A Bandit Approach to Price-Aware Energy Management in Cellular Networks

Xinruo Zhang, Mohammad Reza Nakhai and Wan Nur Suryani Firuz Wan Ariffin

***Abstract*—We introduce a reinforcement learning algorithm inspired by the combinatorial multi-armed bandit problem to minimize the time-averaged energy cost at individual base stations (BSs), powered by various energy markets and local renewable energy sources, over a finite time horizon. The algorithm sustains traffic demands by enabling sparse beamforming to schedule dynamic user-to-BS allocation and proactive energy provisioning at BSs to make ahead-of-time price-aware energy management decisions. Simulation results indicate a superior performance of the proposed algorithm in reducing the overall energy cost, as compared with recently proposed cooperative energy management designs.**

***Index Terms*—Energy management; CMAB; Online Learning;**

I. INTRODUCTION

The rapid rise in energy consumption due to the future ultra-dense deployment of base stations (BSs) to support fast-growing wireless traffic will increase the cost of electricity and may result in the revenue saturation of the network operator [1]. Thereby, powering the BSs with green energy generated from environmental sources has been regarded as a promising technology for the next generation cellular networks. However, the renewable energy generation is naturally uncertain and irregular. Realizing these features and making an impact in sustaining the reliable operation of wireless networks, [2], [3] propose the integration of local renewable energy sources with two-way energy trading via accessing various energy markets. In [2], two-way energy trading between the BSs' in a coordinated multipoint (CoMP) system and the grid is studied based on convex optimization techniques and concluded that the joint management of energy trading by fully cooperative BSs reduces the total energy cost. Partial cooperation based on sparse beamforming is proposed in [3] to account for limited-capacity backhaul links connecting the central processor (CP) and BSs in CoMP systems, whilst two-way energy trading with the grid is performed. In [4], an energy-efficient resource allocation approach which is based on cross-tier interference reduction rather than energy trading is introduced for two-tier macrocell/femtocell networks. However, none of these studies considered the impact of online learning in proactive energy management in wireless networks or provided adaption to the dynamic wireless channel conditions. The authors in [5] first introduce the application of combinatorial multi-armed bandit (CMAB) as an online learning approach to energy management in a simplified network scenario, where wireless channel dynamic is relaxed, the exploration is in single direction and a full exploration CMAB algorithm without an efficient trade-off strategy between the exploration and the exploitation is proposed. In this paper, we account for the wireless channel random dynamism and develop a new CMAB-based online learning algorithm that benefits from an efficient trade-off between the exploration (i.e., online training or learning) and the exploitation (i.e., operational) modes with the goal of minimizing the overall energy cost over a finite time horizon. This goal is achieved by anticipating the amount of energy demand ahead-of-time and purchasing it at a lower rate in the exploration mode and using this purchased energy in the following exploitation mode, so that the spot market energy provisioning at higher rate is minimized.

The authors are with the Centre for Telecommunications Research, Department of Informatics, King's College London, Strand, WC2R 2LS, UK. E-mail:[xinruo.zhang, reza.nakhai, k1206546]@kcl.ac.uk

II. SYSTEM MODEL

Consider a cluster-based CoMP network in the downlink where a set of N BSs partially collaborate to serve K_i user terminals (UTs) over a shared bandwidth. Each BS is equipped with M antennas, whereas each user has a single receiving antenna. Let $\mathcal{L}_b = \{1, \cdots, N\}$ and $\mathcal{L}_i = \{1, \cdots, K_i\}$ denote, respectively, the set of indexes of the BSs and the UTs within a cluster. The CP coordinates all strategies based on perfect knowledge of channel state information and distributes all UTs' data to the corresponding BSs via finite-capacity backhaul links. The energy transmission between the grid and BSs is accomplished via dedicated power lines. The finite time horizon is divided into T discrete time slots indexed as $\mathcal{T} = \{1, \cdots, T\}$.

A. Energy Management Model

At the end of an exploration mode, an amount of $E_n^{[\mathrm{a}]}$ units of energy that can be sustained uniformly over a number of following exploitation time slots is purchased ahead-of-time for the n-th BS, $n \in \mathcal{L}_b$, at a price rate of $\pi^{[\mathrm{a}]}$. Let $E_n^{[\mathrm{a}]}(t)$ denote the ahead-of-time purchased energy allocated to the current time slot t. Let $E_n^{[\mathrm{s}]}(t)$ be the amount of real-time energy required to be purchased at time slot t due to both insufficient $E_n^{[\mathrm{a}]}(t)$ and the available renewable energy $G_n(t)$ at the n-th BS. Note that $E_n^{[\mathrm{s}]}(t)$ should be purchased from the spot market at a higher price rate of $\pi^{[\mathrm{s}]}$, whereas $G_n(t)$ can be obtained locally at much lower rate $\pi^{[\mathrm{g}]}$. The surplus of available energy to a BS, i.e., $S_n(t)$, can be sold back to the grid at a fair rate of $\pi^{[\mathrm{e}]}$. The total energy cost incurred by the n-th BS at the t-th time slot can be written as

$$C_n^{[\mathrm{total}]}(t) = \pi^{[\mathrm{s}]} E_n^{[\mathrm{s}]}(t) + \pi^{[\mathrm{a}]} E_n^{[\mathrm{a}]}(t) + \pi^{[\mathrm{g}]} G_n(t) - \pi^{[\mathrm{e}]} S_n(t). \quad (1)$$

B. Downlink Transmission Model

Let $\mathbf{w}_{ni} \in \mathbb{C}^{M\times 1}$ and $\mathbf{h}_{ni} \in \mathbb{C}^{M\times 1}, n \in \mathcal{L}_b, i \in \mathcal{L}_i$ denote the beamformer and the channel vector from the n-th BS towards the i-th UT, respectively. The signal-to-interference-plus-noise ratio (SINR) at the i-th UT, $i \in \mathcal{L}_i$, is defined as $\text{SINR}_i = \frac{|\sum_{n\in\mathcal{L}_b}(\mathbf{h}_{ni}^H\mathbf{w}_{ni})|^2}{\sum_{j\neq i, j\in\mathcal{L}_i}|\sum_{n\in\mathcal{L}_b}(\mathbf{h}_{ni}^H\mathbf{w}_{nj})|^2+\sigma_i^2}$, where σ_i^2 is the zero-mean circularly symmetric complex Gaussian noise variance. The n-th BS's backhaul capacity consumption is given by

$$B_n^{[\text{backhaul}]} = \sum_{i\in\mathcal{L}_i} \left\| \|\mathbf{w}_{ni}\|_2^2 \right\|_0 R_i, \quad \forall n \in \mathcal{L}_b, \tag{2}$$

where $R_i = \log_2(1+\text{SINR}_i)$ is the achievable data rate (bit/s/Hz) for the i-th UT. The binary indicator function $\left\| \|\mathbf{w}_{ni}\|_2^2 \right\|_0$ illustrates the scheduling choices between the i-th UT and the n-th BS, where $\|\mathbf{w}_{ni}\|_2^2 = 0$ implies that the backhaul link between the CP and the n-th BS is not used for coordinated transmission to the i-th UT.

III. PRICE-AWARE ENERGY MANAGEMENT

As per (1), the total energy cost at the t-th time slot, $\forall t \in \mathcal{T}$, depends on a linear combination of the real-time trading variables, i.e., $E_n^{[\text{s}]}(t)$ and $S_n(t)$, and the ahead-of-time energy purchase, i.e., $E_n^{[\text{a}]}(t)$, given an available amount of renewable energy $G_n(t)$. We aim to minimize the total average energy cost over a finite time horizon via an online-learning assisted convex optimization. The downlink beamformers and the real-time trading parameters, i.e., $E_n^{[\text{s}]}(t)$ and $S_n(t)$, are the variables of the optimization problem. The ahead-of-time energy purchase $E_n^{[\text{a}]}(t)$ is the learning parameter which is proactively determined by the proposed online learning strategy and fedback to the optimization problem. The convex optimization problem is formulated in the current Section and will then be integrated with the online learning strategy, introduced in Section IV, under Algorithm 2.

A. Problem Formulation

In order to minimize the energy cost at each time slot t, the optimization problem is formulated as

$$\min_{\mathbf{w}_{ni}, E_n^{[\text{s}]}(t), S_n(t)} \sum_{n\in\mathcal{L}_b} P_n^{[\text{Tx}]}(t) + \sum_{n\in\mathcal{L}_b} \left\{ E_n^{[\text{s}]}(t) \right\} \tag{3}$$

$$\begin{aligned}
\text{s.t.} \quad & \text{C1}: \text{SINR}_i(t) \geq \gamma_i, \quad \forall i \in \mathcal{L}_i, \\
& \text{C2}: B_n^{[\text{backhaul}]}(t) \leq B_n^{[\text{limit}]}, \quad \forall n \in \mathcal{L}_b, \\
& \text{C3}: P_n^{[\text{Tx}]}(t) + P_n^{[\text{c}]} \leq G_n(t) + E_n^{[\text{a}]}(t) - S_n(t) \\
& \qquad\qquad + E_n^{[\text{s}]}(t), \ \forall n \in \mathcal{L}_b, \\
& \text{C4}: P_n^{[\text{Tx}]}(t) \leq P_n^{[\text{Tmax}]}, \quad \forall n \in \mathcal{L}_b, \\
& \text{C5}: E_n^{[\text{s}]}(t) \geq 0, \qquad \text{C6}: S_n(t) \geq 0, \quad \forall n \in \mathcal{L}_b,
\end{aligned}$$

where $P_n^{[\text{Tx}]}(t) = \sum_{i\in\mathcal{L}_i} ||\mathbf{w}_{ni}||_2^2$ is the total transmit power of the n-th BS at the t-th time slot. C1 indicates the SINR constraint γ_i for the i-th UT and C2 represents the backhaul link capacity restriction, i.e., $B_n^{[\text{limit}]}$, for each BS. C3 emphasises that the individual BS's energy consumption is upper bounded by its energy budget, i.e., $G_n(t)$, $E_n^{[\text{a}]}(t)$, $E_n^{[\text{s}]}(t)$ and $S_n(t)$, where $P_n^{[\text{c}]}$ is the n-th BS's hardware circuit power consumption at the t-th time slot. C4 specifies maximum transmit power, i.e., $P_n^{[\text{Tmax}]}$, at the n-th BS. C5 and C6 indicate, respectively, that the spot market energy provisioning and the excessive energy to be sold back are non-negative.

B. Reweighted ℓ_1-norm and semidefinite programming (SDP)

The intractable constraint C2 in (3) that formulates the sparse beamforming problem as ℓ_0-norm, is handled with *reweighted ℓ_1-norm method* [3], as $B_n^{[\text{backhaul}]}(t) \approx \sum_{i\in\mathcal{L}_i} \left\| [\xi_{ni}\|\mathbf{w}_{ni}\|_2^2] \right\|_1 R_i = \sum_{i\in\mathcal{L}_i} \xi_{ni}\text{tr}(\mathbf{w}_{ni}\mathbf{w}_{ni}^H)R_i$, where the cooperative links between the BSs and the UTs are iteratively removed via alternating between solving optimal beamformer $\mathbf{w}_{ni}^*$ of problem (3) for a given ξ_{ni}, and adjusting the weight $\xi_{ni} = \frac{1}{\text{tr}(\mathbf{w}_{ni}^*\mathbf{w}_{ni}^{*H})+\mu}$ as per backhaul link capacity constraints and the power budgets at the individual BSs. Defining $\mathbf{H}_{ni} = \mathbf{h}_{ni}\mathbf{h}_{ni}^H$ and semidefinite matrices $\mathbf{W}_{ni} = \mathbf{w}_{ni}\mathbf{w}_{ni}^H$, the original problem in (3) can then be transformed to a SDP problem after relaxing the rank-one constraint of $\text{rank}(\mathbf{W}_{ni}) = 1$, as

$$\min_{\substack{\mathbf{W}_{ni}\succeq 0, \\ E_n^{[\text{s}]}(t), S_n(t)}} \sum_{n\in\mathcal{L}_b}\sum_{i\in\mathcal{L}_i} \text{tr}(\mathbf{W}_{ni}) + \sum_{n\in\mathcal{L}_b} \left\{ E_n^{[\text{s}]}(t) \right\} \tag{4}$$

$$\begin{aligned}
\text{s.t.} \quad & \text{C1}: \gamma_i^{-1}\text{tr}\Big(\sum_{n\in\mathcal{L}_b} \mathbf{H}_{ni}\mathbf{W}_{ni}\Big) \geq \\
& \qquad \sum_{j\in\mathcal{L}_i, j\neq i} \text{tr}\Big(\sum_{n\in\mathcal{L}_b} \mathbf{H}_{ni}\mathbf{W}_{nj}\Big) + \sigma_i^2, \forall i \in \mathcal{L}_i, \\
& \text{C2}: \sum_{i\in\mathcal{L}_i} \xi_{ni}\text{tr}(\mathbf{W}_{ni})R_i \leq B_n^{[\text{limit}]}, \ \forall n \in \mathcal{L}_b, \\
& \text{C3}: \sum_{i\in\mathcal{L}_i} \text{tr}(\mathbf{W}_{ni}) \leq G_n(t) + E_n^{[\text{a}]}(t) - P_n^{[\text{c}]} \\
& \qquad\qquad - S_n(t) + E_n^{[\text{s}]}(t), \ \forall n \in \mathcal{L}_b, \\
& \text{C4}: \sum_{i\in\mathcal{L}_i} \text{tr}(\mathbf{W}_{ni}) \leq P_n^{[\text{Tmax}]}, \forall n \in \mathcal{L}_b, \\
& \text{C5}: E_n^{[\text{s}]}(t) \geq 0, \qquad \text{C6}: S_n(t) \geq 0, \quad \forall n \in \mathcal{L}_b.
\end{aligned}$$

IV. PROACTIVE ENERGY MANAGEMENT

Due to the combinatorial nature of distributed energy transmission from the grid to the BSs, the price-aware energy management problem studied in this paper is classified as CMAB problem. The CMAB problem is defined as a system consists of J possible arms, where N arms, $N \subset J$, that form a super arm are played simultaneously and the reward of each arm is observed individually at each trial [6]. The objective is to maximize the long-term accumulated reward via a trade-off between observing the reward of new super arms, known as exploration, and proactively selecting the best-possible super arm for future time slots based on existing knowledge from the previous time slots, known as exploitation. In this paper, each arm corresponds to a discrete ahead-of-time energy package to be selected for a BS and the reward of each arm corresponds to the difference between the energy cost at the t-th time slot and at the initial time slot. Thus, maximizing the accumulated reward is equivalent to minimizing the time-averaged energy cost. Let $\mathcal{K} = \{1, \cdots, K\}$ denote the set of

indexes of the learning trials within a exploration time slot, $\mathcal{J} = \{1, \cdots, J\}$ be the set of indexes associated to J arms, i.e., J ahead-of-time energy packages $\{\mathcal{E}^1, \cdots, \mathcal{E}^J\}$ offered by the grid, where $\mathcal{E}^e = \mathcal{E}^{e-1} + \Delta\mathcal{E}, e \in \mathcal{J}$. At the k-th trial, $k \in \mathcal{K}$, the CP selects a super arm, i.e., N ahead-of-time energy packages for N BSs, for next time slot, denoted by $\mathcal{S}^{[\text{set}]}(k) = \{E_1^{[\text{a}]}(k), \cdots, E_N^{[\text{a}]}(k)\}$. Let the individual reward of the arm $E_n^{[\text{a}]}(k)$ at the k-th trial be defined as

$$\mathcal{R}(E_n^{[\text{a}]}(k)) = C_n^{[\text{total}]}(0) - C_n^{[\text{total}]}(k), \ \forall n \in \mathcal{L}_b, \tag{5}$$

where $C_n^{[\text{total}]}(0)$ and $C_n^{[\text{total}]}(k)$ are the total energy cost of the n-th BS at the initial trial of the initial time slot and the k-th trial of the current time slot, respectively, as per (1). Let $\mathbf{r_n^{[k,t]}} = (r_{n,1}^{[k,t]}, r_{n,2}^{[k,t]}, \cdots, r_{n,J}^{[k,t]})$ be defined as the reward vector of the n-th BS, where $r_{n,e}^{[k,t]}, e \in \mathcal{J}$, is the reward associated to the e-th ahead-of-time energy package in the k-th trial at the t-th time slot averaged over F independent channel realizations. Also let $\hat{\mathbf{r}}_\mathbf{n}^{[\mathbf{t}]} = (\hat{r}_{n,1}^{[t]}, \hat{r}_{n,2}^{[t]}, \ldots, \hat{r}_{n,J}^{[t]})$ and $\bar{\mathbf{r}}_\mathbf{n}^{[\mathbf{t}]} = (\bar{r}_{n,1}^{[t]}, \bar{r}_{n,2}^{[t]}, \ldots, \bar{r}_{n,J}^{[t]})$ denote mean reward vector and adjusted reward vector of individual ahead-of-time energy packages for the n-th BS at the t-th time slot, respectively.

In the sequel, we introduce an online learning algorithm, detailed in Algorithm 1 and 2, to minimize the total energy cost over a finite time horizon. Similar to [7], the proposed algorithm enables smart scheduling that linearly increases the number of exploration with an exponentially increased number of time slots, as presented in Fig. 1 and Table 1, which reduces the exploration overhead in terms of total energy cost over a finite time horizon. The time horizon of T time slots is divided into P periods of increased length growing at a geometric progression, i.e., $T = 2(2^P - 1)$. Let $\mathcal{P} = \{1, \ldots, P\}$ denote the set of indexes of periods. In the p-th period that contains 2^p time slots, $p \in \mathcal{P}$, a total number of p time slots will be randomly selected as exploration whilst the rest time slots are reserved for exploitation. Since the estimation of the super arms' mean reward process is improved for a larger period index, the principle is to reduce the fraction of time slots being selected as exploration with increasing period index.

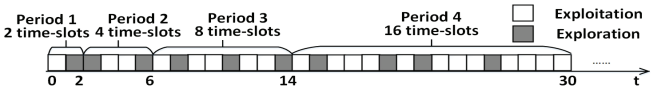


Fig. 1. An exploration-exploitation trade-off model of smart scheduling

TABLE I
PERCENTAGE OF EXPLORATION USING SMART SCHEDULING

Period index	1	2	3	4	5
No. of time slot	2	$2^2 = 4$	$2^3 = 8$	$2^4 = 16$	$2^5 = 32$
No. of exploration	1	2	3	4	5
% of exploration	0.50	0.50	0.429	0.333	0.242

In the exploration mode, Algorithm 1 explores new super arm, i.e., new combination of ahead-of-time energy packages for N BSs, in a two directional way. More specifically, the exploring direction among all possible arms, i.e., forward or backward exploration, will be initially determined as described in steps 9 and 11 of Algorithm 1, respectively, based on the rewards obtained at the current and the previous trials, followed by the super arm exploration for the next trial. The proposed Algorithm 1 guarantees that the individual BSs search in the proper direction towards the optimal arm that associated with the highest reward. Once a given number of K trials are completed, the mean reward for individual energy packages, i.e., $\hat{\mathbf{r}}_\mathbf{n}^{[\mathbf{t}]}$, for the n-th BS at the t-th time slot are estimated and adjusted within a controlled percentage, i.e., $\alpha\hat{\mathbf{r}}_\mathbf{n}^{[\mathbf{t}]}$, respectively, as per step 8 and 9 in Algorithm 2. The adjusted rewards, i.e., $\bar{\mathbf{r}}_\mathbf{n}^{[\mathbf{t}]}$, are first, averaged over all past time slots as per step 13, and then, used to update the index of optimal N arms, to be exploited in the next time slot, as detailed in step 14 and step 4 in Algorithm 2.

Algorithm 1 *Two Directional Super Arm Exploration*

1: **For** $k = 1 : K$
2: Solve problem in (4),
3: Compute $C_n^{[\text{total}]}(k)$ as per (1) and $\mathcal{R}(E_n^{[\text{a}]}(k))$ as per (5),
4: **if** $k = 1$ (initial trial) and $E_n^{[\text{a}]}(k) \neq \mathcal{E}^1$
5: **then** $E_n^{[\text{a}]}(k+1) = E_n^{[\text{a}]}(k) - \Delta\mathcal{E}$,
6: **else if** $k = 1$ (initial trial) and $E_n^{[\text{a}]}(k) = \mathcal{E}^1$
7: **then** $E_n^{[\text{a}]}(k+1) = E_n^{[\text{a}]}(k) + \Delta\mathcal{E}$,
8: **else if** $\mathcal{R}(E_n^{[\text{a}]}(k)) > \mathcal{R}(E_n^{[\text{a}]}(k-1))$,
9: **then** Do **Backward Exploration**,
$E_n^{[\text{a}]}(k+1) = E_n^{[\text{a}]}(k) - \Delta\mathcal{E}$,
10: **else if** $\mathcal{R}(E_n^{[\text{a}]}(k)) < \mathcal{R}(E_n^{[\text{a}]}(k-1))$,
11: **then** Do **Forward Exploration**,
$E_n^{[\text{a}]}(k+1) = E_n^{[\text{a}]}(k) + \Delta\mathcal{E}$,
12: **else** $E_n^{[\text{a}]}(k+1) = E_n^{[\text{a}]}(k), \quad \forall n \in \mathcal{L}_b$,
13: **end if**
14: Compute energy package index as $e = \frac{E_n^{[\text{a}]}(k)}{\Delta\mathcal{E}}, \ n \in \mathcal{L}_b$,
15: Update $r_{n,e}^{[k,t]} = \mathcal{R}(E_n^{[\text{a}]}(k)), \ \forall e \in \mathcal{J}, n \in \mathcal{L}_b$,
16: Update $\mathcal{S}^{[\text{set}]}(k+1) = \{E_1^{[\text{a}]}(k+1), \cdots, E_N^{[\text{a}]}(k+1)\}$.
17: **End for**

V. SIMULATION RESULTS

Consider a downlink system comprises 3 neighbouring 8-antennas BSs with a BS-BS distance of 500 m, transmitting toward 6 single-antenna UTs. The renewable energy supply at individual BSs at each time slot are $G_1 = 1.5$ W, $G_2 = 0.2$ W and $G_3 = 0.05$ W, at a price of $\pi^{[\text{g}]} = £0.05$/W. The other simulation parameters are $\pi^{[\text{a}]} = £0.07$/W, $\pi^{[\text{s}]} = £0.15$/W, $\pi^{[\text{e}]} = £0.02$/W, $\alpha = 0.5$, $P_n^{[\text{c}]} = 30$ dBm, $P_n^{[\text{Tmax}]} = 46$ dBm and $B_n^{[\text{limit}]} = 35$ bits/s/Hz. The performance of the proposed strategy is evaluated with $K = 5$ learning trials averaging over $F = 20$ independent channel realizations for each time slot, $T = 60$ time slots and $J = 30$ possible ahead-of-time energy packages with $\Delta\mathcal{E}$=100 mW, i.e., $\{100, 200, \cdots, 3000\}$ mW.

Fig. 2 compares the normalized total energy cost at $\gamma = 15$ dB target SINR of our proposed strategy against 1) a baseline design in [2] that purchases no ahead-of-time energy, 2) a non-learning design in [3] that always purchases fixed set of ahead-of-time energy packages, i.e., $E_1^{[\text{a}]} = E_2^{[\text{a}]} = E_3^{[\text{a}]} = 700$ mW, 3) a simplified CMAB design in [5] that performs

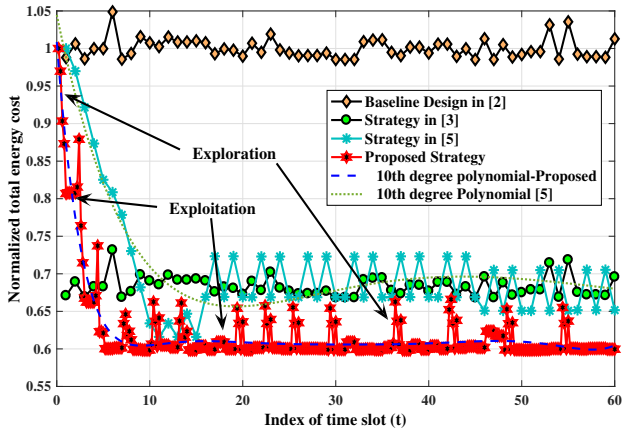
Algorithm 2 *Online Learning Main Algorithm*

1: **For** $t = 1 : T$
2: **if** $t = 1$ (initial time slot)
3: **then** Initialize super arm as $\mathcal{S}^{[\text{set}]}(1) = \{0_1, \cdots, 0_N\}$,
4: **else** Update optimal super arm as
$$\mathcal{S}^{[\text{set}]}(1)^* = \Delta\mathcal{E}[e_1^*, e_2^*, \cdots, e_N^*],$$
5: **end if**
6: **if** t is selected for **Exploration** mode
7: **then** Run Algorithm 1,
8: **Estimation Stage :**
Compute mean reward $\hat{\mathbf{r}}_{\mathbf{n}}^{[\mathbf{t}]} = (\hat{r}_{n,1}^{[t]}, \hat{r}_{n,2}^{[t]}, \ldots, \hat{r}_{n,\mathcal{J}}^{[t]})$,
where $\hat{r}_{n,e}^{[t]} = \frac{\sum_{k=1}^{K} r_{n,e}^{[k,t]}}{K}$, $\forall e \in \mathcal{J}, n \in \mathcal{L}_b$,
9: **Adjustment Stage :**
Adjust $\bar{r}_{n,e}^{[t]} = \hat{r}_{n,e}^{[t]} + [\alpha \hat{r}_{n,e}^{[t]}, \sqrt{\frac{3\ln t}{2\Psi_e}}]^-, \forall e \in \mathcal{J}, n \in \mathcal{L}_b$,
where Ψ_e is number of times the e-th arm has been played,
10: **else if** t is selected for **Exploitation** mode
11: Solve problem in (4),
12: **end if**
13: Average $\bar{\mathbf{r}}_{\mathbf{n}}^{[\mathbf{t}]}$ over accumulated number of time slots, as
$\bar{\mathbf{r}}_{\mathbf{n}} = \frac{\sum_{t'=1}^{t} \bar{\mathbf{r}}_{\mathbf{n}}^{[\mathbf{t}']}}{t} = [\bar{r}_{n,1}, \bar{r}_{n,2}, \cdots, \bar{r}_{n,J}], n \in \mathcal{L}_b$,
14: For the next time slot: find N optimum arm indexes as
$$e_n^* = \underset{e}{\text{argmax}}(\bar{r}_{n,e}), e \in \mathcal{J}, \forall n \in \mathcal{L}_b.$$
15: **End for**

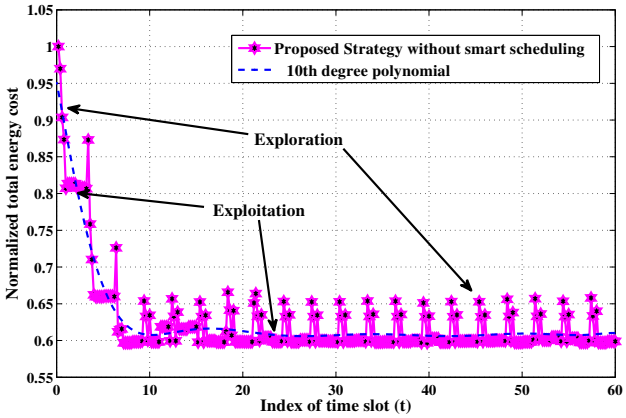
only single directional exploration mode and 4) the proposed strategy without smart scheduling. The total energy cost is normalized with respect to the initial value in the first time slot of the proposed strategy. The bursts and the smooth parts in Fig. 2, respectively, corresponds to the exploration mode and the exploitation mode. Note that the sharp jump in the beginning of an exploration is due to the adjustment stage via perturbation in step 9 of Algorithm 2 that prioritizes the least selected arms for the initial trial of exploration. The fitted 10th-degree-polynomial trend curve to the results of the proposed strategy shows an improvement of approximately 40 percent over the initial state of the system from the 7th time-slot onwards. This is due to reducing significantly the real-time energy cost by ahead-of-time preparation for the future (i.e., real-time) energy demands at lower costs. Furthermore, an average percentage improvement of approximately 40, 8 and 7 per cent can be achieved by the proposed strategy as compared with [2], [3] and [5], respectively, due to the fact that their designs provide no adaption to the time-varying wireless channel conditions. The performance of the proposed strategy without smart scheduling is illustrated in Fig. 2(b), where a fixed trade-off between exploration and exploitation mode is adopted. The trend curve fitted to the results in Fig. 2(b) oscillates around the normalized energy cost of 0.61, as compared to 0.6 of the proposed strategy in Fig. 2(a). This difference in energy cost is due to the fact that the proposed smart-scheduling-enabled strategy reduces the number of high-energy-cost exploration with increasing number of time slots.

VI. CONCLUSION

This paper proposes a CMAB approach to proactive price-aware energy management in cellular network, which adapts to dynamic wireless channel conditions and minimizes the overall energy cost over a finite time horizon. Simulation results confirm that in terms of cost-efficient energy provisioning at BSs, an average performance percentage improvement of 40, 8 and 7 per cent can be achieved by the proposed strategy as compared with three recently proposed designs.



(a) Proposed strategy versus other designs



(b) Proposed strategy without smart scheduling

Fig. 2. Normalized total energy cost at individual time slots at $\gamma = 15$ dB

REFERENCES

[1] A. Fehske and *et al.*, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 55–62, Aug. 2011.

[2] J. Xu and R. Zhang, "Cooperative energy trading in CoMP systems powered by smart grids," *IEEE GLOBECOM*, pp. 2697–2702, Dec. 2014.

[3] W. N. S. F. W. Ariffin and *et al.*, "Sparse beamforming for real-time energy trading in CoMP-SWIPT networks," *IEEE ICC*, May 2016.

[4] R. Estrada and *et al.*, "Energy-efficient resource-allocation model for OFDMA macrocell/femtocell networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3429–3437, Apr. 2013.

[5] W. N. S. F. W. Ariffin, X. Zhang, and M. R. Nakhai, "Combinatorial multi-armed bandit algorithms for real-time energy trading in green C-RAN," *IEEE ICC*, May 2016.

[6] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework, results and applications," *International Conference on Machine Learning*, Jun. 2013.

[7] S. Maghsudi and S. Stanczak, "Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1309–1322, Mar. 2015.