Accepted from Open Call

Multi-Armed Bandits with Application to 5G Small Cells

Setareh Maghsudi and Ekram Hossain

Abstract

Due to the pervasive demand for mobile services, next generation wireless networks are expected to be able to deliver high data rates while wireless resources become more and more scarce. This requires the next generation wireless networks to move toward new networking paradigms that are able to efficiently support resource-demanding applications such as personalized mobile services. Examples of such paradigms foreseen for the emerging 5G cellular networks include very densely deployed small cells and device-to-device communications. For 5G networks, it will be imperative to search for spectrum and energy-efficient solutions to the resource allocation problems that i) are amenable to distributed implementation, ii) are capable of dealing with uncertainty and lack of information, and iii) can cope with users' selfishness. The core objective of this article is to investigate and establish the potential of the MAB framework to address this challenge. In particular, we provide a brief tutorial on bandit problems, including different variations and solution approaches. Furthermore, we discuss recent applications as well as future research directions. In addition, we provide a detailed example of using an MAB model for energy-efficient small cell activation in 5G networks.

Introduction

Due to the ever increasing need for mobile services, we expect a massive growth in demand for wireless services in the years to come. As a result, future mobile networks are expected to accommodate new communications, networking, and energy efficiency concepts, for instance, small cells, device-to-device (D2D) communications, and energy harvesting. In order to realize such novel concepts, system designers face some fundamental challenges. For instance, while in traditional cellular networks some channel state information (CSI) is assumed to be available at some base station (BS) and wireless devices, such information is not likely to be affordable in emerging hierarchical or distributed networks. Centralized resource management in such networks (especially in a very dense deployment scenario) will incur excessive complexity; therefore, distributed approaches will need to be developed, which might trigger competitive user behavior. Moreover, while any planning and scheduling in distributed networks depends heavily on the available energy at network nodes, energy levels are not observable in general. Therefore, it is evident that in the future, efficient and robust wireless communications design needs to deal with inherent uncertainty and lack of information, as well as possible competition for resources, in a distributed manner. As a result, it becomes imperative to search for new mathematical tools to deal with networking problems, which in general involve both uncertainty and conflict.

Multi-armed bandit (MAB) is a class of sequential optimization problems, where, in the most seminal form, given a set of arms (actions), a player pulls an arm at each round in order to receive some reward. The rewards are not known to the player in advance; however, upon pulling any arm, the instantaneous reward of that arm is revealed. In such an unknown setting, at each trial, the player may lose some reward (or incur some cost) due to not selecting the best arm instead of the played arm. This loss is referred to as *regret*. The player decides which arm to pull in a sequence of trials so that its accumulated regret over the horizon is minimized, or its discounted reward over the horizon is maximized.

MAB modeling brings a variety of advantages to the research in the area of wireless communications and networking; some of them are briefly described in the following:

- As mentioned before, every MAB model includes uncertainty, resulting from lack of prior knowledge as well as strictly limited feedback. This corresponds to a primal and natural feature of future distributed wireless networks, where providing nodes with information yields excessive feedback and overhead cost.
- In recent years, multi-agent MAB has attracted great attention, which allows the concept of conflict to be introduced to the seminal single-agent MAB. Multi-agent MAB models can be beneficially applied to solve wireless networking problems such as distributed resource management or interference coordination.
- MAB benefits from many variations in the model (Fig. 1) and therefore is able to accommodate many distinct conditions in wireless networks, for instance, different levels of information availability and different types of randomness.

In this article, we provide a brief tutorial overview of both single-agent and multi-agent MABs,

The authors are with the University of Manitoba.

1536-1284/16/$25.00 © 2016 IEEE

including different variations in the model. Furthermore, we study state-of-the-art applications of MAB models in wireless resource allocation, and discuss future research directions and open problems. Finally, we describe a new bandit-theoretical model for energy-efficient small cell planning in 5G networks.

Single-Agent Multi-Armed Bandits

The single-agent multi-armed bandit (SA-MAB) problem was first introduced in [1]. In the most seminal setting, the problem models an agent facing the challenge of sequentially selecting an arm from a set of arms in order to receive an a priori unknown reward, drawn from some unknown reward generating process. As a result of lack of prior information, at each trial, the player may choose some inferior arm in terms of reward, yielding some *regret*, which is quantified by the difference between the reward that would have been achieved had the player selected the best arm and the actual achieved reward. In such an unknown setting, the player decides which arm to pull in a sequence of trials so that its accumulated regret over the game horizon is minimized. This problem is an instance of an *exploration-exploitation* dilemma, that is, the trade-off between taking actions that yield immediate large rewards on one hand, and taking actions that might result in larger reward only in the future, for instance, activating an inferior arm only to acquire information, on the other hand. Therefore, SA-MAB is mainly concerned with a sequential online decision making problem in an unknown environment. A solution of a bandit problem is thus a decision making strategy called *policy* or *allocation rule* that determines which arm should be played at successive rounds. Policies are in general evaluated based on their regret or discounted reward performance.

As mentioned before, MAB benefits from a wide range of variations in the setting. In the event that arms have different states, the reward depends on arms' states, and the bandit model is referred to as *stateful* (Markovian). Otherwise, the model is *stateless*, which itself can be divided into a few subsets. As stateful and stateless bandits are inherently different, we discuss them separately in the following.

Stateful (Markovian) Bandits

In a stateful bandit model, every arm is associated with some finite state space. Upon being pulled, each arm pays some positive reward that is drawn from some stationary distribution associated with the current state of that arm. Arms' states change over time according to some stochastic model, which is considered to be a Markov process; that is, it satisfies the Markov property. Roughly speaking, a process satisfies the Markov property if its future depends solely on its present state rather than its full history. This type of MAB is also referred to as *Markovian bandit*. Note that in this formulation, after each round, the reward as well as the state of only the played arm is revealed, and the mechanism of state transition is unknown. A Markovian MAB model can thus be defined formally as $\mathfrak{B} := \{\mathcal{M}, \mathcal{S}_m, \pi_{m,s}, \mathbf{T}_m, \mu_m\}$, where $m \in \mathcal{M}$, and $\mathcal{M}$ is the set of arms.

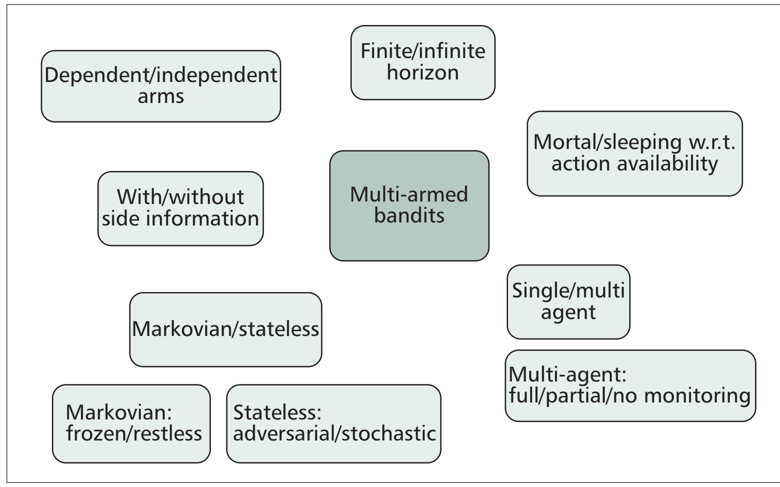

Figure 1. Various types of MABs.

$\mathcal{S}_m$ is the finite set of states of arm m.
$\pi_{m,s}$ is the stationary reward distribution of arm m in state $s \in \mathcal{S}_m$ with mean $\mu_{m,s}$.
$\mathbf{T}_m$ is the state transition matrix of arm m.
μ_m is the average reward of arm m.

For Markovian bandits, the mean reward of arm m, μ_m, is its expected reward under the rewards' stationary distribution in different states. That is, $\mu_m = \Sigma_{s \in \mathcal{S}_m} \mu_{m,s} \times \pi_{m,s}$. Define

$$\mu^* := \max_{m \in \mathcal{M}} \mu_m.$$

Also, let $s_{m,t}$ denote the state of some arm m at time t. The selected arm at each time t is denoted by I_t. Then for Markovian bandit, the cumulative regret up to time n, denoted by $R_{\mathrm{Mar},n}$, is defined as

$$R_{\mathrm{Mar},n} = n\mu - \mathbf{E}\left[\sum_{t=1}^{n} \mu_{I_t, s_{I_t,t}}\right], \tag{1}$$

where the expectation, $\mathbf{E}[\cdot]$, is taken with respect to the random draw of both states and the random actions of the policy. The general problem is to find the optimal policy, where optimality is defined in the sense of minimum regret.

Stateful bandit models are conventionally divided into the two following groups:
- Rested (frozen) bandits, where at each round, only the state of the played arm changes
- Restless bandits, where the state of every arm is subject to change after each round of play

Markovian bandit problems are often solved using *indexing policies*. Roughly speaking, at each round, a real scalar value, referred to as index, is associated with each arm. The index of any arm is counted as a measure of the reward which can be achieved by activating that arm in the current state. The arm with the highest index is played at each round. The optimal indexing policy is not unique for different bandit models; it depends on a variety of factors such as reward generating processes and arms being restless or frozen. An example of an indexing policy can be found in [2].

Note that if the arms' states are entirely observable and the state transition matrices are known, the criterion for action selection is conventionally to maximize the discounted reward

Stochastic MAB is a stateless bandit model, where arms are not associated with states, and the reward generating processes are stochastic. In other words, the series of rewards of each arm are state-independent, drawn from some specific density function, which can be stationary or non-stationary.

instead of minimizing the regret. Such problems belong to bandit models, and are solved by using indexing policies as well, with an example being the Gittins index [3]. However, it should be noted that although the Gittins index has been calculated in closed form for a variety of stochastic reward processes, before using any indexing policy, the indexability of the problem under investigation has to be verified, which is in general not trivial. In addition to the Gittins indices, Whittle's indexing policy [3] is well known and often used to solve restless bandits. This solution does not hold in general, though, and restless bandit problems are intractable in many cases. Important approximation algorithms for some families of restless bandit problems can be found in [4, 5].

Stateless Bandits

In the stateless bandit model, arms do not have any specific state. The arms' reward generating processes, however, are not necessarily stationary. A stateless MAB model is thus formally defined as $\mathfrak{B} := \{\mathcal{M}, \mu_{m,t}\}$, where $m \in \mathcal{M}$, and $\mathcal{M}$ is the set of arms, and $\mu_{m,t}$ is the average reward of arm m at time t.

Based on the nature of the reward generating process, stateless MABs can be divided into a few categories. Before proceeding to describe these categories we provide a notion of regret that is widely used to analyze almost every stateless bandit problem, regardless of the category to which it belongs. Let $g_{m,t}$ and I_t denote the instantaneous reward of arm m and the selected arm, respectively, both at time t. Then the regret of any policy for stateless bandits up to time n, referred to as *external regret*, which is denoted by $R_{\text{Ext},n}$, is defined as

$$R_{\text{Ext},n} = \max_{m\in\mathcal{M}} \mathbf{E}\left[\sum_{t=1}^{n} g_{m,t}\right] - \mathbf{E}\left[\sum_{t=1}^{n} g_{I_t,t}\right], \qquad (2)$$

Given the definition of external regret, the optimal solution of a stateless bandit problem is a policy that minimizes the external regret. Now we are in a position to study two important branches of stateless bandit problems: stochastic and adversarial bandits.

Stochastic Bandits: Stochastic MAB is a stateless bandit model where arms are not associated with states, and the reward generating processes are stochastic. In other words, the series of rewards of each arm are state-independent, drawn from some specific density function, which can be stationary or non-stationary. For the stationary case, $\mu_{m,t} = \mu_m$ for all $m \in \mathcal{M}$. In this case, the first term on the right side of Eq. 2 yields $\nu\mu^*$, where

$$\mu^* := \max_{m\in\mathcal{M}} \mu_m.$$

For the non-stationary case, the first term yields

$$\sum\nolimits_{t=1}^{n} \mu_t^*, \text{ with } \mu_t^* := \max_{m\in\mathcal{M}} \mu_{m,t}.$$

In order to solve the stochastic bandit problem, many methods have been developed so far that are based on the *upper confidence bound* (UCB) policy [6]. The basic idea to deal with the exploration-exploitation dilemma here is to estimate an upper bound of the mean reward of each arm at some fixed confidence level. The arm with the highest estimated bound is then played. Such methods, however, are only applicable when the rewards of each arm are independent and identically distributed, or in other words, the arm is stationary. In order to deal with non-stationarity, algorithms mostly use some statistical test in order to detect changes in the distribution. To elaborate, consider a sequence of independent random variables with some density that depends on some scalar parameter θ. The initial value of this parameter is θ_0, and after an unknown period of time, its value changes to θ_1. *Change-point detection* is performed by using some statistical test (e.g., *generalized likelihood ratio* or *Page-Hinkly* test), which not only identifies the changes, but also estimates the change time. Note that in general it is assumed that not too many of such change points exist. There are also some adapted versions of the seminal UCB algorithm that detect changes in the sample mean reward of arms in order to deal with time-variant statistical characteristics. For example, in *discounted UCB*, the rewards are weighted so that recent outcomes are emphasized when calculating the sample mean rewards, based on which the next arm is selected. In *sliding-window UCB*, the sample mean reward of each arm is calculated only over a fixed-length window of recent rewards, and not by using the entire reward history.

Adversarial (Non-Stochastic) Bandits: Adversarial bandit models are similar to stochastic ones in being stateless, with the difference that the series of rewards of each arm cannot be attributed to any specific distribution; in other words, rewards are non-stochastic. Generally, in an adversarial setting, mixed strategies are used to select an arm. That is, at each time t, the agent selects a probability distribution $\mathbf{P}_t = (p_{1,t}, \ldots, p_{m,t}, \ldots, p_{M,t})$ over arms, and plays arm m with probability $p_{m,t}$. In such a setting, *in addition* to external regret given by Eq. 2, we define another important notion of regret, *internal regret*, as follows:

$$R_{\text{Int},n} = \max_{m,l\in\mathcal{M}} R_{(m\to l),n} = \max_{m,l\in\mathcal{M}} \sum_{t=1}^{n} p_{m,t}(g_{l,t} - g_{m,t}). \qquad (3)$$

By definition, external regret compares the expected reward of the current mixed strategy with that of the best fixed action in hindsight, but fails to compare the rewards achieved by changing actions in a pairwise manner. Internal regret, in contrast, compares actions in pairs. Later, we shall see that the notion of internal regret is in close relation with correlated equilibria in games.

Adversarial bandits are often solved by using *potential-based* or *weighted average* algorithms [6]. In the most seminal form of this approach, at each trial, a mixed strategy is calculated over the set of actions. The selection probability of each action is proportional to its average regret performance in the past, possibly weighted by a specific potential function (e.g., the exponential function). Accordingly, the actions with better past performance (lower average regret) are more likely to become activated in the future, and vice versa.

Other Important Bandit Models

Contextual (Covariate) Bandits: In the basic bandit model, at each round, the agent only observes the reward of the played actions, and hence has to select its future actions only based on its past performance. In *contextual bandits* (also called *bandits with side information* or *covariate bandits*), in contrast, at each round of decision making, some side information is revealed to the learning agent. The agent therefore learns the best mapping of contexts to arms. Note that this type of bandit model is distinguished only based on information availability, and the reward process can still be Markovian, stochastic, or adversarial. In any case, the aforementioned algorithms are adapted to solve contextual bandit problems as well.

Mortal Bandits: While in the basic bandit model it is assumed that all arms are infinitely available, there are also some variants that distinguish the problems where this assumption does not hold. For instance, *mortal bandits* assume that arms are available only for a finite time. The availability time can be either deterministic or stochastic, known or unknown. An example of solution approaches can be found in [7]. The algorithm is called *stochastic with early stopping*, and the main idea behind it is to abandon an arm once it is realized that its maximum possible future reward may not justify its retention. The analysis of such schemes is inherently different from the general bandit model, since in such problems the achievable accumulated reward from each arm is bounded even over an infinite horizon.

Sleeping Bandits: *Sleeping bandits* refer to bandit problems where the action set is time-varying. Clearly, in such models, not only the reward process, but also the arm availability at each round can be Markovian, adversarial, or stochastic. Some solution approaches for different types of this problem can be found in [8]. The core idea of these algorithms is to change the notion of regret from the basic one defined earlier. While the basic regret notion is based on the performance loss with respect to the best action in hindsight, here the best *ordering* of actions serves as the benchmark, as the best action might not be available in some rounds.

Multi-Agent (Game-Theoretical) Multi-Armed Bandits

In MA-MAB, each player $k \in \mathcal{K}$ is assigned an action set $\mathcal{M}_k \subseteq \mathcal{M}$. Similar to the single-agent model, each agent selects an action in successive trials to receive an initially unknown reward; if multiple agents select an arm, the achieved reward is shared in an arbitrary manner. Therefore, in a multi-agent setting, the reward of each agent depends on the joint action profile of all agents. Note that the action set, the played action, and the reward achieved by each agent can be regarded as either private or public information, based on the specific system model and problem formulation. From the viewpoint of each agent k, an MA-MAB model can be seen as a game with two players: the first player is agent k itself, and the second player is the set of all other $K - 1$ agents, whose joint action profile affects the reward achieved by agent k. This game might belong to any conventional category of games; for instance, it can be a potential or zero-sum game.

Given no prior information, every agent needs to interact with the random environment in order to solve the problems that arise in an unknown reactive model, for instance, long-term accumulated reward maximization, or average regret minimization. As a result of competition, which is inherent in non-cooperative multi-agent systems, single-agent bandit models that ignore the possible conflicts among multiple agents do not yield satisfactory outcomes; in particular, equilibrium is not guaranteed to be achieved. This is when game theory and multi-armed bandits meet and complement each other. Thus, equilibrium arises as an asymptotic outcome of repeated interactions in an unknown environment among learning agents with bounded rationality that are provided with strictly limited information and aim at achieving long-term optimality in some sense. The aforementioned problem, that is, efficient and convergent learning in a multi-agent setting, is currently under intensive investigation, mainly in computer science and mathematics. In the theory of wireless communications, there is also some ongoing research on game-theoretical bandits. In the following, we describe some important results in this area, with an emphasis on convergence to equilibrium.

Convergence to Correlated Equilibrium

Recall the definition of internal regret given by Eq. 3. The following (simplified) theorem describes the relation between internal regret and the concept of correlated equilibrium in games.[1]

Theorem 1 [9]: *Consider a multi-agent bandit game with a set of players $\mathcal{K}$, and let $R^{(k)}_{\mathrm{Int},n}$ denote the internal regret of player* k $\in \mathcal{K}$ *at round* n. *If all agents play according to some policy that exhibits per-round vanishing internal regret, that is,* ($\lim_{n\to\infty} 1/n\, R^{(k)}_{\mathrm{Int},n} = 0$, *the game converges to the set of correlated equilibrium in a time-average sense.*

This concept is used, for instance, in [10] to develop joint power control and channel selection strategies in distributed D2D networks.

Besides internal regret, there is another concept, *calibrated forecasting*, which is related to correlated equilibria in games. Before explaining this relation, we briefly describe calibrated forecasting in the following. Consider a random process with a set $\mathcal{D}$ of D outcomes. A forecaster predicts the outcome of the random process sequentially, and is referred to as *calibrated* if, in the limit, the predicted outcome is equal to the true one. Now, consider a game with K players, where each player k selects its action from action set $\mathcal{M}_k$ (pure strategies). Then, for any player k, the joint action profile of its opponent is a random process with its set of outcomes being $\otimes_{i=1,i\neq k}^{K} \mathcal{M}_k$, where $\otimes$ denotes the Cartesian product. In such games, each player might apply a forecaster in order to predict the next joint action profile of its opponents, and uses that prediction to move strategically, for instance, by playing the best response to the predicted

While in the basic bandit model it is assumed that all arms are infinitely available, there are also some variants that distinguish the problems where this assumption does not hold. For instance, mortal bandits assume that arms are available only for a finite time.

[1] The definitions of correlated and Nash equilibria are standard and therefore omitted here for brevity of the article.

The core idea is to combine calibrated forecasting with non-parametric regression, so that after some time, each player has some accurate estimate of the reward process of actions as a function of joint action profile and the next move of opponents.

joint action profile. Note, however, such prediction naturally involves learning, and therefore requires knowledge of the past actions of opponents; in other words, players should be able to observe the actions of each other. In the following theorem, we describe the relation between calibrated forecasting and correlated equilibria.

Theorem 2 [9]: *In any generic game, if every player plays by best responding to a calibrated forecast of its opponents' joint action profile, then the game converges to the set of correlated equilibrium in a time-average sense.*

Note that Theorem 2 declares the convergence for general full information games, where the game matrix is known to all players a priori. The theorem is generalized in [11] to bandit games, and is applied as a basis to develop a convergent solution approach for non-stationary stochastic bandit models. The core idea is to combine calibrated forecasting with non-parametric regression so that after some time, each player has some accurate estimate of (i) the reward process of actions as a function of a joint action profile and (ii) the next move of opponents.

Convergence to Nash Equilibria

While there are only a few solution approaches for multi-agent bandit games that guarantee convergence to correlated equilibria, converging to Nash equilibria seems to be an even more challenging task. There are some convergent approaches for special game classes such as potential games. For instance, in [12], algorithms are proposed for Nash convergence in potential games and games with more general forms of acyclicity. The basic idea therein is to combine Q-learning with stochastic better-reply or stochastic adaptive-play dynamics in order to develop a multi-agent version of Q-learning, which estimates the reward functions using novel forms of the learning policy. Details can be found in [12]. While multi-agent Q-learning converges only for some specific game classes, there are also some algorithms that converge in general games. A concept based on which multiple convergent bandit algorithms are proposed for general games is *regret testing* [9]. Here the basic idea is to perform some sort of exhaustive search, in the sense that each player first selects a mixed strategy according to some predefined protocol, and then checks whether (i) the incurred regret is less than a specific threshold and (ii) the selected mixed strategy is an approximate best response to the others' mixed strategies. Clearly, none of these can be concluded in one round of play; but if the same mixed strategy is played during sufficiently many rounds, each player can simply test the corresponding hypothesis. Different variants of this concept exhibit different convergence characteristics, for instance, convergence to Nash equilibrium in a time-average sense or convergence to the set of ε-Nash equilibria, both for every generic game.

In the area of wireless communications, most works consider a specific game (defined by a specific utility function) and propose a game-theoretical learning algorithm the convergence of which is established by some analysis based on the defined utility function. Such solutions, however, suffer from limited applicability, as they are tailored to converge only for the predefined games.

State of the Art and Future Research Directions

In what follows, we first briefly review the current applications of MABs (in both single-agent and multi-agent settings). Afterward, we discuss some open problems and potential research directions. Note that our focus in this article is on the applications of MAB models in solving resource allocation problems.

State of the Art

Distributed Channel Selection: Distributed channel selection is a widely considered application of MABs. When cast as an MAB, channels are considered as arms, and reward processes are some functions of signal-to-noise ratio (SNR) or signal-to-interference-plus-noise ratio (SINR). Reward functions might evolve as Markovian, stochastic, or adversarial, and any multiple access protocol, orthogonal and non-orthogonal, can be addressed by using bandit models. While the single-user problem has been investigated intensively so far, the multi-user scenario is still under investigation. As described earlier, in a multi-user setting, it is important to address the conflict among users by converging to equilibrium in some sense. This can be done by using multi-agent bandit algorithms, described before. Also, a key point is that the selected channel of users might sometimes be observable. Such information can be used as side information to build contextual bandit models; see [11] for an example.

Opportunistic Spectrum Access: Opportunistic spectrum access in cognitive radio networks can be formulated as a (rested or restless) Bernoulli bandit problem. In doing so, each channel is an arm that is either in state 0, interpreted as busy, with probability p, or in state 1, interpreted as idle, with probability $1 - p$. Since a secondary user is only allowed to transmit through idle channels, a reward is solely associated with state 1. Channel states, as well as parameter p, are unknown, and a secondary user, modeled as an agent, observes the state of the selected channel only. Through transmission time, a secondary user aims at maximizing its reward by selecting some channel that is available with high probability. The problem can then be solved by using indexing policies. An example can be found in [2].

Sensor Scheduling: In sensor networks, the number of sites to be surveyed is very often larger than that of available sensors. For instance, for military surveillance and in smart grids, it is important to find the places where there is attack or failure. Such problems can be addressed by using MABs, where sites are modeled as arms with two possible states, similar to the opportunistic spectrum access problem. Furthermore, if sensors are used for transmission, and the set of residual energy of each sensor is defined on a discrete space, sensor scheduling can be formulated as an MAB problem, aimed at minimizing the energy consumption while maximizing the transmission performance. Reference [13] is an example.

Transmission Mode/Relay Selection: A relay selection problem arises in two-hop (or multihop) transmission, where multiple relays are available to be used in order to improve the transmission performance. To formulate the relay selection problem as an MAB, the reward of each relay, modeled as an arm, is defined as the achievable transmission rate through that relay, which is the minimum achievable rate of first and second hops. Similarly, a problem of transmission mode selection is a natural challenge in distributed hierarchical networks, for example, in a D2D communications system underlaying a cellular infrastructure. In such a hierarchy, any pair (or group) of users can choose to use the traditional cellular transmission mode via a BS, or to establish direct communication link. This problem can be cast and solved by a two-armed bandit model, as proposed in [14].

Power Control: Power control is one of the less explored applications of MA-MABs. In order to model such a problem as a bandit game, a finite discrete set of power levels is considered as the set of arms, and the reward process is defined to be some function of SINR, where, as in channel selection, interference represents the mutual impact of agents. In comparison to channel selection, in a power control problem, the assumption of full or partial monitoring is not realistic. That is, agents might not be able to observe each other's actions (i.e., the transmission power levels). Hence, algorithms should cope with so-called *no monitoring*, where agents' actions are regarded as private. An example can be found in [10].

Energy Harvesting: Energy harvesting is a relatively new concept attracting increasing interest. In a network where nodes are equipped with energy harvesting units, electricity is generated from the surrounding environment (e.g., solar energy, ambient RF signals). Implementing this concept gives rise to some challenges, particularly with respect to scheduling, as the energy arrival is in general not deterministic. Under the assumption that the energy harvesting process is Markovian, MAB models can be used to address the scheduling problem. In the model, similar to the cases before, the reward function is the achievable transmission rate while the unknown variant is the energy harvesting process and/or battery state. See [15] as an example. Application of MAB theory to this area is relatively new, and a variety of open problems exist.

Other Applications: Although in this article we restrict our attention to wireless resource management, it should be noted that MAB models find applications in a variety of other networking problems. An important application is routing, where on the network graph, every end-to-end path consists of multiple edges, and the cost of each edge, sometimes time-varying, is unknown a priori. Moreover, after transmitting through each path, the only feedback is assumed to be the end-to-end cost of the selected path. Such a problem can be modeled and solved using various MAB settings in accordance with the network characteristics. See [16] for an example. Another field of research benefiting from bandit theory is *security*. One example is to combat jamming attacks, where an adversary transmits signals to interfere with normal communications and temporarily disables the network. Under the assumption that the transmitter and receiver do not pre-share any secrets with each other, the channel selection problem can be modeled and solved by using MAB. An example can be found in [17].

Future Research Directions

As described in the previous section, a variety of resource management problems have been investigated by using bandit models; nonetheless, in most research studies, the development of a model and solution is focused on using some infinite-horizon variants of bandit problems, with an emphasis on the type of reward evolution, that is, Markovian, stochastic or adversarial. In other words, models are often distinguished only based on arms' reward generating processes, disregarding many other important and practical aspects such as budget constraints, information availability, number of arms, arms' dependencies, and number of arms to be selected at each round, to name just a few. To clarify this shortcoming, some examples are provided in the following.

Models are often distinguished only based on arms' reward generating processes, disregarding many other important and practical aspects such as budget constraints, information availability, number of arms, arms' dependencies, and number of arms to be selected at each round, to name just a few.

Multimedia Transmission: Most current bandit-theoretical models for wireless networking problems are analyzed asymptotically, assuming that the transmission may continue for infinite time. Therefore, optimality is defined in terms of long-run discounted reward maximization and/or average regret minimization. There are, however, many scenarios, such as multimedia transmission, where transmission has to be completed under strict delay and/or energy constraints, which restricts the number of possible transmission rounds. In such scenarios, algorithms that achieve optimality in some long-run sense are not applicable; indeed, here the goal is merely to perform some task successfully under some budget constraint rather than optimizing the asymptotic performance. Thus, it is imperative to develop new optimality conditions, as well as new solution approaches, for applied bandit models in a finite horizon.

Wireless Sensor Networks: To the best of our knowledge, current research studies, which apply bandit models to solve networking problems, assume that arms, representing resources such as channels or relays, are available infinitely often during networking tasks such as channel selection or routing. This assumption, however, is not always valid. For instance, in a sensor network, a sensor might become unavailable as soon as the energy supply is exhausted. In cognitive radio networks, as another example, a channel becomes busy (unavailable) when a primary user starts to transmit through that channel. In a bandit model, this effect corresponds to restricted arm availability. In other words, each arm might be available only for some limited number of trials. Therefore, new definitions of regret and new algorithms need to be developed.

Correlated Resource Selection: Another practical issue, currently neglected, is arms' dependencies. While in a great majority of models it is assumed that arms are independent, it is not always the case in practice. Indeed, resources, such as relays and channels, might be correlated in many cases. For instance, in a cognitive radio with a fixed number of primary users, primary channels can be considered as dependent since a channel being mostly occupied implies that

Note that the problems mentioned above exist also in full-information case, but become aggravated in the bandit setting, where only the reward of the activated arm is revealed to the player.

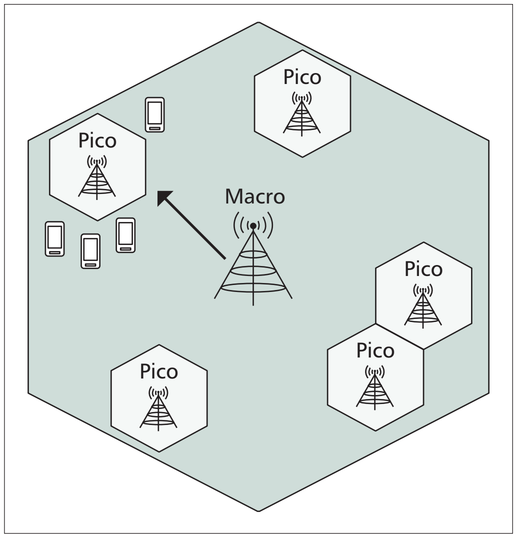

Figure 2. Efficient activation of small cells by macro BS.

other channels might be less busy. Exploiting such dependencies as a source of information about arms yields efficient algorithmic solutions to the bandit and hence the resource management problem.

Multi-Agent Setting: Besides model-related inefficiencies to be alleviated, there are some open problems particularly in multi-agent settings. Although in such settings achieving equilibrium is beneficial for the entire network, there is a variety of anti-equilibrium reasoning that should be addressed. For instance, a general problem is to cope with slow convergence to equilibrium, in particular for large numbers of arms and/or players. Another important problem is equilibrium selection. The problem is concerned with guiding the learning agents to the most efficient equilibrium when multiple equilibria exist. In cases where such convergence is not feasible, an analysis of equilibrium efficiency is highly desired. Moreover, in some scenarios, it is desired to have a fair solution rather than an efficient one. Note that the problems mentioned above exist also in the full information case, but become aggravated in the bandit setting, where only the reward of the activated arm is revealed to the player.

Efficient 5G Small Cells with Combinatorial (Multi-Play) Bandits

Earlier, we briefly studied the state of the art, thereby clarifying the application of MABs to solve a variety of problems that arise in wireless networks. Moreover, we discussed some issues and open problems. In this section, we introduce a new application of MABs in next generation wireless networks to design energy-efficient 5G small cells. Primarily, we consider a single-agent model with independent arms. In contrast to the state of the art, however, in our bandit model, multiple arms are selected at each round. We afterward provide a sketch to generalize the model to a multi-agent scenario, and also to the case where arms are dependent, that is, they affect each other in some specific sense.

Facing an ever increasing influx of data traffic, 5G networks are foreseen to alleviate this problem by deploying dense small cells to underlay the legacy macrocellular networks. This takes advantage of low-power and short-range base stations that operate (preferably) using the same radio spectrum as the macro base stations and offload macrocell traffic [18]. In order to maintain low cost, small cells are desired to be self-organized and energy-efficient. As a result, a small cell is activated only if it improves the overall network performance. Efficient small cell activation can be performed through dynamic small cell on/off by a macrocell. Making smart decisions, however, requires some information, such as the available energy or the number of potential users at each small cell, which are both random in nature. However, such information is notoriously difficult to acquire, specifically when numerous potential small cells exist. Therefore, it is reasonable to search for a solution so that a macrocell activates small cells in an efficient manner given only limited information (Fig. 2).

Formally, consider a macro BS operating in conjunction with a set $\mathcal{M}$ of M potential small cells. Every small cell $m \in \mathcal{M}$ acquires its required energy (at least partially) through energy harvesting. That is, while sleeping, it gains energy from the environment through energy harvesting units, and uses the stored energy to provide services in its active periods. The amount of energy gained during each sleeping period depends on environmental factors (e.g., the weather) and can change adversarially. The amount of energy spent during each servicing period depends on the number of users, which can change as well due to user mobility.[2] Activating a small cell costs energy even if it does not serve any user; thus, at each time (or for every time period), a macro BS aims to activate a set $\mathcal{N} \subseteq \mathcal{M}$ of N small cells. Every activated small cell should be able to afford the required energy to provide services to a large enough number of users so that the initial fixed activation cost diminishes. Moreover, it is desired to select the subset of small cells in the sense that the overall network performance is optimized. The available energy and the number of users in small cells are, however, unknown to the macro BS, which complicates the decision making. We formulate this problem as an SA-MAB by modeling small cells as arms.

We consider a simple model where any small cell $m \in \mathcal{M}$ requires α_m energy units to serve a user with data rate β_m. The cost per energy unit is denoted by r_m. For initial activation, κ_m units of energy are necessary even if no user is served. Let $A_{m,t}$ and $B_{m,t}$ be the available energy and the number of users in a small cell at time t. Define

$$\gamma_{m,t} = \max\left\{ \left\lfloor \frac{A_{m,t}}{\alpha_m} \right\rfloor, B_{m,t} \right\}.$$

Then the utility (reward) of selecting a small cell m at time t is defined as $g_{m,t} = \gamma_{m,t}\beta_m - r_m(\gamma_{m,t}\alpha_m + \kappa_m)$. The total utility of selecting some subset $\mathcal{N}$ therefore yields $g_{\mathcal{N},t} = \Sigma_{m \in \mathcal{N}} g_{m,t}$. Despite similarities, this problem is not identical to the adversarial bandit problem described earlier, due to the following reason. Unlike the standard adversarial bandit problem, here multiple arms, not a single arm, are selected at each round of decision making. This type of bandit problem

[2] Clearly, both harvested energy and number of users can be modeled as stochastic (stationary or non-stationary) variables as well.

is referred to as *combinatorial* or *multiple play* bandits, which stands in contrast to the standard *single play* problems. For combinatorial bandits, the conventional definition of regret given by Eq. 2 does not hold. In fact, in order to calculate the regret where a subset of arms is selected at each round, the reward achieved by the selected subset should be compared to the best *subset* of arms in terms of average reward. The problem can therefore be solved by using algorithms described above, for instance, weighted-average strategy, by considering each set of N actions as a super action, and some other simple adaptations. Nevertheless, such an approach is computationally inefficient, and therefore, new algorithms are developed that are specifically tailored for combinatorial bandit problems. An examples, *EXP3.M* can be found in [19]. However, it should be mentioned that most of algorithms are based on similar concepts and hence exhibit similar performance.

In order to briefly evaluate the proposed model, we consider three networks with $M = 8$, 6, 4 small cells, from which a macro BS selects $N = 4$, 3, 2 to activate, correspondingly. Note that with $M = 8$ and $N = 4$, for example, there are 60 different 4-tuples of small cells that can be selected by the macro BS, which we call *action* and label as 1, …, 60. The number of users and residual energy of each small cell is selected randomly *at each trial*, without assuming any specific density function. Moreover, the macro BS does not have any prior information. To perform selection, the macro BS uses the bandit model described before. The average utility vs. the best possible subset (in an average sense), which is selected through exhaustive search, is shown in Fig. 3. It can be seen that given enough time, the utility of the proposed model converges to that of the best selection.

For $M = 6$ and $N = 3$, Fig. 4 shows the percentage of time in which each one of the 3-tuples is selected by the macro BS in $T = 5 \times 10^5$ trials. From the figure, the best action (here action 19) is played almost all the time. Figures for the other two settings are similar, that is, the best action is played very often. A large fraction of the time spent playing other actions is the exploration time. In general, the number of trials dedicated to exploration depends on the algorithm; nonetheless, in most algorithms, exploration time is mostly determined by an exploration parameter, say η. Roughly speaking, a trial is an exploration trial with probability η and an exploitation trial with $1 - \eta$ probability. Therefore, a larger value of η yields a larger exploration time on average.

The performance evaluation might be improved, for example, by investigating the amount of saved energy (cost) in comparison with the number of users not being served by small cells, both of which depend on the number of activated small cells, N. Thus, N can be regarded as a variable to be optimized according to a system's statistical characteristics instead of being fixed and predefined. The question of whether such optimization can be included in a bandit model or not is an open problem to be investigated in future works.

It is worth mentioning that the complexity of the proposed small cell planning model

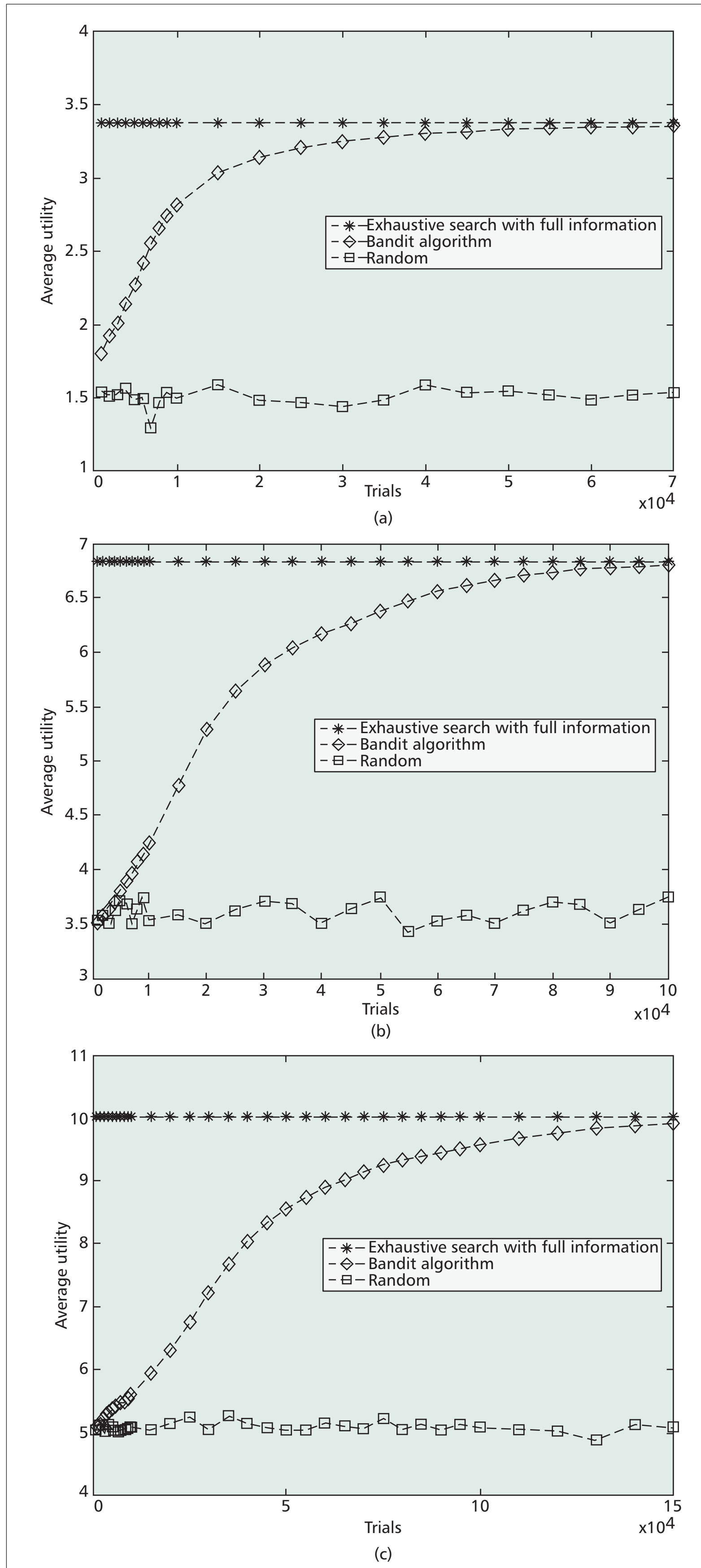

Figure 3. Average utility achieved by bandit model vs. full-information exhaustive search for networks of different scales: a) $M = 4$, $N = 2$; b) $M = 6$, $N = 3$; c) $M = 8$, $N = 4$.

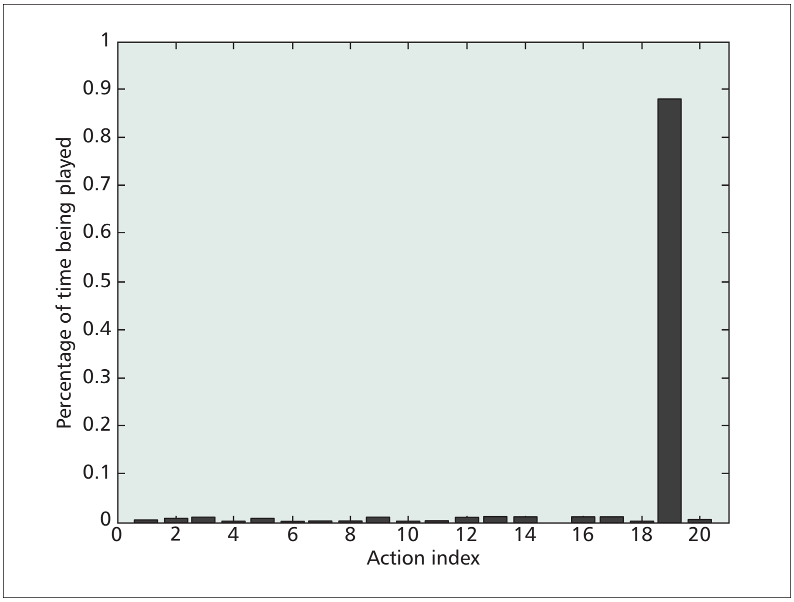

Figure 4. Percentage of time each action is played for $M = 6$, $N = 3$.

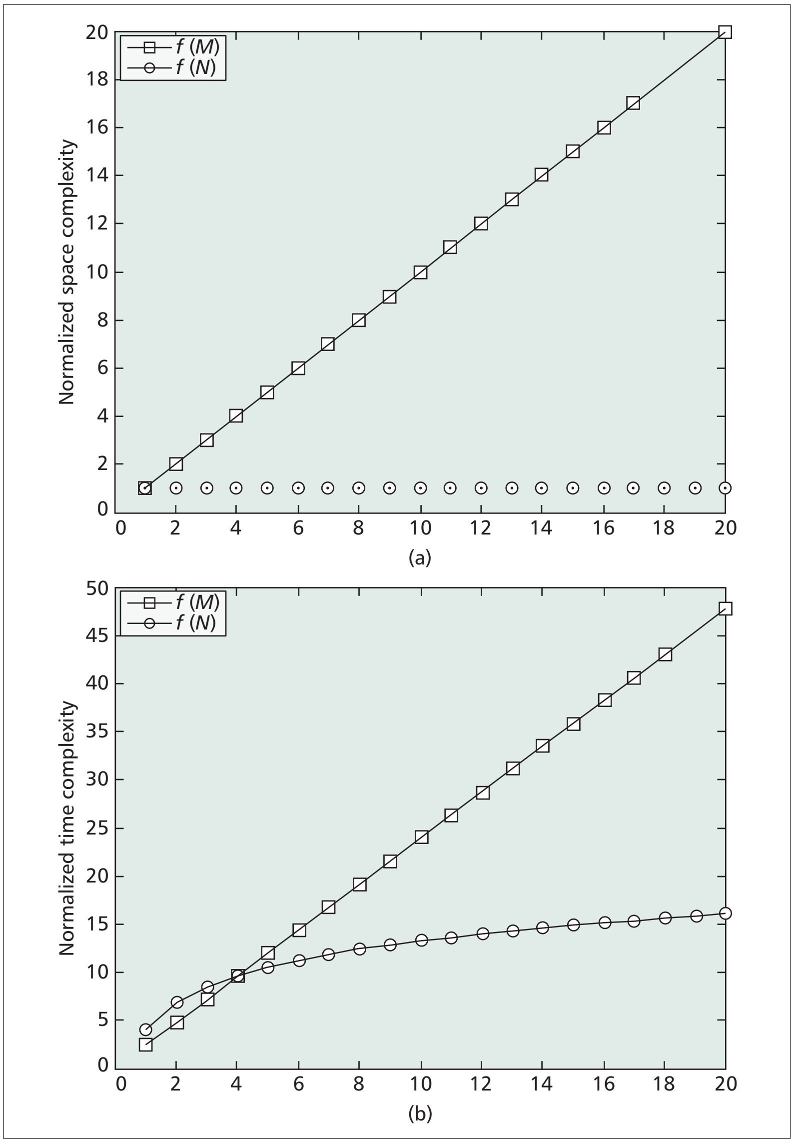

Figure 5. Percentage of time each action is played for $M = 6$, $N = 3$: a) space complexity; b) time complexity.

depends on the algorithm being used to solve the formulated bandit problem. For *EXP3.M* [19], the complexity is of $O(M(\log N + 1))$ and $O(M)$ in time and space, respectively. The growth of complexity is shown in Fig. 5. Due to linear/logarithmic complexity, the proposed approach can be considered to be scalable with the number of small cells; nonetheless, there are scenarios in which distributed solutions to the problem of efficient small cell activation are desired. In such scenarios, each small cell decides whether to start the active mode or to remain passive. Similar to the centralized case, the decision is based on available energy and expected number of users. In case such information is not available at small cells, the problem can be modeled as a two-armed bandit problem, with one safe arm (passive mode) and one risky arm (active mode). While the passive mode results in no reward, the reward of active mode can be modeled as an adversarial or stochastic process, which depends on the energy level as well as number of users arriving in each small cell. The problem might have different solutions depending on the type of interactions among small cells, which can be competitive or cooperative, for instance. In the event small cells are fully decoupled, the problem reduces to M independent two-arm bandit problems. A complete investigation of distributed small cell on/off is beyond the scope of this article and left for our future work.

In addition to considering a distributed model, another direction to improve the current model is to take the dependencies of small cells into account. In the formulated problem, such dependencies can mainly be concluded based on geographical reasons. In fact, nearby small cells are expected to experience similar energy harvesting opportunities (e.g., during sunny weather in a neighborhood), as well as similar user arrival statistics. Therefore, by selecting each small cell and observing the reward, some information can also be derived about its neighboring cells. This information is beneficial to shorten the exploration period, thereby increasing the accumulated reward or achieving optimal average reward in a shorter time. Detailed study of such dependencies is left for our future work.

As the last remark, it should be mentioned that the application of MAB models to 5G small cell goes beyond the efficient small cell activation problem studied here. An instance is the user association problem. While the existence of small cells improves coverage and capacity performance, user association becomes more challenging due to the following reasons. In sharp contrast to conventional cellular networks, small cells are not always active. In addition, the size of each small cell is subject to change as a function of user density, for example. Moreover, a user may associate to a macro BS or a nearby small cell, or even to both a macro BS and a small cell (or even to multiple small cells). The user association problem in a small cell network can also be modeled as a bandit game, in either single-agent (centralized) or multi-agent (distributed) settings. While a centralized approach improves the possibility of interference coordination, a distributed approach reduces the complexity and overhead. Another problem that can be approached by using MAB models is the inter-cell interference coordination problem in small cell networks, which arises due to dense cell deployment of small cells.

Summary and Conclusion

MAB is a class of sequential decision making problems under strictly limited prior information as well as feedback. By providing an overview of MABs, we have argued that a wide range of wireless networking problems, including resource management, security, routing, scheduling, and energy harvesting, can be formulated and solved as a bandit problem. We have also reviewed the state of the art and applications of MABs, with an emphasis on wireless resource allocation. There are a number of open research directions to be explored, which have been briefly outlined. Finally, we have provided a detailed example of an application of MAB in energy-efficient 5G small cells, where the problem of optimal small cell planning is cast as a multi-play (combinatorial) bandit problem. Preliminary performance evaluation results have established the effectiveness of the proposed model and approach.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant.

References

[1] H. Robbins, "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, 1952, pp. 527–35.

[2] H. Liu, K. Liu, and Q. Zhao, "Learning in a Changing World: Restless Multiarmed Bandit with Unknown Dynamics," *IEEE Trans. Info. Theory*, vol. 59, no. 3, Mar. 2013., pp. 1902–16

[3] J. Gittins, K. Glazebrook, and R. Weber, *Multi-Armed Bandit Allocation Indices*, Wiley, 2nd ed., 2011.

[4] S. Guha, K. Munagala, and P. Shi, "Approximation Algorithms for Restless Bandit Problems," *J. ACM*, vol. 58, no. 1, 2010, pp. 3.

[5] D. Bertsimas and J. Nino-Mora, "Restless Bandits, Linear Programming Relaxations, and a Primal-Dual Index Heuristic," *Operations Research*, vol. 48, no. 1, 2000, p. 2000.

[6] S. Bubeck and N. Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, 2012, pp. 1–122.

[7] D. Chakrabarti *et al*., "Mortal Multiarmed Bandits," *Proc. Neural Info. Processing Sys. Conf.*, 2008, pp. 273–80.

[8] R.D. Kleinberg, A. Niculescu-mizil, and Y. Sharma, "Regret Bounds for Sleeping Experts and Bandits," *Proc. Conf. Learning Theory*, 2008, pp. 425–36.

[9] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge Univ. Press, 2006.

[10] S. Maghsudi and S. Stanczak, "Joint Channel Selection and Power Control in Infrastructureless Wireless Networks: A Multi-Player Multiarmed Bandit Framework," *IEEE Trans. Vehic. Tech*., vol. PP, 2014.

[11] S. Maghsudi and S. Stanczak, "Channel Selection for Network-Assisted D2D Communication via No-Regret Bandit Learning with Calibrated Forecasting," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, Mar. 2015, pp. 1309–22.

[12] A.C. Chapman *et al*., "Convergent Learning Algorithms for Unknown Reward Games," *SIAM J. Control and Optimization*, vol. 51, no. 4, 2013, pp. 3154–80.

[13] Y. Chen *et al*., "Transmission Scheduling for Sensor Network Lifetime Maximization: A Shortest Path Bandit Formulation," *Proc. IEEE Int'l. Conf. Acoustics, Speech and Signal Processing*, May 2006, vol. 4, pp. IV–IV.

[14] S. Maghsudi and S. Stanczak, "Transmission Mode Selection for Network-Assisted Device to Device Communication: A Levy-Bandit Approach," *Proc. IEEE Int'l. Conf. Acoustics, Speech and Signal Processing*, May 2014, pp. 7009–13.

[15] P. Blasco and D. Gunduz, "Multi-Access Communications with Energy Harvesting: A Multi-Armed Bandit Model and the Optimality of the Myopic Policy," *IEEE JSAC*, vol. 33, no. 3, March 2015, pp. 585–97.

[16] R. Kleinberg and B. Awerbuch, "Online Linear Optimization and Adaptive Routing," *Elsevier J. Computer and System Sciences*, vol. 74, 2008, pp. 97–104.

[17] Q. Wang *et al.*, "Towards Optimal Adaptive UFH-Based Anti-Jamming Wireless Communication," *IEEE JSAC*, vol. 30, no. 1, Jan. 2012, pp. 16–30.

[18] E. Hossain *et al*., "Evolution toward 5G Multi-Tier Cellular Wireless Networks: An Interference Management Perspective," *IEEE Wireless Commun*., vol. 21, no. 3, June 2014, pp. 118–27.

[19] T. Uchiya, A. Nakamura, and M. Kudo, "Algorithms for Adversarial Bandit Problems with Multiple Plays," *Algorithmic Learning Theory*, Oct 2010, pp. 375–89.

Biographies

Setareh Maghsudi (setarah.maghsudi@umanitoba.ca) received her B.Sc. degree from Iran University of Science and Technology in 2007, her M.Sc. degree from the University of Kiel, Germany, in 2010, and her Ph.D. degree (with distinction) from the Technical University of Berlin, Germany, in 2015, all in electrical engineering. Currently, she is a postdoctoral fellow at the Department of Electrical and Computer Engineering, University of Manitoba, Canada. Her research interests include game theory, multi-armed bandit, and graph theory, with applications to wireless networks.

Ekram Hossain (ekram.hossain@manitoba.ca) [F'15] is a professor in the Department of Electrical and Computer Engineering at the University of Manitoba. He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2001. His current research interests include design, analysis, and optimization of wireless/mobile communications networks, cognitive radio systems, and network economics. He has authored/edited several books in these areas (http://home.cc.umanitoba.ca/~hossaina). He was elevated to an IEEE Fellow "for contributions to spectrum management and resource allocation in cognitive and cellular radio networks." Currently he serves as Editor-in-Chief for *IEEE Communications Surveys & Tutorials* and an Editor for *IEEE Wireless Communications*. Also, he is a member of the IEEE Press Editorial Board. Previously, he served as an Area Editor for *IEEE Transactions on Wireless Communications* in the area of Resource Management and Multiple Access from 2009 to 2011, an Editor for *IEEE Transactions on Mobile Computing* from 2007 to 2012, and an Editor for the *IEEE Journal on Selected Areas in Communications* Cognitive Radio Series from 2011 to 2014. He has won several research awards including the IEEE Communications Society Transmission, Access, and Optical Systems (TAOS) Technical Committee's Best Paper Award at IEEE GLOBECOM 2015, the University of Manitoba Merit Award in 2010 and 2014 (for Research and Scholarly Activities), the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award, and the IEEE Wireless Communications and Networking Conference 2012 Best Paper Award. He was a Distinguished Lecturer of the IEEE Communications Society (2012–2015). He is a registered Professional Engineer in the province of Manitoba, Canada.

We have provided a detailed example of an application of MAB in energy-efficient 5G small cells, where the problem of optimal small cell planning is cast as a multi-play (combinatorial) bandit problem. Preliminary performance evaluation results have established the effectiveness of the proposed model and approach.