{\gamma_2}[(\pi K)^{\frac{\bar{\gamma}_1}{2}} + 2D_2^{\bar{\gamma}_1}]\right)^{\gamma_2} \tau^{-\frac{\gamma_2^2\bar{\gamma}_1}{1+2\gamma_2}},$$

which concludes the proof.

O. Auxiliary Lemma

Lemma 5. *For $\gamma < 0$, $\delta > 0$, the following bound holds for any w_k with $0 \leq w_k \leq 1$ and $\sum_{k=1}^K w_k = 1$:*

$$\sum_{k=1}^{K} \exp(-\delta w_k^\gamma) \leq K \exp(-\delta)$$

Proof. Let $k_{\max} = \max_k w_k$. Then, the following inequalities hold:

$$\begin{aligned}
&\max_{w_k:\sum_{k=1}^K w_k=1,\ 0\leq w_k\leq 1} \sum_{k=1}^{K} \exp\left(-\delta w_k^\gamma\right) \\
&= \max_{w_k:\sum_{k=1}^K w_k=1,\ 0\leq w_k\leq 1} \exp\left(\log\left(\sum_{k=1}^{K} \exp\left(-\delta w_k^\gamma\right)\right)\right) \\
\leq\ & \max_{w_k:\sum_{k=1}^K w_k=1,\ 0\leq w_k\leq 1} \exp\left(\max_{k\in\mathcal{K}}\left(-\delta w_k^\gamma\right) + \log K\right) \\
\leq\ & K \max_{w_k:\sum_{k=1}^K w_k=1,\ 0\leq w_k\leq 1} \exp\left(-\delta w_{k_{\max}}^\gamma\right) \\
\leq\ & K\exp(-\delta)
\end{aligned}$$

where first inequality follows from the following fact

$$\begin{aligned}
&\log\left(\sum_{k=1}^{K} \exp(x_k)\right) \\
&\leq x^* + \log\left(\sum_{k=1}^{K} \exp\left(x_k - x^*\right)\right) \\
&\leq x^* + \log K
\end{aligned}$$

where $x^* = \max_{k\in\mathcal{K}} x_k$. □

REFERENCES

[1] O. Atan, C. Tekin, and M. van der Schaar, "Global multi-armed bandits with hölder continuity," in *Proceedings of 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015, pp. 28–36.

[2] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematic*, vol. 6, no. 1, pp. 4–22, 1985.

[3] A. Mersereau, P. Rusmevichientong, and J. Tsitsiklis, "A structured multiarmed bandit problem and the greedy policy." *Automatic Control, IEEE Transactions on*, vol. 54, pp. 2787–2802, 2009.

[4] T. Lai and H. Robbins, "Adaptive design in regression and control," *Proceedings of the National Academy of Sciences*, vol. 75, no. 2, pp. 586–587, 1978.

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

[6] J. Huang, M. Leng, and M. Parlar, "Demand functions in decision modeling: A comprehensive survey and research directions," *Decision Sciences*, vol. 44, no. 3, pp. 557–609, 2013.

[7] T.-H. Li and K.-S. Song, "On asymptotic normality of nonlinear least squares for sinusoidal parameter estimation," *Signal Processing, IEEE Transactions on*, vol. 56, no. 9, pp. 4511–4515, 2008.

[8] P. Pakrooh, L. L. Scharf, A. Pezeshki, and Y. Chi, "Analysis of fisher information and the cramer-rao bound for nonlinear parameter estimation after compressed sensing," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 6630-6634, Ed., 2013.

[9] R. Iltis, "Density function approximation using reduced sufficient statistics for joint estimation of linear and nonlinear parameters," *Signal Processing, IEEE Transactions on*, vol. 47, no. 8, pp. 2089–2099, 1999.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem." *Machine Learning*, vol. 47, pp. 235–256, 2002.

[11] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," *Journal of Machine Learning Research*, pp. 397–422, 2002.

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

[13] E. Kaufmann, C. O., and A. Garivier, "On bayesian upper confidence bounds for bandit problems," in *Proceedings of 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

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

[15] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012.

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

[17] S. Bubeck and C. Y. Liu, "Prior-free and prior-dependent regret bounds for thompson sampling," in *Advances in Neural Information Processing Systems*, 2013, pp. 638–646.

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

[19] A. Slivkins, "Contextual bandits with similarity information," in *Journal of Machine Learning Research*, vol. 15, 2014, pp. 2533–2568.

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

[21] C. Tekin and M. van der Schaar, "Distributed online learning via cooperative contextual bandits," *Signal Processing, IEEE Transactions on*, vol. 63, no. 14, pp. 3700–3714, 2015.

[22] J. Xu, C. Tekin, S. Zhang, and M. van der Schaar, "Distributed online learning based on global feedback," *Signal Processing, IEEE Transactions on*, vol. 63, no. 9, pp. 2225–2238, 2015.

[23] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *Signal Processing, IEEE Transactions on*, vol. 58, pp. 5547–5567, 2013.

[24] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *Annual Symposium on Foundations of Computer Science*, 1995, pp. 322–331.

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

[26] W. Chu, L. Li, L. Reyzin, and R. E. Schapire, "Contextual bandits with linear payoff functions," in *AISTATS*, vol. 15, 2011, pp. 208–214.

[27] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011, pp. 2312–2320.

[28] P. Rusmevichientong and J. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 5, pp. 395–411, 2010.

[29] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *COLT*, 2008, pp. 355–366.

[30] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari, "Online-to-confidence-set conversions and application to sparse stochastic bandits," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 1–9.

[31] N. Cesa-Bianchi and S. Kakade, "An optimal algorithm for linear bandits," *arXiv preprint arXiv:1110.4322*, 2011.

[32] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework, results and applications," in *International Conference on Machine Learning*, 2013, pp. 151–159.

[33] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *Networking, IEEE/ACM Transactions on (TON)*, vol. 20, no. 5, pp. 1466–1478, 2012.

[34] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *Advances in Neural Information Processing Systems*, 2011, pp. 684–692.
[35] T. Lattimore and R. Munos, "Bounded regret for finite-armed structured bandits," in *Advances in Neural Information Processing Systems*, 2014, pp. 550–558.
[36] D. Russo and B. Van Roy, "An information-theoretic analysis of thompson sampling," *Journal of Machine Learning Research*, 2015.
[37] 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 25th Annual ACM Symposium on the Theory of Computing*, 1993, pp. 382–391.
[38] S. Bubeck and N. Cesa Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Machine Learning*, 2012.
[39] Y. Song, S. Ray, and S. Li, "Structural properties of buyback contracts for price-setting newsvendors," *Manufacturing & Service Operations Management*, vol. 10, no. 1, pp. 1–18, 2008.
[40] E. Chlebus, "An approximate formula for a partial sum of divergent p-series," *Applied Mathematics Letters*, vol. 22, no. 5, pp. 732–737, 2009.
[41] A. B. Tsybakov and V. Zaiats, *Introduction to nonparametric estimation*. Springer, 2009.
[42] P. Rigollet and A. Zeevi, "Nonparametric bandits with covariates," in *Conference on Learning Theory (COLT)*, 2010.