Small Cell Transmit Power Assignment Based on Correlated Bandit Learning

Zhiyang Wang, and Cong Shen, *Senior Member, IEEE*

***Abstract*—Judiciously setting the base station transmit power that matches its deployment environment is a key problem in ultra dense networks and heterogeneous in-building cellular deployments. A unique characteristic of this problem is the tradeoff between sufficient indoor coverage and limited outdoor leakage, which has to be met without explicit knowledge of the environment. In this paper, we address the small base station (SBS) transmit power assignment problem based on stochastic bandit theory. Unlike existing solutions that rely on heavy involvement of RF engineers surveying the target area, we take advantage of the human user behavior with simple coverage feedback in the network, and thus significantly reduce the planned human measurement. In addition, the proposed power assignment algorithms follow the Bayesian principle to utilize the available prior knowledge from system self configuration. To guarantee good performance when the prior knowledge is insufficient, we incorporate the performance *correlation* among similar power values, and establish an algorithm that exploits the correlation structure to recover majority of the degraded performance. Furthermore, we explicitly consider *power switching penalties* in order to discourage frequent changes of the transmit power, which cause varying coverage and uneven user experience. Comprehensive system-level simulations are performed for both single and multiple SBS deployment scenarios, and the resulting power settings are compared to the state-of-the-art solutions. Significant performance gains of the proposed algorithms are observed. Particularly, the correlation structure enables the algorithm to converge much faster to the optimal long-term power than other methods.**

***Index Terms*—Coverage optimization; Transmit power assignment; Heterogeneous Network (HetNet).**

I. Introduction

The massive deployment of distributed low-power low-cost small base stations (SBS) has been viewed as one of the most important solutions to address the challenge of exponential growth of the wireless data traffic, particularly for indoor users [1]. In practice, SBSs may be deployed in drastically different scenarios, from large warehouses and buildings to small residential apartments and single-office enterprises. In addition, the radio frequency (RF) conditions may vary significantly from one site to another. Due to the heterogeneous nature of these deployments, the transmit power assigned to the SBS, which effectively determines the coverage range, cannot be the same but must be decided based on the individual deployment environment, such as the building layout, the RF conditions, and the locations of the base stations. Furthermore, indoor enterprise deployments often have stringent access and security constraints. As a result, judiciously setting the SBS transmit power to automatically match its deployment environment is among the most important challenges for in-building SBS network deployment [2].

To address this challenge, in-building enterprise networks typically rely on RF engineers to carry out extensive measurement and RF survey to determine the transmit power for appropriate coverage and limited leakage. Then, during live network operations, the RF engineers often need to make extra visits to optimize the transmit power for better performance. Clearly, this is a heavy human-in-the-loop model, as the success of the power setting relies on the experience of the seasoned engineers, the result of the RF survey of the engineers' choice, and the planning software. Not only is this approach expensive, inflexible and error-prone, but it also does not scale with the densification of indoor SBS networks [3].

Adaptive, automated and autonomous network optimization is the key principle of the self-organizing networks (SON) paradigm [4], which aims at achieving the optimal network configuration while minimizing the planned human involvement in the deployment, configuration, optimization and maintenance. Self-optimizing the SBS transmit power falls into the framework of SON, and several solutions have already been proposed. Small Cell Forum has defined a common network monitor mode [5], allowing each SBS to periodically measure its surrounding RF environment and adjust its transmit power. This solution relies on an assumed coverage range based on categorization, and the RF measurements are only taken at the SBS location but not over the entire coverage area, which is coarse and may cause RF mismatch [6]. To solve these issues, Supervised Mobile Assisted Range Tuning (SMART) was proposed in [7], which relies on the RF feedback of a technician walking along the sampling routes. The required RF feedback is extensive, including majority of the LTE lower layer quantities such as RSRP RSSI, CQI, etc. These quantities along the measurement routes provide important RF information of the deployment, and a global optimization can be formulated to derive the transmit power that satisfies both coverage and leakage constraints. Unfortunately, this problem is non-convex and the optimal transmit power is difficult to compute [7]. In [8], the authors developed a self-organizing policy for distributed femtocell networks, aiming at minimizing the cell transmit power while satisfying the service requirement. In [9], a heuristic solution was proposed to reliably determine the coverage for the current power level before either increasing or decreasing the power based on user feedback. Solutions from both [8] and [9] have some adaptability but still lack good accuracy when used in differ-

Z. Wang and C. Shen are with the Department of Electronic Engineering and Information Science, School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China. E-mail: wzy43@mail.ustc.edu.cn, congshen@ustc.edu.cn.

ent environments. The authors of [10] modeled SBS power management as a Markov Decision Process problem, focusing on the power control in a time-varying network. Similarly, a downlink transmit power control solution for interference mitigation via reinforcement learning was proposed in [11]. The main objective of [10] and [11], however, is to adjust the transmit power in reaction to the changing circumstance for better quality of service, which makes it more of a power control problem that has to be solved at a fast time scale.

We focus on setting the SBS transmit power of an enterprise network in an unknown deployment environment. We limit our attention to SBS networks with *closed access* mode, which is commonly adopted in the enterprise deployment due to security and management considerations. An adequate power assignment is particularly crucial for the closed access mode, as the transmit power needs to be large enough to provide sufficient coverage for the inside users while small enough to not create significant interference to the outside non-enterprise co-channel users, who cannot be served by the enterprise network. This work proposes to capture this delicate balance between coverage and leakage by a system performance indication function (PIF). If the deployment is known, the optimal power assignment can be obtained by maximizing the PIF.

However, a practical solution needs to be effective in an arbitrarily unknown environment, and prefers minimum human involvement and feedback. Naturally, a good solution must compliment the aforementioned *optimization* problem with an *online learning* approach to remove the uncertainty of the environment, which is a key challenge for efficient transmit power assignment. The SBSs have to balance the immediate gains (selecting a power level that performs best so far) and long-term performance (evaluating other power levels). We thus resort to the theory of multi-armed bandit (MAB) [12] to address the resulting exploration and exploitation tradeoff. However, as opposed to directly applying classical MAB algorithms such as UCB [13], our problem has two unique characteristics that were not exploited. First, SBS transmit power assignment falls into the *self-optimization* category of SON. Generally, a *self-configuration* phase has already taken place before invoking the transmit power assignment algorithm. As a result, there would be some *prior knowledge* of the system that can be utilized. Second, performances of similar power levels are often very similar, which means that 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 a power level not only reveals information about itself, but also nearby power levels that are highly correlated. Such information was not available in classical UCB solutions [12], [13][1], and has not been utilized in SON [7], [8], [10], [11].

In this paper, we leverage these engineering characteristics of the problem, and develop bandit-inspired transmit power assignment algorithms. In the bandit literature, similar models have been studied in [15], [16] and the corresponding bandit algorithms have been proposed. The authors of [15] proposed bandit algorithms with a Bayesian prior on the mean reward that is based on a human decision-making model. [16] further extended the algorithm to focus on the correlation among arms. In our work, we first adopt a Bayesian [17] learning algorithm that incorporates the prior knowledge of the system from the self-configuration phase. The developed *Bayesian Power Assignment* (BPA) algorithm iteratively updates the posterior distribution based on new observations and the prior distribution, and uses the updated posterior distribution to compute the utility function and determine the transmit power level. In addition to utilizing the prior knowledge, we further leverage the correlation structure of the PIF of similar transmit power levels, and a *Correlated Bayesian Power Assignment* (CBPA) algorithm that combines the Bayesian principle with the correlation property is employed. To the authors' best knowledge, this is the first work that incorporates *bandit with correlated arms* into the design of wireless networks. Furthermore, practical deployment often wants to avoid frequent power changes, because it may cause frequent variation of the coverage area and result in uneven user experience. To address this issue, we present a block allocation extension to the proposed BPA and CBPA algorithms which explicitly considers switching cost to discourage frequent changes of power levels. Rigorous analysis of the performance loss with respect to the genie-aided global optimization solution is carried out. A tight upper bound of the performance loss for the most general algorithm (CBPA with switching cost) is derived, and performance characterization of other algorithms can be obtained as special cases. In order to reduce the algorithms' complexity which increases exponentially with the number of SBSs, we further introduce *clustering* based on the prior knowledge, so that the complexity can be drastically reduced without sacrificing much of the accuracy and effectiveness of the algorithms. The performances of all the proposed algorithms are verified by extensive system-level simulations and compared with both the globally optimal power assignment with complete information and the existing state-of-the-art solutions. Not only do the proposed algorithms outperform existing solutions and converge to the globally optimal power assignment quickly, but they also reduce the planned human involvement significantly and only require minimum amount of user feedback (one bit per location), as opposed to the full-blown RF measurement and feedback that is universally required in the existing solutions.

The rest of the paper is organized as follows. The system model and problem formulation can be found in Section II. Section III and IV present the proposed power assignment algorithms without and with switching cost, respectively. Performance analysis for all the algorithms is given in Section V. Complexity issues of the multi-SBS deployment are addressed in Section VI. Simulation results are portrayed in Section VII. Finally, Section VIII concludes the paper.

[1]The authors of [14] studied the continuum-armed bandit with an infinite continuum of strategies, which also captures the dependency among arms. We opt out this approach because in the multi-SBS cases, the continuity of the reward functions may not be guaranteed. The discrete arm setting makes the solutions more effective and flexible for practical adoption.

II. System Model and Problem Formulation

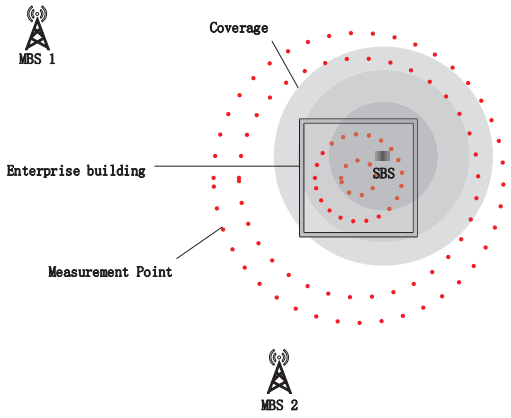
A. Network Model

Both single-SBS and multi-SBS deployments are considered. Note that the former is suitable for modeling single-office enterprises, residential apartments and other small deployments, while the latter mainly applies to large enterprises, for which multiple SBSs are installed to jointly cover the indoor users. The set of SBSs is indexed as $\mathcal{K}_{SBS} = \{1, 2, .., K\}$. Each SBS has a set of candidate pilot[2] power levels, denoted as $\mathcal{P} = \{p_1, p_2, .., p_n\}$. As our focus is on the SBSs with closed access and co-channel with the macro base stations (MBS), we simply assume that the users at the measurement points inside the enterprise building are served by the SBS network, while users at points outside can only be served by one of the MBSs from $\mathcal{K}_{MBS} = \{1, 2, .., K_M\}$, as Fig. 1 illustrates.

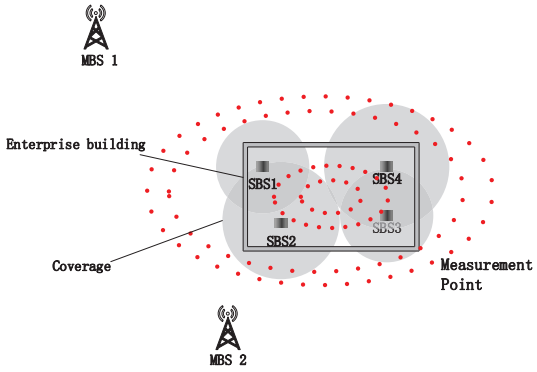
The measurement data come from the customer UE feedback from some inside and outside routes during normal network operations. This is different from the RF survey approach that is carried out during network planning. The detailed mechanism and procedure of obtaining such customer UE feedback are mostly the same as in [9]. However, as opposed to a complete RF feedback required in [9], we only require *one-bit* coverage indication for each inside report. The extended set of RF measurements, such as RSRP, RSSI, and CQI, are not needed in our power assignment algorithm. For non-enterprise UEs, as we only need to know whether the UE is covered at a reporting location, we will rely on the *registration attempt* at the outside location to determine such events. Note that this is a common approach to determine leakage and has been adopted in [18], [7], [19].

In this work, our model and procedure on power assignment follow the common industry SON operations [3]. Specifically, the power assignment policy is executed during the *self-optimization* phase of SON, at the central network controller which is configured to oversee the operation of the entire SBS network. This is a common choice for enterprise cellular networks, as they often have security and privacy constraints which are easier to be satisfied in a centralized architecture. Furthermore, the power assignment algorithm operates in a periodic fashion, which is typical for self-optimization of SON [18]. For each time slot, the SBS first sets the pilot power based on the assignment algorithm. Then the network operates and collects UE feedback from both inside and outside of the intended coverage area. At the end of the current period, a performance measure is computed to evaluate the current pilot power and then used in the assignment algorithm to compute the power level for the next slot. This sequence of operations is illustrated in Fig. 2. Lastly, industry SON operations typically have the *self-optimization* operations follow a *self-configuration* phase, during which a coarse measurement and power calibration are performed [7]. As we will see later, the initial self-configuration, albeit coarse and sometimes inaccurate, offers useful prior knowledge that can be leveraged in the power assignment algorithm.

[2]As the purpose of the long-term power assignment is to determine the appropriate coverage that fits the deployment, we focus on setting the pilot power instead of the power of data and control channels [2].



(a) Single-SBS



(b) Multi-SBS

Fig. 1. An exemplary enterprise SBS network deployment.

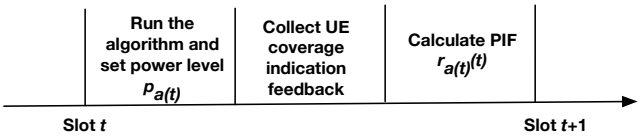


Fig. 2. The power assignment procedure in a time slot t.

B. Problem Formulation

To formulate the power assignment problem, we first need to define the criteria for coverage and leakage. To that end, let us denote the set of measurement points on the inside and outside routes as $N_{in} = \{1, 2, .., n_{in}\}$ and $N_{out} = \{1, 2, .., n_{out}\}$, respectively. The coverage and leakage criteria for a measurement point can be formally defined as:

$$\text{coverage:} \quad \max_{k_S \in \mathcal{K}_{SBS}} \mathsf{SINR}_{k_S,n} > \mathsf{SINR}_{\text{th}}, \text{for } n \in N_{in}, \quad (1)$$

$$\text{leakage:} \quad \max_{k_M \in \mathcal{K}_{MBS}} \mathsf{SINR}_{k_M,n} < \mathsf{SINR}_{\text{th}}, \text{for } n \in N_{out}, (2)$$

where $\mathsf{SINR}_{k_S,n}$ and $\mathsf{SINR}_{k_M,n}$ represent the SINR of the measurement point n inside corresponding to SBS k_S and the SINR of point n outside served by MBS k_M, respectively.

They can be calculated as:

$$\mathsf{SINR}_{k_S,n} = \frac{P^r_{k_S,n}}{\sum_{i=1,i\neq k_S}^{K} P^r_{i,n} + \sum_{j=1}^{K_M} P^{Mr}_{j,n} + N_s},$$

$$\mathsf{SINR}_{k_M,n} = \frac{P^{Mr}_{k_M,n}}{\sum_{i=1}^{K} P^r_{i,n} + \sum_{j=1,j\neq k_M}^{K_M} P^{Mr}_{j,n} + N_s},$$

where $P^r_{i,n}$ and $P^{Mr}_{j,n}$ represent the received power at point n from SBS i and MBS j, respectively, N_s denotes the uncontrolled noise and interference, and $\mathsf{SINR}_{\text{th}}$ is the SINR threshold.

With the definition at each measurement point, the overall system coverage and leakage are defined as the percentage of measurement points which satisfy the coverage condition (1) and leakage condition (2), respectively. If we denote the number of measurement points that satisfy the corresponding conditions as n_{cov} and n_{lea}, then the coverage percentage and leakage percentage can be computed as $\eta_{in} = n_{cov}/n_{in} \times 100\%$ and $\eta_{out} = n_{lea}/n_{out} \times 100\%$. Note that a larger pilot transmit power may simultaneously increase the indoor coverage percentage and the outdoor leakage percentage. Hence, the system performance indication function (PIF) associated with each candidate power level must balance coverage and leakage. In this work, we adopt a simple linear PIF as

$$r = \alpha\eta_{in} - (1-\alpha)\eta_{out}, \tag{3}$$

where $\alpha \in [0,1]$ is a control parameter and can be tuned to weigh differently between coverage and leakage. Note that PIF (3) is chosen as an example to illustrate the proposed power assignment algorithms. Other meaningful PIFs that capture the tradeoff between coverage and leakage can be used in place of (3). The objective of a power assignment algorithm is to find the optimal solution $p^* \in \mathcal{P}$ that maximizes the PIF (3).

Strictly speaking, the function r in (3) is a random variable for a given pilot power level. This is due to the random channel effect such as shadowing, fast fading and other disturbance in the deployment environment. We focus on a probabilistic model with *Gaussian* random fluctuation around the mean. As we will see in Sec. VII, Gaussian distribution indeed is a very good approximation for the actual PIF. Furthermore, we evaluate the proposed algorithms in settings with non-Gaussian PIF distributions, and the empirical results suggest that algorithms developed based on the Gaussian assumption are very effective.

III. Power Assignment Algorithms based on Bayesian Bandit Learning

A. Stochastic Bandit Model

The necessity of balancing the short-term performance and long-term learning has motivated us to take a stochastic multi-armed bandit approach to the power assignment problem. Specifically, we model the set of candidate pilot power values $\mathcal{P} = \{p_1, p_2, .., p_n\}$ as n arms, denoted by $\mathcal{N}_{pow} = \{1, 2, .., n\}$. At the beginning of each time slot $t = 1, 2, .., T$, a power value $p_{a(t)} \in \mathcal{P}, a(t) \in \mathcal{N}_{pow}$ is selected. At the end of the time slot t, the SBS observes a performance feedback $r_{a(t)}(t)$ based on UE measurement reports, which corresponds to $reward$ in the bandit theory. As discussed in Sec. II, we model the random PIF associated with *each* power value as a Gaussian random variable. The objective is to develop an efficient power assignment solution to maximize the cumulative PIF for any given time horizon T. For the multi-SBS case, each arm corresponds to a set of power levels of all K SBSs, and other definitions remain the same.

In multi-armed bandit theory, a quantity termed as *expected cumulative regret* [12] is often used to characterize the algorithm performance, which represents the cumulative difference between the reward of the arms chosen and the maximum expected reward, which is attainable by a "genie" who knows the expected reward of all arms. We comment that minimizing the expected cumulative regret is equivalent to maximizing the expected accumulated reward, which is the objective of the power assignment problem. This is because the maximum expected reward is independent of the adopted learning algorithm and the regret is equivalent to the performance loss of any power assignment problem due to learning.

Formally, we denote

$$G_T = \sum_{t=1}^{T} r_{a(t)}(t) \tag{4}$$

as the cumulative PIF up to a given time horizon $T > 0$, and we define the cumulative PIF loss *due to learning* as

$$R_T = G^*_T - G_T = \max_{i=1,..,n} \left(\sum_{t=1}^{T} r_i(t) \right) - \sum_{t=1}^{T} r_{a(t)}(t), \tag{5}$$

which corresponds to the definition of cumulative regret. Here the optimal power level can be obtained by a genie-aided solution, e.g. a global optimization of the expected PIF with complete RF information from the technician survey. We are interested in finding efficient algorithms that maximize the cumulative PIF (4). Equivalently, the goal is to minimize the PIF loss of the system (5) for any given time horizon T. The expected PIF loss can be written as:

$$\begin{aligned} \mathbb{E}[R_T] &= T\mu^* - \mathbb{E}\sum_{t=1}^{T} \mu_{a(t)} \\ &= \left(\sum_{i=1}^{n} \mathbb{E}[N_i(T)] \right) \mu^* - \mathbb{E}\sum_{i=1}^{n} N_i(T)\mu_i \\ &= \sum_{i=1}^{n} \Delta_i \mathbb{E}[N_i(T)], \end{aligned} \tag{6}$$

where $\mu^* = \max_{i=1,..,n} \mu_i$ is the true mean PIF of the optimal power level and $\Delta_i = \mu^* - \mu_i$ measures the mean PIF gap between the chosen power level and the optimum. $N_i(T)$ represents the number of times power level p_i is selected. According to the ground-breaking work of Lai and Robbins [20], if the expected loss $\mathbb{E}[R_T]$ of our proposed algorithms can be upper bounded[3] by $\mathcal{O}(\log T)$, an asymptotically optimal performance is achieved in the sense that the convergence rate is of the same order as the optimum.

[3] $\log(\cdot)$ represents natural logarithm if the base is not specified.

B. Bayesian Power Assignment Algorithm

The first algorithm utilizes the *prior* knowledge of the PIF estimation *before* the algorithm is invoked. In practice, the most common form for the prior knowledge comes from the self-configuration phase of SON, which is performed during network initialization. This phase can provide us with some prior estimation of the PIFs as it typically tries different power levels before settling on one. However, all practical SON solutions have certain requirements on the elapsed time of the self-configuration operations. This is because self-configuration affects the boot-up time, and thus must be carefully controlled. As a result, massive measurement during self-configuration is typically out of the question and we often encounter coarse initial setup. Another possibility is that as the proposed power assignment algorithm is recursive over time, it also progressively collects PIF estimations for each selected power level. This can be used iteratively to update the prior knowledge. The quality of the prior depends on the detailed process in self-configuration phase, e.g. the time duration, mechanisms for large power settings, which is uncontrollable and out of scope of this paper. However, it is worth noting that the proposed algorithms also work with inaccurate prior or even without any prior knowledge, at the expense of slower convergence.

We first consider the power assignment algorithm without considering the correlation between power levels. We adopt the well-known Bayesian principle [17] that integrates the prior distribution and quantiles of the posterior distribution. The proposed *Bayesian Power Assignment (BPA)* algorithm, which adopts the deterministic *upper credible limit (UCL)* principle in [15], is given in Algorithm 1. In this algorithm, $\{\mu_i^0, \sigma_0^2\}$ denotes the prior knowledge of the Gaussian distribution for PIF. The utility function defined in step 2 is composed of an estimated performance term and a measure of uncertainty, which reflects the tradeoff between exploration and exploitation. More specifically, $\Phi^{-1} : (0,1) \to \mathbb{R}$ is the inverse cumulative distribution function (CDF) for a standard Gaussian random variable. We use the quantile function to indicate: $\mathbb{P}(\mu_i \leq Q_i^{\text{BPA}}(t)) = 1 - 1/(\sqrt{2\pi}et^2)$. Asymptotically, the true mean PIF μ_i is more likely to be less than the estimation Q_i^{BPA}, which leads to the convergence to the optimal power level.

If the prior knowledge is not available, the BPA algorithm can be slightly modified to address this issue. Specifically, the estimated PIF term and uncertainty measurement have to be updated simultaneously in each time slot. This philosophy leads to the following utility function:

$$Q_i^{\text{UiPA}}(t) = \bar{r}_i(t) + \sqrt{\frac{\sum_{\tau=1}^{t} r_i^2(\tau) - \bar{r}_i^2(t) N_i(t)}{(N_i(t)-1)N_i(t)}} \Phi^{-1}(1 - 1/(\sqrt{2\pi}et^2)). \tag{7}$$

The *Uninformative Power Assignment (UiPA)* algorithm thus can be obtained by replacing the utility function in step 2 of Algorithm 1 with (7), while removing the prior input at the beginning and estimation state update in step 6.

Algorithm 1 The Bayesian Power Assignment (BPA) Algorithm

Input: Prior estimation of PIF mean $\{\mu_i^0\}_{i=1}^n$, variance σ_0^2

Initialize: $N_i(t) = 0, \bar{r}_i(t) = 0, Q_i^{BPA}(t) = 0, \hat{\mu}_i(1) = \mu_i^0, \hat{\sigma}_i(1) = \sigma_0$ for all $i \in \mathcal{N}_{pow}$, $t \in 1, .., T$.

1: **for** $t \in 1, 2, .., T$ **do**

2: For each arm $i \in \mathcal{N}_{pow}$ update the utility function: $Q_i^{\text{BPA}}(t) = \hat{\mu}_i(t) + \hat{\sigma}_i(t)\Phi^{-1}(1 - 1/(\sqrt{2\pi}et^2))$,

3: Select a power value $p_{a(t)}$ according to: $a(t) = \arg\max\{Q_i^{\text{BPA}}(t) | i \in \mathcal{N}_{pow}\}$,

4: Observe the PIF $r_{a(t)}(t)$,

5: Update the average PIF and the selected times of $p_{a(t)}$: $\bar{r}_{a(t)}(t+1) = \frac{N_{a(t)}(t)\bar{r}_{a(t)}(t) + r_{a(t)}(t)}{N_{a(t)}(t)+1}$, $N_{a(t)}(t+1) = N_{a(t)}(t) + 1$,

6: Update the estimated mean and variance of PIF of power level $p_{a(t)}$: $\hat{\mu}_{a(t)}(t+1) = \frac{\mu_{a(t)}^0 + N_{a(t)}(t+1)\bar{r}_{a(t)}(t+1)}{N_{a(t)}(t+1)+1}$, $\hat{\sigma}_{a(t)}(t+1) = \frac{\sigma_0}{\sqrt{N_{a(t)}(t+1)+1}}$.

7: **end for**

C. Correlated Bayesian Power Assignment Algorithm

In the BPA algorithm, $\{\mu_i^0, \sigma_0^2\}$ is used as our prior knowledge of performance for each power level. If the PIFs of different arms are independent, then utilizing individual Gaussian distributions is sufficient in our framework. However, for the considered power assignment problem, the PIFs of similar transmit power levels are generally correlated due to the slow and continuous changing nature of RF propagation. In other words, a stronger PIF correlation exists between adjacent power levels than distant pairs, and leveraging the full covariance matrix of the joint distribution may provide significant performance boost compared to the BPA algorithm. Intuitively, if a transmit power level results in a bad PIF with respect to the balance of coverage and leakage, then an intelligent algorithm may not need to waste much exploration on its immediate neighboring power levels, as they are highly likely to be bad as well.

We formally present the *Correlated Bayesian Power Assignment (CBPA)* algorithm in Algorithm 2. Let $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ be a correlated prior assumption while Σ_0 is a positive definite matrix, we define $\{\boldsymbol{\phi}_t \in \mathbb{R}^n\}_{t \in \{1,..,T\}}$ as the indicator vector to reveal the currently selected power value $p_{a(t)}$, i.e.,

$$(\boldsymbol{\phi}_t)_k = \begin{cases} 1 & k = a(t), \\ 0 & \text{otherwise}, \end{cases}$$

where $(\boldsymbol{\phi}_t)_k$ represents the k-th entry of $\boldsymbol{\phi}_t$. The estimation of the mean PIFs and correlation structure of the PIF $(\boldsymbol{\mu}_t, \Sigma_t)$

Algorithm 2 The Correlated Bayesian Power Assignment (CBPA) Algorithm

Input: Prior estimation of joint Gaussian distribution of the PIFs: $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$;

Initialize: $N_i(t) = 0, \bar{r}_i(t) = 0, Q_i^{CBPA}(t) = 0, \hat{\mu}_i(1) = \mu_i^0, \hat{\Sigma}_1 = \Sigma_0$ for all $i \in \mathcal{N}_{pow}$ and $t \in 1, .., T$.

1: **for** $t \in 1, 2, .., T$ **do**

2: For each $i \in \mathcal{N}_{pow}$ update the utility function $Q_i^{\text{CBPA}}(t) = \hat{\mu}_i(t) + \hat{\sigma}_i(t)\sqrt{\sum_{j=1}^{n} \rho_{ij}^2(t)}\Phi^{-1}(1 - 1/(\sqrt{2\pi e}t^2))$, where $\rho_{ij}(t)$ is the correlation coefficient between power value i and j at time t, which is obtained from $\hat{\Sigma}_t$; $\hat{\mu}_i(t)$,$\hat{\sigma}_i^2(t)$ is the i-th entry of $\hat{\boldsymbol{\mu}}_t$ and diagonal entry of $\hat{\Sigma}_t$.

3: Select a power value $p_{a(t)}$ according to: $a(t) = \arg\max\{Q_i^{\text{CBPA}}(t) | i \in \mathcal{N}_{pow}\}$,

4: Collect the performance function $r_{a(t)}(t)$,

5: Update the average performance and the selected time of $p_{a(t)}$: $\bar{r}_{a(t)}(t+1) = \frac{N_{a(t)}(t)\bar{r}_{a(t)}(t) + r_{a(t)}(t)}{N_{a(t)}(t)+1}$, $N_{a(t)}(t+1) = N_{a(t)}(t) + 1$,

6: Update the estimation state: $\hat{\boldsymbol{\mu}}_{t+1} = (\Sigma_0^{-1} + P(t+1)^{-1})^{-1}(P(t+1)^{-1}\bar{\mathbf{r}}_{t+1} + \Sigma_0^{-1}\boldsymbol{\mu}_0)$, $\hat{\Sigma}_{t+1}^{-1} = \Sigma_0^{-1} + P(t+1)^{-1}$.

7: **end for**

is updated following the Bayesian principle [16]:

$$\mathbf{q}_t = \frac{r_t\boldsymbol{\phi}_t}{\sigma_0^2} + \hat{\Lambda}_{t-1}\hat{\boldsymbol{\mu}}_{t-1}, \quad \hat{\Lambda}_t = \frac{\boldsymbol{\phi}_t\boldsymbol{\phi}_t^T}{\sigma_0^2} + \hat{\Lambda}_{t-1},$$
$$\hat{\Sigma}_t = \hat{\Lambda}_t^{-1}, \qquad \hat{\boldsymbol{\mu}}_t = \hat{\Sigma}_t\mathbf{q}_t = \hat{\Lambda}_t^{-1}\mathbf{q}_t,$$

where r_t is the PIF observed at time slot t. To derive a general expression of the estimation, we introduce a diagonal matrix $P(t)$ with entries $\sigma_0^2/N_i(t), i \in \mathcal{N}_{pow}$, and $\bar{\mathbf{r}}_t$ is the vector of $\bar{r}_i(t), i \in \mathcal{N}_{pow}$. We first rewrite the expression of $\hat{\Lambda}_t$ as:

$$\begin{aligned}\hat{\Lambda}_t &= \frac{\boldsymbol{\phi}_t\boldsymbol{\phi}_t^T}{\sigma_0^2} + \frac{\boldsymbol{\phi}_{t-1}\boldsymbol{\phi}_{t-1}^T}{\sigma_0^2} + \hat{\Lambda}_{t-2} \\ &= \frac{\boldsymbol{\phi}_t\boldsymbol{\phi}_t^T}{\sigma_0^2} + \frac{\boldsymbol{\phi}_{t-1}\boldsymbol{\phi}_{t-1}^T}{\sigma_0^2} + \ldots + \frac{\boldsymbol{\phi}_1\boldsymbol{\phi}_1^T}{\sigma_0^2} + \Lambda_0 \\ &= \frac{1}{\sigma_0^2}\begin{pmatrix} N_1(t) & & & \\ & N_2(t) & & \\ & & \ddots & \\ & & & N_n(t) \end{pmatrix} + \Lambda_0 \\ &= P(t)^{-1} + \Lambda_0.\end{aligned} \tag{8}$$

Then, $\hat{\boldsymbol{\mu}}_t$ can be derived based on (8):

$$\begin{aligned}\hat{\boldsymbol{\mu}}_t &= \hat{\Lambda}_t^{-1}\mathbf{q}_t = \hat{\Lambda}_t^{-1}\left(\frac{r_t\boldsymbol{\phi}_t}{\sigma_0^2} + \hat{\Lambda}_{t-1}\hat{\Sigma}_{t-1}\boldsymbol{q}_{t-1}\right) \\ &= \hat{\Lambda}_t^{-1}\left(\frac{r_t\boldsymbol{\phi}_t}{\sigma_0^2} + \frac{r_t\boldsymbol{\phi}_t}{\sigma_0^2} + \ldots + \frac{r_t\boldsymbol{\phi}_t}{\sigma_0^2} + \Lambda_0\boldsymbol{\mu}_0\right) \\ &= \hat{\Lambda}_t^{-1}\left(\begin{pmatrix} \frac{N_1(t)}{\sigma_0^2}\bar{r}_1(t) & & \\ & \ddots & \\ & & \frac{N_n(t)}{\sigma_0^2}\bar{r}_n(t) \end{pmatrix} + \Lambda_0\boldsymbol{\mu}_0\right) \\ &= (\Lambda_0 + P(t)^{-1})^{-1}(P(t)^{-1}\bar{\mathbf{r}}_t + \Lambda_0\boldsymbol{\mu}_0).\end{aligned} \tag{9}$$

Finally, combining equation (8) and (9), the estimation at time slot t can be written as:

$$\hat{\Lambda}_t = P(t)^{-1} + \Lambda_0,$$
$$\hat{\boldsymbol{\mu}}_t = (\Lambda_0 + P(t)^{-1})^{-1}(P(t)^{-1}\bar{\mathbf{r}}_t + \Lambda_0\boldsymbol{\mu}_0),$$

which is used in Algorithm 2.

IV. Power Assignment with Switching Cost

A. Problem Formulation with Switching Cost

In practice, it is very critical for any practical cellular deployment to avoid frequent power changes. In a cellular network, coverage variation due to the change of transmit power often results in poor user experience (call drop, low data rate, frequent handover, etc.), which in turn degrades the network performance significantly. To address this problem, we explicitly add a switching cost when the power level changes. In this way, a good power assignment policy will determine the optimal power value while minimizing frequent switches. We adopt a general switching loss function $s_{ij} = f(|p_i - p_j|)$, which is a bounded non-decreasing function of the difference between the two power values with $f(0) = 0$. s_{ij} is incurred whenever SBS changes its pilot power value between p_j and p_i. The cumulative switching cost up to T can be written as:

$$\mathsf{SC}(T) = \sum_{t=2}^{T} s_{a(t)a(t-1)} = \sum_{t=2}^{T} f(|p_{a(t)} - p_{a(t-1)}|).$$

Thus the cumulative PIF in this problem can be expressed as:

$$G_T^S = G_T - \mathsf{SC}(T).$$

In a multi-SBS deployment, the switching cost is defined as the sum of individual switching costs of all SBSs.

B. The Power Assignment Algorithm with Switching Cost

We extend the preceding algorithms to a *block allocation* scheme to address switching costs. Block allocation schemes, such as the one in [21], determine specific intervals of time over which the selection is consistent. A power value is selected at the beginning of each interval. The construction of the intervals should ensure the expected number of switches scales at most logarithmically in time to guarantee good performance. This idea is graphically presented in Fig. 3. We first divide time into frames whose last time slot is denoted as $L_f, f \in \{1, 2...l\}, l = \lceil\sqrt{\log_2 T}\rceil$. Each frame is then subdivided into $b_f = \lceil(2^{f^2} - 2^{(f-1)^2})/f\rceil$ blocks

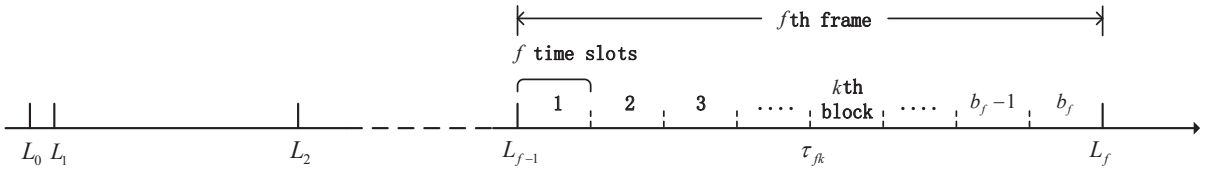


Fig. 3. The block allocation scheme used in BPA-SC and CBPA-SC.

each of which contains f time slots. Each block is identified by $(f,k), f \in \{1,2,..,l\}, k \in \{1,2,..,b_f\}$, with f and k representing the frame number and block number within the frame respectively. The beginning time slot of block k in the f-th frame is denoted as τ_{fk}. Note that the key element in selecting the blocking length is to only incur $o(\log T)$ switching cost. In this way, the $\mathcal{O}(\log T)$ regret of the standard algorithm still dominates the total regret.

Algorithm 3 The Power Assignment with Switching Cost Algorithm

Input: Prior estimation of PIF mean: $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$;

Initialize: $N_i(t) = 0, \bar{r}_i(t) = 0, Q_i(t) = 0, \hat{\mu}_i(1) = \mu_i^0, \hat{\Sigma}_1 = \Sigma_0$ for all $i \in \mathcal{N}_{pow}$, $t \in 1,..,T$.

1: **for** $f \in \{1,2,..,l\}$ **do**
2: **for** $k \in \{1,2,..,b_f\}$ **do**
3: The beginning time slot of k-th block in the f-th frame $\tau_{fk} = L_{f-1} + 1 + f(k-1)$,
4: For each $i \in \mathcal{N}_{pow}$ update the utility function Q_i,
5: Select a power value p_{a^*} according to: $a^* = \arg\max\{Q_i | i \in \mathcal{N}_{pow}\}$,
6: Keep SBS on power value p_{a^*} for the next $(n_f - 1)$ slots,
7: Collect the performance function $r_{a^*}(t)$, possibly excluding a switching loss $s_{a(t)a(t-1)}$, $t \in \{\tau_{fk}, \tau_{fk}+1,..,T_e\}$, $T_e = \tau_{fk}+f-1, n_f = f$ if $\tau_{fk}+f-1 \leqslant T$, otherwise $T_e = T, n_f = T - \tau_{fk} + 1$;
8: Update the average performance and the selected time of p_{a^*}: $\bar{r}_{a^*} = \frac{N_{a^*}\bar{r}_{a^*} + \sum_{t=\tau_{fk}}^{T_e}(r_{a^*}(t) - s_{a(t)a(t-1)})}{N_{a^*}+n_f}$, $N_{a^*} = N_{a^*} + n_f$,
9: Update the estimation state.
10: **end for**
11: **end for**

The *Power Assignment with Switching Cost* algorithm is formally presented in Algorithm 3. The *Uninformative* (UiPA-SC), *Bayesian* (BPA-SC) and *Correlated Bayesian Power Assignment with Switching Cost* (CBPA-SC) algorithms can be similarly obtained, by replacing Q_i with Q_i^{UiPA}, Q_i^{BPA} and Q_i^{CBPA} respectively. Note that in BPA-SC, the prior estimation state Σ_0 becomes a diagonal matrix with entries σ_0^2, while there is no prior input in UiPA-SC. At the beginning of each block, a power value is selected and the SBS locks on this power value in each of the next f time slots in the block. The estimation update in step 9 also follows step 6 in Algorithm 1 and step 6 in Algorithm 2.

There are two key ideas of Algorithm 3. The first is that since the switching cost results in a penalty in performance, the algorithm needs to "explore in bulk". This is done by grouping time slots and not switching within these slots. The second is that as time goes by, the algorithm has more information about the optimal power value, and hence the block size should increase to take advantage of the better knowledge.

V. Performance Analysis of the Proposed Algorithms

So far, we have presented two sets of power assignment algorithms (without and with switching cost), each of which further consists of components that have different assumptions on the prior knowledge and the correlation structure. In this section, we will provide a *unified* performance analysis framework that can be applied to *all* of the developed algorithms. We focus on the *finite-time* analysis where, for a given stopping time T, the cumulative PIF loss and the convergence speed will be characterized. In this way, we can shed important light on the fundamental differences of these algorithms, and how these differences impact their performances.

We start with the expected cumulative PIF loss defined in Sec. III-A. For BPA and CBPA, the expected PIF loss can be written as (6). When the switching cost is considered, equation (5) and (6) should be rewritten as:

$$\begin{aligned} R_T^{SC} &= G_T^* - G_T^S \\ &= \max_{i=1,..,n}\left(\sum_{t=1}^{T} r_i(t)\right) - \sum_{t=1}^{T} r_{a(t)}(t) + \sum_{t=2}^{T} s_{a(t)a(t-1)}, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[R_T^{SC}] &= T\mu^* - \mathbb{E}\left(\sum_{t=1}^{T}\mu_{a(t)} - \mathsf{SC}(T)\right) \\ &= \sum_{i=1}^{n}\Delta_i \mathbb{E}[N_i(T)] + \mathbb{E}[\mathsf{SC}(T)], \end{aligned}$$

respectively.

A. Upper Bound Analysis

In order to derive the unified framework that applies to all the algorithms, we focus on analyzing CBPA-SC as it is the

most general algorithm consisting of all the key components. As we have discussed, the expected cumulative PIF loss should grow sub-linearly with T in order to achieve the optimal performance, which indicates that $\lim_{T\to\infty} R_T/T = 0$. We have the following theorem to bound the expected cumulative PIF loss of CBPA-SC.

Theorem 1. *The expected cumulative PIF loss $\mathbb{E}[R_T^{SC}]$ of CBPA-SC is bounded above as:*

$$\begin{aligned}\mathbb{E}[R_T^{SC}] &\leqslant \sum_{i=1,i\neq i^*}^{n} \Delta_i \mathbb{E}[N_i(T)] + \mathbb{E}[\mathsf{SC}(t)] \\ &\leqslant \sum_{i=1,i\neq i^*}^{n} \Big(\Delta_i (C_1^i \log T + C_2^i) + (\tilde{s}_i^{max} + \tilde{s}_{i^*}^{max})\mathbb{E}[S_i(T)]\Big) \\ &\quad + \tilde{s}_{i^*}^{max} \\ &\leqslant \sum_{i=1,i\neq i^*}^{n} \Delta_i (C_1^i \log T + C_2^i) + \sum_{i=1,i\neq i^*}^{n} (\tilde{s}_i^{max} + \tilde{s}_{i^*}^{max}) \\ &\left(\log 2C_1^i \sqrt{\log_2 T} + (C_2^i + \log 2C_1^i)\left(1 + \frac{\pi^2}{6}\right)\right) + \tilde{s}_{i^*}^{max},\end{aligned}$$

where

$$C_1^i = \frac{16\sigma_0^2}{\Delta_i^2} + \frac{\log 2}{2}\left(e^{\frac{3M_{i^*}^2}{2\sigma_0^2}} + e^{\frac{3M_i^2}{2\sigma_0^2}}\right),$$

$$C_2^i = \frac{4\sigma_0^2}{\Delta_i^2}\log\sqrt{2\pi e} + \left(e^{\frac{M_{i^*}^2}{3\sigma_0^2}} + e^{\frac{M_i^2}{3\sigma_0^2}}\right),$$

$\delta_i^2 = \sigma_0^2/\sigma_{i-cond}^2$, and $\sigma_{i-cond}^2 = \sigma_0^2 - \sigma_i(0)\Sigma_{\sim i}^{-1}(0)\sigma_i^T(0)$. $M_i = \sigma_0^2\sqrt{1+\delta_i^2}\sum_{j=1}^{n}\sum_{k=1}^{n}|\lambda_{kj}^0||\mu_j^0 - \mu_j|$ measures the accuracy of the prior knowledge, where $\Sigma_{\sim i}$ is the submatrix of Σ_0, which excludes the i-th column and i-th row and λ_{kj}^0 is the component of Λ_0. $\tilde{s}_i^{max} = \max_{j=1,..,n}\mathbb{E}[s_{ij}]$ is the maximum expected switching loss when SBS changes power to p_i.

Proof. See Appendix A. □

Theorem 1 provides an $\mathcal{O}(\log T)$ upper bound for CBPA-SC, which guarantees that its cumulative PIF will converge to that of the global optimum power value at a rate of $\mathcal{O}(\log T/T)$. Furthermore, this upper bound applies to any finite time T and any general function of switching loss $f(|p_i - p_j|)$ as long as f is a non-decreasing and finite function.

Theorem 1 is a powerful result as it gives an $\mathcal{O}(\log T)$ bound for the most general algorithm CBPA-SC. We can now derive similar results for all the other proposed algorithms. First, when $s_{ij} = 0, \forall i, j \in \mathcal{N}_{pow}$, Theorem 1 can be applied to CBPA. Formally, we have the following corollary.

Corollary 2. *The expected cumulative PIF loss $\mathbb{E}[R_T]$ of CBPA is bounded above as:*

$$\mathbb{E}[R_T] \leq \sum_{i=1,i\neq i^*}^{n} \Delta_i \left(\lceil \frac{4\sigma_0^2}{\Delta_i^2}(\log 2\pi e + 4\log T) - 1\rceil + \hat{N}_i\right),$$

where

$$\hat{N}_i = e^{\frac{M_{i^*}^2}{3\sigma_0^2}} + e^{\frac{M_i^2}{3\sigma_0^2}} + \frac{9}{2}\left(e^{\frac{3M_{i^*}^2}{2\sigma_0^2}} + e^{\frac{3M_i^2}{2\sigma_0^2}}\right).$$

Proof. See Appendix B. □

As Corollary 2 shows, the $\mathcal{O}(\log T)$ upper bound of the cumulative PIF loss still holds for the CBPA algorithm. Thus, adding switching cost into the problem does not change the optimal scaling of the cumulative PIF loss. However, the algorithm that deals with the switching cost (CBPA-SC) is considerably more complicated than the one without the switching cost (CBPA).

Next, we note that the difference between BPA and CBPA lies in the correlation structure. We can further remove the correlation component in Corollary 2 to analyze BPA.

Corollary 3. *The expected cumulative PIF loss $\mathbb{E}[R_T]$ of BPA is bounded above as:*

$$\begin{aligned}\mathbb{E}[R_T] \quad \leq \quad & \sum_{i=1,i\neq i^*}^{n} \Delta_i \Big(\lceil \frac{4\sigma_0^2}{\Delta_i^2}(\log 2\pi e + 4\log T) - 1\rceil \\ & + e^{\frac{\Delta m_{i^*}^2}{3\sigma_0^2}} + e^{\frac{\Delta m_i^2}{3\sigma_0^2}} + \frac{9}{2}e^{\frac{3\Delta m_{i^*}^2}{2\sigma_0^2}} + \frac{9}{2}e^{\frac{3\Delta m_i^2}{2\sigma_0^2}}\Big),\end{aligned}$$

where $\Delta m_i = \mu_i - \mu_i^0$ measures the accuracy of the prior knowledge of the mean PIF.

Proof. See Appendix C. □

Finally, because the UiPA algorithm does not use any prior knowledge, its utility function $Q_i^{UiPA}(t)$ is similar to the *UCB1-NORMAL* algorithm in [13]. Thus, the upper bound of the expected PIF loss can be derived analogously.

Theorem 4. *The expected cumulative PIF loss $\mathbb{E}[R_T]$ of UiPA is bounded above as:*

$$\begin{aligned}\mathbb{E}[R_T] \quad \leq \quad & \sum_{i=1,i\neq i^*}^{n} \Delta_i \Big(\frac{16\sigma_0^2}{\Delta_i^2}(\log 2\pi e + 4\log T) \\ & + ((2\pi e)^{-1/4} + 2)\log T + \frac{\log 2\pi e}{2} + \frac{2}{\sqrt{2\pi e}}\Big),\end{aligned}$$

Proof. See Appendix D. □

We can see that even though the constant terms in the upper bounds of CBPA and BPA may possibly be larger than the ones of UiPA, with a much smaller coefficient of $\log T$, the performance turns out to be better. Moreover, if the prior knowledge is accurate in BPA and CBPA, the upper bounds for both will become:

$$\mathbb{E}[R_T] \leq \sum_{\substack{i=1\\ i\neq i^*}}^{n} \Delta_i \left(\lceil \frac{4\sigma_0^2}{\Delta_i^2}(\log 2\pi e + 4\log T) - 1\rceil + \frac{4}{\sqrt{2\pi e}}\right),$$

which can be easily derived from the corollaries.

VI. Reducing Complexity in Multi-SBS

A practical problem in a multi-SBS deployment may arise due to the "curse of dimensionality". As the set of arms consists of the combinations of different power levels at all SBSs, it leads to n^K arms and incurs exponential time and space complexity for the proposed algorithms. Plus, the number of available power levels for each SBS n can be large. Note that in the CBPA and CBPA-SC algorithms, we

need matrix calculations when updating the estimated state, which calls for $\mathcal{O}(n^{3K})$ time complexity and $\mathcal{O}(n^{2K})$ space complexity [22]. This severely limits the applicability of the proposed algorithms in large enterprise networks.

To reduce the complexity, we first explore a practical constraint that has not been utilized in the proposed algorithms. In real-world deployment, the neighboring SBSs are generally not allowed to have vastly different pilot power levels. This is because otherwise they may result in significantly different coverage areas and therefore lead to very uneven load distributions. Thus, utilizing this practical constraint, we only need to consider the combinations of power levels in which neighboring SBS power levels are different by no more than a certain threshold P_{th}.

Even with the power difference threshold, the size of set is still exponential in K. To further reduce the complexity, we notice that the performance space of all set of arms exhibits certain "clustering" effect that can be utilized. For two power settings that differ only slightly (e.g., $\{0, 3, 5\}$ and $\{0, 4, 4\}$ dBm for $K = 3$), the performances may be very similar. Thus, if we can carefully group the power settings into a few clusters, and only use the cluster center as the representative power setting, we can achieve a good tradeoff between complexity and performance for the algorithms.

We propose to perform a clustering operation to address the complexity issue. The clustering operation is done after the self-configuration phase to leverage the prior knowledge, but before invoking the power assignment algorithm. We adopt the *K-medoids* clustering [23] because, different from the well-known K-means clustering, K-medoids is based on the most central object instead of the centroids in K-means, each of which is the mean point of all objects in the cluster. Therefore, the medoids in each cluster can be seen as the representative power settings. We note that the choice of the number of clusters N plays a critical role in the overall performance. If it is too large, the global optimum power setting may be a medoid with high probability, which contributes to high accuracy for the power assignment process but also increases the complexity and leads to low efficiency, and vice versa.

We further note that there is a $\mathcal{O}(n^K N)$ time complexity for K-medoids clustering [23], but as clustering is done prior to the self-optimization phase, the process can be handled offline. Thus, time complexity is less of a concern.

VII. Simulation Results

A. Simulation setup

We resort to numerical simulations to verify the effectiveness of the developed transmit power assignment algorithms. A system-level heterogeneous network simulator is developed considering both indoor SBS and outdoor MBS. We consider a large warehouse with $K = 1, 2, 4$ SBSs deployed inside and a MBS outside with a fixed transmit power setting. The measurement points constitute two routes inside and outside respectively which we assume to follow concentric circle or ellipse pattern. 100 measurement points are set uniformly on each route. At each time slot, the measurement points feedback their own coverage condition, determined by the respective SINR which is naturally decided by the current SBS power setting. We set the total time horizon as $T = 3000$ slots and iterate each simulation setting for 50 times to average out the randomness. The size of the warehouse and the SBS locations are given in Table I. Here we set the center of the warehouse as origin. The PIF r under each power value can be calculated following the procedure in Sec. II.

We obtain the received power from SBS or MBS using the indoor femto channel model of urban deployment from [24] as follows.

- indoor UE to MBS:

$$PL(d)[\text{dB}] = 15.3 + 37.6 \times \log_{10}(d) + L_{ow} + X_{\sigma_{dB}}, \quad (10)$$

- outdoor UE to MBS:

$$PL(d)[\text{dB}] = 15.3 + 37.6 \times \log_{10}(d) + X_{\sigma_{dB}}, \quad (11)$$

- indoor UE to SBS:

$$PL(d)[\text{dB}] = 38.46 + 20 \times \log_{10}(d) + X_{\sigma'_{dB}}, \quad (12)$$

- outdoor UE to SBS:

$$\begin{aligned} PL(d)[\text{dB}] = \max\{&15.3 + 37.6 \log_{10}(d), \\ &38.46 + 20 \log_{10}(d)\} + L_{ow} + X_{\sigma'_{dB}}. \end{aligned} \quad (13)$$

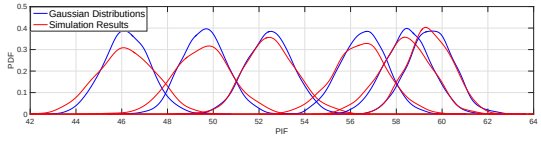
Note that (10) and (12) are for indoor routes while (11) and (13) are for outdoor routes; d represents the separation between a BS and the measurement point; L_{ow} is the penetration loss of an outdoor wall, which indoor user suffers when receiving power from outdoor MBS and outdoor user receiving from indoor SBS; $X_{\sigma_{dB}}$ and $X_{\sigma'_{dB}}$ stand for shadow fading. Other important simulation parameters are summarized in Table I.

TABLE I
Simulation Parameters

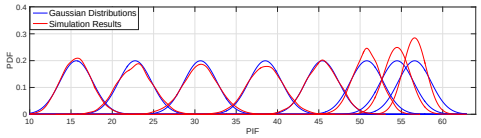
Parameters	**Value**
SBS transmit power	[-10dBm, 20dBm]
MBS transmit power	40dBm
Thermal noise density	-174dBm/Hz
Bandwidth	20MHz
Carrier frequency	2GHz
Penetration loss (L_{ow})	20dB
Shadowing effect	log-normal with $\sigma = 8$dB, $\sigma' = 4$dB
d_0	1m
α	0.7
Enterprise Size	K=1 30m×30m K=2 40m×40m K=4 50m×40m
SBS location	K=1 (12m,8m) K=2 (16m,17m), (-15m,-11m) K=4 (20m,18m), (11m,-19m) (-11m,18.5m), (-10.5m,-19m)

B. Evaluation of the PIF Gaussian Distribution

We first study the empirical distribution of the PIF r in $K = 1$ SBS with the set of power levels $\mathcal{P} = \{-10, -5, .., 15, 20\}$



(a) MBS-SBS distance is 150m.

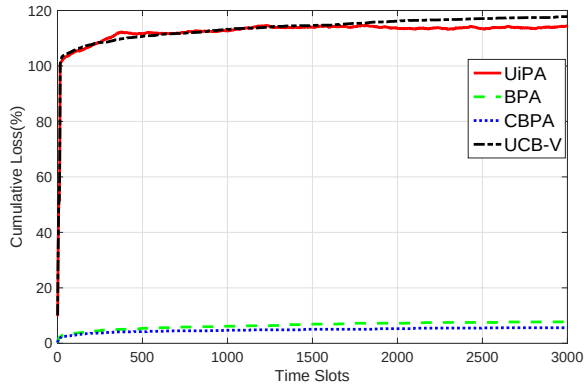


(b) MBS-SBS distance is 70m.

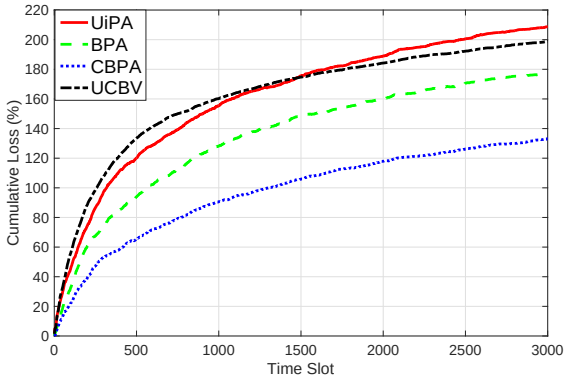
Fig. 4. A $40 \times 30m^2$ warehouse, with elliptic routes inside and outside.

dBm. We present the comparison of empirical and Gaussian distributions in two representative scenarios in Fig. 4(a) and 4(b). As we can see, the assumption on Gaussian distributed PIFs matches well with the empirical distributions.

To further verify the dependency on the Gaussian distribution, we study the performance of the proposed algorithms compared with a well-behaved UCB extended algorithm which makes no assumptions on the distribution of the rewards, e.g. UCB-V in [25] under non-Gaussian reward distributions. More specifically, two well-adopted distributions in wireless communications, *uniform* and *Rayleigh*, are considered. We can see from Fig. 5 that performances under non-Gaussian distributions are still very good, particularly for BPA and CBPA. This observation indicates that Gaussianness is not a fundamental assumption that must be met to guarantee the effectiveness of the algorithms.



(a) Uniform

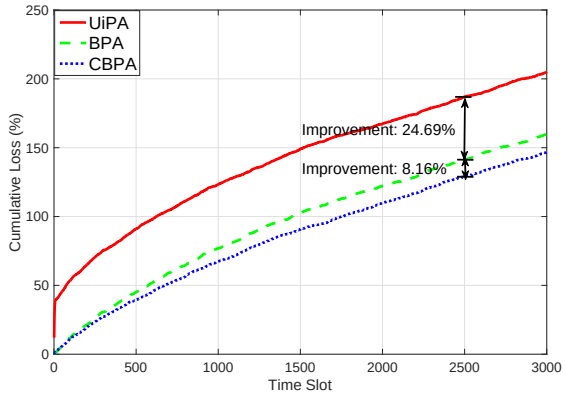


(b) Rayleigh

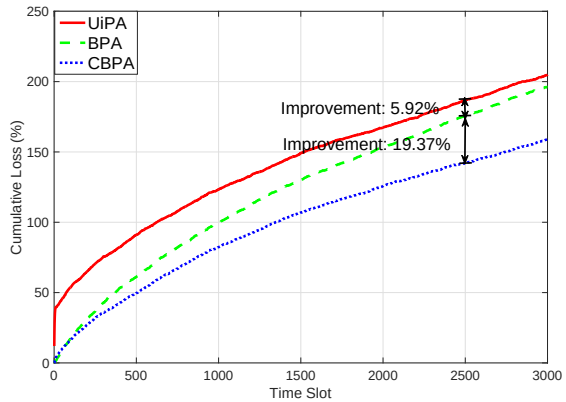
Fig. 5. Verifications of algorithms under non-Gaussian distributed rewards.

C. System Performance

In the simulation setting for $K = 1$, we deploy an outside MBS at $[100m, 100m]$. The set of power levels for SBS is $\mathcal{P} = \{-10, -8, .., 18, 20\}$ dBm while other settings follow Table I. The inside and outside routes have the concentric circle pattern, whose radiuses are (2, 13) meters for the two indoor routes, and (24, 30) meters for the two outdoor routes. The cumulative loss over time is used to evaluate the performance, and we use the optimal power achieved by the global optimization of the expected PIF with complete RF information as the genie-aided optimum.



(a) Good quality (with estimated mean)



(b) Poor quality (with uniform distribution)

Fig. 6. Cumulative loss comparison of prior knowledge with different qualities in a single-SBS deployment with $\alpha = 0.7$.

We first compare the performance of UiPA, BPA and CBPA algorithms with different quality of priors. Fig. 6(a) reports the

cumulative loss over time for all three algorithms when the prior knowledge is of good quality, i.e. the estimated mean from the empirical distribution is used. Fig. 6(b) shows the same simulation but with a poor prior knowledge, which uses a uniform prior distribution with each element $\mu_0 = 50$. A few important observations can be made from these simulations. First of all, we see that all three algorithms can converge to the optimal power value asymptotically, but with different speed. To further evaluate the convergence speed, we plot the empirical CDF of the convergence time for all three algorithms in Fig. 7. It becomes clear that leveraging both the prior knowledge and the correlation structure significantly accelerates the convergence of CBPA. In terms of minimizing the total PIF loss, CBPA also outperforms BPA which performs better than UiPA. Second, degradation of the quality of the priors degrades the performance of BPA and CBPA. Particularly, performance of the BPA is getting close to UiPA with poor prior knowledge. It is interesting to note that even with poor prior, CBPA still converge faster than other algorithms with good prior. This is because when the prior knowledge is inaccurate, CBPA recovers some of the PIF degradation by leveraging its correlation structure. Lastly, as UiPA does not leverage the prior knowledge, changing its quality does not affect the convergence speed.

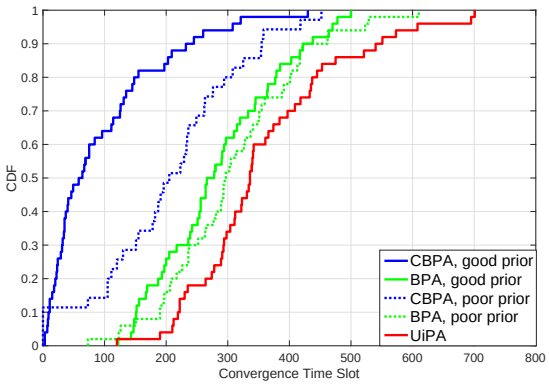
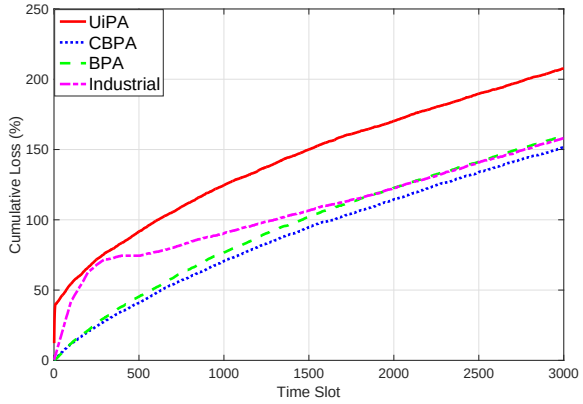


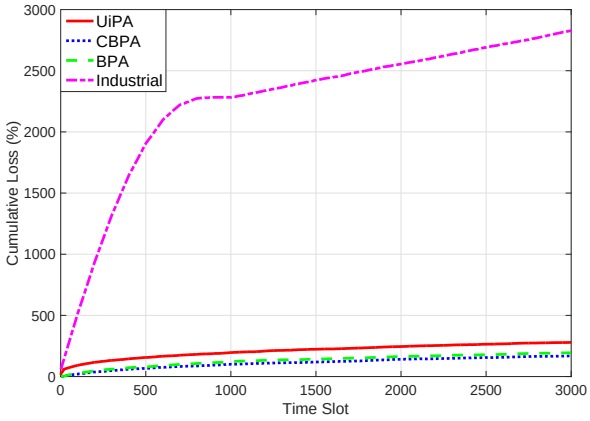
Fig. 7. Convergence speed comparison of prior knowledge with different qualities in a single-SBS deployment with $\alpha = 0.7$. "Good prior" corresponds to using the estimated mean while "poor prior" uses a uniform distribution.

Next, we compare the proposed algorithms with the industry solution. The heuristic solution [9] keeps a power value long enough to obtain a near-perfect PIF estimation, and then it either increases or decreases the power value by a fixed step size. Clearly, this method trades off fast convergence for certainty. Fig. 8 reports the numerical comparison with a maximum 20dBm and step size 2dB. We can see that the industrial solution adapts poorly to different deployments, while our algorithms are stable thanks to online learning.

For $K = 2$ and $K = 4$, an outside MBS is deployed at $[100m, 100m]$. The power value difference threshold is $P_{th} = 5$dB. The power value for each SBS is selected from $\mathcal{P} = \{-10, -5, .., 15, 20\}$dBm. It results in $n = 19$ for $K = 2$ without any clustering, which may be acceptable in terms of complexity. The cumulative PIF loss with respect to the optimal power setting is shown in Fig. 9(a). For $K = 4$ case,



(a) Large warehouse



(b) Small single-office enterprise

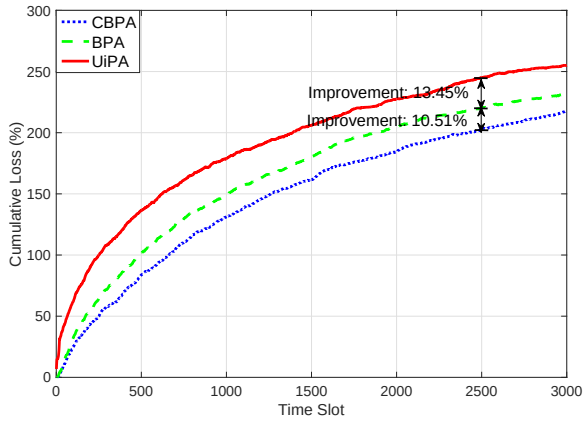
Fig. 8. Comparison of the industrial solution to UiPA, BPA and CBPA in different scenarios.

however, there are $n = 149$ power settings. We thus employ the clustering strategy in Sec. VI and study two cases where the number of clusters is either $N = 20$ or $N = 40$. The PIF loss normalized by time is shown in Fig. 9(b). We can see that all algorithms exhibit a decaying loss per slot. As for the effect of N, there exists an initial period when larger cluster number results in worse performance for all the algorithms. This is because during the initial slots, more power settings lead to more exploration and thus sub-optimal power settings are selected more. As time goes by, the algorithms have more knowledge about the optimal power setting. While a larger cluster number means one of the selected clustering medoids is closer to the globally optimal power setting, a larger N results in a better performance. Detailed coverage and leakage results under optimal selections are reported in Table II.

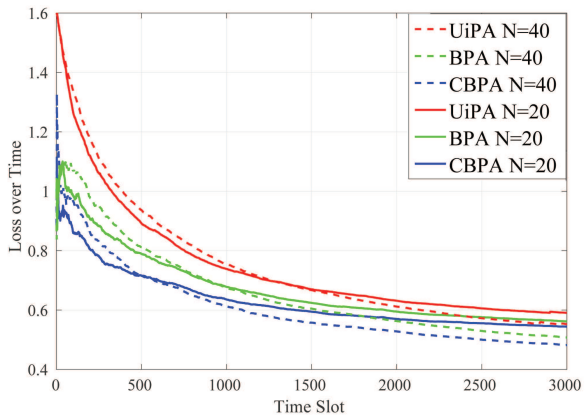
Fig. 10(a) and 10(b) study the impact of power switching cost for $K = 2$ and $K = 4$, respectively. Here we adopt a simple linear function of switching loss as $s_{ij} = \gamma|p_i - p_j|$, where γ is a tunable parameter for different scenarios and we set as 0.2. We can see that the additional performance loss occurring whenever a SBS changes its power value increases the overall performance loss in all algorithms. However, the

TABLE II
MULTI-SBS SIMULATION RESULTS

Metric	K=2	K=4
Globally optimal power [dBm]	(0, 5)	(0, 5, 10, 15)
Coverage percentage	91.506%	96.548%
Leakage percentage	5.691%	28.725%
Simulation output power [dBm]	(0, 5)	(-5, 0, 5, 10) when $N = 20$ (0, 5, 10, 15) when $N = 40$

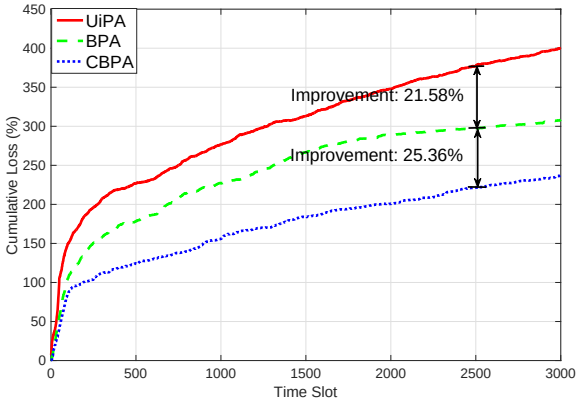


(a) Two SBSs deployed in a 40m × 40m enterprise

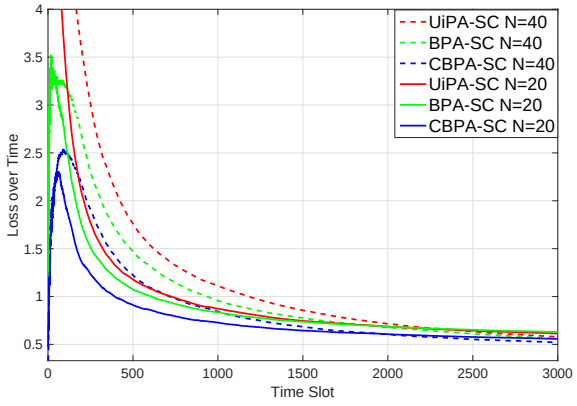


(b) Four SBSs deployed in a 50m × 40m enterprise

Fig. 9. Simulation results for two and four SBSs in different scenarios.



(a) Cumulative loss, $K = 2$



(b) Per-slot loss, $K = 4$

Fig. 10. Performance with switching cost factor $\gamma = 0.2$.

algorithms can still converge to the optimal power settings asymptotically in a sub-linear fashion, matching the regret analysis in Sec. III. In Fig. 10(b), the performances of different cluster numbers also comply with our previous analysis.

VIII. CONCLUSION

We have studied the pilot power assignment problem associated with indoor enterprise closed-access SBS networks, in which the focus is on achieving optimal balance between providing sufficient coverage for the indoor users and suppressing leakage that causes interference to outdoor MBS users. We modeled power assignment as an online learning problem, and adopted a Bayesian approach that leverages the prior information of the Gaussian distribution. We proposed bandit-inspired power assignment algorithms that utilize different levels of the statistical information. The CBPA algorithm makes use of both prior knowledge of the mean and variance of each arm as well as the dependency of PIFs across different power values. In contrast, the BPA algorithm only uses the prior knowledge but not the correlation information, and its performance is worse than CBPA but better than the UiPA algorithm that does not use either prior or correlation. Furthermore, we explicitly took into account the power switching cost,

and enhanced the power assignment algorithms with a block allocation scheme to reduce frequent power-switchings. A sub-linear upper bound for performance loss was proved for all the algorithms. Furthermore, for the multi-SBS deployment, we proposed to use K-medoids clustering to reduce the complexity while maintaining the performance. When the cluster number becomes large, the algorithms can approach the globally optimal power setting for all K SBSs.

As a possible future direction, the *spectral bandits* method proposed in [26] offers a new perspective to efficiently handle a large number of arms while capturing the correlation structure. This can be an interesting alternative for the enterprise transmit power assignment problem. In particular, complexity and performance comparison with the algorithms of this paper may shed light into its feasibility.

Appendix A
Proof of Theorem 1

We start by proving for the case $L_{l-1} < T \leqslant L_l$. Note that

$$\begin{aligned} N_i(T) &= \sum_{t=1}^{T} \mathcal{I}(p_{a(t)} = i) \leqslant \sum_{t=1}^{T} \mathcal{I}(Q_i^t \geqslant Q_{i^*}^t) \\ &\leqslant \eta_i + \sum_{t=1}^{T} \mathcal{I}(Q_i^t \geqslant Q_{i^*}^t, N_i(t-1) \geqslant \eta_i) \qquad (14) \\ &\leqslant \eta_i + \sum_{f=1}^{l} \sum_{k=1}^{b_f} f\mathcal{I}(Q_i^t \geqslant Q_{i^*}^t, N_i(\tau_{fk}) \geqslant \eta_i), \qquad (15) \end{aligned}$$

where $i^* = \arg\max_{i=1,..,n} \mu_i$, η_i is a positive integer, and $\mathcal{I}(x)$ is the indicator function. At any time t, sub-optimal i is selected only when $Q_{i^*}^t \leqslant Q_i^t$, which is true as long as one of the following inequalities holds:

$$\hat{\mu}_{i^*}(\tau_{fk}) \leqslant \mu_{i^*} - U_{i^*}(\tau_{fk}), \qquad (16a)$$

$$\hat{\mu}_i(\tau_{fk}) \geqslant \mu_i + U_i(\tau_{fk}), \qquad (16b)$$

$$\mu_{i^*} < \mu_i + 2U_i(\tau_{fk}), \qquad (16c)$$

where $U_i(\tau_{fk}) = \hat{\sigma}_i(\tau_{fk}) \sqrt{\sum_{j=1}^{n} \rho_{ij}^2(\tau_{fk})} \Phi^{-1}(1 - 1/\sqrt{2\pi e}\tau_{fk}^2)$. Define the bias $\boldsymbol{e}$ and covariance $\bar{\Sigma}$ of the estimate $\hat{\boldsymbol{\mu}}(t)$, with e_i and $\bar{\sigma}_i$ representing the i-th entry of $\boldsymbol{e}$ and the diagonal of $\bar{\Sigma}$, and we have $\hat{\boldsymbol{\mu}}(t) \sim \mathcal{N}(\boldsymbol{e}(t) + \boldsymbol{\mu}, \bar{\Sigma}(t))$, with $e_i(t) = \sum_{j=1}^{n} \sum_{k=1}^{n} \hat{\sigma}_{ik}(t) \lambda_{kj}^0 (\mu_j^0 - \mu_j)$.

We now separately analyze (16a), (16b), and (16c). First, if $N_{i^*}(\tau_{fk}) = 0$, then (16a) is false if [16, Lemma 7]

$$U_{i^*}(\tau_{fk}) > \sigma_{i*-cond} \sqrt{3 \log \tau_{fk}} \geqslant \frac{M_{i^*}}{\sqrt{1 + \delta_{i^*}^2}} \geqslant |e_{i^*}(\tau_{fk})|$$

or equivalently,

$$\tau_{fk} > e^{\frac{M_{i^*}^2 \delta_{i^*}^2}{3\sigma_0^2(1+\delta_{i^*}^2)}}. \qquad (17)$$

Otherwise, if $N_{i^*}(\tau_{fk}) \geqslant 1$, we have

$$\begin{aligned} &\mathbb{P}\{\hat{\mu}_{i^*}(\tau_{fk}) \leqslant \mu_{i^*} - U_{i^*}(\tau_{fk})\} \\ \leqslant\ &\mathbb{P}\left\{ z \geqslant \Phi^{-1}(1 - 1/\sqrt{2\pi e}\tau_{fk}^2) - \frac{M_{i^*}}{\sigma_0} \right\} \\ \leqslant\ &\mathbb{P}\left\{ z \geqslant \sqrt{3 \log \tau_{fk}} - \frac{M_{i^*}}{\sigma_0} \right\}, \qquad (18) \end{aligned}$$

where z is a standard Gaussian random variable. This indicates that $\sqrt{3 \log \tau_{fk}} - \frac{M_{i^*}}{\sigma_0} \geqslant 0$. Thus we have $\tau_{fk} > e^{M_{i^*}^2/3\sigma_0^2} = \tau_1$. For $\tau_{fk} > \tau_1$, we have

$$\begin{aligned} \mathbb{P}\{\text{(16a) holds}\} &\leqslant \frac{1}{2} exp\left(-\frac{1}{2} \left(\sqrt{3 \log \tau_{fk}} - \frac{M_{i^*}}{\sigma_0} \right)^2 \right) \\ &\leqslant \frac{1}{2} exp\left(-\frac{1}{2} \left(\frac{9}{4} \log \tau_{fk} - 3\frac{M_{i^*}^2}{\sigma_0^2} \right) \right) \\ &= \frac{1}{2} e^{\frac{3M_{i^*}^2}{2\sigma_0^2}} \tau_{fk}^{-\frac{9}{8}}. \qquad (19) \end{aligned}$$

Inequality (19) is deduced using [16, Lemma 2].

Similarly, we can deduce that if $N_i(\tau_{fk}) > \eta_i$ and $\tau_{fk} \geqslant \tau_2 := e^{M_i^2/3\sigma_0^2}$, then

$$\mathbb{P}\{\text{(16b) holds}\} \leqslant \frac{1}{2} e^{\frac{3M_i^2}{2\sigma_0^2}} \tau_{fk}^{-\frac{9}{8}}. \qquad (20)$$

For inequality (16c), it holds if

$$\begin{aligned} &\mu_{i^*} - \mu_i < 2U_i(\tau_{fk}) \\ \Longrightarrow\ &\Delta_i < \frac{2\sigma_0}{\sqrt{1 + N_i(\tau_{fk})}} \Phi^{-1}(1 - 1/\sqrt{2\pi e}\tau_{fk}^2) \\ \Longrightarrow\ &N_i(\tau_{fk}) < \frac{4\sigma_0^2}{\Delta_i^2}(\log 2\pi e + 4 \log T) - 1. \end{aligned}$$

Thus we have that (16c) does not hold if

$$N_i(\tau_{fk}) \geqslant \frac{4\sigma_0^2}{\Delta_i^2}(\log 2\pi e + 4 \log T) - 1. \qquad (21)$$

Setting $\eta_i = \lceil \frac{4\sigma_0^2}{\Delta_i^2}(\log 2\pi e + 4 \log T) - 1 \rceil$ and combining (17), (19) and (20), the inequality (15) can be written as

$$\mathbb{E}[N_i(T)] \leqslant \eta_i + \tau_1 + \tau_2 + \frac{1}{2} \left(e^{\frac{3M_{i^*}^2}{2\sigma_0^2}} + e^{\frac{3M_i^2}{2\sigma_0^2}} \right) \sum_{f=1}^{l} \sum_{k=1}^{b_f} f \tau_{fk}^{-\frac{9}{8}}. \qquad (22)$$

We now focus on $\sum_{f=1}^{l} \sum_{k=1}^{b_f} f\tau_{fk}^{-\frac{9}{8}}$. With $\tau_{fk} = L_{f-1} + 1 + (k-1)f$ and $2^{f^2} \leqslant L_f \leqslant 2^{f^2} + f^2$, we have

$$\begin{aligned} \sum_{k=1}^{b_f} f\tau_{fk}^{-\frac{9}{8}} &\leqslant \sum_{k=1}^{b_f} f(2^{(f-1)^2} + (f-1)^2 + 1 + (r-1)f)^{-9/8} \\ &\leqslant \sum_{k=1}^{b_f} \frac{f}{2^{(f-1)^2} + (f-1)^2 + 1 + (r-1)f} \\ &\leqslant \int_1^{b_f} \frac{f}{2^{(f-1)^2} + (f-1)^2 + 1 + (r-1)f} dr \\ &= \log \frac{2^{f^2} + (f-1)^2 + 1}{2^{(f-1)^2} + (f-1)^2 + 1} \\ &\leqslant \log \frac{2^{f^2}}{2^{(f-1)^2}}, \end{aligned}$$

and

$$\sum_{f=1}^{l} \sum_{k=1}^{b_f} f\tau_{fk}^{-\frac{9}{8}} \leqslant \sum_{f=1}^{l} \log \frac{2^{f^2}}{2^{(f-1)^2}} = l^2 \log 2 \leqslant \log 2T.$$

Therefore (22) yields

$$\begin{aligned}\mathbb{E}[N_i(T)] &\leqslant \eta_i+\tau_1+\tau_2+\frac{1}{2}\left(e^{\frac{3M_{i^*}^2}{2\sigma_0^2}}+e^{\frac{3M_i^2}{2\sigma_0^2}}\right)\log 2T\\ &\leqslant C_1^i\log T+C_2^i,\\ C_1^i &= \frac{16\sigma_0^2}{\Delta_i^2}+\frac{\log 2}{2}\left(e^{\frac{3M_{i^*}^2}{2\sigma_0^2}}+e^{\frac{3M_i^2}{2\sigma_0^2}}\right),\\ C_2^i &= \frac{4\sigma_0^2}{\Delta_i^2}\log\sqrt{2\pi e}+\left(e^{\frac{M_{i^*}^2}{3\sigma_0^2}}+e^{\frac{M_i^2}{3\sigma_0^2}}\right).\end{aligned}$$

We then establish the expected number of switches to a sub-optimal arm i from a different arm. We have

$$\begin{aligned}S_i(T) &\leqslant 1+\sum_{f=1}^{l}\frac{N_i(L_f)-N_i(L_{f-1})}{f}\\ &= 1+\sum_{f=1}^{l}\frac{N_i(L_f)}{f}-\sum_{f=0}^{l-1}\frac{N_i(L_{f-1})}{f+1}\\ &= \frac{N_i(L_l)}{l}+\sum_{f=1}^{l-1}N_i(L_f)\left(\frac{1}{f}-\frac{1}{f+1}\right)\\ &\leqslant \frac{N_i(L_l)}{l}+\sum_{f=1}^{l-1}\frac{1}{f^2},\end{aligned}$$

using the same argument as [21]. Then it follows that

$$\mathbb{E}[S_i(T)]\leqslant\frac{\mathbb{E}[N_i(L_l)]}{l}+\sum_{f=1}^{l-1}\frac{\mathbb{E}[N_i(L_f)]}{f^2}. \tag{23}$$

With the upper bound on $\mathbb{E}[N_i(T)]$ and $L_f\leqslant 2^{f^2}+f^2\leqslant 2^{f^2+1}$, (23) can be further deduced as

$$\begin{aligned}&\mathbb{E}[S_i(T)]\leqslant\frac{C_1^i\log L_l+C_2^i}{l}+\sum_{f=1}^{l-1}\frac{C_1^i\log L_f+C_2^i}{f^2}\\ \leqslant\ &\frac{C_2^i}{l}+\sum_{f=1}^{l-1}\frac{C_2^i}{f^2}+\frac{C_1^i\log 2^{l^2+1}}{l}+\sum_{f=1}^{l-1}\frac{C_1^i\log 2^{f^2+1}}{f^2}\\ \leqslant\ &C_2^i\left(1+\frac{\pi^2}{6}\right)+\log 2C_1^i\left(l+\frac{\pi^2}{6}\right)\\ \leqslant\ &\log 2C_1^i\sqrt{\log_2 T}+(C_2^i+\log 2C_1^i)\left(1+\frac{\pi^2}{6}\right).\end{aligned}$$

Finally, the cumulative switching cost can be bounded as

$$\begin{aligned}\mathsf{SC}(T) &\leqslant \sum_{i=1,i\neq i^*}^{n}\tilde{s}_i^{max}\mathbb{E}[S_i(T)]+\tilde{s}_{i^*}^{max}\mathbb{E}[S_{i^*}(T)]\\ &\leqslant \sum_{i=1,i\neq i^*}^{n}(\tilde{s}_i^{max}+\tilde{s}_{i^*}^{max})\mathbb{E}[S_i(T)]+\tilde{s}_{i^*}^{max}.\end{aligned}$$

Appendix B
Proof of Corollary 2

For the CBPA algorithm, (14) still holds. Hence, the argument from (16a) to (21) equally applies to any time slot $t=1,2,..,T$. The proof is complete by rewriting (14) as

$$\begin{aligned}\mathbb{E}[N_i(T)] &\leqslant \eta_i+\tau_1+\tau_2+\frac{1}{2}\left(e^{\frac{3M_{i^*}^2}{2\sigma_0^2}}+e^{\frac{3M_i^2}{2\sigma_0^2}}\right)\sum_{t=1}^{T}t^{-\frac{9}{8}}\\ &\leqslant \lceil\frac{4\sigma_0^2}{\Delta_i^2}(\log 2\pi e+4\log T)-1\rceil+\hat{N}_i,\\ \hat{N}_i &= e^{\frac{M_{i^*}^2}{3\sigma_0^2}}+e^{\frac{M_i^2}{3\sigma_0^2}}+\frac{9}{2}\left(e^{\frac{3M_i^2}{2\sigma_0^2}}+e^{\frac{3M_i^2}{2\sigma_0^2}}\right).\end{aligned}$$

Appendix C
Proof of Corollary 3

In the BPA algorithm, inequalities (14)(16a)(16b)(16c) still hold for any time slot $t=1,2,..,T$, with $U_i(t)=\frac{\sigma_0}{\sqrt{1+N_i(t)}}\Phi^{-1}(1-1/\sqrt{2\pi e}t^2)$. The estimated mean $\hat{\mu}_i(t)$ is a Gaussian random variable with mean $\frac{\mu_i^0+N_i(t)\mu_i}{1+N_i(t)}$ and variance $\frac{N_i(t)\sigma_0^2}{(1+N_i(t))^2}$. The proof then follows the similar steps as Appendix A, with inequality (18) written as

$$\begin{aligned}&\mathbb{P}\{\hat{\mu}_{i^*}(t)\leqslant\mu_{i^*}-U_{i^*}(t)\}\leqslant\\ &\mathbb{P}\left\{z\geqslant\sqrt{\frac{N_{i^*}+1}{N_{i^*}}}\Phi^{-1}\left(1-\frac{1}{\sqrt{2\pi e}\tau_{fk}^2}\right)-\frac{\Delta m_{i^*}}{\sigma_0\sqrt{N_{i^*}(t)}}\right\}.\end{aligned}$$

Thus, inequalities (19) and (20) become

$$\begin{aligned}&\mathbb{P}\{\text{(16a) holds}, t>\tau_1\}\geqslant\frac{1}{2}e^{\frac{3\Delta m_{i^*}^2}{2\sigma_0^2}}t^{-\frac{9}{8}},\quad \tau_1=e^{\frac{\Delta m_{i^*}^2}{3\sigma_0^2}}\\ &\mathbb{P}\{\text{(16b) holds}, t>\tau_2\}\geqslant\frac{1}{2}e^{\frac{3\Delta m_i^2}{2\sigma_0^2}}t^{-\frac{9}{8}},\quad \tau_2=e^{\frac{\Delta m_i^2}{3\sigma_0^2}}.\end{aligned}$$

This leads to

$$\begin{aligned}&\mathbb{E}[N_i(T)]\leqslant\eta_i+\tau_1+\tau_2+\frac{1}{2}\left(e^{\frac{3\Delta m_{i^*}^2}{2\sigma_0^2}}+e^{\frac{3\Delta m_i^2}{2\sigma_0^2}}\right)\sum_{t=1}^{T}t^{-\frac{9}{8}}\\ &\leqslant\lceil\frac{4\sigma_0^2}{\Delta_i^2}(\log 2\pi e+4\log T)-1\rceil+e^{\frac{\Delta m_{i^*}^2}{3\sigma_0^2}}+e^{\frac{\Delta m_i^2}{3\sigma_0^2}}+\\ &\qquad\frac{9}{2}e^{\frac{3\Delta m_{i^*}^2}{2\sigma_0^2}}+\frac{9}{2}e^{\frac{3\Delta m_i^2}{2\sigma_0^2}},\end{aligned}$$

which completes the proof.

Appendix D
Proof of Theorem 4

According to the Lemma 1 in [16], the utility function Q_i^{UiPA} can be written as

$$Q_i^{UiPA}(t)\leqslant\bar{r}_i(t)+U_i(t),$$

with

$$U_i(t)\doteq\sqrt{\frac{\sum_{\tau=1}^{t}r_i^2(\tau)-\bar{r}_i^2(t)N_i(t)}{(N_i(t)-1)N_i(t)}(\log 2\pi e+4\log t)}.$$

Then, we can use [13, Theorem 4] to bound the expected loss. We have (24), shown at the top of the next page, for all $N_i(t)\geqslant\log 2\pi e/2+2\log t$. Furthermore, $\mathbb{P}\{\hat{\mu}_{i^*}(t)\geqslant$

$$\mathbb{P}\{\hat{\mu}_i(t) \geqslant \mu_i + U_i(t)\} = \mathbb{P}\left\{ \frac{\hat{\mu}_i(t) - \mu_i}{\sqrt{(\sum_{\tau=1}^{t} r_i^2(\tau) - \bar{r}_i^2(t)N_i(t))/(N_i(t)(N_i(t)-1))}} \geqslant \sqrt{\log 2\pi e + 4\log t} \right\} \leqslant 1/\sqrt{2\pi e} t^{-2} \quad (24)$$

$\mu_{i^*} + U_{i^*}(t)\}$ can be similarly bounded. Lastly, using the Chi-squared distribution, we have

$$\begin{aligned}
&\mathbb{P}\{\mu_{i^*} < \mu_i + 2U_i(t)\} = \\
&\mathbb{P}\left\{ \frac{\sum_{\tau=1}^{t} r_i^2(\tau) - \bar{r}_i^2(t)N_i(t)}{\sigma_0^2} > \frac{(N_i(t)-1)\Delta_i^2 N_i(t)}{4\sigma_0^2(\log 2\pi e + 4\log t)} \right\} \\
&\leqslant \mathbb{P}\left\{ \frac{\sum_{\tau=1}^{t} r_i^2(\tau) - \bar{r}_i^2(t)N_i(t)}{\sigma_0^2} > 4(N_i(t)-1) \right\} \\
&\leqslant e^{-N_i(t)/2} \\
&\leqslant (2\pi e)^{-1/4} t^{-1},
\end{aligned} \quad (25)$$

and

$$N_i(t) \geqslant \max\left\{ \frac{16\sigma_0^2}{\Delta_i^2}, \frac{1}{2} \right\} (\log 2\pi e + 4\log T). \quad (26)$$

Combining (24)(25)(26), $N_i(t)$ can be bounded as

$$\begin{aligned}
N_i(T) &\leqslant \max\left\{ \frac{16\sigma_0^2}{\Delta_i^2}, \frac{1}{2} \right\} (\log 2\pi e + 4\log T) + \\
&\sum_{t=1}^{T} \left(2/\sqrt{2\pi e} t^{-2} + (2\pi e)^{-1/4} t^{-1} \right) \\
&\leqslant \frac{16\sigma_0^2}{\Delta_i^2} (\log 2\pi e + 4\log T) + \\
&((2\pi e)^{-1/4} + 2)\log T + \frac{\log 2\pi e}{2} + \frac{2}{\sqrt{2\pi e}}.
\end{aligned}$$

This completes the proof.

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020," February 2016.

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

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

[4] O. Aliu, A. Imran, M. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 336–361, 2013.

[5] Small Cell Forum, "Interference management in UMTS femtocells: topic brief," February 2014.

[6] 3GPP, "FDD Home NodeB RF Requirements," TR 25.967 v9.0.0.

[7] S. Nagaraja *et al.*, "Downlink transmit power calibration for enterprise femtocells," in *IEEE VTC*, 2011.

[8] D. López-Pérez, X. Chu, A. Vasilakos, and H. Claussen, "Minimising cell transmit power: Towards self-organized resource allocation in OFDMA femtocells," in *ACM SIGCOMM*, August 2011, pp. 410–411.

[9] S. Nagaraja *et al.*, "Transmit power self-calibration for residential UMTS/HSPA+ femtocells," in *IEEE WiOpt*, May 2011, pp. 451–455.

[10] J. Kim, P.-Y. Kong, N.-O. Song, J.-K. K. Rhee, and S. Al-Araji, "MDP based dynamic base station management for power conservation in self-organizing networks," in *IEEE WCNC*, April 2014, pp. 2384–2389.

[11] 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, July 2013.

[12] 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, pp. 1–122, 2012.

[13] 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.

[14] R. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem," in *NIPS*, 2004.

[15] P. Reverdy, V. Srivastava, and N. Leonard, "Modeling human decision-making in generalized Gaussian multi-armed bandits," *Proceedings of the IEEE*, pp. 1–23, February 2014.

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

[17] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in *International Conference on Artificial Intelligence and Statistics*, April 2012, pp. 592–600.

[18] H. Claussen, L. T. W. Ho, and L. G. Samuel, "Self-optimization of coverage for femtocell deployments," in *Wireless Telecommunications Symposium*, April 2008, pp. 278–285.

[19] 3GPP, "User Equipment (UE) procedures in idle mode and procedures for cell reselection in connected mode," TR 25.304.

[20] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[21] 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, October 1988.

[22] D. Knuth, *The Art of Computer Programming*. Addison-Wesley, 1997.

[23] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications*, vol. 36, pp. 3336–3341, October 2009.

[24] 3GPP, "Evolved Universal Terrestrial Radio Access; Further advancements for E-UTRA physical layer aspects," TR 36.814.

[25] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876 – 1902, 2009.

[26] M. Valko, R. Munos, B. Kveton, and T. Kocák, "Spectral bandits for smooth graph functions," in *ICML*, 2014, pp. 46–54.