Distributed Foresighted Energy Management in Smart-Grid-Powered Cellular Networks

Xinruo Zhang, Mohammad Reza Nakhai, Gan Zheng, Sangarapillai Lambotharan, and Jonathon A. Chambers

***Abstract*—This paper studies energy management in a smart grid-powered cellular network consisting of an independent system operator (ISO) and multiple geographically distributed aggregators. The aggregators have energy storage devices and can purchase energy from the electric grid via the ISO to serve their users. To account for the uncertainty of the renewable energy supply as well as the impacts of multiple aggregators on the electric grid and energy prices, a foresighted strategy combined with the adaptive ϵ-greedy method is developed for the aggregators to distributively and adaptively minimize the long-term overall cost of the system, based on the ahead-of-time decision making of the storage pre-charging amount. Simulation results validate that the proposed strategy surpasses a recent learning-based storage management design and a myopic design.**

***Index Terms*—Distributed management, energy management, online learning.**

I. Introduction

The enormous energy consumption in the next generation of dense wireless networks has been considered as one of the most challenging issues from both technical and economic perspectives. The integration of renewable energy generated from natural sources with the conventional electric grid is viewed as a promising approach towards achieving the targets of green communications and is implemented in the smart grid infrastructure [1]. The stochastic nature of renewable energy, nevertheless, causes significant uncertainty in energy generation and fluctuations of the electricity price. Such uncertainty could lead to abrupt ramping in the power plants or the adoption of peaker plants that use non-renewable sources of energy to compensate for random variations in energy generation, which is one of the most expensive operations or even may not be technically feasible for some energy generators [2]. Hence, efficient new control mechanisms are advocated to improve the flexibility and robustness of smart grid networks, so that the independent system operator (ISO) can maintain its reliable and cost-efficient operation under such variability. Furthermore, the deployment of energy storage units in the demand side has also been proposed as a viable solution to tackle these concerns [2], as it can not only compensate for the real-time energy shortage, but also be used to minimize the long-term energy consumption cost via being proactively pre-charged either from the power grid at a lower price or by the excessive renewable energy generation. Most approaches in the literature [3]–[5] propose myopic demand side energy management designs based on an optimization method and seek for instantaneous/short-term cost minimization without taking into account either the deployment of energy storage units or the possibility of learning the system dynamics during the operation. The authors in [6] concentrate on the interaction between a single aggregator and customers and develop an online learning algorithm for stochastic storage management based on the Markov decision process. Using stochastic optimization rather than online learning, the authors in [7] propose a dynamic energy management design for the smart-grid-powered coordinated multipoint system. However, the stochastic optimization based approaches usually require a model or the statistics of the system dynamics to be known upfront. Furthermore, [6]–[8] focus merely on a single aggregator, while the impacts of the decisions of the aggregators on the power network have been neglected.

Accounting for the intermittent renewable energy generation and the impacts of aggregators on the power network, this paper focuses on designing an online energy management strategy to minimize the long-term average energy consumption cost of the ISO in a decentralized manner. This is challenging because the ISO and the distributed aggregators that are in geographically different locations have no access to the local information of one another. Hence, the conventional bandit approaches typically designed for solving centralized problems can not tackle such a distributed problem with strictly limited signalling information. Furthermore, the statistics of the system dynamics such as renewable energy generation are unknown *a priori*, and the storage pre-charging decisions have strong temporal correlations, which render the problem intractable for traditional stochastic optimization based approaches. Hence, the novelty of this paper is the introduction of a distributed online foresighted energy management strategy that requires no upfront knowledge and can optimize the long-term average energy cost while learning via distributively alternating between two decision making processes. The first process minimizes the current cost at each aggregator based on the storage pre-charging amount and local user demand using convex optimization. The second process designs the ahead-of-time storage pre-charging strategies at distributed aggregators via online learning.

II. System Model

As illustrated in Fig. 1, this paper considers a smart grid powered cellular network with one ISO and a set of geographically distributed aggregators with central processing units (A-CPUs), indexed by $\mathcal{L}_a = \{1, \ldots, N\}$. Each aggregator is associated with a cloud radio access network (C-RAN) of N_b base stations (BSs) serving N_i users, indexed as $\mathcal{L}_i = \{1, \ldots, N_i\}$, using a shared spectrum. Featuring smart grid operations, the individual A-CPUs are deployed with on-site energy storage units and can exploit renewable energy sources with harvesting facilities. Furthermore, the A-CPUs have access to ancillary grid energy markets at various prices via energy trading mechanisms. The ISO operates the system through receiving energy purchase requests from the A-CPUs and dispatching the power grid based on its operational status. In particular, the ISO only has access to the status information of the grid as well as the energy purchase requests from the A-CPUs, whilst each A-CPU only has the status information of its associated C-RAN cluster and its storage units.

Manuscript received May 8, 2018; revised July 11, 2018 and January 21, 2019; accepted February 8, 2019. Date of publication February 18, 2019; date of current version April 16, 2019. This work was supported in part by the UK EPSRC under Grant EP/N007840/1 and in part by the Leverhulme Trust under Grant RPG-2017-129. The review of this paper was coordinated by Prof. Y. Guo. *(Corresponding author: Gan Zheng.)*

X. Zhang, G. Zheng, and S. Lambotharan are with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, LE11 3TU Loughborough, U.K. (e-mail: x.zhang@lboro.ac.uk; g.zheng@lboro.ac.uk; s.lambotharan@lboro.ac.uk).

M. R. Nakhai is with the Centre for Telecommunications Research, King's College London, WC2B 4BG London, U.K. (e-mail: reza.nakhai@kcl.ac.uk).

J. A. Chambers is with the Department of Engineering, University of Leicester, LE1 7RH Leicester, U.K. (e-mail: jonathon.chambers@leicester.ac.uk).

Digital Object Identifier 10.1109/TVT.2019.2899464

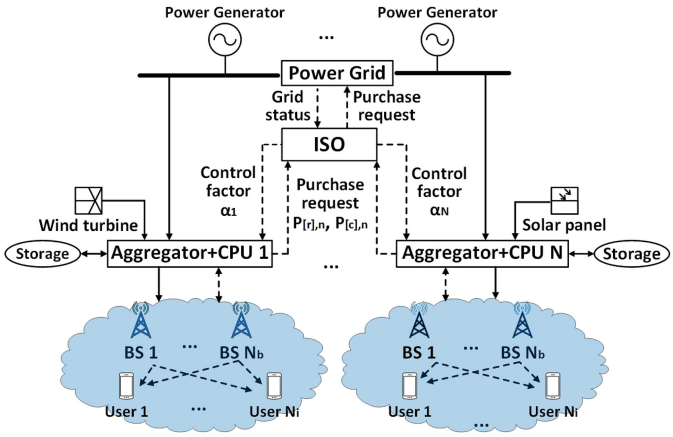


Fig. 1. Illustration of system scenario. The information flow is denoted by dashed lines and the energy flow is denoted by solid lines.

Let the time horizon T be divided into discrete time slots and indexed as $\mathcal{T} = \{1, \ldots, T\}$. Assume that the renewable energy generation varies across time slots but remains invariant within each time slot. For convenience, the slot duration is normalized to unity, thus the terms 'energy' and 'power' are used synonymously throughout the paper. Let $P^t_{[\mathrm{g}],n}$ and $P^t_{[\mathrm{r}],n}$ denote, respectively, the amount of renewable energy generation and the amount of energy shortage to be supplied by the grid in real-time at time slot t, $t \in \mathcal{T}$, to the n^{th} A-CPU. Let $P^t_{[\mathrm{s}],n}$ be the initial amount of energy stored in the storage device at time slot t and $P^t_{[\mathrm{c}],n}$ denote the amount of pre-charged energy stored in the n^{th} A-CPU prior to the actual time of energy demand at the beginning of the t^{th} time slot. The total energy consumption cost incurred by the n^{th} A-CPU at time slot t, i.e., C^t_n, can be described as

$$C^t_n = c^t_{[\mathrm{r}]} P^t_{[\mathrm{r}],n} + c^t_{[\mathrm{c}]} P^t_{[\mathrm{c}],n} + c^t_{[\mathrm{g}]} P^t_{[\mathrm{g}],n} + c^t_{[\mathrm{s}]} P^t_{[\mathrm{s}],n}, \tag{1}$$

where $c^t_{[\mathrm{r}]}$ and $c^t_{[\mathrm{c}]}$ are, respectively, the per unit energy prices for $P^t_{[\mathrm{r}],n}$ and $P^t_{[\mathrm{c}],n}$ that will be updated by the ISO. $c^t_{[\mathrm{g}]}$ and $c^t_{[\mathrm{s}]}$ are the per unit equivalent annual cost of renewable harvesters and storage devices, respectively. It is assumed that the grid status and the real-time energy price $c^t_{[\mathrm{r}]}$ are under the influence of the real-time energy production level in the previous time slot [9], i.e., $\{|P^{t-1}_{[\mathrm{r}],n}|\}$. Similar to [9], the real-time energy price of the grid can be modelled as

$$c^{t+1}_{[\mathrm{r}]} \propto \underbrace{0.5 \sum\nolimits_{n=1}^{N} |P^t_{[\mathrm{r}],n}|}_{\text{generation price}} + \underbrace{0.1 \left(\sum\nolimits_{n=1}^{N} \left[|P^t_{[\mathrm{r}],n}| - |P^{t-1}_{[\mathrm{r}],n}| \right]^+ \right)^2 \Big/ \sum\nolimits_{n=1}^{N} |P^t_{[\mathrm{r}],n}|}_{\text{ramping price}}. \tag{2}$$

Let us denote by $P^t_{[\mathrm{Tx}],n} = \sum_{i \in \mathcal{L}_i} ||\mathbf{w}^t_{n,i}||^2$ and $P^t_{[\mathrm{h}],n}$ the total transmit power consumption of the n^{th} A-CPU at the t^{th} time slot and the hardware circuit power consumption, respectively, where $\mathbf{w}^t_{n,i}$ is the beamforming vector from all BSs in the n^{th} A-CPU to the i^{th} user. Per time slot t, the total energy consumption of the n^{th} A-CPU, $n \in \mathcal{L}_a$, is bounded as

$$\xi P^t_{[\mathrm{Tx}],n} + P^t_{[\mathrm{h}],n} \le P^t_{[\mathrm{g}],n} + P^t_{[\mathrm{s}],n} + P^t_{[\mathrm{c}],n} + P^t_{[\mathrm{r}],n}, \tag{3}$$

where $\xi > 0$ is the power amplifier efficiency. We simplify the constraints of the storage devices to the capacity limit only and denote by $P^t_{[\mathrm{cap}],n}$ the finite capacity of the n^{th} storage device. As shown in Fig. 2,

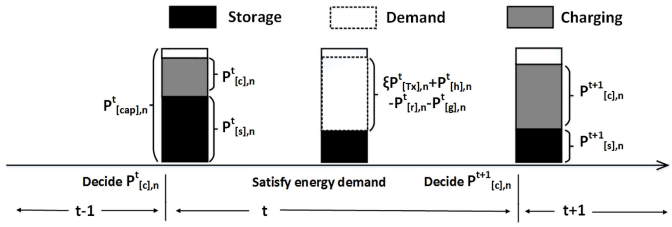


Fig. 2. Illustration of storage charging and discharging processes.

the initial amount of stored energy of the n^{th} A-CPU at time slot t is constrained as follows

$$P^t_{[\mathrm{s}],n} = \min \Big\{ [P^{t-1}_{[\mathrm{g}],n} + P^{t-1}_{[\mathrm{s}],n} + P^{t-1}_{[\mathrm{c}],n} + P^{t-1}_{[\mathrm{r}],n} - P^{t-1}_{[\mathrm{Tx}],n} - P^t_{[\mathrm{h}],n}]^+, P^t_{[\mathrm{cap}],n} \Big\}. \tag{4}$$

III. Foresighted Energy Management Strategy

A. Problem Formulation

The ISO aims to minimize the long-term overall energy consumption cost of the entire network, whilst the individual A-CPUs need to satisfy the signal-to-interference-plus-noise ratio (SINR) requirements of their users in the presence of variability in wireless channels as well as in the power grid, e.g., uncertain renewable energy generation and real-time energy price [9]. Hence, the problem of interest covering the entire cost control in the power network and the beamforming design in the cellular network can be formulated as

$$\min_{\{P^t_{[\mathrm{c}],n}, P^t_{[\mathrm{r}],n}, \mathbf{w}^t_{n,i}\}} \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} C^t_n \right\}$$

$$\text{s.t.} \quad \mathrm{C1}: \mathrm{SINR}_i(\{\mathbf{w}^t_{n,i}\}) \ge \gamma_i, \quad \forall i \in \mathcal{L}_i, n \in \mathcal{L}_a, t \in \mathcal{T},$$

$$\mathrm{C2}: \xi P^t_{[\mathrm{Tx}],n} + P^t_{[\mathrm{h}],n} - P^t_{[\mathrm{g}],n} - P^t_{[\mathrm{s}],n} - P^t_{[\mathrm{c}],n} \le P^t_{[\mathrm{r}],n}, \quad \forall n \in \mathcal{L}_a, t \in \mathcal{T},$$

$$\mathrm{C3}: P^t_{[\mathrm{s}],n} = \min \Big\{ [P^{t-1}_{[\mathrm{g}],n} + P^{t-1}_{[\mathrm{s}],n} + P^{t-1}_{[\mathrm{c}],n} + P^{t-1}_{[\mathrm{r}],n} - P^{t-1}_{[\mathrm{Tx}],n} - P^t_{[\mathrm{h}],n}]^+, P^t_{[\mathrm{cap}],n} \Big\}, \ \forall n \in \mathcal{L}_a, t \in \mathcal{T}, \tag{5}$$

where the constraint C1 indicates the minimum SINR requirements $\{\gamma_i\}$ for the users. Let us denote by C^0_n the total cost incurred by the n-th A-CPU at time slot 0, where no learning or battery pre-charging has yet been applied in the system. Then, we can consider C^0_n as a constant reference point and define the reward for the n^{th} A-CPU at time slot t, as

$$\mathcal{R}^t_n = C^0_n - C^t_n, \ \forall n \in \mathcal{L}_a, t \in \mathcal{T}. \tag{6}$$

Then, the problem of minimizing long-term overall energy cost in (5) is equivalent to the following optimization problem that maximizes the time-averaged accumulated reward, as

$$\max_{\{P^t_{[\mathrm{c}],n}, P^t_{[\mathrm{r}],n}, \mathbf{w}^t_{n,i}\}} \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{n=1}^{N} \mathcal{R}^t_n \right\}$$

$$\text{s.t.} \quad \mathrm{C1–C3}. \tag{7}$$

B. Optimization-Assisted Sequential Decision Making Problem

The stochastic optimization problem in (7) is difficult to solve directly since we aim to minimize the long-term average system cost while the statistics of the system dynamics are unknown in advance. In particular, the time-coupling storage constraints C2 and C3 render the problem intractable for traditional solvers. Furthermore, both the ISO and the A-CPUs only have access to their respective local information. For instance, the storage status is only available to its local A-CPU whilst the energy purchase decisions of the individual A-CPUs are unknown to the others and will affect the overall cost of the system. Hence, the problem cannot be solved in a centralized manner in either the ISO or the A-CPUs through the traditional stochastic optimization based approaches.

Thus, we employ the bandit approach that requires no upfront knowledge of the system dynamics to account for the time-coupled variables, and consider (7) as an optimization-assisted sequential decision making problem that alternates between two decision making processes and indirectly affects the decisions of the A-CPUs slot-by-slot, so as to distributively and asymptotically maximize the average reward in the long run. More specifically, the online learning part of the proposed strategy aims at proactive one-slot-ahead decision making of the storage pre-charging amount for time slot t, i.e., $P^t_{[\mathrm{c}],n}$, prior to any possible random variations in the renewable generation or real-time energy-shortage at time slot t. During time slot t and based on the one-slot-ahead decision of $P^t_{[\mathrm{c}],n}$, the individual A-CPUs will then make decisions on the real-time energy request, i.e., $P^t_{[\mathrm{r}],n}$, for satisfying energy demand of the users via solving the following optimization problem distributively in real-time, as

$$\begin{aligned}
\min_{\{\mathbf{w}^t_{n,i}, P^t_{[\mathrm{r}],n}\geq 0\}} \quad & P^t_{[\mathrm{Tx}],n} + \alpha_n\ P^t_{[\mathrm{r}],n} \\
\text{s.t.} \quad & \mathrm{SINR}^t_i = \frac{|\mathbf{h}_i^H \mathbf{w}^t_{n,i}|^2}{\sum_{j\neq i, j\in\mathcal{L}_i} |\mathbf{h}_i^H \mathbf{w}^t_{n,j}|^2 + \sigma_i^2} \geq \gamma_i,\ \forall i \in \mathcal{L}_i, \\
& \xi P^t_{[\mathrm{Tx}],n} + P^t_{[\mathrm{h}],n} \leq P^t_{[\mathrm{g}],n} + P^t_{[\mathrm{s}],n} + P^t_{[\mathrm{c}],n} + P^t_{[\mathrm{r}],n}, \qquad (8)
\end{aligned}$$

where $\mathbf{h}_i$ represents the channel vector from all BSs in the n^{th} A-CPU to the i^{th} user, $i \in \mathcal{L}_i$, and $P^t_{[\mathrm{s}],n}$ can be updated in the light of (4). The objective function of problem (8) seeks an optimal schedule for beamforming vectors $\{\mathbf{w}^t_{n,i}\}$ and the real-time purchase of energy shortage $P^t_{[\mathrm{r}],n}$, in order to minimize the total cost C^t_n incurred by the n^{th} A-CPU, given the one-slot-ahead decision of $P^t_{[\mathrm{c}],n}$. In particular, the penalty factor α_n in the objective function in (8) is assigned and updated by the ISO according to $\alpha_n \propto \frac{P^{t-1}_{[\mathrm{r}],n}}{\min_{n\in\mathcal{L}_a}\{P^{t-1}_{[\mathrm{r}],n}\}}$ at the beginning of time slot t. This is to account for the grid status at the previous time slot and to control the impacts of the A-CPUs on the grid energy prices at the current time slot. For instance, a larger α_n emphasizes more on the minimization of peak-time energy purchases for those A-CPUs with larger real-time energy shortage in the previous time slot, in order to bring the higher real-time electricity price under control. Furthermore, α_n will also be employed in the proposed online learning algorithm to penalize the reward and indirectly affect the one-slot-ahead decisions of the A-CPUs, i.e., $\{P^t_{[\mathrm{c}],n}\}$.

To solve problem in (8), let us define the rank-one semidefinite matrices $\mathbf{W}^t_{n,i} = \mathbf{w}^t_{n,i}\mathbf{w}^{tH}_{n,i}$ and $\mathbf{H}_i = \mathbf{h}_i\mathbf{h}_i^H$. The problem in (8) can be transformed to a tractable form after relaxing the rank-one constraint

Algorithm 1: Foresighted Energy Management Algorithm.

1: **Initialize**: $\epsilon^0 = 1$, $P^1_{[\mathrm{s}],n} = 0$, $P^0_{[\mathrm{c}],n} = 0$, estimated mean reward $\bar{\mathbf{r}}^t_n = [\bar{r}^t_{n,1}, \ldots, \bar{r}^t_{n,E}] = \mathbf{0}$, temporary reward $\mathbf{r}^{[\mathbf{k},\mathbf{t}]}_{\mathbf{n}} = [r^{[k,t]}_{n,1}, \ldots, r^{[k,t]}_{n,E}] = \mathbf{0}$, total numbers of actions E with step of Δ, time slots T and learning trials K.
2: **REPEAT**
3: With the probability of ϵ^t: t is for **Exploration**
4: **While** $k \leq K$
5: **if** $k = 1$
-with probability of ϵ^t, randomly select an action $P^{t,1}_{[\mathrm{c}],n}$,
-with probability of $1-\epsilon^t$, select $P^{t,1}_{[\mathrm{c}],n}$ as
$$P^{t,1}_{[\mathrm{c}],n} = \operatorname*{argmax}_e(\bar{\mathbf{r}}^t_n),\ e \in \mathcal{E}.$$
end if
6: Solve (9) and compute $\mathcal{R}(P^{t,k}_{[\mathrm{c}],n})$ as per (10).
7: **if** $k = 2$ or $\mathcal{R}(P^{t,k}_{[\mathrm{c}],n}) > \mathcal{R}(P^{t,k-1}_{[\mathrm{c}],n})$
$$P^{t,k+1}_{[\mathrm{c}],n} = P^{t,k}_{[\mathrm{c}],n} + \Delta.$$
8: **else if** $\mathcal{R}(P^{t,k}_{[\mathrm{c}],n}) < \mathcal{R}(P^{t,k-1}_{[\mathrm{c}],n})$
$$P^{t,k+1}_{[\mathrm{c}],n} = P^{t,k}_{[\mathrm{c}],n} - \Delta.$$
9: **else** $P^{t,k+1}_{[\mathrm{c}],n} = P^{t,k}_{[\mathrm{c}],n}$.
end if
10: Update the temporary reward matrix $\mathbf{r}^{[\mathbf{k},\mathbf{t}]}_{\mathbf{n}}$, as
$$r^{[k,t]}_{n,e} = \mathcal{R}(P^{t,k}_{[\mathrm{c}],n}),\ e = \tfrac{P^{t,k}_{[\mathrm{c}],n}}{\Delta} \in \mathcal{E}.$$
11: Update $k = k + 1$.
End While
12: With the probability of $1-\epsilon^t$: t is for **Exploitation**
13: Select best action $P^t_{[\mathrm{c}],n} = \operatorname{argmax}_e(\bar{\mathbf{r}}^t_n),\ e \in \mathcal{E}$.
14: Solve problem (9) and compute $\mathcal{R}(P^t_{[\mathrm{c}],n})$ as per (10).
15: Compute the estimated mean reward vector, as
$$\bar{\mathbf{r}}^t_n = \frac{\sum_{t'=1}^{t}\left(\frac{\sum_{k=1}^{K} r^{[k,t']}_{n,1}}{K}, \frac{\sum_{k=1}^{K} r^{[k,t']}_{n,2}}{K}, \ldots, \frac{\sum_{k=1}^{K} r^{[k,t']}_{n,E}}{K}\right)^{\beta(t-t')}}{t},$$
where β is the discount factor.
16: Update $t = t + 1$.
17: **UNTIL** $t = T$

of $\mathrm{rank}(\mathbf{W}^t_{n,i}) = 1$, as

$$\begin{aligned}
\min_{\{\mathbf{W}^t_{n,i}\succeq 0, P^t_{[\mathrm{r}],n}\geq 0\}} \quad & \sum_{i\in\mathcal{L}_i} \mathrm{tr}(\mathbf{W}^t_{n,i}) + \alpha_n\ P^t_{[\mathrm{r}],n} \\
\text{s.t.} \quad & \gamma_i^{-1}\mathrm{tr}(\mathbf{H}_i\mathbf{W}^t_{n,i}) \geq \sum_{j\in\mathcal{L}_i, j\neq i} \mathrm{tr}(\mathbf{H}_i\mathbf{W}^t_{n,j}) + \sigma_i^2, \forall i \in \mathcal{L}_i, \\
& \xi\sum_{i\in\mathcal{L}_i}\mathrm{tr}(\mathbf{W}^t_{n,i}) + P^t_{[\mathrm{h}],n} - P^t_{[\mathrm{s}],n} - P^t_{[\mathrm{c}],n} - P^t_{[\mathrm{g}],n} \leq P^t_{[\mathrm{r}],n}. \qquad (9)
\end{aligned}$$

It can be proved that the optimal solutions to problem (9) satisfy the rank-one constraint and therefore are also optimal for (8). The proof is similar to that in [4] and thus omitted.

C. The Proposed Foresighted Algorithm

Next we turn to the online learning part of the proposed strategy to decide the one-slot-ahead storage pre-charging amount at individual A-CPUs. We employ the multi-armed bandit model [10], where the agents are the A-CPUs and a total number of E actions correspond to the E discrete amounts of energy with step of Δ that can be charged as storage within an A-CPU prior to the occurrence of any possible energy

shortage. Let us denote by $\mathcal{E} = \{1, \ldots, E\}$ and $\mathcal{A} = \{A_1, \ldots, A_E\}$ the indexes and the set of E possible actions, respectively. The proposed foresighted algorithm aims to distributively and sequentially determine the one-slot-ahead storage pre-charging amount, i.e., select the best-possible action $P^{t*}_{[\mathrm{c}],n} \in \mathcal{A}$, before experiencing an energy shortage at time slot t, and adaptively react with the foresighted best response via solving the problem in (9). To accelerate the learning process, we implement K learning trials within a time slot, indexed by $\mathcal{K} = \{1, \ldots, K\}$. Let the action selected to pre-charge the storage unit of the n^{th} A-CPU in the k^{th} trial of the t^{th} time slot, $n \in \mathcal{L}_a, k \in \mathcal{K}, t \in \mathcal{T}$, be denoted by $P^{t,k}_{[\mathrm{c}],n}$. The instantaneous reward of $P^{t,k}_{[\mathrm{c}],n}$, can be defined as

$$\mathcal{R}\left(P^{t,k}_{[\mathrm{c}],n}\right) = \tilde{C}^{0,1}_n - \tilde{C}^{t,k}_n, \ \forall n \in \mathcal{L}_a, t \in \mathcal{T}, \tag{10}$$

where $\tilde{C}^{t,k}_n$ is the penalized total cost incurred by the n^{th} A-CPU and is defined as

$$\tilde{C}^{t,k}_n = \alpha_n c^t_{[\mathrm{r}]} P^{t,k}_{[\mathrm{r}],n} + c^t_{[\mathrm{c}]} P^{t,k}_{[\mathrm{c}],n} + c^t_{[\mathrm{g}]} P^t_{[\mathrm{g}],n} + c^t_{[\mathrm{s}]} P^{t,k}_{[\mathrm{s}],n}. \tag{11}$$

The penalized reward function in (10) urges the A-CPUs with higher real-time energy shortage in the current round to pre-charge more energy into their storages in order to be prepared for the next round.

The proposed foresighted algorithm is governed by a trade-off between exploring other actions that may yield a better accumulated reward in the presence of uncertainties on the renewable energy supply, and exploiting current knowledge to make the empirically best decisions among a set of actions. The details of the proposed foresighted procedures to be executed at the individual A-CPUs are described in Algorithm 1, where the individual time slots can either be allocated as an exploration cycle with the probability of ϵ^t, or an exploitation cycle with the probability of $1 - \epsilon^t$, as explained below:

- Exploration: With the probability of ϵ^t, a perturbation procedure is applied to explore actions uniformly at random, as the starting point of learning, i.e., $P^{t,1}_{[\mathrm{c}],n} \in \mathcal{A}$. With the probability of $1 - \epsilon^t$, the action associated with the highest mean reward so far will be selected as an initial learning point and will be gradually improved until a total number of K learning trials is attained.
- Exploitation: The individual A-CPUs select the best-possible action that yields the highest estimated mean reward so far, based on the observed knowledge up to the $(t-1)^{\text{th}}$ time slot, and then feedback it to solve (9).

Instead of the traditional hand-tuning of ϵ^t [8], we modified the value-difference based exploration method [11], such that the exploration-exploitation control parameter ϵ^t, $0 \leq \epsilon^t \leq 1$, is adaptive to the uncertainty in the learning progress, i.e.,

$$\epsilon^{t+1} = \delta \cdot \frac{1 - e^{\frac{-|\mathcal{R}^{t+1}_n - \frac{1}{t}\sum_{t'=1}^{t} \mathcal{R}^{t'}_n|}{\sigma}}}{1 + e^{\frac{-|\mathcal{R}^{t+1}_n - \frac{1}{t}\sum_{t'=1}^{t} \mathcal{R}^{t'}_n|}{\sigma}}} + (1 - \delta) \cdot \epsilon^t, \tag{12}$$

where σ is a positive constant indicating the inverse sensitivity and $\delta \in [0, 1)$ [11]. More specifically, a time-decayed exploration rate can be adopted in a relative static environment, where the estimation of the mean reward process of the actions is improved with time. On the contrary, a relative high exploration rate can be employed when a sudden change in the environment or the reward is observed.

IV. Simulation Results

Consider an ISO with 3 A-CPUs, where each aggregator consists of a C-RAN of 3 BSs and 6 users. The renewable energy generation at each time slot varies as $P^t_{[\mathrm{g}],1} \in [0\ 0.5]$ W, $P^t_{[\mathrm{g}],2} \in [0.3\ 1.0]$ W and $P^t_{[\mathrm{g}],3} \in [0.5\ 1.5]$ W, respectively. The other simulation parameters are described as follows: $E = 20$ with $\Delta = 300$ mW, $\beta = 0.95$, $c^1_{[\mathrm{r}]} = £0.15/\mathrm{W}, c^t_{[\mathrm{c}]} = £0.07/\mathrm{W}, c^t_{[\mathrm{g}]} = £0.05/\mathrm{W}, c^t_{[\mathrm{s}]} = £0.01/\mathrm{W}$, $P^t_{[\mathrm{cap}],n} = 40$ dBm, $P^t_{[\mathrm{h}],n} = 33$ dBm, $\xi = 1$, $\sigma = 5$ and $\delta = \frac{1}{20}$. The proposed strategy is evaluated with $K = 7$ learning trials and $T = 60$ time slots. Two designs are chosen as benchmark schemes: an adaptive storage management design in [8] that only considers a single A-CPU with fixed exploration-exploitation trade-off, and a baseline myopic design in [5] that myopically minimizes the current energy cost.

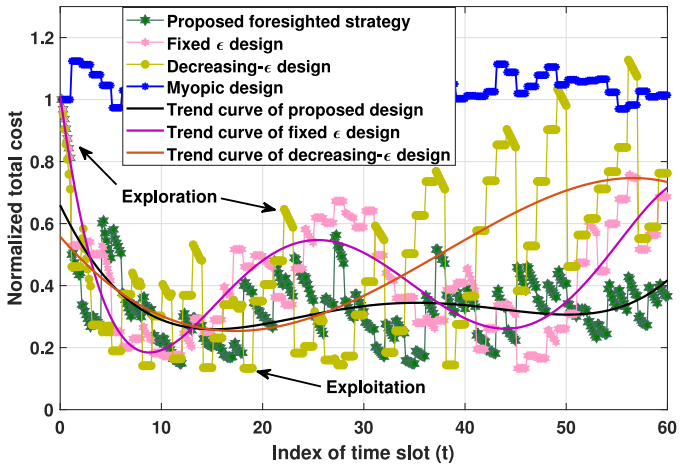
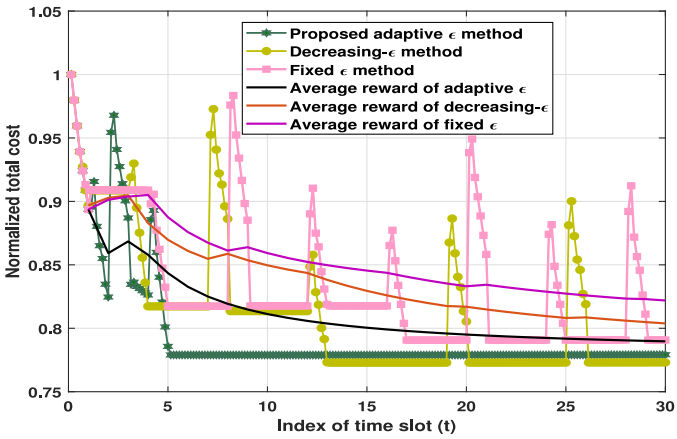


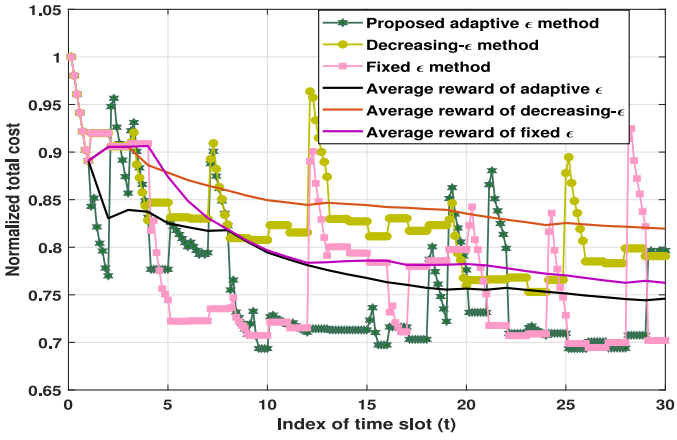
Fig. 3. Normalized total cost at individual time slots at $\gamma = 15$ dB.

Fig. 3 presents the comparison of the normalized total costs when $\gamma = 15$ dB. For fair comparison, identical constraints have been applied to all strategies and the overall cost is normalized to the cost at the initial time of the proposed strategy. The bursts at the beginning of each exploration cycle appear as a result of highly uncertain renewable energy generation as well as the exploration perturbation in step 5 of Algorithm 1. As seen in Fig. 3, the proposed strategy outperforms the baseline myopic design in [5]. Furthermore, the proposed strategy indicates smaller variations and better average reward with increasing time, as compared to that of the designs in [8] with fixed and with decreasing ϵ, respectively. This is due to the negligence of the coupling effects among individual A-CPUs in the nature of the design in [8], which provides poorer adaptation to the variations in the renewable energy generation as well as the real-time energy price in a decentralized scenario.

Fig. 4 illustrates the comparison of the proposed adaptive ϵ-greedy method against a traditional hand-tuning fixed-ϵ-greedy method in [8] and a decreasing-ϵ method [12]. In order to demonstrate the advantage of the proposed adaptive method, all strategies are implemented with fixed real-time price of $c^t_{[\mathrm{r}]} = £0.15/\mathrm{W}$ and identical constraints in two extreme cases, i.e., a near-static environment and a highly uncertain environment. One can conclude from the average reward curves in Fig. 4(a) that the proposed adaptive method outperforms the fixed ϵ-greedy method in the near-static environment. This is due to the fact that the probability of high-cost exploration in the proposed method is gradually decreased with the agent's increasing knowledge of the environment. Whilst as can be observed from Fig. 4(b), the average accumulative reward of the proposed method achieves an approximately 9% improvement as compared to the decreasing-ϵ method. This is because the latter method fails to adjust the exploration-exploitation trade-off according to the true learning progress in the highly uncertain



(a)



(b)

Fig. 4. Comparison of various ϵ-greedy methods at $\gamma = 15$ dB in (a) a near-static environment, (b) a highly uncertain environment.

environment, e.g., intermittent and highly irregular renewable energy generation.

V. Conclusion

The variability of renewable sources introduces large ramps in the energy supply which can lead to increased overall cost as well as grid stability issues. A foresighted energy management strategy has been proposed in this paper for the ISO to minimize the long-term overall cost of the network in the presence of uncertain renewable energy generation. The proposed strategy accounts for the impacts of the individual A-CPUs on the power network and reacts adaptively with the storage pre-charging amount in a distributed way. Simulation results confirm the effectiveness of the proposed foresighted strategy in achieving a significant performance gain over a recently proposed learning-based storage management design and a baseline myopic design.

References

[1] D. Li, W.-Y. Chiu, H. Sun, and H. V. Poor, "Multiobjective optimization for demand side management program in smart grid," *IEEE Trans. Ind. Inform.*, vol. 14, no. 4, pp. 1482–1490, Apr. 2018.

[2] Z. Fan *et al.*, "Smart grid communications: Overview of research challenges, solutions, and standardization activities," *IEEE Commun. Surv. Tut.*, vol. 15, no. 1, pp. 21–38, Firstquarter 2013.

[3] Z. Zhu *et al.*, "A game theoretic optimization framework for home demand management incorporating local energy resources," *IEEE Trans. Ind. Inform.*, vol. 11, no. 2, pp. 353–362, May 2015.

[4] W. N. S. F. Wan Ariffin *et al.*, "Sparse beamforming for real-time resource management and energy trading in green C-RAN," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 2022–2031, Jul. 2017.

[5] J. Xu and R. Zhang, "Cooperative energy trading in CoMP systems powered by smart grids," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2142–2153, Apr. 2016.

[6] Y. Zhang and M. van der Schaar, "Structure-aware stochastic storage management in smart grids," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 6, pp. 1098–1110, Dec. 2014.

[7] X. Wang, Y. Zhang, T. Chen, and G. B. Giannakis, "Dynamic energy management for smart-grid-powered coordinated multipoint systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1348–1359, May 2016.

[8] X. Zhang, M. R. Nakhai, and W. N. S. F. Wan Ariffin, "Adaptive energy storage management in green wireless networks," *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 1044–1048, May 2017.

[9] Y. Xao and M. van der Schaar, "Distributed demand side management among foresighted decision makers in power networks," *in Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2013, pp. 1383–1387.

[10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge MA, USA: MIT Press, 2017.

[11] M. Tokic, "Adaptive ϵ-greedy exploration in reinforcement learning based on value differences," *in KI 2010: Advances in Artificial Intelligence*. Berlin, Germany: Springer, 2010.

[12] S. Maghsudi *et al.*, "Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1309–1322, Mar. 2015.