Generalized Global Bandit and Its Application in Cellular Coverage Optimization

Cong Shen, *Senior Member, IEEE*, Ruida Zhou, Cem Tekin, *Member, IEEE*, and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—Motivated by the engineering problem of cellular coverage optimization, we propose a novel multiarmed bandit model called generalized global bandit. We develop a series of greedy algorithms that have the capability to handle nonmonotonic but decomposable reward functions, multidimensional global parameters, and switching costs. The proposed algorithms are rigorously analyzed under the multiarmed bandit framework, where we show that they achieve *bounded* regret, and hence, they are guaranteed to converge to the optimal arm in *finite* time. The algorithms are then applied to the cellular coverage optimization problem to achieve the optimal tradeoff between sufficient small cell coverage and limited macroleakage without prior knowledge of the deployment environment. The performance advantage of the new algorithms over existing bandits solutions is revealed analytically and further confirmed via numerical simulations. The key element behind the performance improvement is a more efficient "trial and error" mechanism, in which any trial will help improve the knowledge of *all* candidate power levels.**

***Index Terms*—Multi-armed bandit, online learning, regret analysis, coverage optimization.**

I. Introduction

RECENT years have witnessed a significant growth of small base stations (SBS), such as pico and femto, that are massively deployed to address the capacity and coverage challenge of wireless networks [1]. In practice, SBSs may be deployed in drastically different scenarios, with different target coverage objectives. In addition, the radio frequency (RF) conditions may vary significantly from one deployment to another. Due to the heterogeneous nature of these deployments, setting an appropriate transmit power of each deployed SBS, which effectively determines the coverage, becomes an important task that must be decided based on the specific deployment scenario. If the transmit power is too small, the resulting coverage may not sufficiently cover the intended area. On the other hand, if the transmit power is too large, the SBS coverage will leak into macrocells and cause unnecessary interference, especially if the SBS operates in a close-access mode.

Traditional approaches to cell coverage optimization rely on RF engineers to carry out on-the-spot field measurements to effectively "learn" the specific deployment environment, and optimize the coverage and leakage using RF planning tools. This approach, however, becomes increasingly infeasible for SBS deployment as it does not scale with the significant increase of network nodes (high density), multiple layers of nodes (heterogeneity), and multiple radio access technologies (3G/4G/Wifi) [2]. Furthermore, non-stationarity of the environment, such as dynamic user behavior and RF footprint variations, may cause the previously optimal configuration to become highly sub-optimal and lead to performance degradation [3].

Applying *online learning* algorithms to cellular coverage optimization is an important means to address the aforementioned challenges, as they allow for adaptive, automated and autonomous coverage adjustment while minimizing the planned human involvement. A good coverage learning solution has to balance the immediate gains (selecting a coverage that is the best based on current knowledge) and long-term performance (evaluating other coverage levels). We thus resort to the theory of multi-armed bandit (MAB) [4] to address the resulting exploration and exploitation tradeoff. It is worth noting that MAB-inspired algorithms have been adopted in various other similar tasks, such as power calibration [5], mobility management [6]–[8], and channel selection [9].

However, a direct application of standard MAB algorithms (such as UCB [10][1]) to the coverage optimization problem, albeit feasible, ignores the inherent structure and hence cannot fully exploit the characteristics of the underlying communication model. First, unlike the standard MAB model where different arms are independent, coverage performances of similar transmit power levels are often very similar, which means that

Manuscript received July 14, 2017; revised October 30, 2017; accepted January 15, 2018. Date of publication January 25, 2018; date of current version February 16, 2018. The work of C. Shen and R. Zhou was supported by the National Natural Science Foundation of China under Grant 61572455 and Grant 61631017. The work of C. Tekin was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under 3501 Program Grant 116E229. The work of M. van der Schaar was supported by the National Science Foundation under Grant 1407712 and Grant 1533983. The guest editor coordinating the review of this paper and approving it for publication was Prof. H. Vincent Poor. *(Corresponding author: Cong Shen.)*

C. Shen and R. Zhou are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: congshen@ustc.edu.cn; zrd127@mail.ustc.edu.cn).

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

M. van der Schaar is with the Oxford-Man Institute of Quantitative Finance (OMI) and the Department of Engineering Science, University of Oxford, Oxford OX1 2JD, U.K., and also with the Electrical Engineering Department, University of California, Los Angeles (UCLA), Los Angeles, CA 90095 USA (e-mail: mihaela.vanderschaar@oxford-man.ox.ac.uk).

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/JSTSP.2018.2798164

[1]Throughout this paper, UCB specifically refers to `UCB1` in [10].

if we adopt the MAB model, nearby arms are highly *correlated*. Intuitively, such correlation can be used to accelerate the convergence to the optimal selection, because any sampling of one arm not only reveals information about itself, but also nearby arms that are highly correlated. Second, the correlated coverage performance of different power levels fundamentally originates from the fact that they all follow the same physical RF propagation law, which has been captured in various standard models (e.g., 3GPP model [11]).

In this work, motivated by this engineering problem, we first propose a novel MAB model, called Generalized Global Bandit (GGB), that is a non-trivial extension of Global Bandit (GB) in [12], [13]. In GB, the expected reward of each arm is a (possibly non-linear) function of a single parameter, and different arms are correlated through this global parameter. Furthermore, this function is required to be monotonic. As we will see, the original GB model and the resulting algorithms cannot be directly applied to coverage optimization due to three unique features. First, cellular coverage optimization needs to balance sufficient coverage within the intended area and limited leakage to outside macro users. As a result, the reward function will not be monotonic.[2] Second, the reward function may have multiple unknown parameters. The GB model in [12], however, cannot be trivially extended to handle more than one global parameter. Lastly, a practical coverage optimization solution needs to avoid frequent power changes, as it may cause frequent variation of the coverage area and result in uneven user experience. Hence, the solution should explicitly consider switching cost to discourage frequent changes to the coverage area.

We address these three new challenges in the GGB model. The reward function of each arm is allowed to be non-monotonic but decomposable, which fits well with the considered coverage optimization problem. Multi-dimensional global parameters and switching cost are also considered in the GGB model. We then present the *ad-greedy* policy, which can simultaneously maximize the accumulated rewards and estimate the unknown parameters via an updated weight average on different arms. Rigorous regret analysis is carried out for the proposed policy and its variants, where we show that *bounded* regret is achievable, and hence, the policy is guaranteed to converge to the optimal arm in *finite* time. In other words, the one-step regret approaches zero asymptotically. The algorithms are then applied to the cell coverage optimization problem to achieve the optimal tradeoff between sufficient SBS coverage and limited macro leakage without prior knowledge of the deployment. Numerical simulation results are provided to demonstrate the performance advantage of the new algorithms over the existing bandit solutions.

The main contributions of this work are summarized as follows.

- Motivated by the practical constraints of the cellular coverage optimization problem, we propose a generalized global bandit model, which can handle non-monotonic but decomposable reward functions, multi-dimensional global parameters, and switching costs.
- We develop the *ad-greedy* policy for the considered GGB model, and rigorously analyze its regret. We show that the (total) regret is bounded, and hence, the one-step regret diminishes asymptotically.
- We apply the GGB model and the ad-greedy policy and its variants to the cellular coverage optimization problem, and illustrate how the proposed variants fit to this engineering problem. We further verify the advantages of the new algorithms via numerical simulations. Furthermore, we also numerically evaluate the algorithm performance in a non-stationary environment, when the MBS signal strength slowly changes over time.

The rest of the paper is organized as follows. Related literature is discussed in Section II. The GGB formulation, the ad-greedy policies, and the corresponding regret analysis are given in Section III. In Section IV, we describe how the GGB model can be applied to the cellular coverage optimization problem, and present the numerical simulation results. Finally, Section V concludes the paper.

II. Related Work

A. MAB With Arm Correlations

MAB is a powerful tool to model sequential decision problems with an intrinsic exploration-exploitation tradeoff. In the classic stochastic MAB model, each arm, if played, generates an instantaneous reward that is independently and identically (i.i.d.) drawn from a fixed distribution, which is unknown to the forecaster a priori. The design objective is to maximize the total expected reward accumulated through a sequence of T plays, which can be equivalently formulated as to minimize the regret between the total expected reward from always playing the arm with the highest expected reward and that from the learning algorithm.

The fundamental regret lower bound for stochastic MAB was developed by Lai and Robbins in [14], and a matching upper bound is achieved by the celebrated Upper Confidence Bound (UCB) algorithm [10]. Using UCB, at each round the player simply pulls the arm that has the highest sample mean reward plus an uncertainty term that is inversely proportional to the number of times the arm has been played. There is a rich body of literature on MABs, which we will not survey comprehensively. Interested readers are referred to [4] and the references therein.

In the MAB literature, the most relevant work to our GGB model is the study on MAB with arm correlations. Existing research on this topic can be divided into two categories: Bayesian model [15], [16] and parameterized model [17]–[19]. In the Bayesian model, arm correlation is captured by stochastic measures such as mean and covariance matrix. This approach and the corresponding bandit algorithms have been studied in [15], [16]. The authors of [15] propose bandit algorithms with a Bayesian prior on the mean reward that is based on a human decision-making model. The authors of [16] further extend the algorithm to focus on the correlation among arms. Linear bandit [17] is a primary example of the parameterized model, in which the

[2]It is worth noting that the monotonicity requirement is fundamental to the WAGP algorithm in [12], which is one of the two key assumptions in [12, Sec. III].

expected reward of each arm is a *linear* function of a global parameter. For this model, [18] proves a regret bound that scales linearly with the dimension of the parameter. The authors of [19] establish a lower bound for an arbitrary policy for the multi-dimensional linear bandit, and then provide a matching upper bound through a policy that alternates between exploration and exploitation. The GGB setting is more general than these above models as it allows for non-linear non-monotonic reward functions with multi-dimensional parameters. For the special case when the expected sub-function rewards are all linear in multi-dimensional parameters, our setting reduces to the linear bandit model.

B. Non-Linear Parameter Estimation

Another line of relevant work is in the area of non-linear parameter estimation [20]–[22]. In [21], the author studies the non-linear parameter estimation problem with additive Gaussian noise. The authors of [20] prove that the nonlinear least-square estimator is able to asymptotically attain the Cramer-Rao lower bound under additive Gaussian noise, even when there is a mismatch of noise distribution. The authors of [22] focus on the impact of compressed sensing on Fisher information and the Cramer-Rao bound. The main difference to our work is that these papers do not need to consider the exploration and exploitation tradeoff that is fundamental to the MAB problems. They only care about estimating the parameter as accurately as possible, while we aim at maximizing the long-term reward of a bandit policy.

C. Cellular Coverage Optimization

Coverage optimization is an important task in cellular network deployment. Under the self-organizing networking (SON) framework, this task is captured in the Capacity and Coverage Optimization (CCO) feature, which is part of the 3GPP SON deliverables [23]. In practice, coverage optimization has been implemented and deployed in commercial SON products such as Cisco SON [24] and Qualcomm UltraSON [25]. In academia, coverage optimization is an active research topic [26], [27]. Existing studies have focused on optimizing different system parameters, such as antenna tilt [28]–[30], and downlink transmit power [5], [31], [32].

In [28], the impact of half-power beamwidths and downtilt antenna angle to the overall network performance is studied. A cell coverage optimization problem for uplink massive MIMO is studied in [29], which is based on optimizing the tilt-adjustable antennas at SBS. A general framework incorporating both downlink and uplink coverage, while requiring very sparse system knowledge, is proposed in [30].

Besides antenna parameter optimization, another line of study focuses on adjusting the SBS transmit power so that the resulting coverage balances maximizing intended coverage and minimizing undesirable leakage. Our application of coverage optimization also falls into this category. Claussen *et al.* [31] have proposed a method that uses information on mobility events of outdoor and indoor users to optimize the transmit power. This approach is further enhanced in [32], where a systematic study of indoor enterprise SBS networks is carried out. Both of these works require knowledge of the deployment, such as intended area and co-channel macrocell footprint, which is not assumed in this work. Alternatively, adjusting the SBS transmit power without deployment knowledge has recently been considered in [5], where the MAB model is applied and the correlation of different power levels is captured using a Bayesian framework. However, it does not fully utilize the available structural information of the system.

III. Generalized Global Bandit Model and Greedy Policies

In this section, we first present the common baseline formulation, and then discuss three generalizations to the underlying model: non-monotonic decomposable reward functions, multi-dimensional global parameters, and switching costs. For each of these generalizations, we will present the greedy policies and analyze their regrets.

A. The Baseline GGB Formulation

We consider a stochastic MAB formulation with K arms, indexed by $\mathcal{K} = \{1, \ldots, K\}$. A forecaster can choose and play exactly one arm at each time slot. Arm $k \in \mathcal{K}$, if played, will offer a bounded reward that is drawn from a distribution ν_k with a finite support, and we denote its mean as μ_k. We use $X_{k,t}$ to denote the random reward of arm k at time slot t, which is independently drawn from other arms. Without loss of generality, we assume that the rewards are bounded within the unit interval $[0, 1]$. The forecaster has no prior knowledge of either ν_k or μ_k, $\forall k \in \mathcal{K}$. The forecaster's goal is to design an arm selection policy that maximizes the total reward it obtains over time.

Within the framework of global bandits [12], there exists a global parameter θ_*, which is associated to the expected rewards of all arms $\mu_k = \mu_k(\theta_*) = \mathbb{E}_{\nu_k}[X_{k,t}]$, where $\mathbb{E}_{\nu_k}[\cdot]$ denotes expectation with respect to distribution ν_k. The parameter θ_*, unknown to the forecaster, belongs to a parameter set Θ, which again is normalized to be the unit interval for simplicity.

The forecaster knows the reward function $\mu_k(\theta)$ for each $k \in \mathcal{K}$, but not the true global parameter θ_*. At each time slot, the forecaster only observes the random reward from the chosen arm, and her goal is to maximize the cumulative reward up to any given time T. Obviously, if the global parameter is perfectly known to the forecaster, she will always select the optimal arm $k^*(\theta_*) = \arg\max_{k \in \mathcal{K}} \mu_k(\theta_*)$, with the corresponding optimal expected reward $\mu^*(\theta_*) = \max_{k \in \mathcal{K}} \mu_k(\theta_*)$. When θ_* is clear from the context, we use k^* and μ^* instead of $k^*(\theta_*)$ and $\mu^*(\theta_*)$. For simplicity of exposition and without loss of generality, throughout the paper it is assumed that there exists a unique best arm for θ_*. We define the one-step (pseudo) regret at time t as $r_{I_t}(\theta_*) \doteq \mu^*(\theta_*) - \mu_{I_t}(\theta_*)$, where I_t is the selected arm by the forecaster's policy at time t. The total regret [4] by time T is given as

$$\mathsf{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_{I_t}(\theta_*)\right]. \tag{1}$$

B. Non-Monotonic Decomposable Reward Functions

1) Model and Algorithm: A significant constraint of [12] is that $\mu_k(\theta)$ must be an invertible function of θ. Hence the reward function must be monotonic. If the monotonicity condition is totally removed, the GB problem becomes difficult to study. Our approach in this work, nevertheless, is to exploit the structure of the cellular coverage optimization problem and relax the monotonicity constraint to a certain degree such that it not only fits our problem setting but also is tractable.

Fortunately, for the cellular coverage optimization problem, the objective function is generally defined as a linear combination of two (or more) conflicting functions (see [5, eq. (3)] for an example). As a result, the overall function is not monotonic with respect to θ, but the individual sub-functions are, and they are monotonic in the *opposite* direction. For instance, the specific Performance Indication Function (PIF) $f_k(\theta)$ of [5] (equivalent to the expected reward function $\mu_k(\theta)$ in GB), can be decomposed into the linear combination of two sub-functions. One of them denotes the *coverage*, which increases monotonically with the coverage radius d, while the other one denotes the *leakage*, which decreases monotonically with d.

Formally, the expected reward function $\mu_k(\theta)$ can be decomposed into J continuous sub-functions:

$$\mu_k(\theta) = \sum_{j=1}^{J} \alpha_j \mu_{j,k}(\theta). \tag{2}$$

These sub-functions and their weights are assumed to be known to the forecaster, but not the true parameter θ_*. For each $j \in \mathcal{J} = \{1, \ldots, J\}$ and $k \in \mathcal{K} = \{1, \ldots, K\}$, the sub-function $\mu_{j,k}(\theta)$ satisfies the Hölder continuity and monotonicity assumptions as [12, Assumption 1]. More specifically, the following assumptions are made.

Assumption 1:

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

$$|\mu_{j,k}(\theta) - \mu_{j,k}(\theta')| \leq D_{2,j,k} |\theta - \theta'|^{\gamma_{2,j,k}}. \tag{3}$$

2) **(Sub-function monotonicity)** For each $j \in \mathcal{J}, k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$, there exists $D_{1,j,k} > 0$ and $1 < \gamma_{1,j,k}$, such that:

$$|\mu_{j,k}(\theta) - \mu_{j,k}(\theta')| \geq D_{1,j,k} |\theta - \theta'|^{\gamma_{1,j,k}}. \tag{4}$$

We want to emphasize that Assumption 1 is mild. For the application of cellular coverage optimization, the sub-function monotonicity has been discussed, and the Hölder continuity condition can also be met when the coverage/leakage function changes smoothly with the intended coverage area. This point will become more clear when we discuss the application of GGB to the cellular coverage problem in Section IV.

We further assume that whenever an arm k is played, the forecaster receives the sub-function reward realizations $\{X_{k,t}^j\}_{j \in \mathcal{J}}$. Sub-function reward realizations are independent between arms, and i.i.d. over time. Receiving $\{X_{k,t}^j\}_{j \in \mathcal{J}}$ is a reasonable assumption for some practical scenarios, such as the considered coverage optimization problem where the coverage events are reported by SBS users through the measurement reports and mobility protocols, and the leakage events are reported by macro users through registration attempts, respectively. Such approach has been adopted in previous papers, e.g., [32], [33], and has been successfully adopted in practical transmit power assignment solution, e.g., [25].

With Assumption 1, we have the following proposition.

Proposition 1: Define

$$\begin{aligned} D_2 &= \max\{D_{2,j,k} | j \in \mathcal{J}, k \in \mathcal{K}\}, \\ \gamma_1 &= \max\{\gamma_{1,j,k} | j \in \mathcal{J}, k \in \mathcal{K}\}, \\ \gamma_2 &= \min\{\gamma_{2,j,k} | j \in \mathcal{J}, k \in \mathcal{K}\}, \\ \underline{\mu}_{j,k} &= \min_{\theta \in \Theta} \mu_{j,k}(\theta),\ \overline{\mu}_{j,k} = \max_{\theta \in \Theta} \mu_{j,k}(\theta) \end{aligned}$$

and

$$\bar{\gamma}_1 = \frac{1}{\gamma_1},$$

$$\bar{D}_1 = \max\left\{ \left(\frac{1}{D_{1,j,k}} \right)^{\frac{1}{\gamma_{1,j,k}}} | j \in \mathcal{J}, k \in \mathcal{K} \right\}.$$

The following statements hold:

1) For each $j \in \mathcal{J}$, $k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$,

$$|\mu_{j,k}(\theta) - \mu_{j,k}(\theta')| \leq D_2 |\theta - \theta'|^{\gamma_2}. \tag{5}$$

2) For each $j \in \mathcal{J}$, $k \in \mathcal{K}$ and $y, y' \in [\underline{\mu}_{j,k}, \overline{\mu}_{j,k}]$,

$$|\mu_{j,k}^{-1}(y) - \mu_{j,k}^{-1}(y')| \leq \bar{D}_1 |y - y'|^{\bar{\gamma}_1}. \tag{6}$$

Proof: Inequality (5) is a direct application of (3) of Assumption 1 and the definitions of D_2 and γ_2. For the proof of inequality (6), we first note that the reward sub-functions are invertible due to the sub-function monotonicity part of Assumption 1. Then, inequality (6) is directly obtained by plugging the inverse functions in (4) and applying the definitions of $\bar{\gamma}_1$ and $\bar{D}_1$. ■

We present a greedy policy, called the *ad-greedy* policy, which can effectively handle non-monotonic decomposable reward functions. The pseudocode of the *ad-greedy* policy is given in Algorithm 1.

As the name suggests, the *ad-greedy* policy is a greedy procedure at its core, with the capability to adaptively update the parameter estimate using all the observed reward realizations from sub-functions. Other than the initial time slot where no prior information is available, where an arm I_1 is uniformly chosen among all arms, the arm selection always chooses I_t with the highest estimated reward:

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

This is the same as the greedy policy for classic MAB problems, which is well-known [4], [10] to be strictly sub-optimal and cannot achieve $\log(t)$ order of regret. However, as we will see later in the regret analysis, this simple greedy policy suffices to achieve bounded regret in our GGB model due to the global informativeness.

In addition to the greedy arm selection, the policy also carries out an update on the global parameter estimation using

Algorithm 1: The *ad-greedy* policy.

Input : $\mu_{j,k}(\theta)$ and $\mu_k(\theta)$ for each $k \in \mathcal{K}$ and $j \in \mathcal{J}$;
Initialize: $\omega_k(0) = 0, \hat{\theta}^j_{k,0} = 0, \hat{X}^j_{k,0} = 0, N_k(0) = 0$ for each $k \in \mathcal{K}$ and $j \in \mathcal{J}$, $t = 1$;
while $t \geq 1$ **do**
  if $t = 1$ **then**
    Select arm I_1 uniformly at random from set $\mathcal{K}$;
  else
    Select arm $I_t = \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_{t-1})$;
  end
  Observe sub-rewards $\{X^j_{I_t,t}\}_{j\in\mathcal{J}}$;
  $\hat{X}^j_{k,t} = \hat{X}^j_{k,t-1}$ for $k \in \mathcal{K}\backslash I_t$, $j \in \mathcal{J}$;
  $\hat{X}^j_{I_t,t} = \frac{N_{I_t}(t-1)\hat{X}^j_{I_t,t-1} + X^j_{I_t,t}}{N_{I_t}(t-1)+1}$ for $j \in \mathcal{J}$;
  $\hat{\theta}^j_{k,t} = \arg\min_{\theta\in\Theta} |\mu_{j,k}(\theta) - \hat{X}^j_{k,t}|$ for $k \in \mathcal{K}$ and $j \in \mathcal{J}$;
  $N_k(t) = N_k(t-1)$ for $k \in \mathcal{K}\backslash I_t$;
  $N_{I_t}(t) = N_{I_t}(t-1) + 1$;
  $\omega_k(t) = \frac{N_k(t)}{t}$ for $k \in \mathcal{K}$;
  $\hat{\theta}_t = \sum_{k\in\mathcal{K}} \omega_k(t) \frac{\sum_{j\in\mathcal{J}} \hat{\theta}^j_{k,t}}{J}$;
  $t = t + 1$;
end

estimates from individual sub-functions of individual arms, and weighing them differently. The weights are updated according to the number of times the arm is played.

2) Regret Analysis: The optimality region Θ_k for any arm k is defined as

$$\Theta_k = \{\theta \in \Theta | k \in k^*(\theta)\}. \tag{7}$$

We then define δ as the smallest Euclidean distance between θ_* and the boundary of Θ_{k^*}. Since there is a unique best arm for θ_* and since the reward functions are continuous in θ, we have $\delta > 0$. The total regret up to time T can be written as the sum of one-step regrets $\mathsf{Reg}(T) = \mathbb{E}[\sum_{t=1}^T r_{I_t}(\theta_*)]$. Thanks to the normalization, the one-step regret for $t > 1$ can be bounded by

$$\mathbb{E}[r_{I_t}(\theta_*)] \leq 1 \cdot \Pr\{I_t \neq k^*(\theta_*)\} = \Pr\left\{\hat{\theta}_{t-1} \in \Theta\backslash\Theta_{k^*}\right\}.$$

Theorem 1: The regret of the ad-greedy policy for a finite time horizon T is upper bounded by

$$\mathsf{Reg}(T) \leq 1 + 2JK\frac{e^{-\alpha} - Te^{-\alpha T} + (T-1)e^{-\alpha(T+1)}}{(1-e^{-\alpha})^2},$$

where $\alpha = 2(\frac{\delta}{K\bar{D}_1})^{2\gamma_1} > 0$. Furthermore, the infinite time horizon regret of the ad-greedy policy is finite, i.e.,

$$\mathsf{Reg}(\infty) \leq 1 + 2JK\frac{e^{-\alpha}}{(1-e^{-\alpha})^2}.$$

Proof: Before deriving a bound of the gap between the parameter estimate at time slot t and the true parameter, we let $\tilde{\mu}^{-1}_{j,k}(y) \doteq \arg\min_{\theta\in\Theta} |\mu_{j,k}(\theta) - y|$ for $y \in [0,1]$. By the monotonicity of $\mu_{j,k}(\cdot)$ and Proposition 1, we have $|\tilde{\mu}^{-1}_{j,k}(y) - \tilde{\mu}^{-1}_{j,k}(y')| \leq \bar{D}_1|y - y'|^{\bar{\gamma}_1}$ for all $y, y' \in [0,1]$. Then, we have

$$\begin{aligned}
&|\hat{\theta}_t - \theta_*| \\
=& \left|\sum_{k\in\mathcal{K}} \omega_k(t) \frac{\sum_{j\in\mathcal{J}} \hat{\theta}^j_{k,t}}{J} - \theta_*\right| \\
\leq& \frac{1}{J}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \left|\omega_k(t)\hat{\theta}^j_{k,t} - \omega_k(t)\theta_*\right| \\
\leq& \frac{1}{J}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \omega_k(t) \left|\tilde{\mu}^{-1}_{j,k}(\hat{X}^j_{k,t}) - \tilde{\mu}^{-1}_{j,k}(\mu_{j,k}(\theta_*))\right| \\
\leq& \frac{1}{J}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \omega_k(t)\bar{D}_1|\hat{X}^j_{k,t} - \mu_{j,k}(\theta_*)|^{\bar{\gamma}_1}.
\end{aligned}$$

Next, we analyze the event that the gap $|\hat{\theta}_t - \theta_*|$ is no smaller than δ. Note that when the gap is smaller than δ, $I_{t+1} = k^*$. This may not hold when the gap is larger than or equal to δ.

$$\begin{aligned}
&\left\{\delta \leq |\hat{\theta}_t - \theta_*|\right\} \\
\subset& \left\{\delta \leq \frac{1}{J}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \omega_k(t)\bar{D}_1|\hat{X}^j_{k,t} - \mu_{j,k}(\theta_*)|^{\bar{\gamma}_1}\right\} \\
=& \left\{\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \frac{\delta}{JK} \leq \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \frac{\omega_k(t)\bar{D}_1}{J}|\hat{X}^j_{k,t} - \mu_{j,k}(\theta_*)|^{\bar{\gamma}_1}\right\} \\
\subset& \bigcup_{k\in\mathcal{K}}\bigcup_{j\in\mathcal{J}} \left\{\frac{\delta}{JK} \leq \frac{\omega_k(t)\bar{D}_1}{J}|\hat{X}^j_{k,t} - \mu_{j,k}(\theta_*)|^{\bar{\gamma}_1}\right\} \\
=& \bigcup_{k\in\mathcal{K}}\bigcup_{j\in\mathcal{J}} \left\{\left(\frac{\delta}{K\bar{D}_1\omega_k(t)}\right)^{\gamma_1} \leq |\hat{X}^j_{k,t} - \mu_{j,k}(\theta_*)|\right\}.
\end{aligned} \tag{8}$$

Define $\bar{X}^j_{k,s}$ as the empirical mean of the first s observations of sub-reward j of arm k, which are i.i.d. based on the assumption that sub-rewards of the same arm are i.i.d. over time. The one-step regret is bounded as follows:

$$\begin{aligned}
&\Pr\{I_{t+1} \neq k^*\} = \Pr\left\{\hat{\theta}_t \in \Theta\backslash\Theta_{k^*}\right\} \\
&\leq \Pr\left\{\delta \leq |\hat{\theta}_t - \theta_*|\right\} \\
&\overset{(a)}{\leq} \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \Pr\left\{\left(\frac{\delta}{K\bar{D}_1\omega_k(t)}\right)^{\gamma_1} \leq |\hat{X}^j_{k,t} - \mu_{j,k}(\theta_*)|\right\} \\
&= \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}\left[\mathbb{1}\left\{\left(\frac{\delta t}{K\bar{D}_1 N_k(t)}\right)^{\gamma_1} \leq |\hat{X}^j_{k,t} - \mu_{j,k}(\theta_*)|\right\}\right] \\
&\overset{(b)}{\leq} \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \mathbb{E}\left[\mathbb{1}\left\{\exists s \in \{1,2,\ldots,t\}, \left(\frac{\delta t}{K\bar{D}_1 s}\right)^{\gamma_1}\right.\right. \\
&\qquad\qquad \left.\left. \leq |\bar{X}^j_{k,s} - \mu_{j,k}(\theta_*)|\right\}\right]
\end{aligned}$$

$$\overset{(c)}{\leq} \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{1}\left\{\left(\frac{\delta t}{K\bar{D}_1 s}\right)^{\gamma_1} \leq |\bar{X}^j_{k,s}-\mu_{j,k}(\theta_*)|\right\}\right]$$

$$\overset{(d)}{\leq} \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{s=1}^{t} 2\left[\exp\left(-2\left(\frac{\delta t}{K\bar{D}_1 s}\right)^{2\gamma_1} s\right)\right]$$

$$= 2JK\sum_{s=1}^{t}\left[\exp\left(-2\left(\frac{\delta}{K\bar{D}_1}\right)^{2\gamma_1}\left(\frac{s}{t}\right)^{1-2\gamma_1} t\right)\right]$$

$$\overset{(e)}{\leq} 2JKt\exp\left(-2\left(\frac{\delta}{K\bar{D}_1}\right)^{2\gamma_1} t\right), \tag{9}$$

where (a) is from (8) and the union bound, (b) follows from the fact that

$$\left\{(\frac{\delta t}{K\bar{D}_1 N_k(t)})^{\gamma_1} \leq |\hat{X}^j_{k,t}-\mu_{j,k}(\theta_*)|\right\}$$

$$\subset \left\{\exists s\in\{1,2,\ldots,t\},\ \left(\frac{\delta t}{K\bar{D}_1 s}\right)^{\gamma_1} \leq |\bar{X}^j_{k,s}-\mu_{j,k}(\theta_*)|\right\},$$

(c) is again from the union bound, (d) is obtained via Hoeffding's inequality, and (e) is based on the fact that $\left(\frac{s}{t}\right)^{1-2\gamma_1} \geq 1$, which is true by Assumption 1 and Proposition 1.

Finally, with the one-step regret bound (9), the total regret $\mathsf{Reg}(T)$ can be bounded by summing (9) over $t = 1,\ldots,T$ as follows:

$$\mathsf{Reg}(T) \leq \sum_{t=1}^{T}\Pr\{I_t \neq k^*(\theta_*)\}$$

$$\leq\ 1+\sum_{t=1}^{T-1} 2JKt\exp\Big(-2(\frac{\delta}{K\bar{D}_1})^{2\gamma_1} t\Big)$$

$$=\ 1+2JK\frac{e^{-\alpha}-Te^{-\alpha T}+(T-1)e^{-\alpha(T+1)}}{(1-e^{-\alpha})^2}.$$

Letting T go to infinity gives

$$\mathsf{Reg}(\infty) \leq 1+2JK\frac{e^{-\alpha}}{(1-e^{-\alpha})^2}.$$

■

Theorem 1 is important as it states that the regret of the ad-greedy policy is *bounded*. This also implies that the ad-greedy policy converges to the optimal arm k^* with probability one.

C. Multi-Dimensional Global Parameters

1) Model and Algorithm: The model in Section III-B and the original GB model of [12] only consider a *scalar* parameter θ. In this section, the GB model and the ad-greedy policy are extended to the *multi-dimensional* case for $\vec{\theta}$. In reality, practical problems often have multiple parameters that affect the system performance. For example, cellular coverage optimization is dependent on many environmental variables, such as deployment area, macrocell footprints, target SINR, etc.

Increasing the parameter dimension brings non-trivial technical difficulty to the GGB model. To highlight the contribution and for the ease of illustration, we study the case where the global parameter is 2-dimensional. Extensions to higher dimensions can be done with the same philosophy, but the resulting analysis is much more complicated. Furthermore, we note that this section still considers the non-monotonic reward functions as in Section III-B.

To accommodate for the vector form of GGB, we re-define some of the previous notations and introduce new notations.

- $\vec{\theta}_* = [\theta_{1*}, \theta_{2*}]$ denotes the true unknown 2-dimensional global parameter. $\vec{\theta} = [\theta_1, \theta_2] \in \Theta$ denotes any parameter vector that is in the parameter set Θ. We normalize Θ such that $||\vec{\theta}-\vec{\theta}'|| \leq 1$ for any $\vec{\theta}, \vec{\theta}' \in \Theta$, where $||\cdot||$ denotes the Euclidean norm.
- $k^* = k^*(\vec{\theta}_*)$ denotes the true best arm. $k^*(\vec{\theta})$ denotes the set of best arm(s) when the global parameter is $\vec{\theta}$.
- $\Theta_k = \{\vec{\theta} \in \Theta | k \in k^*(\vec{\theta})\}$.
- δ is the Euclidean distance between $\vec{\theta}_*$ and the boundary of Θ_{k^*}.
- $\mu_k(\vec{\theta}) \in [0,1]$ is the reward function that is composed of J sub-functions:

$$\mu_k(\vec{\theta}) = \sum_{j=1}^{J}\alpha_j\mu_{j,k}(\vec{\theta}).$$

- $\Psi_{j,k}(X) \subset \Theta$, is the contour of $\mu_{j,k}(\vec{\theta})$ to X, i.e.,

$$\Psi_{j,k}(X) = \left\{\vec{\theta}\in\Theta | \mu_{j,k}(\vec{\theta}) = X\right\}. \tag{10}$$

Furthermore, the following assumptions are imposed for the multi-dimensional GGB problem.

Assumption 2:

1) $J \geq 2$.
2) For $\vec{\theta}_* \in \Theta$ and $k \in \mathcal{K}$, there exists a J-dimensional cube with center $(\mu_{1,k}(\vec{\theta}_*), \mu_{2,k}(\vec{\theta}_*), \ldots, \mu_{J,k}(\vec{\theta}_*))$ and the edge length $2\lambda_k(\vec{\theta}_*)$ such that, for any $j, j' \in \mathcal{J}$, $j \neq j'$, $X \in [\mu_{j,k}(\vec{\theta}_*) - \lambda_k(\vec{\theta}_*), \mu_{j,k}(\vec{\theta}_*) + \lambda_k(\vec{\theta}_*)]$, and $X' \in [\mu_{j',k}(\vec{\theta}_*) - \lambda_k(\vec{\theta}_*), \mu_{j',k}(\vec{\theta}_*) + \lambda_k(\vec{\theta}_*)]$, two contours $\Psi_{j,k}(X)$ and $\Psi_{j',k}(X')$ have exactly one intersection. Denote $\lambda = \min_{k\in\mathcal{K}}(\lambda_k(\vec{\theta}_*))$.
3) For $j, j' \in \mathcal{J}$, $j' \neq j$, $k \in \mathcal{K}$, and $\vec{\theta}, \vec{\theta}' \in \Psi_{j,k}(X)$, there exists $D_{1,j,j',k,X} > 0$ and $0 < \gamma_{1,j,j',k,X} \leq 1$, $D_{2,j,j',k,X} > 0$ and $1 < \gamma_{2,j,j',k,X}$, such that:

$$|\mu_{j',k}(\vec{\theta}) - \mu_{j',k}(\vec{\theta}')| \geq\ D_{1,j,j',k,X}||\vec{\theta}-\vec{\theta}'||^{\gamma_{1,j,j',k,X}}_{j,k,X}$$

$$|\mu_{j',k}(\vec{\theta}) - \mu_{j',k}(\vec{\theta}')| \leq\ D_{2,j,j',k,X}||\vec{\theta}-\vec{\theta}'||^{\gamma_{2,j,j',k,X}}_{j,k,X}$$

where $||\vec{\theta}-\vec{\theta}'||_{j,k,X}$ is the rectification of contour $\Psi_{j,k}(X)$ between $\vec{\theta}$ and $\vec{\theta}'$.

The first assumption is made so that the forecaster can estimate $\vec{\theta}_*$ by using pairs of $\Psi_{j,l}(\hat{X}^j_{l,t})$, $j \in \mathcal{J}$ sets at each play. The second assumption guarantees that as the estimation of X is sufficiently close to the true value, contours of different sub-functions intersect exactly once. This is similar to the Hölder continuity and monotonicity conditions for the scalar parameter case. As we will see in Section IV-A, the objective function in coverage optimization satisfies this requirement. The last assumption is the 2-dimensional counterpart to Assumption 1.

Algorithm 2: The *ad-greedy-2D* policy.

Input : $\mu_{j,k}(\vec{\theta})$ and $\mu_k(\vec{\theta})$, $\forall k \in \mathcal{K}$, $\forall j \in \mathcal{J}$;
Initialize: $N_k(0) = 0$ and $\hat{X}^j_{k,0} = 0$, $\forall k \in \mathcal{K}$, $\forall j \in \mathcal{J}$, and $\vec{\theta}^k_{i,j}(0) = 0$, $\forall k \in \mathcal{K}$, $\forall i, j \in \mathcal{J}, i \neq j$, $t = 1$;
while $t \geq 1$ **do**
  if $t = 1$ **then**
    Select arm I_1 uniformly at random from set $\mathcal{K}$;
  else
    Select arm $I_t = \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\vec{\theta}}^l_{t-1})$;
  end
  Observe sub-rewards $\{X^j_{I_t,t}\}_{j\in\mathcal{J}}$;
  $\hat{X}^j_{k,t} = \hat{X}^j_{k,t-1}$ for $k \in \mathcal{K}\backslash I_t$, $j \in \mathcal{J}$;
  $\hat{X}^j_{I_t,t} = \frac{N_{I_t}(t-1)\hat{X}^j_{I_t,t-1} + X^j_{I_t,t}}{N_{I_t}(t-1)+1}$ for $j \in \mathcal{J}$;
  $N_k(t) = N_k(t-1)$ for $k \in \mathcal{K}\backslash I_t$;
  $N_{I_t}(t) = N_{I_t}(t-1) + 1$;
  $l = \arg\max_{k\in\mathcal{K}} N_k(t)$;
  Construct $\Psi_{i,l}(\hat{X}^i_{l,t})$ for $i \in \mathcal{J}$ from (10);
  Set $\mathcal{A}_t$ as 1;
  for $i, j \in \mathcal{J}$ *and* $i \neq j$ **do**
    if $|\Psi_{i,l}(\hat{X}^i_{l,t}) \cap \Psi_{j,l}(\hat{X}^j_{l,t})| \neq 1$ **then**
      Set $\mathcal{A}_t$ as 0;
      break;
    else
      $\vec{\theta}^l_{i,j}(t) = \Psi_{i,l}(\hat{X}^i_{l,t}) \cap \Psi_{j,l}(\hat{X}^j_{l,t})$;
    end
  end
  if $\mathcal{A}_t = 1$ **then**
    $\hat{\vec{\theta}}^l_t = \frac{1}{J(J-1)} \sum_{i\neq j\in\mathcal{J}} \vec{\theta}^l_{i,j}(t)$;
  else
    Randomly select $\hat{\vec{\theta}}^l_t \in \Theta_l$;
  end
  $t = t + 1$;
end

While these assumptions are necessary in the regret analysis, the proposed policy works well in practice, even when these assumptions do not hold exactly.

With these assumptions, we have the following proposition. The proof is similar to that of Proposition 1 and is omitted.

Proposition 2: For any $k \in \mathcal{K}$, $j, j' \in \mathcal{J}$, $j \neq j'$, and $X \in [0, 1]$, define $\gamma = \frac{1}{\max(\gamma_{1,j,j',k,X})}$ and $D = 2(\frac{1}{\min(D_{1,j,j',k,X})})^\gamma$. Then for any $\vec{\theta}, \vec{\theta}' \in \Psi_{j,k}(X)$, we have

$$||\vec{\theta} - \vec{\theta}'|| \leq \frac{D}{2}|X_{j',k} - X'_{j',k}|^\gamma$$

with $X_{j',k} = \mu_{j',k}(\vec{\theta})$ and $X'_{j',k} = \mu_{j',k}(\vec{\theta}')$.

Now, we are in the position to present the *ad-greedy-2D* policy in Algorithm 2, which enhances the *ad-greedy* policy to handle the 2-dimensional $\vec{\theta}$. Nevertheless, the basic principle remains the same: choose the best arm at time t based on the highest estimated reward, and update the estimated parameter by using the parameter estimates from all sub-functions and all arms. A naive approach would be to estimate the parameter for each dimension *separately*, but this method ignores the intrinsic relationship between the dimensions. The *ad-greedy-2D* policy jointly estimates the parameter over all dimensions.

2) Regret Analysis: We analyze the regret of the *ad-greedy-2D* for a 2-dimensional GGB model. Let $l_t = \arg\max_{k\in\mathcal{K}} N_k(t)$. We will drop the subscript in l_t, when the time slot is clear from the context. Also let

$$\mathcal{G}_t = \bigcap_{j\in\mathcal{J}} \{\hat{X}^j_{l,t} \in [\mu_{j,l}(\vec{\theta}_*) - \lambda_l(\vec{\theta}_*), \mu_{j,l}(\vec{\theta}_*) + \lambda_l(\vec{\theta}_*)]\}$$

denote the *good* event in which the sub-function reward estimates of arm l_t are accurate. By Assumption 2-(2), $\mathcal{A}_t = 1$ when $\mathcal{G}_t$ happens.

First we establish two lemmas that will be used in the proof of the main result in Theorem 2.

Lemma 1: $\mathbb{1}(\mathcal{G}_t)||\hat{\vec{\theta}}^l_t - \vec{\theta}_*|| \leq \frac{1}{J}\sum_{j\in\mathcal{J}} D|\hat{X}^j_{l,t} - X_{j,l}|^\gamma$, where $X_{j,l} = \mu_{j,l}(\vec{\theta}_*)$.

Proof: The inequality is trivial if $\mathbb{1}(\mathcal{G}_t) = 0$, so we only consider $\mathbb{1}(\mathcal{G}_t) = 1$ in the following. Note that a unique $\vec{\theta}^l_{i,j} = \Psi_{i,l}(\hat{X}^i_{l,t}) \cap \Psi_{j,l}(\hat{X}^j_{l,t})$ exists when $\mathcal{G}_t$ is true. On the other hand, $\vec{\theta}_* = \Psi_{j,l}(\mu_{j,l}(\vec{\theta}_*)) \cap \Psi_{i,l}(\mu_{i,l}(\vec{\theta}_*))$ because of Assumption 2-(2). Define $\vec{\theta}^{l,*}_{i,j} = \Psi_{j,l}(\mu_{j,l}(\vec{\theta}_*)) \cap \Psi_{i,l}(\hat{X}^i_{l,t})$. Note that the uniqueness of $\vec{\theta}^{l,*}_{i,j}$ is also guaranteed due to Assumption 2-(2) when $\mathbb{1}(\mathcal{G}_t) = 1$. Thus, when $\mathbb{1}(\mathcal{G}_t) = 1$, the following series of inequalities can be proven using the triangle inequality and Hölder continuity condition given in Proposition 2.

$$\begin{aligned}||\vec{\theta}^l_{i,j} - \vec{\theta}_*|| &\leq ||\vec{\theta}^l_{i,j} - \vec{\theta}^{l,*}_{i,j}|| + ||\vec{\theta}^{l,*}_{i,j} - \vec{\theta}_*|| \\ &\leq \frac{D}{2}|\hat{X}^j_{l,t} - X_{j,l}|^\gamma + \frac{D}{2}|\hat{X}^i_{l,t} - X_{i,l}|^\gamma. \quad (11)\end{aligned}$$

With (11), further derivation leads to

$$\begin{aligned}&\mathbb{1}(\mathcal{G}_t)||\hat{\vec{\theta}}_t - \vec{\theta}_*|| \\ &\leq \frac{1}{J(J-1)} \sum_{i\neq j\in\mathcal{J}} |\vec{\theta}^l_{i,j}(t) - \vec{\theta}_*| \\ &\leq \frac{1}{J(J-1)} \sum_{i\neq j\in\mathcal{J}} \left[\frac{D}{2}|\hat{X}^i_{l,t} - X_{i,l}|^\gamma + \frac{D}{2}|\hat{X}^j_{l,t} - X_{j,l}|^\gamma\right] \\ &= \frac{1}{J}\sum_{j\in\mathcal{J}} D|\hat{X}^j_{l,t} - X_{j,l}|^\gamma. \quad (12)\end{aligned}$$

■

Lemma 2:

$$\{\delta \leq ||\hat{\vec{\theta}}^l_t - \vec{\theta}_*||\} \subset \bigcup_{j\in\mathcal{J}} \left\{\sigma \leq |\hat{X}^j_{l,t} - X_{j,l}|\right\} \quad (13)$$

where $\sigma = \min((\frac{\delta}{D})^{\frac{1}{\gamma}}, \lambda)$ and $X_{j,l} = \mu_{j,l}(\vec{\theta}_*)$.

Proof:

$$\begin{aligned}&\{\delta \leq ||\hat{\vec{\theta}}^l_t - \vec{\theta}_*||\} \\ =\ &\left\{\mathcal{G}_t \bigcap \{\delta \leq ||\hat{\vec{\theta}}^l_t - \vec{\theta}_*||\}\right\} \bigcup \left\{\bar{\mathcal{G}}_t \bigcap \{\delta \leq ||\hat{\vec{\theta}}^l_t - \vec{\theta}_*||\}\right\}\end{aligned}$$

$$
\begin{aligned}
&\subset \left\{\mathcal{G}_t \bigcap \{\delta \le ||\hat{\vec{\theta}}_t^l - \vec{\theta}_*||\}\right\} \bigcup \bar{\mathcal{G}}_t \\
&\subset \left\{\delta \le \mathbb{1}(\mathcal{G}_t)||\hat{\vec{\theta}}_t^l - \vec{\theta}_*||\right\} \bigcup \bar{\mathcal{G}}_t \\
&\subset \left\{\delta \le \frac{1}{J}\sum_{j\in\mathcal{J}} D|\hat{X}_{l,t}^j - X_{j,l}|^\gamma\right\} \bigcup_{j\in\mathcal{J}} \left\{\lambda \le |\hat{X}_{l,t}^j - X_{j,l}|\right\} \\
&\subset \bigcup_{j\in\mathcal{J}} \left\{\frac{\delta}{J} \le \frac{1}{J} D|\hat{X}_{l,t}^j - X_{j,l}|^\gamma\right\} \bigcup_{j\in\mathcal{J}} \left\{\lambda \le |\hat{X}_{l,t}^j - X_{j,l}|\right\} \\
&= \bigcup_{j\in\mathcal{J}} \left\{\min\left(\left(\frac{\delta}{D}\right)^{\frac{1}{\gamma}}, \lambda\right) \le |\hat{X}_{l,t}^j - X_{j,l}|\right\}.
\end{aligned}
$$

■

The regret bound for the *ad-greedy-2D* policy is given in the following theorem.

Theorem 2: The regret of the *ad-greedy-2D* policy for a finite time horizon T is upper bounded by

$$
\mathsf{Reg}(T) \le 1 + 2J(K-1)\frac{e^{-\beta} - Te^{-\beta T} + (T-1)e^{-\beta(T+1)}}{\left(1-e^{-\beta}\right)^2}, \tag{14}
$$

where $\beta = \frac{2\sigma^2}{K}$. Furthermore, the infinite time horizon regret is upper bounded by a constant:

$$
\mathsf{Reg}(\infty) \le 1 + 2J(K-1)\frac{e^{-\beta}}{\left(1-e^{-\beta}\right)^2}. \tag{15}
$$

Proof: Similar to the previous proof of Theorem 1, the one-step regret is analyzed first, and then the (total) regret is bounded. Using Lemma 1 and 2, and Hoeffing's inequality, the one-step regret is bounded as follows:

$$
\begin{aligned}
&\Pr\left\{I_{t+1} \ne k^*(\vec{\theta}_*)\right\} = \Pr\left\{\hat{\vec{\theta}}_t^l \in \Theta \backslash \Theta_{k^*}\right\} \\
\le\ &\Pr\left\{\delta \le |\hat{\vec{\theta}}_t^l - \theta_*|\right\} \\
\overset{(f)}{\le}\ &\sum_{j\in\mathcal{J}} \Pr\left\{\sigma \le |\hat{X}_{l,t}^j - X_{j,l}|\right\} \\
=\ &\sum_{j\in\mathcal{J}} \mathbb{E}\left[\mathbb{1}\left(\sigma \le |\hat{X}_{l,t}^j - X_{j,l}|\right)\right] \\
\overset{(g)}{\le}\ &\sum_{j\in\mathcal{J}} \mathbb{E}\Big[\mathbb{1}\Big\{\exists s \in \{\lceil t/K\rceil, \ldots, t\},\ \exists k \in \mathcal{K}, \\
&\qquad \sigma \le |\bar{X}_{k,s}^j - X_{j,k}|\Big\}\Big] \\
\overset{(h)}{\le}\ &\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{s=\lceil *t/K\rceil}^{t} \mathbb{E}\left[\mathbb{1}\left\{\sigma \le |\bar{X}_{k,s}^j - X_{j,k}|\right\}\right] \\
\overset{(i)}{\le}\ &\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}}\sum_{s=\lceil t/K\rceil}^{t} 2\exp\left(-2\sigma^2 s\right) \\
\overset{(j)}{\le}\ &2J(K-1)t\exp\left(-2\sigma^2\frac{t}{K}\right)
\end{aligned} \tag{16}
$$

where (f) is from Lemma 2 and the union bound, (g) is from the fact that

$$
\begin{aligned}
&\left\{\sigma \le |\hat{X}_{l,t}^j - X_{j,l}|\right\} \\
&\subset \left\{\exists s \in \{\lceil t/K\rceil, \ldots, t\},\ \exists k \in \mathcal{K},\ \sigma \le |\bar{X}_{k,s}^j - X_{j,k}|\right\},
\end{aligned}
$$

(h) is again from the union bound, (i) is obtained via Hoeffding's inequality, and (j) is based on the fact that $s \ge \frac{t}{K}$.

Finally, with the one-step regret bound (16), the total regret $\mathsf{Reg}(T)$ can be bounded by summing (16) over $t = 1, \ldots, T$ as follows:

$$
\begin{aligned}
\mathsf{Reg}(T) &\le 1 + \sum_{t=2}^{T} \Pr\left\{I_t \ne k^*(\theta_*)\right\} \\
&\le\ 1 + \sum_{t=1}^{T-1} 2J(K-1)t\exp\left(-2\sigma^2\frac{t}{K}\right) \\
&=\ 1 + 2J(K-1)\frac{e^{-\beta} - Te^{-\beta T} + (T-1)e^{-\beta(T+1)}}{\left(1-e^{-\beta}\right)^2}.
\end{aligned}
$$

Letting T go to infinity gives

$$
\mathsf{Reg}(\infty) \le 1 + 2J(K-1)\frac{e^{-\beta}}{\left(1-e^{-\beta}\right)^2}.
$$

■

D. Switching Costs

1) Model and Algorithm: One of the important challenges in practice is how to learn the environment without frequent arm changes. This is especially critical for the coverage optimization problem, as changing coverage frequently may cause unnecessary service interruptions such as call drop or temporary service outage. As a result, it is desirable to have a learning policy for coverage optimization that minimizes the changes over time. In the bandit setting, this requirement can be captured by imposing a *switching cost*. More specifically, if the selected arm changes from time t to $t+1$, a switching cost C_{t+1} will be subtracted from the observed reward in $t+1$.

Since the proposed ad-greedy policy has *bounded* regret without considering the switching cost, it is easy to see that directly applying the ad-greedy policy can still result in bounded regret even with switching cost. This holds because the best arm is guaranteed to be found in finite time, and thus, the total switching cost will also be bounded. However, this does not mean that the ad-greedy policy will have the best performance when facing switching cost. Typically, due to the additional penalty of switches, a good bandit algorithm needs to "explore in block". This is done by grouping time slots and not switching during these slots. The proposed block ad-greedy policy that follows this design philosophy is given in Algorithm 3. In order to focus on block exploration, here we present a version of the block ad-greedy policy for the baseline GGB. Later in Section IV-B, a version extended to handle multi-dimensional global parameter is compared against the ad-greedy-2D policy. Thanks to the block exploration structure, we show that the regret due to switching cost is smaller for the block ad-greedy policy.

Algorithm 3: The *block ad-greedy* policy.

Input : $\mu_k(\theta)$, $\forall k \in \mathcal{K}$;
Initialize: $b = 0$, $h(0) = 0$, $\tau = 0$, and $R_k = []$, $N_k = 0$, $\hat{X}_k = 0$, $\forall k \in \mathcal{K}$;
while $b \geq 0$ **do**
  if $b = 0$ **then**
    Select arm I_0 uniformly at random from set $\mathcal{K}$;
  else
    Select arm $I_b = \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_{b-1})$;
  end
  while $\tau < h(b)$ **do**
    Play arm I_b, receive a reward $X_{I_b,t}$ and store in R_{I_b};
    $\tau = \tau + 1$;
  end
  $\hat{X}_{I_b}$ is set to the sample mean of R_{I_b};
  $N_{I_b} = N_{I_b} + h(b) - h(b-1)$;
  $k_b = \arg\max_{k\in\mathcal{K}} N_k$;
  $\hat{\theta}_b = \arg\min_{\theta\in\Theta} |\mu_{k_b}(\theta) - \hat{X}_{k_b}|$;
  $b = b + 1$;
  $\tau = 0$;
end

Another important note regarding the block ad-greedy policy is the choice of the block length $h(b)$, which has not been specified. In the classic MAB problem with switching cost, such as the one considered in [34], the block length is controlled to be *exponentially* increasing over time. This is because, as time goes by, the algorithm has more information about the true values of arms and hence the "block" size should increase to take advantage of the better arm. This construction of block sizes makes sure that the switching cost scales as $o(\log T)$ while the reward without cost still scales as $\mathcal{O}(\log T)$. In our GGB model, however, an exponentially increasing block size $h(b) = 2^b$ may not be necessarily the best choice, as sampling the suboptimal arms still provides useful information in estimating the global parameter and hence helps determine the best arm. In the following regret analysis, we derive regret upper bound for both *exponentially* increasing block length $h(b) = 2^b$ and *linearly* increasing block length $h(b) = bT_c$, where $T_c > 1$ is an integer.

For the regret analysis, we consider a constant switching cost $C_t = C$ for simplicity. We also impose analogues of Assumption 1 and Proposition 1 for $\mu_k(\theta)$: (i) $|\mu_k(\theta) - \mu_k(\theta')| \leq D_{2,k}|\theta - \theta'|^{\gamma_{2,k}}$, (ii) $|\mu_k(\theta) - \mu_k(\theta')| \geq D_{1,k}|\theta - \theta'|^{\gamma_{1,k}}$ for all $k \in \mathcal{K}$, which implies that (i) $|\mu_k(\theta) - \mu_k(\theta')| \leq D_2|\theta - \theta'|^{\gamma_2}$ and (ii) $|\mu_k^{-1}(y) - \mu_k^{-1}(y')| \leq D|y - y'|^{\bar{\gamma}_1}$, where $D_2 = \max_{k\in\mathcal{K}} D_{2,k}$, $\gamma_2 = \min_{k\in\mathcal{K}} \gamma_{2,k}$, $\bar{\gamma}_1 = 1/\gamma_1$, $\gamma_1 = \max_{k\in\mathcal{K}} \gamma_{1,k}$ and $D = \max_{k\in\mathcal{K}} (1/D_{1,k})^{1/\gamma_{1,k}}$.

2) Regret Analysis for the Exponential Block ad-Greedy Policy ($h(b) = 2^b$): The total regret from time $t = 1$ to $T = 2^B - 1$, i.e., the regret incurred in the first B blocks, can be written as

$$\mathsf{Reg}(T) = \sum_{b=0}^{B-1} \mathbb{E}[r_{I_b}(\theta_*)] \tag{17}$$

where $\mathbb{E}[r_{I_b}(\theta_*)]$ denotes the "one-block" regret incurred in block b. For $b > 0$, this can be upper bounded as follows:

$$\mathbb{E}[r_{I_b}(\theta_*)] \leq 2^b \cdot 1 \cdot \Pr\{I_b \neq k^*\} + C \cdot \Pr\{I_{b+1} \neq I_b\}. \tag{18}$$

We start by bounding the first term in (18). Similar to the proof of Theorem 1, we let $\tilde{\mu}_k^{-1}(y) \doteq \arg\min_{\theta\in\Theta} |\mu_k(\theta) - y|$ for $y \in [0,1]$, for which we have $|\tilde{\mu}_k^{-1}(y) - \tilde{\mu}_k^{-1}(y')| \leq D|y - y'|^{\bar{\gamma}_1}$ for all $y, y' \in [0,1]$. We have

$$\begin{aligned}
\{I_b \neq k^*\} &\subset \left\{\delta \leq |\hat{\theta}_{b-1} - \theta_*|\right\} \\
&= \left\{\delta \leq |\tilde{\mu}_{k_{b-1}}^{-1}(\hat{X}_{k_{b-1}}) - \tilde{\mu}_{k_{b-1}}^{-1}(\mu_{k_{b-1}}(\theta_*))|\right\} \\
&\subset \left\{\delta \leq D|\hat{X}_{k_{b-1}} - \mu_{k_{b-1}}(\theta_*)|^{\bar{\gamma}_1}\right\} \\
&= \left\{\left(\frac{\delta}{D}\right)^{\gamma_1} \leq |\hat{X}_{k_{b-1}} - \mu_{k_{b-1}}(\theta_*)|\right\}.
\end{aligned} \tag{19}$$

Following steps similar to the proof of Theorem 1 and using Hoeffding's inequality, we obtain

$$\begin{aligned}
\Pr\{I_b \neq k^*\} &\leq \Pr\left\{\left(\frac{\delta}{D}\right)^{\gamma_1} \leq |\hat{X}_{k_{b-1}} - \mu_{k_{b-1}}(\theta_*)|\right\} \\
&\leq 2\exp\left(-2\left(\frac{\delta}{D}\right)^{2\gamma_1} \frac{2^{b-1}}{K}\right)
\end{aligned} \tag{20}$$

where (20) follows from the fact that $N_{k_{b-1}} \geq 2^{b-1}/K$.

Let $\eta = 2\left(\frac{\delta}{D}\right)^{2\gamma_1}/K$. Next, the second item in (18) can be bounded as follows:

$$\begin{aligned}
&\Pr\{I_{b+1} \neq I_b\} \leq \Pr\{I_{b+1} \neq k^*\} + \Pr\{I_b \neq k^*\} \\
&\leq 2\exp\left(-\eta 2^b\right) + 2\exp\left(-\eta 2^{b-1}\right).
\end{aligned} \tag{21}$$

Finally, plugging (20) and (21) back to (18) and (17), we obtain

$$\begin{aligned}
\mathsf{Reg}(T) &\leq 1 + \sum_{b=1}^{B-1} \left(2^{b+1} + 2C\right) \exp\left(-\eta 2^{b-1}\right) \\
&\quad + 2C \sum_{b=1}^{B-1} \exp\left(-\eta 2^b\right) \\
&\leq 1 + 4(C+1)\frac{e^{-\eta}}{(1-e^{-\eta})^2}.
\end{aligned}$$

3) Regret Analysis for the Linear Block ad-Greedy Policy ($h(b) = bT_c$): The total regret from time $t = 1$ to $T = 1 + T_c(B-1)B/2$, i.e., the regret incurred in the first B blocks can be written as

$$\mathsf{Reg}(T) = \sum_{b=0}^{B-1} \mathbb{E}[r_{I_b}(\theta_*)] \tag{22}$$

where $\mathbb{E}[r_{I_b}(\theta_*)]$ denotes the "one-block" regret in block b. This can be upper bounded as follows:

$$\mathbb{E}[r_{I_b}(\theta_*)] \leq T_c \cdot b \cdot \Pr\{I_b \neq k^*\} + C \cdot \Pr\{I_{b+1} \neq I_b\} \tag{23}$$

for $b > 0$.

The first item in (23) can be further bounded as follows. First, we have

$$\begin{aligned}\{I_b \neq k^*\} &\subset \left\{\delta \leq |\hat{\theta}_{b-1} - \theta_*|\right\} \\ &= \left\{\delta \leq |\tilde{\mu}_{k_{b-1}}^{-1}(\hat{X}_{k_{b-1}}) - \mu_{k_{b-1}}^{-1}(\tilde{\mu}_{k_{b-1}}(\theta_*))|\right\} \\ &\subset \left\{\delta \leq D|\hat{X}_{k_{b-1}} - \mu_{k_{b-1}}(\theta_*)|^{\bar{\gamma}_1}\right\} \\ &= \left\{\left(\frac{\delta}{D}\right)^{\gamma_1} \leq |\hat{X}_{k_{b-1}} - \mu_{k_{b-1}}(\theta_*)|\right\}. \end{aligned} \tag{24}$$

Applying Hoeffding's inequality, we obtain

$$\begin{aligned}\Pr\{I_b \neq k^*\} &\leq \Pr\left\{\left(\frac{\delta}{D}\right)^{\gamma_1} \leq |\hat{X}_{k_{b-1}} - \mu_{k_{b-1}}(\theta_*)|\right\} \\ &\leq 2\exp\left(-\left(\frac{\delta}{D}\right)^{2\gamma_1}\frac{(b-1)^2 T_c}{K}\right) \end{aligned} \tag{25}$$

where (25) follows from the fact that $N_{k_{b-1}} \geq (b-1)^2 T_c/(2K)$.

Let $\kappa = (\frac{\delta}{D})^{2\gamma_1}\frac{T_c}{K}$. Next, the second term in (23) can be bounded as follows.

$$\begin{aligned}&\Pr\{I_{b+1} \neq I_b\} \leq \Pr\{I_{b+1} \neq k^*\} + \Pr\{I_b \neq k^*\} \\ &\leq 2\exp\left(-\kappa b^2\right) + 2\exp\left(-\kappa(b-1)^2\right). \end{aligned} \tag{26}$$

Finally, plugging (25) and (26) back to (23) and (22), the regret is upper bounded as:

$$\begin{aligned}\mathsf{Reg}(T) &\leq 1 + \sum_{b=1}^{B-1}(2T_c b + 2C)\exp\left(-\kappa(b-1)^2\right) \\ &\quad + 2C\sum_{b=1}^{B-1}\exp\left(-\kappa b^2\right) \\ &\leq 1 + 2T_c\sum_{b=1}^{B-2} b e^{-\kappa b^2} + (2T_c + 4C)\sum_{b=0}^{B-2} e^{-\kappa b^2} \\ &\leq 1 + 2T_c\frac{e^{-\frac{1}{2}}}{\sqrt{2\kappa}} + 2T_c\int_0^{B-2} z e^{-\kappa z^2}\,\mathrm{d}z \\ &\quad + (2T_c + 4C)\sum_{b=0}^{B-2} e^{-\kappa b} \\ &\leq 1 + 2T_c\frac{e^{-\frac{1}{2}}}{\sqrt{2\kappa}} + T_c\frac{1}{\kappa} + (2T_c + 4C)\frac{1}{e^{\kappa} - 1}. \end{aligned}$$

Although both linear and exponential block size can achieve a bounded regret for block ad-greedy with switching cost, the analysis here only reflects the upper bounds of the regret, not necessarily the actual performance. We will evaluate their performances in the coverage optimization problems and report the simulation results in Section IV-B.

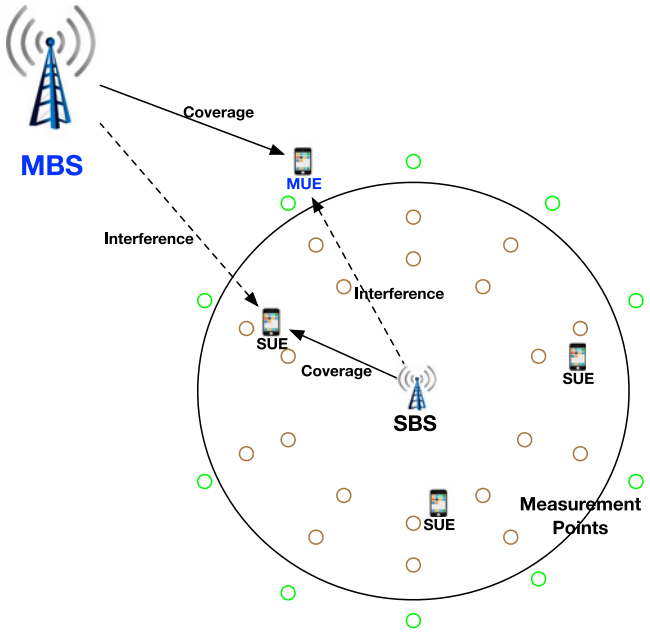


Fig. 1. Illustration of a co-channel deployment of MBS and SBS with overlapping coverage. MBS causes interference to SBS users (SUE), while SBS creates leakage to MBS users (MUE) that are close to the SBS coverage but cannot be served by the SBS.

IV. Application of GGB to Cellular Coverage Optimization

In this section, we describe how to apply the greedy policies developed in Section III to the cellular coverage optimization problem, and evaluate their performances via numerical simulations.

A. Coverage Optimization Problem Formulation

We focus on the SBS deployment that is co-channel with an overlaid macro base station (MBS) coverage, as illustrated in Fig. 1. The design objective is to set the SBS transmit power such that: (1) it provides sufficient coverage to the intended coverage area (e.g., a warehouse or an office room), which is not known a priori; and (2) it limits the "leakage" to users outside the intended coverage area. If coverage area and RF footprints are known to the algorithm, this problem can be solved by formulating an optimization problem that maximizes the PIF which balances coverage and leakage [33]. When such information is entirely unavailable, it can be formulated as an online learning problem similar to [5], which is a general approach that relies on limited assumptions about the deployment. However, the lack of structure to the problem modeling in [5] also sacrifices the algorithm performance when it is indeed known to the system designer [33].

For simplicity, the intended SBS coverage area is approximated by a circle of radius d, which is unknown to the algorithm. The set of measurement points for SBS coverage is denoted as N_{in} with cardinality n_{in}, and the set of measurement points outside the SBS coverage for SBS leakage is denoted as N_{out} with cardinality n_{out}. Both n_{in} and n_{out} are fixed irrespective of

the deployment. Furthermore, we assume that the measurement points have uniformly distributed distances to the SBS, for both inside and outside routes. Such uniform spacing has been similarly adopted in [35] for evaluation of the area spectral efficiency. In practice, choosing the measurement points for coverage estimation and optimization has been studied in [33], which has argued that uniform sampling of the area offers the least bias to the algorithm. Practical methods to collect such measurement reports without repeated measurements have also been proposed in [33]. Furthermore, we note that this assumption on measurement point placement is not crucial to our algorithm because it only affects the specific format of the objective function. In other words, other reasonable setup for the measurement points can be adopted and it will only result in a change of the objective function as described in (27).

We consider maximizing the total spectral efficiency of the measurement points under a proportional fairness constraint. This has been proved to be equivalent to maximizing the sum of logarithms of the user rate [36]. Formally, we have

$$f_k(d, P_m) = \alpha \sum_{i \in N_{\text{in}}} g(R_{\text{SBS},i}(d, P_m)) + (1-\alpha) \sum_{i \in N_{\text{out}}} g(R_{\text{MBS},i}(d, P_m)), \quad (27)$$

where d denotes the radius of the intended coverage area, P_m denotes the average MBS received signal power, $R_{\text{SBS}}(d, P_m)$ and $R_{\text{MBS}}(d, P_m)$ denote the rate function for SBS-served and MBS-served users, respectively, and α is a weight coefficient that balances coverage and leakage. A large α suggests that the design favors having sufficient SBS coverage over leakage that affects MBS users, and vice versa. Subscript k indicates that SBS adopts transmit power $P_k \in \{P_1, \ldots, P_K : P_1 < \cdots < P_K\}$. We note that the reward function (27) is defined for each individual SBS if a distributed deployment is considered, in which (27) can be different across SBSs.

For evaluation of the proposed solutions, in the following we focus on some specific system configurations. Since we have assumed a uniform placement of measurement points for N_{in} and N_{out}, we can re-write (27) as

$$\begin{aligned} f_k(d, P_m) &= \alpha \sum_{i=1}^{n_{\text{in}}} \log\left(R_{\text{SBS}}\left(\frac{i}{n_{\text{in}}} d, P_m\right)\right) \\ &\quad + (1-\alpha) \sum_{i=1}^{n_{\text{out}}} \log\left(R_{\text{MBS}}\left(\left(1 + \frac{i}{n_{\text{in}}}\right) d, P_m\right)\right) \\ &\doteq \alpha f_k^{(1)}(d, P_m) + (1-\alpha) f_k^{(2)}(d, P_m). \end{aligned} \quad (28)$$

Denoting the pathloss function as $\mathsf{PL}(d)$, the received signal power at distance d_1 from the SBS with transmit power P_k can be written as

$$P_r(d_1)[\text{dB}] = P_k[\text{dB}] - \mathsf{PL}(d_1)[\text{dB}] + \delta, \quad (29)$$

where δ denotes the shadowing fading in the dB domain. Note that $\mathsf{PL}(d)$ can be any reasonable pathloss model that fits the environment. The corresponding SINR at distance d_1 is

$$\mathsf{SINR}_{\text{SBS}}(d_1, P_m) = \frac{P_r(d_1)}{P_m + N_0}, \quad (30)$$

where N_0 denotes the uncontrolled noise and interference. Finally, we apply the Shannon capacity formula for the SBS and MBS user rate:

$$R_{\text{SBS}}(d_1, P_m) = \log\left(1 + \frac{P_r(d_1)}{P_m + N_0}\right), \quad (31)$$

$$R_{\text{MBS}}(d_2, P_m) = \log\left(1 + \frac{P_m}{P_r(d_2) + N_0}\right). \quad (32)$$

To see that the GGB model can be used in this problem, we note that each power level P_k can be viewed as an arm. The average reward of arm k can be written as $\mu_k = \bar{f}_k(d, P_m)$, which is a function of two parameters, d and P_m. The function $\bar{f}_k(d, P_m)$ can be written as $\alpha \bar{f}_k^{(1)}(d, P_m) + (1-\alpha)\bar{f}_k^{(2)}(d, P_m)$. Note that the first sub-function is decreasing while the second sub-function is increasing with respect to both d and P_m. Hence the problem formulation satisfies the prerequisite of GGB, and we will evaluate the performance of the proposed algorithm in the next section.

TABLE I
SIMULATION PARAMETERS

Parameters	Value
n_{in}	50
n_{out}	50
Noise density	-174 dBm/Hz
Bandwidth	20 MHz
Carrier frequency	2.1 GHz
Time horizon	1000 time slots
P_m	[-90, -70] dBm
L_w	5 dB
P_k	[-15, 10] dBm
d	[10, 50] m
α	0.5
d_0	10 m

B. Numerical Simulations

We resort to numerical simulations to verify the effectiveness of the developed ad-greedy policies in the coverage optimization problem. The simulated deployment scenario is the same as in Section IV-A. The objective is to maximize the sum of logarithms of the user rates, as in (27). In the simulations, we use the same feedback mechanism as [33]: at each time slot, UEs report measured SINRs of their serving BSs at the corresponding measurement points. We adopt the standard 3GPP dual-strip pathloss model for urban deployment, which has been recommended for system simulations of small cells and heterogeneous networks [11]:

$$\mathsf{PL}(d)[\text{dB}] = 38.46 + 20\log_{10}(d) + 0.7d + L_w, d \geq d_0. \quad (33)$$

Other important simulation parameters are summarized in Table I.

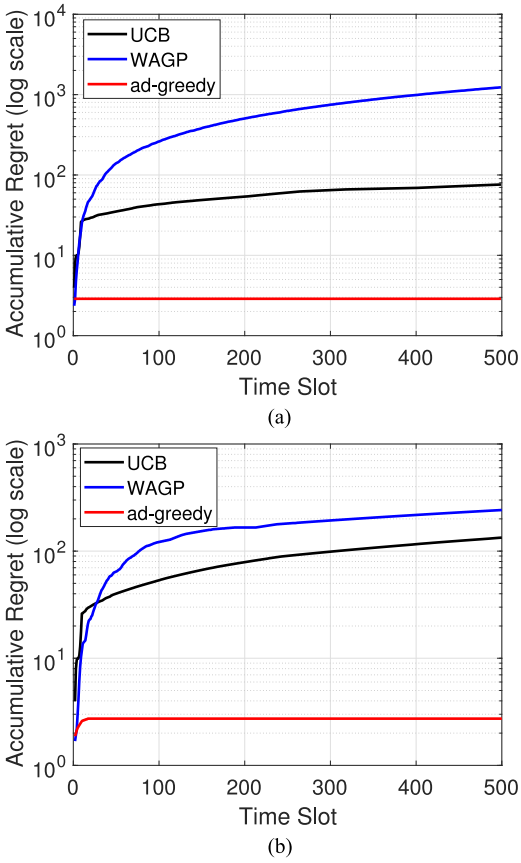


Fig. 2. Accumulative regret comparison of ad-greedy, WAGP and UCB, with P_m (a) and d (b) as the single global parameter.

In the simulations, we focus on evaluating the developed ad-greedy policy and compare its performance with two alternatives: WAGP algorithm that was proposed for the original GB in [12], and the celebrated UCB algorithm [10] for stochastic MAB. Note that UCB is not designed for parametric bandit models as GB or GGB, and the numerical comparison is only meant to demonstrate the improvement thanks to exploiting the structure of GGB. WAGP, on the other hand, is designed only for single-parameter monotonic reward functions, and the numerical comparison will shed light into its effectiveness in the considered coverage optimization problem.

In the first set of simulations, we fix either P_m or d, and let the other parameter be the single global parameter. This will satisfy the single-parameter requirement of WAGP. Fig. 2 reports the simulation results for both cases. When the single global parameter is chosen to be P_m (d), the corresponding d (P_m) is set as 30 m (-85 dBm). As can be seen from the plots, the ad-greedy policy significantly outperforms UCB and WAGP. In addition, WAGP performs even worse compared to UCB, which does not exploit the parametric structure of the reward functions. This is because our average reward functions are non-monotonic, and WAGP is designed only for monotonic reward functions. This model mismatch results in worse performance than not exploiting the structure at all.

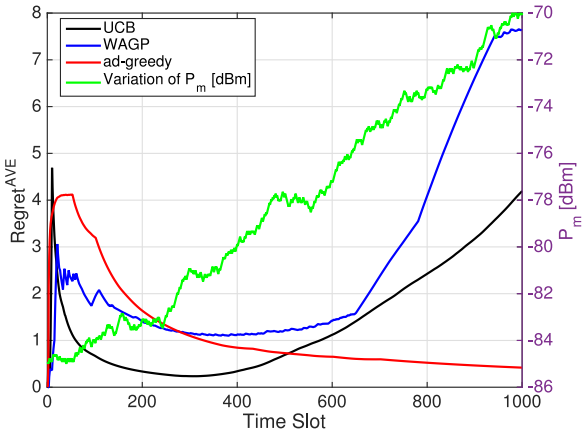


Fig. 3. Average regret versus time with non-stationary P_m. For the understanding purpose, the variation of P_m is also plotted.

Next, we evaluate the regret performance of these three algorithms when the parameters are *non-stationary*. For example, if the coverage environment experiences some changes, the average MBS received signal power may be different. Fig. 3 reports the numerical comparison under non-stationary P_m. In this simulation, the variation of P_m models the change from a cell edge (small P_m) to a cell site (large P_m). It is worth noting that we plot the *average* regret because the reward function is time-varying. We see from Fig. 3 that the ad-greedy policy initially suffers from the non-stationarity as the parameter estimation is not accurate due to both the change of P_m and the insufficient estimation at the beginning, but gradually converges to the true estimate and catches up with the non-stationarity, while WAGP again suffers from the drawback of non-monotonicity of the reward functions.

Having verified the performance improvement of the ad-greedy policy with non-monotonic reward functions, we now turn our attention to the simulations with a 2-dimensional global parameter setting as in (27). Again, we compare the proposed ad-greedy-2D policy with WAGP and UCB. Note that WAGP cannot take 2-dimensional parameters, and thus we either fix d or P_m and use the other one as the scalar parameter. From the simulation results reported in Fig. 4, we can clearly see the benefit of the ad-greedy-2D policy when dealing with 2-dimensional global parameter (d, P_m), as it has the lowest regret throughout the simulations. A closer look at the one-step regret in Fig. 4(b) further reveals the advantage of ad-greedy-2D: it only suffers at the beginning and then quickly converges to the optimal arm, while all other methods have higher one-step regret. Such "bounded regret" behavior has been theoretically analyzed in Theorem 2, and is now numerically verified in Fig. 4.

We further plot the individual sub-rewards, i.e., coverage and leakage, as a function of time slot in Fig. 5(a) and (b), respectively. As expected, both coverage and leakage functions fluctuate around the optimal values during some initial period, when the ad-greedy-2D algorithm tries to learn the deployment while simultaneously maintaining good initial performance. The algorithm converges to the optimal coverage and leakage values,

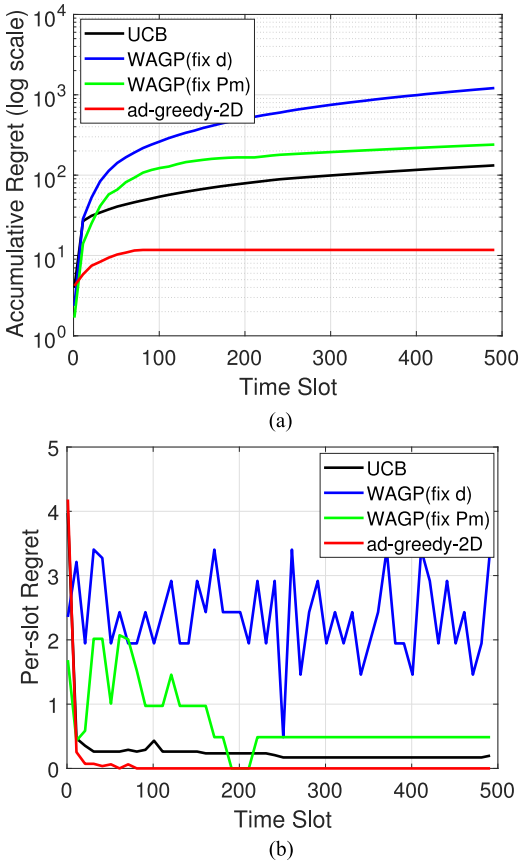


Fig. 4. Regret comparison of ad-greedy, WAGP and UCB with (d, P_m) as the 2-dimensional global parameter. WAGP can be used when we fix either d or P_m. (a) Accumulative regret vs. time. (b) One-step regret vs. time.

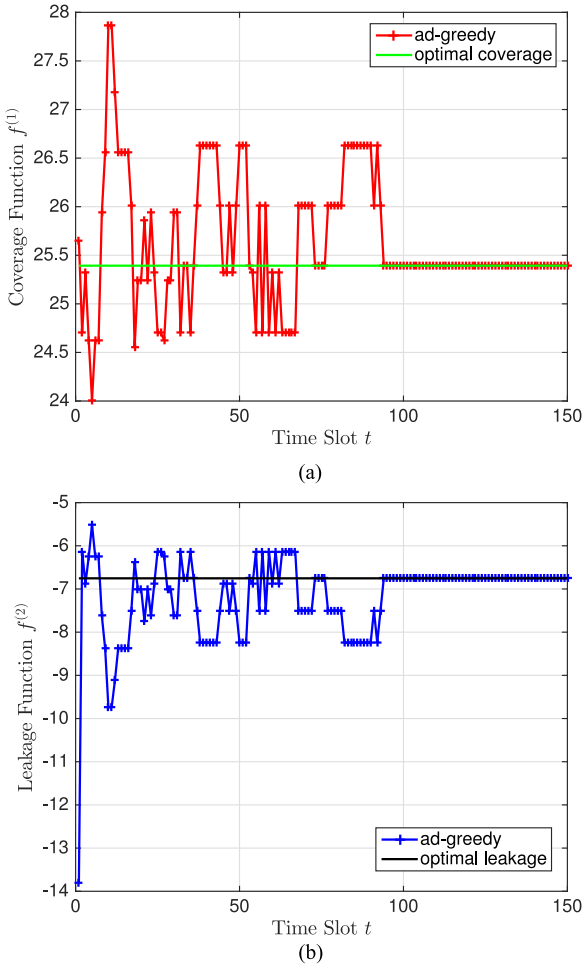


Fig. 5. Coverage and leakage sub-function evolution with time slot t. (a) Coverage vs. time. (b) Leakage vs. time.

via setting the optimal transmit power, at around 100 time slots. This is also consistent with the regret performance reported in Fig. 4.

Finally, we evaluate the proposed algorithm for switching cost. The results are reported in Fig. 6 for the 2-dimensional global parameter (d, P_m). We set $C = 25$ to penalize the change of coverage areas, which is significantly higher than the (normalized) one-step regret and hence highlights the importance of addressing switching cost in the algorithm. We compare the block ad-greedy policy in Algorithm 3 (using linear block size) with the ad-greedy policy in Algorithm 2 which does not consider the switching cost, block UCB, and WAGP with either d or P_m fixed. It is worth noting that in order to compare these algorithm fairly, we have adopted the same blocking philosophy for UCB so that it can also handle the switching cost. Clearly, algorithms that do not consider the severe penalty of switching cost incur significantly higher regret. Furthermore, we can see from Fig. 6(b) that both the block ad-greedy policy and block UCB have very small one-step regret. This is further verified from Fig. 6(c), where we plot the total number of arm switchings of all algorithms. It is evident that the benefit of both the block ad-greedy policy and block UCB is due to the blocking structure that reduces switches, and a careful examination of the simulation results shows that the block ad-greedy policy outperforms block UCB. The reason for this performance improvement is that even though the block ad-greedy policy may stuck in sub-optimal arms for certain durations because of the block structure, such periods are not wasted as it can still estimate the global parameter effectively, and as a result when the block ends, the algorithm will have a more accurate estimation and hence a better choice of the next arm to play. The ad-greedy policy, on the other hand, suffers from larger initial regret, because it does not take switching cost into consideration. However, this loss becomes negligible as times goes by, which can be seen in Fig. 6(b). This is because the ad-greedy policy is guaranteed to find the optimal arm in finite time, and once this happens, there will be no further arm changes. It is worth noting that this behavior is very different to the standard stochastic MAB with switching cost [34], where the goal is simply to control the number of arm switches to scale as $o(\log T)$ whereas the optimal regret scales as $\mathcal{O}(\log T)$. In the GGB model, the regret of Algorithm 2 is already proven to be *finite* in time, and thus even the ad-greedy policy, which does not consider switching cost incurs bounded regret.

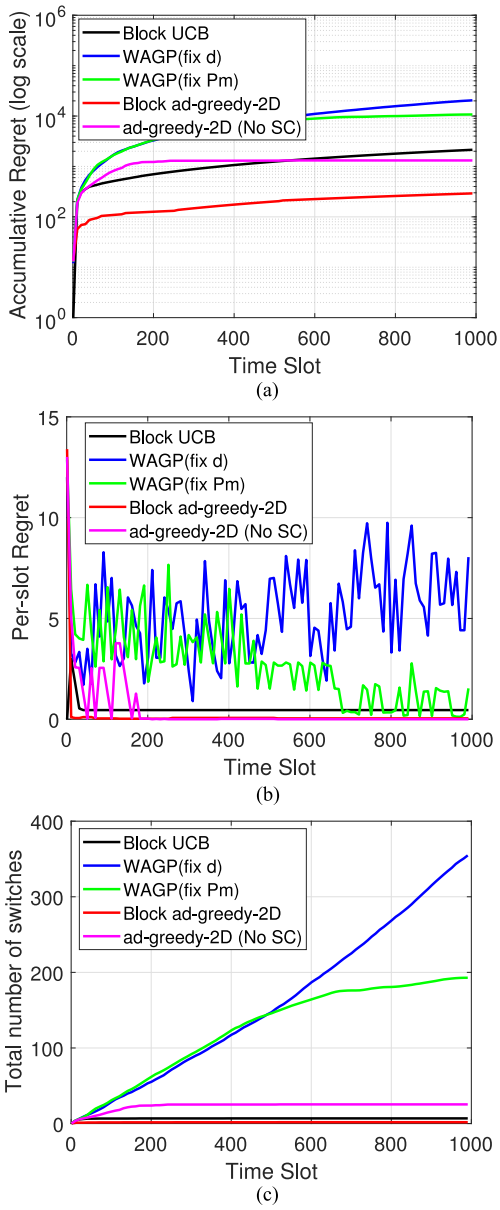


Fig. 6. Regret and switches comparison of block ad-greedy, ad-greedy, WAGP and block UCB with global parameter (d, P_m) and switching cost. (a) Accumulative regret vs. time. (b) One-step regret vs. time. (c) Total number of switches vs. time.

V. Conclusion

We have extended the global bandit model to a more general setting, allowing for non-monotonic decomposable reward functions with multi-dimensional global parameters and switching costs. Such extensions are technically non-trivial and we have developed the *ad-greedy* policies to achieve bounded regret for the generalized global bandit model. This is intuitively reasonable because although accumulative reward may suffer when a sub-optimal arm is played, the algorithm still gains from a better estimation of the global parameter.

The motivation behind the GGB model was to address the cellular coverage optimization problem, which we used as the case study and demonstrated the advantages of the ad-greedy policies over existing solutions via numerical simulations. However, the GGB model and the proposed algorithms are very general and can be applied to other problems, such as interference mitigation [37], load balancing [38], energy-efficient wireless networks [39], and cognitive radio [40], [41]. Applications of the GGB model and ad-greedy policies to these engineering problems are an interesting future research direction.

References

[1] Cisco, "Cisco visual networking index: Global mobile data Traffic forecast update, 2015–2020," San Jose, CA, USA, Feb. 2016.

[2] T. Quek, G. de la Roche, I. Guvenc, and M. Kountouris, *Small Cell Networks: Deployment, PHY Techniques, and Resource Allocation*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[3] J. Ramiro and K. Hamied, *Self-Organizing Networks (SON): Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE*. New York, NY, USA: Wiley, Nov. 2011.

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

[5] Z. Wang and C. Shen, "Small cell transmit power assignment based on correlated bandit learning," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 4, pp. 1–16, Apr. 2017.

[6] M. Simsek, M. Bennis, and I. Guvenc, "Context-aware mobility management in HetNets: A reinforcement learning approach," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2015, pp. 1536–1541.

[7] C. Shen, C. Tekin, and M. van der Schaar, "A non-stochastic learning approach to energy efficient mobility management," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3854–3868, Dec. 2016.

[8] C. Shen and M. van der Schaar, "A learning approach to frequent handover mitigations in 3GPP mobility protocols," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2017, pp. 1–6.

[9] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proc. IEEE Symp. New Frontiers Dyn. Spectr.*, Apr. 2010, pp. 1–9.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2/3, pp. 235–256, May 2002.

[11] 3GPP, "Evolved universal terrestrial radio access; Further advancements for E-UTRA physical layer aspects," 3GPP, Sophia Antipolis, France, TR 36.814, 2010.

[12] O. Atan, C. Tekin, and M. Schaar, "Global multi-armed bandits with Hölder continuity," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, San Diego, CA, USA, May 2015, pp. 28–36. [Online]. Available: http://proceedings.mlr.press/v38/atan15.html

[13] O. Atan, C. Tekin, and M. van der Schaar, "Global bandits," arXiv:1503.08370, 2017.

[14] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, pp. 4–22, 1985.

[15] P. Reverdy, V. Srivastava, and N. Leonard, "Modeling human decision-making in generalized Gaussian multiarmed bandits," *Proc. IEEE*, vol. 102, no. 4, pp. 544–571, Apr. 2014.

[16] V. Srivastava, P. Reverdy, and N. Leonard, "Correlated multiarmed bandit problem: Bayesian algorithms and regret analysis," ArXiv e-prints, Jul. 2015.

[17] 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.

[18] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 2312–2320.

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

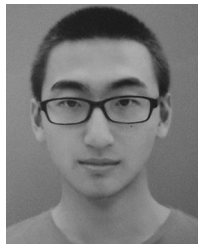
[20] T. H. Li and K. S. Song, "On asymptotic normality of nonlinear least squares for sinusoidal parameter estimation," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4511–4515, Sep. 2008.

[21] 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.

[22] 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6630–6634.

[23] 3GPP, "Evolved universal terrestrial radio access; Self-configuring and self-optimizing network (SON) use cases and solutions," 3GPP, Sophia Antipolis, France, TR 36.902, 2010.
[24] Cisco, San Jose, CA, USA, Cisco SON for Small Cells, White Paper, 2015.
[25] Qualcomm Research, San Diego, CA, USA, Cost-effective Enterprise Small Cell Deployment with UltraSON, Apr. 2016, White Paper.
[26] G. Hampel, K. L. Clarkson, J. D. Hobby, and P. A. Polakos, "The tradeoff between coverage and capacity in dynamic optimization of 3G cellular networks," in *Proc. IEEE 58th Veh. Technol. Conf.*, Oct. 2003, vol. 2, pp. 927–932.
[27] A. Engels, M. Reyer, X. Xu, R. Mathar, J. Zhang, and H. Zhuang, "Autonomous self-optimization of coverage and capacity in LTE cellular networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 1989–2004, Jun. 2013.
[28] O. N. C. Yilmaz, S. Hamalainen, and J. Hamalainen, "Analysis of antenna parameter optimization space for 3GPP LTE," in *Proc. IEEE 70th Veh. Technol.—Fall*, Sep. 2009, pp. 1–5.
[29] S. Jin, J. Wang, Q. Sun, M. Matthaiou, and X. Gao, "Cell coverage optimization for the multicell Massive MIMO uplink," *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5713–5727, Dec. 2015.
[30] S. Berger, M. Simsek, A. Fehske, P. Zanier, I. Viering, and G. Fettweis, "Joint downlink and uplink tilt-based self-organization of coverage and capacity under sparse system knowledge," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2259–2273, Apr. 2016.
[31] H. Claussen, L. T. W. Ho, and L. G. Samuel, "Self-optimization of coverage for femtocell deployments," in *Proc. Wireless Telecommun. Symp.*, Apr. 2008, pp. 278–285.
[32] S. Nagaraja *et al.*, "Transmit power self-calibration for residential UMTS/HSPA+ femtocells," in *Proc. Int. Symp. Modeling Optim. Mobile, Ad Hoc, and Wireless Netw.*, May 2011, pp. 451–455.
[33] S. Nagaraja *et al.*, "Downlink transmit power calibration for enterprise femtocells," in *Proc. IEEE Veh. Technol. Conf.*, 2011, pp. 1–5.
[34] R. Agrawal, M. V. Hegde, and D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost," *IEEE Trans. Autom. Control*, vol. 33, no. 10, pp. 899–906, Oct. 1988.
[35] M. S. Alouini and A. Goldsmith, "Area spectral efficiency of cellular mobile radio systems," in *Proc. IEEE Veh. Technol. Conf.*, May 1997, vol. 2, pp. 652–656.
[36] J.-Y. Le Boudec, "Rate adaptation, congestion control and fairness: A tutorial," 2005, unpublished manuscript. [Online]. Available: http://ica1www.epfl.ch/PS_files/LEB3132.pdf
[37] M. Bennis, S. Perlaza, P. Blasco, Z. Han, and H. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3202–3212, Jul. 2013.
[38] I. Koutsopoulos and L. Tassiulas, "Joint optimal access point selection and channel assignment in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 15, no. 3, pp. 521–532, Jun. 2007.
[39] N. Mastronarde and M. van der Schaar, "Fast reinforcement learning for energy-efficient wireless communication," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6262–6266, Dec. 2011.
[40] H. P. Shiang and M. van der Schaar, "Delay-sensitive resource management in multi-hop cognitive radio networks," in *Proc. 3rd IEEE Symp. New Frontiers Dyn. Spectr. Access Netw.*, Oct. 2008, pp. 1–12.
[41] M. S. Greco, F. Gini, and P. Stinco, "Cognitive radars: Some applications," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2016, pp. 1077–1082.

Cong Shen (S'01–M'09–SM'15) received the B.S. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China., in 2002 and 2004, respectively He received the Ph.D. degree from the Electrical Engineering Department, UCLA, Los Angeles, CA, USA, in 2009. From 2009 to 2014, he was with Qualcomm Research, San Diego, CA, USA, where he focused on various cutting-edge research topics including cognitive radio, TV white space, heterogeneous and ultradense networks. In 2015, he returned to academia and joined the School of Information Science and Technology, University of Science and Technology of China (USTC) as the 100 Talents Program Professor. His general research interests include communication theory, wireless networks, and machine learning. Currently, he serves as an editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.



Ruida Zhou is currently a Senior Undergraduate Student at the School of Information Science and Technology, University of Science and Technology of China (USTC), Hefei, China. His research interests include machine learning, information theory, and statistics.

Cem Tekin (M'13) received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, in 2008, 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 an Assistant Professor with the Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey. From February 2013 to January 2015, he was a Postdoctoral Scholar with the University of California, Los Angeles. His research interests include machine learning, multiarmed bandit problems, data mining, multiagent systems, and smart healthcare. He received the University of Michigan Electrical Engineering Departmental Fellowship in 2008, and the Fred W. Ellersick Award for the best paper in MILCOM 2009.

Mihaela van der Schaar (F'09) is currently a Man Professor of quantitative finance with the Oxford–Man Institute of Quantitative Finance (OMI), Oxford, U.K. and the Department of Engineering Science, University of Oxford, Oxford, U.K., a Fellow of Christ Church College, and a Faculty Fellow of the Alan Turing Institute, London, UK. She is also a Chancellor's Professor of electrical engineering with the University of California, Los Angeles, CA, USA. Her current research interests include machine learning, data science and decisions for medicine, education, and finance. She was a Distinguished Lecturer of the Communications Society, the Editor-in-Chief of the IEEE TRANSACTIONS ON MULTIMEDIA, a Senior Editorial Board member of the Editorial Board of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING (JSTSP) and the IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS (JETCAS). She received an NSF CAREER Award (2004), the Best Paper Award from IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS for Video Technology (2005), the Okawa Foundation Award (2006), the IBM Faculty Award (2005, 2007, 2008), the Most Cited Paper Award from the *EURASIP: Image Communications Journal* (2006), the Gamenets Conference Best Paper Award (2011), and the 2011 IEEE Circuits and Systems Society Darlington Award Best Paper Award. She played a lead role in the MPEG video compression and streaming international standardization activities for which she received 3 ISO Awards and for which she holds 33 granted US patents.