Global Bandits

Onur Atan, Cem Tekin, *Member, IEEE*, and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*— Multiarmed bandits (MABs) 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 works on MAB assume 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* (GB). 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 sublinear, with an exponent that depends on the informativeness of the arms. We also propose a variant of the greedy policy that achieves $\tilde{\mathcal{O}}(\sqrt{T})$ worst case and $\mathcal{O}(1)$ parameter-dependent regret. Finally, we perform experiments on dynamic pricing and show that the proposed algorithms achieve significant gains with respect to the well-known benchmarks.**

***Index Terms*— Bounded regret, informative arms, multiarmed bandits (MABs), online learning, regret analysis.**

I. Introduction

MULTIARMED bandits (MABs) 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 that maximize the total reward, which turns out to be equivalent to minimizing the regret, where the regret is defined as the difference between the total expected reward obtained by an oracle that always selects the best arm based on complete knowledge of arm reward distributions, and that of the learner, who does not know the expected arm rewards beforehand. Classical K-armed MAB [1] 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 an MAB model, which we call 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. For this problem, 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 sublinear 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 [2], 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 nonlinear function of a single global parameter. On the technical side, nonlinear expected reward functions significantly complicate 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 nonlinear, 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 need to be incorporated to ensure that this bias asymptotically converges to 0.

Many applications can be formalized as GBs. Examples include: 1) clinical trials involving similar drugs (e.g., drugs with a similar chemical composition) or treatments that may have similar effects on the patients and 2) dynamic pricing with the objective of maximizing revenue over a finite time horizon.

Manuscript received April 13, 2017; revised December 21, 2017; accepted March 1, 2018. Date of publication April 12, 2018; date of current version November 16, 2018. The work of O. Atan and M. van der Schaar was supported by the NSF under Grant 1533983, Grant 1407712, and Grant 1462245. This paper was presented at the 2015 International Conference on Artificial Intelligence and Statistics (AISTATS), San Diego, CA, USA, May 2015. *(Corresponding author: Onur Atan.)*

O. Atan and M. van der Schaar are with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: oatan@ucla.edu; mihaela@ee.ucla.edu).

C. Tekin is with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: cemtekin@ee.bilkent.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2818742

[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 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 1: Let y_t be the dosage level of the drug for patient t and x_t be the response of patient t. The relationship between the drug dosage and patient response is modeled in [3] as $x_t = M(y_t; \theta_*) + \epsilon_t$, 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 ϵ_t 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.

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 [4]. At instance t, the agent first selects a price $p_t \in \mathcal{P}$ and then observes the amount of sales at time t, which is denoted by $S(p_t; \theta_*)$. We have $S(p_t; \theta_*) = F(p_t; \theta_*) + \epsilon_t$, where $F(.)$ is the modulating function, θ_* is the market size, and ϵ_t is the noise term with zero mean. The modulating function is equal to the purchase probability of an item of price p_t given the market size θ_*. Examples of commonly used modulating functions can be found in [5]. The revenue is then given by $R(p_t; \theta_*) = p_t F(p_t; \theta_*) + p_t \epsilon_t$. In this example, the market size is the unknown global parameter which needs to be learned online by setting prices and observing the related revenues. In Section IX, we illustrate the use of methods proposed in this paper on this dynamic pricing example.

In addition to the above examples, GBs can also be applied in any setting in which the parameters of a system that depends on the rewards in a nonlinear way need to be estimated in order to learn the optimal arms. At this point, it is important to note that our work differs from the existing works on nonlinear parameter estimation [6]–[8], because its focus is to maximize the total reward by using the estimates of the parameter to decide which arms to select.

The remainder of this paper is organized as follows. Contribution and the key results are summarized in Section II. Related work is discussed in Section III. Problem formulation is given in Section IV. A greedy policy is proposed in Section V and its regret is analyzed in Section VI. An improved algorithm that combines the greedy policy with an upper confidence bound policy is proposed in Section VII. Learning under time-varying global parameter is considered in Section VIII. Numerical results are given in Section IX, followed by the concluding remarks given in Section X. All proofs are given in the Appendix.

II. Contribution and Key Results

This paper is an extended version of [9], adding the following contributions. First, it provides two new theoretical results on *weighted-arm greedy policy* (WAGP): mean-squared convergence of the estimated global parameter and a lower bound on the regret. Second, it provides two new algorithms: 1) *Best of UCB and WAGP* (BUW) that switches between the UCB1 and WAGP in order to achieve optimal parameter-dependent and worst case regrets and 2) nonstationary WAGP that tracks the time-varying global parameter to take optimal actions. Third, it provides an illustration of the use of the proposed algorithms on the dynamic pricing example. In addition, this paper has extended introduction and related work sections, and includes proofs of all theorems. Our main contributions can be summarized as follows.

1) We propose a nonlinear parametric model for MABs, which we refer to as GBs, and a greedy policy, referred to as WAGP, which achieves bounded regret.
2) We define the concept of *informativeness*, which measures how well one can estimate the expected reward of an arm by using rewards observed from the other arms, and then, prove a sublinear in time worst case regret bound for WAGP that depends on the informativeness.
3) We also propose another learning algorithm called the BUW, which fuses the decisions of the UCB1 [10] and WAGP in order to achieve $\tilde{\mathcal{O}}(\sqrt{T})$[2] worst case and $\mathcal{O}(1)$ parameter-dependent regrets.
4) We study a nonstationary version of GB, where the global parameter slowly changes over time. For this case, we prove a bound on the time-averaged regret that depends on the speed of change of the global parameter.
5) We simulate our algorithms on a synthetic dynamic pricing data set and show that they beat other state-of-the-art MAB algorithms.

III. Related Work

There is a wide strand of literature on MABs including finite-armed stochastic MAB [1], [10]–[12], the 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: noninformative, group informative, and globally informative MABs.

A. Noninformative MAB

We call an MAB as *noninformative* if the reward observations of any arm do not reveal any information about the rewards of the other arms. Examples of noninformative MABs include finite-armed stochastic [1], [10] and nonstochastic [24] MABs. Lower bounds derived for these settings point out to the impossibility of bounded regret.

B. Group-Informative MAB

We call an MAB as *group-informative* if the reward observations from an arm provides information about a group of other arms. Examples include linear contextual bandits [25], [26], multidimensional 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 at least logarithmically many times in order to learn the unknown parameters.

Another related work [34] studies a setting that interpolates between the bandit (partial feedback) and experts (full feedback) settings. In this setting, the decision maker obtains not

[2] $\mathcal{O}(\cdot)$ is the Big O notation and $\tilde{\mathcal{O}}(\cdot)$ is the same as $\mathcal{O}(\cdot)$ except it hides terms that have polylogarithmic growth.

TABLE I

COMPARISON WITH RELATED WORKS. $\gamma \leq 1$ REPRESENTS THE INFORMATIVENESS, WHICH IS GIVEN IN DEFINITION 1

	GB (our work)	[27]–[31]	[2]	[37]
Parameter dimension	Single	Multi	Single	Multi
Reward functions	Non-linear	Linear	Linear	Generalized linear
Worst-case regret	BUW: $\tilde{\mathcal{O}}(\sqrt{T})$, WAGP: $\mathcal{O}(T^{1-\frac{\gamma}{2}})$	$\tilde{\mathcal{O}}(\sqrt{T})$	$\mathcal{O}(\sqrt{T})$	$\tilde{\mathcal{O}}(\sqrt{T})$
Parameter dependent regret	BUW: $\mathcal{O}(1)$, WAGP: $\mathcal{O}(1)$	$\mathcal{O}\left(\log T\right)$	$\mathcal{O}(1)$	$\mathcal{O}\left((\log T)^3\right)$

only the reward of the selected arm but also an unbiased estimate of the rewards of a subset of the other arms, where this subset is determined by a graph. This is not possible in our setting due to the nonlinear reward structure and bandit feedback.

C. Globally Informative MAB

We call a MAB problem as *globally informative* if the reward observations from an arm provide information about the rewards of all the arms [2], [35]. GB belongs to the class of globally informative MAB and includes the linearly parametrized MAB [2] as a subclass. Hence, our results reduce to the results of [2] 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 through a global parameter results in bounded parameter-dependent and a sublinear worst case regrets.

Table I summarizes our model and theoretical results, and compares them with the existing literature in the parametric MAB models. Although GB is more general than the model in [2], both WAGP and BUW achieve bounded parameter-dependent regret, and BUW is able to achieve the same worst case regret as the policy in [2]. On the other hand, although the linear MAB models are more general than GB, it is not possible to achieve bounded regret in these models.

IV. PROBLEM FORMULATION

A. Arms, Reward Functions, and Informativeness

There are K arms indexed by the set $\mathcal{K} := \{1, \ldots, K\}$. The 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 a distribution $\nu_k(\theta_*)$ with support $\mathcal{X}_k \subseteq [0, 1]$. The expected reward of arm 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. This is formalized in the following assumption.

Assumption 1: We assume the following:

1) For each $k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$ there exists $D_{1,k} > 0$ and $1 < \gamma_{1,k}$, such that
$$|\mu_k(\theta) - \mu_k(\theta')| \geq D_{1,k}|\theta - \theta'|^{\gamma_{1,k}}.$$

2) 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 ensures that the reward functions are monotonic and the second assumption, which is also known Hölder continuity, ensures that the reward functions are smooth. These assumptions imply that the reward functions are invertible and the inverse reward functions are also Hölder continuous. Moreover, they generalize the model proposed in [2], and allow us to model real-world scenarios described in Examples 1 and 2, and propose algorithms that achieve bounded regret.

Some examples of the reward functions that satisfy Assumption 1 are: 1) exponential functions, such as $\mu_k(\theta) = a\exp(b\theta)$, where $a > 0$; 2) linear and piecewise linear functions; and 3) sublinear and superlinear functions in θ, which are invertible in Θ, such as $\mu_k(\theta) = a\theta^\gamma$, where $\gamma > 0$ and $\Theta = [0, 1]$.

Proposition 1: Define $\underline{\mu}_k = \min_{\theta\in\Theta} \mu_k(\theta)$ and $\overline{\mu}_k = \max_{\theta\in\Theta} \mu_k(\theta)$. Under Assumption 1, the following are true: 1) for all $k \in \mathcal{K}$, $\mu_k(\cdot)$ is invertible and 2) for all $k \in \mathcal{K}$ and $x, x' \in [\underline{\mu}_k, \overline{\mu}_k]$:
$$\left|\mu_k^{-1}(x) - \mu_k^{-1}(x')\right| \leq \bar{D}_{1,k}|x - x'|^{\bar{\gamma}_{1,k}}$$
where $\bar{\gamma}_{1,k} = (1/\gamma_{1,k})$ and $\bar{D}_{1,k} = (1/D_{1,k})^{(1/\gamma_{1,k})}$.

Invertibility of the reward functions allows us to use the rewards obtained from an arm to estimate the expected rewards of other arms. Let $\bar{\gamma}_1$ and γ_2 be the minimum exponents and $\bar{D}_1$, D_2 be the maximum constants, that is
$$\bar{\gamma}_1 = \min_{k\in\mathcal{K}} \bar{\gamma}_{1,k}, \quad \gamma_2 = \min_{k\in\mathcal{K}} \gamma_{2,k}$$
$$\bar{D}_1 = \max_{k\in\mathcal{K}} \bar{D}_{1,k}, \quad D_2 = \max_{k\in\mathcal{K}} D_{2,k}.$$

Definition 1: The informativeness of arm k is defined as $\gamma_k := \bar{\gamma}_{1,k}\gamma_{2,k}$. The informativeness of the GB instance is defined as $\gamma := \bar{\gamma}_1\gamma_2$.

The informativeness of arm k measures the extent of information that can be obtained about the expected rewards of other arms from the rewards observed from arm k. As we will show later, when the informativeness is high, one can form better estimates of the expected rewards of other arms by using the rewards observed from arm k.

B. Definition of the Regret

The learner knows $\mu_k(\cdot)$ for all $k \in \mathcal{K}$ but does not know θ_*. At each time t, it selects one of the arms, denoted by I_t, and receives the random reward $X_{I_t,t}$. The learner's goal is to maximize its cumulative reward up to any time T.

Let $\mu^*(\theta) := \max_{k\in\mathcal{K}} \mu_k(\theta)$ be the maximum expected reward and $\mathcal{K}^*(\theta) := \{k \in \mathcal{K} : \mu_k(\theta) = \mu^*(\theta)\}$ be the optimal set of arms for parameter θ. In addition, let $k^*(\theta)$ denote an arm that is optimal for parameter θ. We refer the policy that selects one of the arms in $\mathcal{K}^*(\theta_*)$ as the *oracle* policy. The learner incurs a regret (loss) at each time it deviates from the oracle policy. We define the one-step regret at time t as the difference between the expected reward of the oracle policy and the learner, which is given by $r_t(\theta_*) := \mu^*(\theta_*) - \mu_{I_t}(\theta_*)$.

Based on this, the cumulative regret of the learner by time T (also referred to as the regret hereafter) is defined as

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

Maximizing the reward is equivalent to minimizing the regret. In the seminal work by Lai and Robbins [3], it is shown that the regret becomes infinite as T grows for the classical K-armed bandit problem. On the other hand, $\lim_{T\to\infty} \mathrm{Reg}(\theta_*, T) < \infty$ will imply that the learner deviates from the oracle policy only finitely many times. In Section V, we prove that this holds for GB.

V. Weighted-Arm Greedy Policy

In this section, we propose a greedy policy called WAGP. The pseudocode of WAGP is given in Algorithm 1. The WAGP consists of two phases: arm selection phase and parameter update phase.

Algorithm 1 WAGP

```
Inputs: $\mu_k(\cdot)$ for each arm $k$
Initialization: $w_k(0) = 0, \hat{\theta}_{k,0} = 0, \hat{X}_{k,0} = 0, N_k(0) = 0$ for all $k \in \mathcal{K}$, $t = 1$
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 $k \in \mathcal{K} \setminus I_t$
    $\hat{X}_{I_t,t} = \frac{N_{I_t}(t-1)\hat{X}_{I_t,t-1}+X_{I_t,t}}{N_{I_t}(t-1)+1}$
    $\hat{\theta}_{k,t} = \arg\min_{\theta\in\Theta} |\mu_k(\theta) - \hat{X}_{k,t}|$ for all $k \in \mathcal{K}$
    $N_{I_t}(t) = N_{I_t}(t-1) + 1$
    $N_k(t) = N_k(t-1)$ 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
```

Let $N_k(t)$ denote the number of times arm k is selected until time t, $\hat{X}_{k,t}$ denote the reward estimate, $\hat{\theta}_{k,t}$ denote the global parameter estimate, and $w_k(t)$ denote the weight of arm k at time t. Initially, all the counters and estimates are set to zero. In the arm selection phase at time $t > 1$, the WAGP selects the arm 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 estimate of the global parameter calculated at the end of time $t-1$.[3,4]

In the parameter update phase, the WAGP updates: 1) the estimated reward of selected arm I_t, denoted by $\hat{X}_{I_t,t}$; 2) the global parameter estimate of the selected arm I_t, denoted by $\hat{\theta}_{I_t,t}$; 3) the global parameter estimate $\hat{\theta}_t$; and 4) the counters $N_k(t)$. The reward of estimate of arm I_t is updated as

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

The reward estimates of the other arms are not updated. The WAGP constructs estimates of the global parameter from the rewards of all the arms and combines their estimates using a weighted sum. 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., $\hat{\theta}_{I_t,t} = \arg\min_{\theta\in\Theta} |\mu_{I_t}(\theta) - \hat{X}_{I_t,t}|$. Then, 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.

VI. Regret Analysis of the WAGP

A. Preliminaries for the Regret Analysis

In this section, 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

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

Note that Θ_k can be written as union of intervals in each of which arm k is optimal. Each such interval is called *optimality interval*. 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 $k \in \mathcal{K}$. The *suboptimality gap* of arm $k \in \mathcal{K}$ given global parameter $\theta_* \in \Theta$ is defined as $\delta_k(\theta_*) := \mu^*(\theta_*) - \mu_k(\theta_*)$. The *minimum suboptimality gap* given global parameter $\theta_* \in \Theta$ is defined as $\delta_{\min}(\theta_*) := \min_{k\in\mathcal{K}\setminus\mathcal{K}^*(\theta_*)} \delta_k(\theta_*)$.

Let $\Theta^{\mathrm{sub}}(\theta_*)$ be the suboptimality region of the global parameter θ_*, which is defined as the subset of the parameter space in which none of the arms in $\mathcal{K}^*(\theta_*)$ is optimal, that is

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

We will show that as time proceeds, the global parameter estimate will converge to θ_*. However, if θ_* lies close to $\Theta^{\mathrm{sub}}(\theta_*)$, the global parameter estimate may fall into the suboptimality region for a large number of times, thereby resulting in a large regret. In order to bound the expected number of times this happens, we define the *suboptimality distance* as the smallest distance between the global parameter and the suboptimality region.

[3] The ties are broken randomly.

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

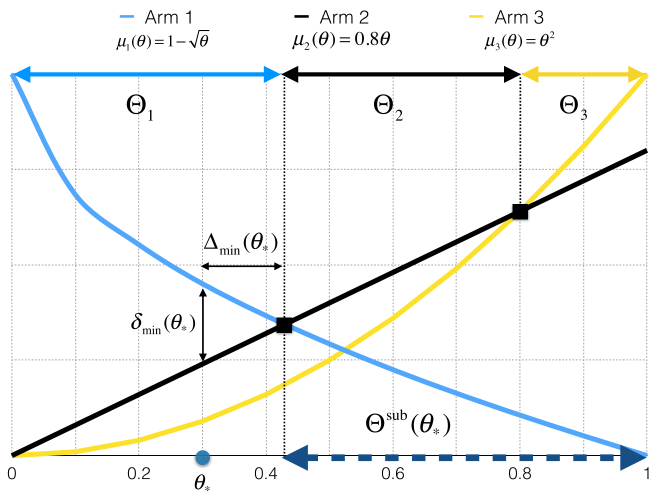


Fig. 1. Illustration of the minimum suboptimality gap and the suboptimality distance.

TABLE II
FREQUENTLY USED NOTATIONS IN REGRET ANALYSIS

$\mathcal{K}^*(\theta_*)$	set of optimal arms for θ_*
$\mu^*(\theta_*)$	expected reward of optimal arms
I_t	selected arm at time t
$\hat{\theta}_t$	global parameter estimate at time t
$\delta_* = \delta_{\min}(\theta_*)$	minimum suboptimality gap
$\Delta_* = \Delta_{\min}(\theta_*)$	minimum suboptimality distance
Θ_k	optimality region of arm k
$\Theta^{\text{sub}}(\theta_*)$	suboptimality region of θ_*
γ	informativeness of the arms

Definition 2: For a given global parameter θ_*, the *suboptimality distance* is defined as

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

From the definition of the suboptimality distance, it is evident that the proposed policy always selects an optimal arm in $\mathcal{K}^*(\theta_*)$ when $\hat{\theta}_t$ is within $\Delta_{\min}(\theta_*)$ of θ_*. For notational brevity, we also use $\Delta_* := \Delta_{\min}(\theta_*)$ and $\delta_* := \delta_{\min}(\theta_*)$. An illustration of the suboptimality gap and the suboptimality distance is given in Fig. 1 for the case with three arms and reward functions $\mu_1(\theta) = 1 - \sqrt{\theta}$, $\mu_2(\theta) = 0.8\theta$, and $\mu_3(\theta) = \theta^2$, $\theta \in [0, 1]$.

The notations frequently used in the regret analysis are highlighted in Table II.

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

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

Theorem 2: Under Assumption 1, we have for WAGP $\mathbb{E}[r_t(\theta_*)] \leq \mathcal{O}(t^{-(\gamma/2)})$.

Theorem 2 proves that the expected one-step regret of the WAGP converges to zero.[5] This is a *worst case* bound in the sense that it holds for any θ_*. Using this result, we derive the following worst case regret bound for the WAGP.

Theorem 3: Under Assumption 1, the worst case regret of WAGP is

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

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 informativeness γ. When the reward functions are linear or piecewise linear, we have $\gamma = 1$, which is an extreme case of our model; hence, the worst case regret is $\mathcal{O}(\sqrt{T})$, which matches with: 1) the worst case regret bound of the standard MAB algorithms in which a linear estimator is used [38] and 2) the bounds obtained for the linearly parametrized bandits [2].

C. Parameter-Dependent Regret Bounds for the WAGP

In this section, we bound the parameter-dependent regret of the WAGP. First, we introduce several constants that will appear in the regret bound.

Definition 3: $C_1(\Delta_*)$ is the smallest integer τ such that $\tau \geq (\bar{D}_1\ K/\Delta_*)^{(2/\bar{\gamma}_1)}(\log(\tau)/2)$ and $C_2(\Delta_*)$ is the smallest integer τ such that $\tau \geq (\bar{D}_1\ K/\Delta_*)^{(2/\bar{\gamma}_1)}\log(\tau)$.

Closed-form expressions for these constants can be obtained in terms of the *glog* function [39], for which the following equivalence holds: $y = \text{glog}(x)$ if and only if $x = (\exp(y)/y)$. Then, we have

$$C_1(\Delta_*) = \left\lceil \frac{1}{2}\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \text{glog}\left(\frac{1}{2}\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}}\right)\right\rceil$$

$$C_2(\Delta_*) = \left\lceil \left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \text{glog}\left(\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}}\right)\right\rceil.$$

Next, we define the expected regret incurred between time steps T_1 and T_2 given θ_* as $R_{\theta_*}(T_1, T_2) := \sum_{t=T_1}^{T_2} \mathbb{E}[r_t(\theta_*)]$. The following theorem bounds the parameter-dependent regret of the WAGP.

Theorem 4: Under Assumption 1, the regret of the WAGP is bounded as follows.

1) For $1 \leq T < C_1(\Delta_*)$, the regret grows sublinearly in time, that is

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

where S_1 and S_2 are constants that are independent of the global parameter θ_*, whose exact forms are given in Appendix F.

[5] 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.

2) For $C_1(\Delta_*) \le T < C_2(\Delta_*)$, the regret grows logarithmically in time, that is

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

3) For $T \ge C_2(\Delta_*)$, the growth of the regret is bounded, that is

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

Thus, we have $\lim_{T\to\infty} \text{Reg}(\theta_*, T) < \infty$, i.e., $\text{Reg}(\theta_*, T) = \mathcal{O}(1)$.

Theorem 4 shows that the regret is inversely proportional to the suboptimality distance Δ_*, which depends on θ_*. The regret bound contains three regimes of growth: initially, the regret grows sublinearly until time threshold $C_1(\Delta_*)$. After this, it grows logarithmically until time threshold $C_2(\Delta_*)$. Finally, the growth of the regret is bounded after time threshold $C_2(\Delta_*)$. In addition, 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. It is also possible to 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 $\gamma = 1$ and $\mathcal{O}(T^{1-\gamma})$ when $\gamma < 1$ (see [9]).

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

Theorem 5 implies that a suboptimal arm is selected by the WAGP only finitely many times. This is the major difference between GB and the classical MAB [1], [10], [36], in which every arm needs to be selected infinitely many times asymptotically by any *good* learning algorithm.

Remark 1: Assumption 1 ensures that the parameter-dependent regret is bounded. When this assumption is relaxed, bounded regret may not be achieved, and the best possible regret becomes logarithmic in time. For instance, consider the case when the reward functions are constant over the global parameter space, i.e., $\mu_k(\theta_*) = m_k$ for all $\theta_* \in [0, 1]$, where m_k is a constant. This makes the reward functions noninvertible. In this case, the learner cannot use the rewards obtained from the other arms when estimating the rewards of arm k. Thus, it needs to learn m_k of each arm separately, which results in logarithmic in time regret when a policy, such as UCB1 [10], is used. This issue still exists even when there are only finitely many possible solutions to $\mu_k(\theta_*) = x$ for some x, in which case some of the arms should be selected at least logarithmically many times to rule out the incorrect global parameters.

D. Lower Bound on the Worst Case Regret

Theorem 3 shows that the worst case regret of the WAGP is $\mathcal{O}(T^{1-\frac{\gamma}{2}})$, which implies that the regret decreases with γ. In this section, we give lower bounds on the parameter-dependent and the worst case regrets.

Theorem 6: For $T \ge 8$ and any policy, the parameter-dependent regret is lower bounded by $\Omega(1)$ and the worst case regret is lower bounded by $\Omega(\sqrt{T})$.

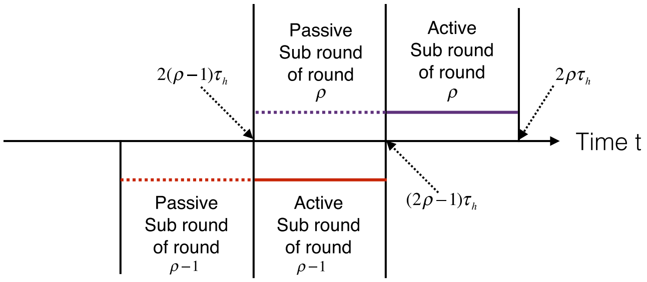


Fig. 2. Operation of the nonstationary WAGP.

The above-mentioned theorem raises a natural question: can we achieve both $\tilde{\mathcal{O}}(\sqrt{T})$ worst case regret (such as the UCB-based MAB algorithms [10]) and bounded parameter-dependent regret by using a combination of UCB and WAGP policies? We answer this question in the affirmative in Section VII.

VII. Best of the UCB and the WAGP

In this section, we propose the BUW, which combines the UCB1 and the WAGP to achieve bounded parameter-dependent and $\mathcal{O}(\sqrt{T})$ worst case regrets. In the worst case, the WAGP achieves $\mathcal{O}(T^{1-\gamma/2})$ regret, which is weaker than $\tilde{\mathcal{O}}(\sqrt{T})$ worst case regret of UCB1. On the other hand, the WAGP achieves bounded parameter-dependent regret, whereas UCB1 achieves a logarithmic parameter-dependent regret. In this section, we propose an algorithm which combines these two algorithms and achieves both $\tilde{\mathcal{O}}(\sqrt{T})$ 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 $O(T^{1-\gamma/2})$ regret when $1 < T < C_1(\Delta_*)$. If the BUW could follow the recommendations of UCB1 when $T < C_1(\Delta_*)$ and the recommendations of the WAGP when $T \ge C_1(\Delta_*)$, then it will achieve a worst case regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$ and bounded parameter-dependent regret. The problem with this approach is that the 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 data-dependent estimate $\tilde{\Delta}_t$ is given as

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

where

$$\hat{\Delta}_t = \Delta_{\min}(\hat{\theta}_t) = \begin{cases} \inf_{\theta' \in \Theta^{\text{sub}}(\hat{\theta}_t)} |\hat{\theta}_t - \theta'| & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) \neq \emptyset \\ 1 & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) = \emptyset. \end{cases}$$

The pseudocode for the BUW is given in Fig. 2. The regret bounds for the BUW are given in Theorem 7.

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

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

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

Algorithm 2 BUW

```
Inputs: $T$, $\mu_k(\cdot)$ for each arm $k$.
Initialization: Select each arm once for $t = 1, 2, \dots, K$, compute $\hat{\theta}_{k,K}$, $N_k(K)$, $\hat{\mu}_k$, $\hat{X}_{k,K}$ for all $k \in \mathcal{K}$, and $\hat{\theta}_K$, $\hat{\Delta}_K$, $\tilde{\Delta}_K$, $t = K + 1$
while $t \geq K + 1$ do
    if $t < C_2\left(\max\left(0, \tilde{\Delta}_{t-1}\right)\right)$ then
        $I_t \in \arg\max_{k \in \mathcal{K}} \hat{X}_{k,t-1} + \sqrt{\frac{2\log(t-1)}{N_k(t-1)}}$
    else
        $I_t \in \arg\max_{k \in \mathcal{K}} \mu_k(\hat{\theta}_{t-1})$
    end if
    Update $\hat{X}_{I_t,t}$, $N_k(t)$, $w_k(t)$, $\hat{\theta}_{k,t}$, $\hat{\theta}_t$ as in the WAGP
    Solve
        $\hat{\Delta}_t = \begin{cases} \inf_{\theta' \in \Theta^{\text{sub}}(\hat{\theta}_t)} |\hat{\theta}_t - \theta'| & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) \neq \emptyset \\ 1 & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) = \emptyset \end{cases}$
    $\tilde{\Delta}_t = \hat{\Delta}_t - \bar{D}_1 \; K \left(\frac{\log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}$
end while
```

1) For $1 \leq T < C_2(\Delta_*/3)$, the regret grows logarithmically in time, that is

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

2) For $T \geq C_2(\Delta_*/3)$, the growth of the regret is bounded, that is

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

The BUW achieves the lower bound given in Theorem 6, that is, $\mathcal{O}(1)$ parameter-dependent regret and $\tilde{\mathcal{O}}(\sqrt{T})$ worst case regret.

VIII. Extension: Learning Under Time-Varying Global Parameter

In this section, we consider the case when the global parameter slowly changes over time.

A. Time-Varying Global Parameter

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

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

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

In the static global parameter model, we were able to bound the parameter-dependent regret with a finite constant number (independent of time horizon T) and the worst-case regret with a sublinear function of time. However, when the global parameter is changing, it is not possible to obtain these bounds. Therefore, we focus on the average regret, which is given as

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

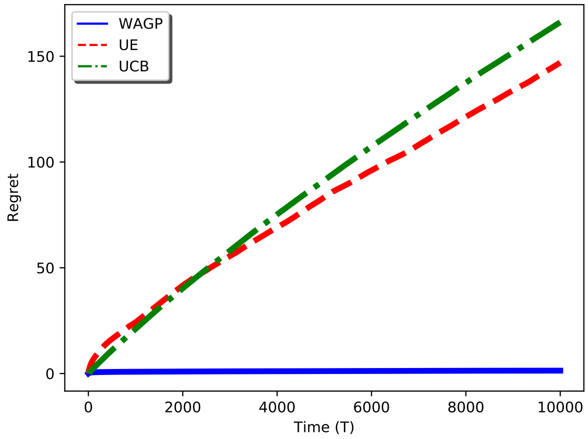


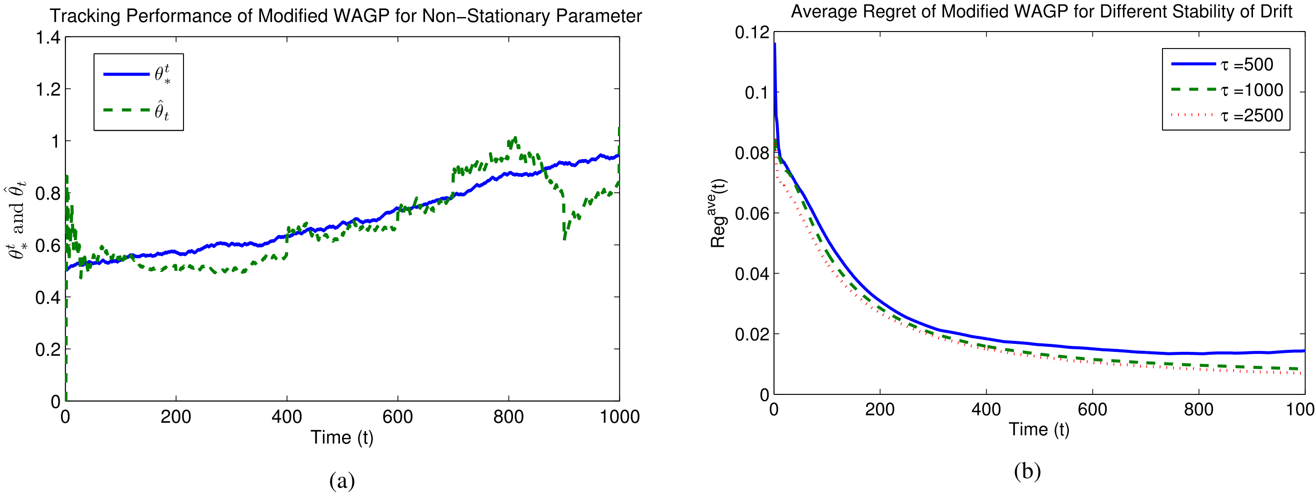
Fig. 3. Comparison of UCB1, UE, and the WAGP for dynamic pricing example on 10 000 samples.

The WAGP needs to be modified to handle the nonstationary global parameter since the optimal arms $\mathcal{K}^*(\theta_*^t)$ may change over time.

B. Description and Regret of the Nonstationary WAGP

The nonstationary WAGP uses only a recent past window of reward observations when estimating the global parameter [40]. 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 nonstationary WAGP groups the time steps into rounds $\rho = 1, 2, \dots$, 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 subrounds. The first subround is called passive subround, whereas the second one is called the active subround. The first round, $\rho = 0$, is an exception where it is both an active and a passive subround.

A different instance of the modified WAGP is run in each round. 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 subround 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 set 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.



(a)

(b)

Fig. 4. Performance of the modified WAGP for a nonstationary global parameter. (a) Tracking performance of modified WAGP. (b) Expected regret of modified WAGP.

Similar to static parameter case, the WAGP selects the arm with the highest estimated reward. Let $\hat{\theta}_{k,\rho,t}$ denote the parameter estimate from arm k at round ρ at time t, which is given as $\arg\min_{\theta\in\Theta} |\mu_k(\theta) - \hat{X}_{k,\rho,t}|$.

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

Theorem 8: Under Assumptions 1 and 2, when the half window length of the nonstationary WAGP is set to $\tau_h = \lceil \tau^{(\gamma_2)/((\gamma_2+0.5))} \rceil$, the average regret is $\mathrm{Reg}^{\mathrm{ave}}(T) \leq \mathcal{O}(\tau^{(-\gamma\gamma_2)/((2\gamma_2+1))})$.

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. Note that the tracking performance of nonstationary WAGP depends on the informativeness, because it is directly related to learning rate of the global parameter.

IX. Illustrative Results: Dynamic Pricing Example

To the best of our knowledge, there are currently no public benchmarks to test bandit algorithms on real world data. This is because the real world data does not contain the rewards of the arms that are not selected in the real time—the counterfactuals. Hence, bandit algorithms are generally tested on synthetic data sets [2], [35], [37].

A. Synthetic Dynamic Pricing Data

We perform experiments on synthetic data inspired by the dynamic pricing example formulated in Section I. We assume that the expected sales $S_{p,t}$ at time t under price p are of the form $\mathbb{E}[S_{p,t}] = (1 - p\theta_*)^2$, where θ_* characterizes the market size, and is set to 0.4. Note that this is the linear-power demand model used in [5] and [41]. The expected revenue is $\mathbb{E}[R_{p,t}] = p(1 - p\theta_*)^2$. Note that the reward function is $\mu_p = \mu_p(\theta_*) = p(1 - p\theta_*)^2$ 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., $R_{p,t} \sim \mathrm{Beta}(1, (1 - \mu_p)/\mu_p)$, and hence, $\mathbb{E}[R_{p,t}] = \mu_p$. We set the arms to be $\{0.4, 0.45, \ldots, 0.95\}$ so $K = 12$.

B. Results

1) Experiment 1 (Comparison): We compare our algorithm with two different benchmarks: UCB1 [10] and uncertainty ellipsoid (UE) [28]. UCB1 treats each arm independently and learn their expected rewards by exploration. UE is proposed for linearly parametrized reward structure with high-dimensional parameter space. In our setting, UE can be used by setting an arm vector $u_p = [p, p^2, p^3]$ in order to fit a polynomial with order 3 for the expected rewards. We generate rewards according to the above-mentioned setting and average the results over 100 iterations. Fig. 3 shows that the WAGP significantly outperforms UCB1 by exploiting the correlations between the arms. The significant performance advantage obtained by the WAGP as compared to UCB1 is due to the fact that the WAGP is able to focus on good arms early on whereas UCB1 learns each arm separately. The WAGP selects arm 10 (the best arm) at 81.7% of time, arm 9 (the second best arm) at 16.4% of time, and the rest of the arms at 1.9% of time. UE outperforms UCB1 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, UE needs to select three linearly independent arms in order to learn about the parameter.

2) Experiment 2 (Effect of the Suboptimality Distance): Table III shows the regret of the WAGP for different θ_* and hence different Δ_*. From this, it can be seen that the regret of the WAGP is indeed decreasing with the suboptimality distance as predicted by Theorem 4.

3) Experiment 3 (Nonstationary Parameter): In this section, we show the performance of the proposed methods for a nonstationary setting. The expected revenue for price p at time t is given by $\mathbb{E}[R_{p,t}] = p(1 - p\theta_*^t)^2$. We assume that $\theta_1^* = 0.5$ and $\theta_*^t = \theta_*^{t-1} + Y_t/\tau$, where Y_t is a random variable with

TABLE III

REGRET OF THE WAGP FOR DIFFERENT VALUES OF θ_* ON 10 000 SAMPLES

θ_*	0.2	0.1	0.3	0.8	0.5
Δ_*	0.17	0.1	0.07	0.02	0.01
Regret	0.3	0.65	0.72	2.02	2.47

TABLE IV

REGRETS OF WAGP, UCB1, AND UE ON 10 000 SAMPLES FOR DIFFERENT λ VALUES

$\lambda/Algorithm$	WAGP	UCB1	UE
$\lambda = 0.01$	1.58	164.85	162.18
$\lambda = 0.05$	10.07	169.47	291.40
$\lambda = 0.1$	32.68	164.38	326.77

$\Pr(Y_t = 1) = 0.6$ and $\Pr(Y_t = -1) = 0.4$ and $\tau > 0$. Hence

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

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

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

4) Experiment 4 (Nonideal Model): We show the performance of the WAGP when the revenue of the price p deviates from the expected revenue from the model due to unobserved/unmeasured covariates or unexpected events. Let $R_{p,t} \sim \text{Beta}(1, (1 - \tilde\mu_p(\theta_*))/\tilde\mu_p(\theta_*))$, where $\tilde\mu_p(\theta_*) = \mu_p(\theta_*) + b_p$ and $b_p \sim \text{Uniform}[-\lambda, \lambda]$ denotes the shift from the model due to some unobserved covariates. Table IV shows the regret for different values of λ averaged over 100 different iterations, where the model is regenerated in each iteration. As seen from the table, the WAGP outperforms UCB1 and UE algorithms by exploiting (nonideal) structure in the model.

X. CONCLUSION

In this paper, we introduce a new class of MAB problems called GB. This general class encompasses the previously introduced linearly parametrized bandits as a special case. We proved that the regret of the GB has three regimes, which we had 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 worst-case regret bound, which grows 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.

APPENDIX

A. Preliminaries

In all the proofs given below, 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 $\boldsymbol{w}(t) = \boldsymbol{N}(t)/t$. 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

The following arguments hold.

1) Let k and $\theta \neq \theta'$ be arbitrary. Then, by Assumption 1
$$|\mu_k(\theta) - \mu_k(\theta')| \ge D_{1,k}|\theta - \theta'|^{\gamma_{1,k}} > 0$$
and hence $\mu_k(\theta) \neq \mu_k(\theta')$.
2) Suppose $x = \mu_k(\theta)$ and $x' = \mu_k(\theta')$ for some arbitrary θ and θ'. Then, by Assumption 1
$$|x - x'| \ge D_{1,k}|\mu_k^{-1}(x) - \mu_k^{-1}(x')|^{\gamma_{1,k}}.$$

C. Preliminary Results

Lemma 1: For the WAGP the following relation between $\hat\theta_t$ and θ_* holds with probability one: $|\hat\theta_t - \theta_*| \le \sum_{k=1}^K w_k(t)\bar D_1|\hat X_{k,t} - \mu_k(\theta_*)|^{\bar\gamma_1}$.

Proof: Before deriving a bound of gap between the global parameter estimate and true global parameter at time t, we let $\tilde\mu_k^{-1}(x) = \arg\min_{\theta\in\Theta}|\mu_k(\theta) - x|$. By monotonicity of $\mu_k(\cdot)$ and Proposition 1, we have $|\tilde\mu_k^{-1}(x) - \tilde\mu_k^{-1}(x')| \le \bar D_1|x - x'|^{\bar\gamma_1}$. Then

$$\begin{aligned}|\theta_* - \hat\theta_t| &= \left|\sum_{k=1}^K w_k(t)\hat\theta_{k,t} - \theta_*\right| = \sum_{k=1}^K w_k(t)|\theta_* - \hat\theta_{k,t}| \\ &\le \sum_{k=1}^K w_k(t)\left|\tilde\mu_k^{-1}(\hat X_{k,t}) - \tilde\mu_k^{-1}(\tilde\mu_k(\theta_*))\right| \\ &\le \sum_{k=1}^K w_k(t)\bar D_1|\hat X_{k,t} - \mu_k(\theta_*)|^{\bar\gamma_1}\end{aligned}$$

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

Lemma 2: The one-step regret of the WAGP is bounded by $r_t(\theta_*) = \mu^*(\theta_*) - \mu_{I_t}(\theta_*) \le 2D_2|\theta_* - \hat\theta_{t-1}|^{\gamma_2}$ with probability one, for $t \ge 2$.

Proof: Note that $I_t \in \arg\max_{k\in\mathcal{K}} \mu_k(\hat\theta_{t-1})$. Therefore, we have

$$\mu_{I_t}(\hat\theta_{t-1}) - \mu_{k^*(\theta_*)}(\hat\theta_{t-1}) \ge 0. \qquad (1)$$

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

$$\begin{aligned}&\mu^*(\theta_*) - \mu_{I_t}(\theta_*) \\ &\quad= \mu_{k^*(\theta_*)}(\theta_*) - \mu_{I_t}(\theta_*) \\ &\quad\le \mu_{k^*(\theta_*)}(\theta_*) - \mu_{I_t}(\theta_*) + \mu_{I_t}(\hat\theta_{t-1}) - \mu_{k^*(\theta_*)}(\hat\theta_{t-1}) \\ &\quad= \mu_{k^*(\theta_*)}(\theta_*) - \mu_{k^*(\theta_*)}(\hat\theta_{t-1}) + \mu_{I_t}(\hat\theta_{t-1}) - \mu_{I_t}(\theta_*) \\ &\quad\le 2D_2|\theta_* - \hat\theta_{t-1}|^{\gamma_2}\end{aligned}$$

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

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.

Lemma 3: For WAGP, we have

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

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

Proof: Observe that

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

where the first inequality follows from Lemma 1. Then

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

□

D. Proof of Theorem 1

Using Lemma 1, the mean-squared error can be bounded as

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

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

$$\begin{aligned}&\mathbb{E}[|\theta_* - \hat{\theta}_t|^2]\\ &\leq K\bar{D}_1^2\mathbb{E}\left[\sum_{k=1}^{K} w_k^2(t)\mathbb{E}\left[|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1}|\boldsymbol{w}(t)\right]\right]\\ &\leq K\bar{D}_1^2\mathbb{E}\left[\sum_{k=1}^{K} w_k^2(t)\int_{x=0}^{\infty}\Pr(|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1} \geq x|\boldsymbol{w}(t))dx\right] \qquad (3)\end{aligned}$$

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

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

where $\Gamma(\cdot)$ is gamma function. Then, we have

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

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 $\bar{\gamma}_1 \leq 1$.

E. Proof of Theorem 2

By Lemma 2 and Jensen's inequality, we have

$$\mathbb{E}[r_{t+1}(\theta_*)] \leq 2D_2\mathrm{E}[|\theta_* - \hat{\theta}_t|]^{\gamma_2}. \qquad (4)$$

Also by Lemma 1 and Jensen's inequality, we have

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

where $\mathbb{E}[\cdot|\cdot]$ denotes the conditional expectation. Using Hoeffding's inequality, we have for each $k \in \mathcal{K}$

$$\begin{aligned}&\mathbb{E}[|\hat{X}_{k,t} - \mu_k(\theta_*)|\ |\boldsymbol{w}(t)]\\ &\quad= \int_{x=0}^{1}\Pr(|\hat{X}_{k,t} - \mu_k(\theta_*)| > x|\boldsymbol{w}(t))\mathrm{d}x\\ &\quad\leq \int_{x=0}^{\infty} 2\exp(-2x^2 N_k(t))\,\mathrm{d}x \leq \sqrt{\frac{\pi}{2N_k(t)}}. \qquad (6)\end{aligned}$$

Combining (5) and (6), we get

$$\mathbb{E}[|\theta_* - \hat{\theta}_t|] \leq \bar{D}_1\left(\tfrac{\pi}{2}\right)^{\frac{\bar{\gamma}_1}{2}}\frac{1}{t^{\frac{\bar{\gamma}_1}{2}}}\mathbb{E}\left[\textstyle\sum_{k=1}^{K} w_k(t)^{1-\frac{\bar{\gamma}_1}{2}}\right]. \qquad (7)$$

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 $\mathbb{E}[\sum_{k=1}^{K} w_k(t)^{1-(\bar{\gamma}_1/2)}] \leq K^{(\bar{\gamma}_1/2)}$. Then, combining (4) and (7), we have

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

F. Proof of Theorem 3

This bound is a consequence of Theorem 2 and the inequality given in bound, where for $\gamma > 0$ and $\gamma \neq 1$, $\sum_{t=1}^{T} 1/t^{\gamma} \leq 1 + \frac{(T^{1-\gamma}-1)}{1-\gamma}$, that is

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

G. Proof of Theorem 4

We need to bound the probability of the event that $I_t \notin \mathcal{K}^*(\theta_*)$. Since at time $t+1$, the arm with the highest $\mu_k(\hat{\theta}_t)$ is selected by the WAGP, $\hat{\theta}_t$ should lie in $\Theta \setminus \Theta_{k^*(\theta_*)}$ for a suboptimal arm to be selected. Therefore, we can write

$$\{I_{t+1} \notin \mathcal{K}^*(\theta_*)\} = \{\hat{\theta}_t \in \Theta \setminus \Theta_{k^*(\theta_*)}\} \subseteq \mathcal{G}_{\theta_*,\hat{\theta}_t}(\Delta_*). \quad (8)$$

By Lemma 3 and (8), we have

$$\begin{aligned}
&\Pr(I_{t+1} \notin \mathcal{K}^*(\theta_*)) \\
&\quad \leq \sum_{k=1}^{K} \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left(\mathcal{F}^k_{\theta_*,\hat{\theta}_t}\left(\left(\frac{\Delta_*}{w_k(t)\bar{D}_1 K}\right)^{\frac{1}{\bar{\gamma}_1}}\right)\right)|\boldsymbol{N}(t)\right]\right] \\
&\quad \leq \sum_{k=1}^{K} 2\mathbb{E}\left[\exp\left(-2\left(\frac{\Delta_*}{w_k(t)\bar{D}_1 K}\right)^{\frac{2}{\bar{\gamma}_1}} w_k(t)t\right)\right] \\
&\quad \leq 2K\exp\left(-2\left(\frac{\Delta_*}{\bar{D}_1 K}\right)^{\frac{2}{\bar{\gamma}_1}} t\right) \qquad (9)
\end{aligned}$$

where $\mathbb{I}(\cdot)$ is an indicator function which is 1 if the statement is correct and 0 otherwise, the first inequality follows from a union bound, the second inequality is obtained by using the Chernoff–Hoeffding bound, and the last inequality is obtained by using Lemma 4. We have $\Pr(I_{t+1} \notin \mathcal{K}^*(\theta_*)) \leq 1/t$ for $t > C_1(\Delta_*)$ and $\Pr(I_{t+1} \notin \mathcal{K}^*(\theta_*)) \leq 1/t^2$ 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 (9) from $C_1(\Delta_*)$ to T and $C_2(\Delta_*)$ to T, respectively.

H. 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 $\omega \in N^c$, then $\lim_{t\to\infty} I_t(\omega) \in \mathcal{K}^*(\theta_*)$. 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 $I_t \in \mathcal{K}^*(\theta_*)$ for all but a finite number of t.

I. 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 positive integer and rewards are Bernoulli distributed with $X_{1,t} \sim \mathrm{Ber}(\mu_1(\theta))$ and $X_{2,t} \sim \mathrm{Ber}(\mu_2(\theta))$. 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. We can show that

$$\begin{aligned}
|\mu_k(\theta) - \mu_k(\theta')| &\leq D_2|\theta - \theta'| \\
|\mu_k^{-1}(x) - \mu_k^{-1}(x')| &\leq \bar{D}_1|x - x'|^{1/\gamma}.
\end{aligned}$$

Let $\theta^* = 2^{-\frac{1}{\gamma}}$. Consider the 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 lower bounded by

$$\begin{aligned}
&(\theta^* + \Delta)^\gamma - (1 - (\theta^* + \Delta)^\gamma) \\
&\quad = 2(\theta^* + \Delta)^\gamma - 1 \\
&\quad = 2\left((\theta^*)^\gamma + \binom{\gamma}{1}(\theta^*)^{\gamma-1}\Delta + \binom{\gamma}{2}(\theta^*)^{\gamma-2}\Delta^2 + \ldots\right) - 1 \\
&\quad \geq 2\gamma\, 2^{\frac{1-\gamma}{\gamma}}\Delta.
\end{aligned}$$

Similarly, in the second case, the loss due to choosing arm 1 is $2\gamma\, 2^{(1-\gamma/\gamma)}\Delta + \sum_{i=2}^{\gamma}\binom{\gamma}{i}(\theta^*)^{(\gamma-i)}(-\Delta)^i$. Let $A_1(\Delta) = 2\gamma\, 2^{(1-\gamma/\gamma)}\Delta + \sum_{i=2}^{\gamma}\binom{\gamma}{i}(\theta^*)^{(\gamma-i)}(-\Delta)^i$.

Define two processes $\nu_1 = \mathrm{Ber}(\mu_1(\theta^* + \Delta)) \otimes \mathrm{Ber}(\mu_2(\theta^* + \Delta))$ and $\nu_2 = \mathrm{Ber}(\mu_1(\theta^* - \Delta)) \otimes \mathrm{Ber}(\mu_2(\theta^* - \Delta))$, where $x \otimes y$ denotes the product distribution of x and y. Let $\Pr_\nu$ denote the probability associated with distribution ν. Then, the following holds:

$$\begin{aligned}
&\mathrm{Reg}(\theta^* + \Delta, T) + \mathrm{Reg}(\theta_* - \Delta, T) \\
&\qquad \geq A_1(\Delta)\sum_{t=1}^{T}\left(\Pr_{\nu_1^{\otimes t}}(I_t = 2) + \Pr_{\nu_2^{\otimes t}}(I_t = 1)\right) \qquad (10)
\end{aligned}$$

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 [42], we have

$$\begin{aligned}
&\mathrm{Reg}(\theta^* + \Delta, T) + \mathrm{Reg}(\theta^* - \Delta, T) \qquad (11) \\
&\qquad \geq A_1(\Delta)\sum_{t=1}^{T}\exp\left(-\mathrm{KL}\left(\nu_1^{\otimes t}, \nu_2^{\otimes t}\right)\right) \qquad (12)
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{KL}(\nu_1^{\otimes t}, \nu_2^{\otimes t}) = t(&\mathrm{KL}(\mathrm{Ber}(\mu_1(\theta^* + \Delta)), \mathrm{Ber}(\mu_1(\theta^* - \Delta)) \\
&+\mathrm{KL}(\mathrm{Ber}(\mu_2(\theta^* + \Delta)), \mathrm{Ber}(\mu_2(\theta^* - \Delta))). \qquad (13)
\end{aligned}$$

Define $A_2 = (1 - \exp(-4\, D_2^2\Delta^2\, T/(\theta^* - \Delta)^\gamma(1 - (\theta^* - \Delta)^\gamma)))(\theta^* - \Delta)^\gamma(1 - (\theta^* - \Delta)^\gamma)$. By using the fact $\mathrm{KL}(p, q) \leq ((p - q)^2/q(1 - q))$ [43], we can further bound (12) by

$$\begin{aligned}
&\mathrm{Reg}(\theta^* + \Delta, T) + \mathrm{Reg}(\theta^* - \Delta, T) \\
&\quad \geq A_1(\Delta)\sum_{t=1}^{T}\exp\left(-\frac{4D_2 t\Delta^2}{(\theta^* - \Delta)^\gamma(1 - (\theta^* - \Delta)^\gamma)}\right) \\
&\quad \geq A_1(\Delta)\frac{A_2}{4D_2\Delta^2}
\end{aligned}$$

where $A_2 \in (0, 1)$ for any $\Delta \in (0, \max(\theta^*, 1 - \theta^*))$. Hence, the lower bound for the parameter-dependent regret is $\Omega(1)$. In order to show the lower bound for the worst case regret, observe that

$$\begin{aligned}
&\mathrm{Reg}(\theta^* + \Delta, T) + \mathrm{Reg}(\theta^* - \Delta, T) \\
&\qquad \geq \frac{2\gamma\, 2^{\frac{1-\gamma}{\gamma}} A_2}{4D_2\Delta} + \sum_{i=2}^{\gamma}\binom{\gamma}{i}(-\Delta)^{i-2}(\theta^*)^{\gamma-i}.
\end{aligned}$$

By choosing $\Delta = 1/\sqrt{T}$, we can show that for a large T, $A_2 = 0.25(1 - \exp(-16\, D_2^2))$. Hence, worst case lower bound is $\Omega(\sqrt{T})$.

J. Proof of Theorem 7

Without loss of generality, we assume that a unique arm is optimal for $\hat{\theta}_t$ and θ_*. First, we show that $|\hat{\theta}_t - \theta_*| = \epsilon$ implies $|\hat{\Delta}_t - \Delta_*| \leq \epsilon$. There are four possible cases for $\hat{\Delta}_t$.

1) θ_* and $\hat{\theta}_t$ lie in the same optimality interval of the optimal arm, and Δ_* and $\hat{\Delta}_t$ are computed with respect to the same endpoint of that interval.
2) θ_* and $\hat{\theta}_t$ lie in the same optimality interval, and Δ_* and $\hat{\Delta}_t$ are computed with respect to the different endpoints of that interval.
3) θ_* and $\hat{\theta}_t$ lie in adjacent optimality intervals.
4) θ_* and $\hat{\theta}_t$ lie in nonadjacent optimality intervals.

In the first case, $|\hat{\theta}_t - \theta_*| = |\hat{\Delta}_t - \Delta_*| = \epsilon$. In the second case, $\hat{\Delta}_t$ cannot be larger than $\Delta_* + \epsilon$, since in that case, $\hat{\theta}_t$ would be computed with respect to the same endpoint of that interval. Similarly, $\hat{\Delta}_t$ cannot be smaller than $\Delta_* - \epsilon$, since in that case, θ_* would be computed with respect to the same endpoint 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 $|\hat{\Delta}_t - \Delta_*| < \bar{D}_1(2\ K \log t/t)^{\bar{\gamma}_1/2}$ holds with high probability

$$\begin{aligned}
&\Pr\left(|\hat{\Delta}_t - \Delta_*| \geq \bar{D}_1 \left(\frac{K \log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}\right) \\
&\quad \leq \Pr\left(|\hat{\theta}_t - \theta_*| \geq \bar{D}_1 \left(\frac{K \log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}\right) \\
&\quad \leq \sum_{k=1}^{K} 2\mathbb{E}\left[\exp\left(-2\left(\frac{\bar{D}_1\ K\left(\frac{\log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}}{\bar{D}_1\ K w_k(t)}\right)^{\frac{2}{\bar{\gamma}_1}} N_k(t)\right) \mid N_k(t)\right] \\
&\quad \leq \sum_{k=1}^{K} 2\mathbb{E}\big[\exp\big(-2w_k(t)^{1-\frac{2}{\bar{\gamma}_1}} \log t\big)\big|w_k(t)\big] \\
&\quad \leq 2Kt^{-2}
\end{aligned} \tag{14}$$

where the second inequality follows from Lemma 3 and Chernoff–Hoeffding inequality and third inequality by Lemma 4. Then, at time t, with probability at least $1-2\ Kt^{-2}$, the following holds:

$$\Delta_* - 2\bar{D}_1 K\left(\frac{\log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}} \leq \tilde{\Delta}_t. \tag{15}$$

Also, note that if $t \geq C_2(\Delta_*/3)$, then $2\bar{D}_1\ K(\log t/t)^{(\bar{\gamma}_1/2)} \leq (2\Delta_*/3)$. Thus, for $t \geq C_2(\Delta_*/3)$, we have $\Delta_*/3 \leq \tilde{\Delta}_t$. Note that the BUW follows UCB1 only when $t < C_2(\tilde{\Delta}_t)$. From the above, we know that $C_2(\tilde{\Delta}_t) \leq C_2(\Delta_*/3)$ when $t \geq C_2(\Delta_*/3)$ with probability at least $1 - 2\ Kt^{-2}$. This implies that the BUW follows the WAGP with probability at least $1 - 2\ Kt^{-2}$ when $t \geq C_2(\Delta_*/3)$.

We also know from Theorem 4 that the WAGP selects an optimal action with probability at least $1 - 1/t^2$ when $t > C_2(\Delta_*)$. Since $C_2(\Delta_*/3) > C_2(\Delta_*)$, when the BUW follows the WAGP, it will select an optimal action with probability at least $1 - 1/t^2$ when $t > C_2(\Delta_*/3)$.

Let I_t^g denote the action that selected by algorithm $g \in \{\text{BUW}, \text{WAGP}, \text{UCB1}\}$, $r_t^g(\theta_*) = \mathbb{E}[\mu^*(\theta_*) - \mu_{I_t^g}(\theta_*)]$ denote the one-step regret, and $R_{\theta_*}^g(T_1, T_2)$ denote the cumulative regret incurred by algorithm g from T_1 to T_2. Then, when $T < C_2(\Delta_*/3)$, the regret of the BUW can be written as

$$\begin{aligned}
R_{\theta^*}^{\text{BUW}}(1, T) &\leq \sum_{t=1}^{T} r_t^{\text{UCB1}}(\theta_*) + 2\ Kt^{-2} \\
&\leq R_{\theta^*}^{\text{UCB1}}(1, T) + \frac{2\ K\pi^2}{3}.
\end{aligned}$$

Moreover, when $T \geq C_2(\Delta_*/3)$, we have

$$\begin{aligned}
&R_{\theta^*}^{\text{BUW}}(C_2(\Delta_*/3), T) \\
&\quad \leq \sum_{t=C_2(\Delta_*/3)}^{T} r_t^{\text{WAGP}}(\theta_*) + 2\ Kt^{-2} \\
&\quad \leq R_{\theta^*}^{\text{WAGP}}(C_2(\Delta_*/3), T) + \frac{2\ K\pi^2}{3}.
\end{aligned}$$

This concludes the parameter-dependent regret bound.

The worst case bound can be proven by replacing $\delta_k = \mu^* - \mu_k = 1/\sqrt{TK\log T}$ for all $k \notin \mathcal{K}^*(\theta_*)$ for the regret bound given above.

K. Proof of Theorem 8

When the round is clear from the context, we use $\hat{\theta}_t$ to represent $\hat{\theta}_{\rho,t}$. By Lemma 2 and Jensen's inequality, we have

$$\mathbb{E}\big[r_{t+1}(\theta_*^{t+1})\big] \leq 2D_2\mathbb{E}\big[\big|\theta_*^{t+1} - \hat{\theta}_t\big|\big]^{\gamma_2} \tag{16}$$

where $\hat{\theta}_t = (\sum_{k=1}^{K} N_{k,\rho}(t)\tilde{\mu}_k^{-1}(\hat{X}_{k,\rho,t})/\tau_\rho(t))$ and $\sum_{k=1}^{K} N_{k,\rho}(t) = \tau_\rho(t)$. Then, by using Lemma 1, we have

$$\begin{aligned}
&\mathbb{E}\big[\big|\hat{\theta}_t - \theta_*^{t+1}\big|\big] \\
&\quad \leq \frac{\sum_{k=1}^{K} \bar{D}_1\mathbb{E}\big[N_{k,\rho}(t)\mathbb{E}\big[\big|\hat{X}_{k,\rho,t} - \mu_k\big(\theta_*^{t+1}\big)\big||N_{k,\rho}(t)\big]^{\bar{\gamma}_1}\big]}{\tau_\rho(t)}.
\end{aligned}$$

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

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

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

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

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

$$
\begin{aligned}
&\mathbb{E}\big[\big|\hat{X}_{k,\rho,t} - \mu_k(\theta_*^{t+1})\big| | N_{k,\rho}(t)\big]\\
&\quad\le \mathbb{E}\left[\left|\frac{\sum_{t'\in\mathcal{S}_{k,\rho,t}^{\tau_h}} \left(X_{k,t'} - \mu_k\left(\theta_*^{t'}\right)\right)}{N_{k,\rho}(t)}\right|\right]\\
&\qquad + \frac{\sum_{t'\in\mathcal{S}_{k,\rho,t}^{\tau_h}} \left|\mu_k\left(\theta_*^{t'}\right) - \mu_k\left(\theta_*^{t+1}\right)\right|}{N_{k,\rho}(t)}\\
&\quad\le \mathbb{E}\left[\left|\frac{\sum_{t'\in\mathcal{S}_{k,\rho,t}^{\tau_h}} \tilde{X}_{k,t'}}{N_{k,\rho}(t)}\right|\right] + \frac{\sum_{t'\in\mathcal{S}_{k,\rho,t}^{\tau_h}} 2D_2\left|\theta_*^{t'} - \theta_*^{t+1}\right|^{\gamma_2}}{N_{k,\rho}(t)}
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbb{E}\Big[\Big|\frac{\sum_{t'\in\mathcal{S}_{k,\rho,t}^{\tau_h}} \tilde{X}_{k,t'}}{N_{k,\rho}(t)}\Big|\Big]\\
&\quad= \int_{x=0}^{\infty} \Pr\left(\left|\frac{\sum_{t'\in\mathcal{S}_{k,\rho,t}^{\tau_h}} \tilde{X}_{k,t'}}{N_{k,\rho}(t)}\right| > x\right) \mathrm{d}x\\
&\quad\le \int_{x=0}^{\infty} 2\exp(-x^2\ N_{k,\rho}(t))\,\mathrm{d}x = \sqrt{\frac{\pi}{N_{k,\rho}(t)}} \qquad (17)
\end{aligned}
$$

where the inequality follows from the Chernoff–Hoeffding bound and:

$$\left|\theta_*^{t+1} - \theta_*^{t'}\right| \le (2\tau_h/\tau) \qquad (18)$$

since for all t, t' in the same round $|t - t'| \le 2\tau_h$. Then, using (17) and (18), the expected gap between θ_*^{t+1} and $\hat{\theta}_t$ can be bounded as

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

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 $k \in \mathcal{K}$, where $\tau_\rho(t)/K$ is assumed to be an integer without loss of generality, and the fourth inequality follows from the fact that $\tau_\rho(t) \ge \tau_h$. By choosing $\tau_h = \lceil\tau\rceil^b$, we get the optimal $b = (\gamma_2/0.5 + \gamma_2)$. Then, cumulative regret at time T can be bounded as

$$
\begin{aligned}
&\mathrm{Reg}^{\mathrm{ave}}(T)\\
&\qquad \le \tau^{-\frac{\gamma_2}{0.5+\gamma_2}} + \left(2D_2\bar{D}_1^{\gamma_2}[(\pi K)^{\frac{\bar{\gamma}_1}{2}} + 2D_2^{\bar{\gamma}_1}]\right)^{\gamma_2} \tau^{-\frac{\gamma_2^2\bar{\gamma}_1}{1+2\gamma_2}}
\end{aligned}
$$

which concludes the proof.

L. Auxiliary Lemma

Lemma 4: For $\gamma < 0$, $\delta > 0$, the following bound holds for any w_k with $0 \le w_k \le 1$ and $\sum_{k=1}^{K} w_k = 1$:

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

Proof: Let $k_{\max} = \arg\max_k w_k$. Then

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

□

References

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

[2] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis, "A structured multiarmed bandit problem and the greedy policy," *IEEE Trans. Autom. Control*, vol. 54, no. 12, pp. 2787–2802, Dec. 2009.

[3] T. L. Lai and H. Robbins, "Adaptive design in regression and control," *Proc. Nat. Acad. Sci. USA*, vol. 75, no. 2, pp. 586–587, 1978.

[4] Y. Chen and V. F. Farias, "Simple policies for dynamic pricing with imperfect forecasts," *Oper. Res.*, vol. 61, no. 3, pp. 625–643, 2013.

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

[6] T. H. Li and K. S. Song, "On asymptotic normality of nonlinear least squares for sinusoidal parameter estimation," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4511–4515, Sep. 2008.

[7] 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 *Proc. ICASSP*, May 2013, pp. 6630–6634.

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

[9] O. Atan, C. Tekin, and M. van der Schaar, "Global multi-armed bandits with Hölder continuity," in *Proc. AISTATS*, 2015, pp. 28–36.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.

[11] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *J. Mach. Learn. Res.*, vol. 3, pp. 397–422, Nov. 2002.

[12] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. COLT*, 2011, pp. 359–376.

[13] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in *Proc. AISTATS*, 2012, pp. 592–600.

[14] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, pp. 285–294, Dec. 1933.

[15] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Proc. COLT*, 2012, pp. 39.1–39.26.

[16] N. Korda, E. Kaufmann, and R. Munos, "Thompson sampling for 1-dimensional exponential family bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1448–1456.

[17] S. Bubeck and C.-Y. Liu, "Prior-free and prior-dependent regret bounds for thompson sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 638–646.

[18] J. Langford and T. Zhang, "The Epoch-Greedy algorithm for contextual multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 817–824.

[19] A. Slivkins, "Contextual bandits with similarity information," *J. Mach. Learn. Res.*, vol. 15, pp. 2533–2568, Jan. 2014.

[20] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *Proc. ICML*, 2013, pp. 127–135.

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

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

[23] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[24] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *Proc. 36th Annu. Symp. Found. Comput. Sci.*, Oct. 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 *Proc. 19th Int. Conf. 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 *Proc. AISTATS*, vol. 15. 2011, pp. 208–214.

[27] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2312–2320.

[28] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Math. Oper. Res.*, vol. 35, no. 2, pp. 395–411, 2010.

[29] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proc. 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 *Proc. Int. Conf. Artif. Intell. Stat.*, 2012, pp. 1–9.

[31] N. Cesa-Bianchi and S. Kakade. (2011). "An optimal algorithm for linear bandits." [Online]. Available: https://arxiv.org/abs/1110.4322

[32] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework, results and applications," in *Proc. ICML*, 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," *IEEE/ACM Trans. Netw*, vol. 20, no. 5, pp. 1466–1478, Oct. 2012.

[34] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 684–692.

[35] T. Lattimore and R. Munos, "Bounded regret for finite-armed structured bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 550–558.

[36] D. Russo and B. Van Roy, "An information-theoretic analysis of thompson sampling," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2442–2471, 2015.

[37] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári, "Parametric bandits: The generalized linear case," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 586–594.

[38] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

[39] D. Kalman, "A generalized logarithm for exponential-linear equations," *College Math. J.*, vol. 32, no. 1, pp. 2–14, 2001.

[40] A. Garivier and E. Moulines. (2008). "On upper-confidence bound policies for non-stationary bandit problems." [Online]. Available: https://arxiv.org/abs/0805.3415

[41] Y. Song, S. Ray, and S. Li, "Structural properties of buyback contracts for price-setting newsvendors," *Manuf. Service Oper. Manage.*, vol. 10, no. 1, pp. 1–18, 2008.

[42] A. B. Tsybakov and V. Zaiats, *Introduction to Nonparametric Estimation*. New York, NY, USA: Springer-Verlag, 2009.

[43] P. Rigollet and A. Zeevi, "Nonparametric bandits with covariates," in *Proc. COLT*, 2010, pp. 1–18.

Onur Atan received the B.Sc. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2013, and the M.S. degree in electrical engineering from the University of California at Los Angeles, Los Angeles, CA, USA, in 2014, where he is currently pursuing the Ph.D. degree in electrical engineering.

His current research interests include online learning, multiarmed bandit problems, off-policy optimization and their applications in healthcare informatics.

Mr. Atan received the UCLA Electrical Engineering Outstanding M.S. Thesis Award in 2015 and the UCLA Dissertation Year Fellowship in 2017.

Cem Tekin (M'13) received the B.Sc. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2008, and the M.S.E. degree in electrical engineering: systems, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering: systems from the University of Michigan, Ann Arbor, MI, USA, in 2010, 2011, and 2013, respectively.

He is currently an Assistant Professor with the Electrical and Electronics Engineering Department, Bilkent University, Ankara. His current research interests include machine learning, multiarmed bandit problems, data mining, and multiagent systems.

Dr. Tekin received the University of Michigan Electrical Engineering Departmental Fellowship in 2008 and the Fred W. Ellersick Award for the best paper in Military Communications 2009.

Mihaela van der Schaar (F'10) is currently a Chancellor Professor of electrical engineering with the University of California at Los Angeles, Los Angeles, CA, USA. She holds 33 granted U.S. patents. Her current research interests include machine learning, data science, and decisions for medicine, education, and finance.

Prof. van der Schaar received the NSF CAREER Award in 2004, the Best Paper Award from the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2005, the Okawa Foundation Award in 2006, the IBM Faculty Award in 2005, 2007, and 2008, the Most Cited Paper Award from the EURASIP: Image Communications Journal in 2006, the Gamenets Conference Best Paper Award in 2011, and the 2011 IEEE Circuits and Systems Society Darlington Award Best Paper Award. She also received three ISO awards for her contributions to the MPEG video compression and streaming international standardization activities. She is a member of the Editorial Board of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING and an Editor-in-Chief of IEEE TRANSACTIONS ON MULTIMEDIA. She was a Distinguished Lecturer of the Communications Society from 2011 to 2012.