Learning-Based Demand Response for Privacy-Preserving Users

Amir Ghasemkhani, *Student Member, IEEE*, Lei Yang, *Member, IEEE*, and Junshan Zhang, *Fellow, IEEE*

***Abstract*—Demand response (DR), as a vital component of smart grid, plays an important role in shaping the load profiles in order to improve system reliability and efficiency. Incentive-based DR has been used in many DR programs by incentivizing customers to adapt their loads to supply availability. Note that users' behavior patterns can be easily identified from fine-grained power consumption when interacting with the load serving entity (LSE), giving rise to serious privacy concerns. One common approach to address the privacy threats is to incorporate perturbations in users' load measurements. Although it can protect the users' privacy, yet the usage data modification would degrade the LSE's performance in achieving an optimal incentive strategy due to unknown characteristics of the augmented perturbations. In this paper, we cast the incentive-based DR problem as a stochastic Stackelberg game. To tackle the challenge induced by users' privacy protection behaviors, we propose a two-timescale reinforcement learning algorithm to learn the optimal incentive strategy under users' perturbed responses. The proposed algorithm computes the expected utility cost to mitigate the impacts of the random characteristics of the augmented perturbations and then updates the incentive strategy based on the perceived expected utility costs. We derive the conditions under which the proposed incentive scheme converges almost surely to an ϵ-optimal strategy. The efficacy of the proposed algorithm is demonstrated using extensive numerical simulation using real data.**

***Index Terms*—Differential privacy, incentive-based demand response (DR) program, privacy-preserving demand response, two-timescale reinforcement learning algorithm.**

I. Introduction

THE share of advanced control and communication technologies are steadily growing as the traditional power grids are being transformed into smart grids with a resulting surge in integration of renewable energy resources [1]. Demand response (DR), as a vital component of smart grid, plays a key role in reducing the peak load and incorporating renewables into the grid by providing incentives for users to adapt their electricity demand to supply availability. Existing DR programs can be generally categorized into price-based DR [2]–[5] and incentive-based DR programs [6]–[10]. Under the price-based scheme, users are encouraged to individually and voluntarily manage their loads by receiving dynamic prices from the load serving entity (LSE), whereas the incentive-based scheme induces DR on mandatory [9] or voluntary [7] basis by offering incentives to the users.

However, many DR programs are designed on the basis of the assumption that users' response functions are available or predictable at the LSE side [5]–[8]. Users' privacy protection behaviors are ignored when designing the DR programs. Since users' specific activities or behavior patterns can be easily identified from the highly accurate profiles of energy usage [11]–[13], privacy has become a major concern of users, and various privacy-preserving approaches have been developed [12], [14]–[18]. Therefore, the response of privacy-aware users would be deviated from the designed target, which calls for a new design of DR scheme accounting for these privacy-preserving users.

In this paper, we study LSE's incentive strategy by proposing a voluntary incentive-based framework for privacy-preserving users. Currently, to protect privacy, each user often leverages energy storage devices (e.g., rechargeable battery) and uses the charging and discharging mechanisms to perturb the real energy consumption level [16]. Consequently, the LSE cannot infer users' real consumption. Besides, any mischaracterization of the users' responses by assuming predefined response functions in the DR problem could lead to higher system costs due to divergent behaviors of the users. To tackle these challenges, a learning algorithm is developed to jointly learn the divergent habituates of users and update the LSE's incentive strategy, while taking into account users' privacy protection behaviors in the incentive scheme.

A. Summary of Major Contributions

Our main contributions are summarized as follows:

1) We formulate the incentive-based DR problem as a stochastic Stackelberg game. A key challenge is how to determine the incentive rates to incentivize users to adapt their loads to supply availability, as users would perturb

Manuscript received August 8, 2018; revised January 15, 2019; accepted February 2, 2019. Date of publication February 11, 2019; date of current version September 3, 2019. This work was supported in part by the U.S. National Science Foundation under Grant IIA-1301726, Grant EEC-1801727, Grant IIS-1838024; in part by the DTRA under Grant HDTRA1-13-1-0029; and in part by Open Project of State Key Laboratory of Industrial Control Technology under Grant ICT1800373. Paper no. TII-18-2064. *(Corresponding author: Amir Ghasemkhani.)*

A. Ghasemkhani and L. Yang are with the Department of Computer Science and Engineering, University of Nevada Reno, Reno, NV 89557 USA (e-mail: aghasemkhani@nevada.unr.edu; leiy@unr.edu).

J. Zhang is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: junshan.zhang@asu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2019.2898462

their power usage to protect their privacy. In particular, users' responses are stochastic and time varying; even if the same incentive signal is received, different responses will be observed by the LSE. This is because of the fact that the users have idiosyncratic behaviors and, at the same time, perturb their actual responses using randomized algorithms to protect their privacy. Therefore, the LSE needs to learn from users' responses to adjust its incentive strategy adaptively. Along this line, we cast the incentive-based DR problem for the privacy-preserving users as a learning problem, such that the best incentive strategy can be obtained by learning the users' randomized behaviors.

2) Under the stochastic Stackelberg game, the LSE and users are playing their best responses so that no one can improve its payoff by switching their own strategy. However, the proposed stochastic game would not converge to a stable mixed strategy equilibrium using the classic best response dynamics, since the LSE learns over the users' perturbed behaviors [19], [20]. To tackle this challenge, we adopt the notation of the smooth best response as a tool to overcome difficulties in converging to the optimal incentive strategy for the perturbed observations. Moreover, a two-timescale reinforcement learning algorithm, consisting of a fast and a slow timescale learning processes, is proposed to deal with users' perturbed measurements. The proposed learning algorithm enables the LSE to neutralize the effect of the perturbations by calculating the expected utility cost and learning the aggregated behaviors of the users in order to find the optimal incentive strategy.
3) We show that the proposed learning algorithm converges almost surely to an ϵ-optimal strategy, while a moderate parameter β is chosen to strike a balance between exploration and exploitation of the learning algorithm. Besides, we evaluate the performance of the proposed algorithm in comparison with a naive reinforcement learning algorithm. Using real data, we corroborate the superior performance of the proposed algorithm via numerical simulations. Moreover, we quantify the effect of the users' privacy level on the performance of the proposed algorithm.

B. Related Works

DR commonly refers to the process of managing the consumption of energy to optimize available and planned generation resources [21]. According to the U.S. Department of Energy, DR is defined as "actions taken on the customer's side of the meter to change the amount or timing of energy consumption." Effective DR depends on fine granularity power consumption data to predict load, provide future pricing information, and show the consumer the cost of his or her consumption. With these highly accurate profiles of energy usage, however, it is possible to identify consumers' specific activities or behavior patterns, thereby giving rise to serious privacy concerns [11]–[13]. To resolve the security and privacy concerns, cryptography-based approaches are proposed. Li *et al.* [22] presented a distributed incremental data aggregation approach to protect user's privacy, using homomorphic encryption. Seo *et al.* [23] proposed a secure and efficient power management mechanism to securely gather the power demands of users, by leveraging a homomorphic data aggregation and capability-based power distribution. Li *et al.* [24] employed the homomorphic encryption mechanism to achieve a privacy-preserving DR scheme. Lu *et al.* [25] proposed a privacy-preserving aggregation scheme for secure and efficient smart grid communications in order to realize multidimensional data aggregation approach. Although cryptography-based approaches can prevent adversaries from eavesdropping the communication between consumers and the LSE and identifying the consumers' electricity demand, yet the consumers' personal identities still can be exploited and shared with third parties by the LSE.

To preserve users' privacy, various privacy-preserving approaches have been developed [12], [14]–[18]. The main idea of these approaches is to perturb users' responses to the pricing signals such that users' specific activities or behavior patterns cannot be identified. However, this renders a challenge for the design of DR programs, as privacy-aware users would not actively respond to incentive or price signals and thereby make it challenging to effectively adapt the electricity demand to supply availability. Recently, Maharjan *et al.* [26] proposed a game-theoretical framework to model the interactions among multiple LSEs and multiple users. They have considered a mathematical model for the users' cost functions to obtain an optimal amount of power to demand from the LSEs. However, mathematical models may be error prone and cannot accurately characterize the divergent behaviors of users.

The problem of using reinforcement learning for DR program has been studied in [27]–[30]. The authors in [27] and [28] proposed an energy management system approach that uses a reinforcement learning algorithm to learn the users' behaviors in order to dynamically adapt energy scheduling to future uncertain energy prices. Note that these two works consider the DR problem from users' point of view. However, Lu *et al.* [29] has incorporated reinforcement learning into a dynamic price-based DR framework in order to adaptively determine the retail price of energy from the LSE's point of view. A reinforcement-learning-based DR program was proposed in [30] considering the LSE objectives without assuming any specific forms of users' response functions. However, none of these studies have considered the privacy-preserving behaviors of the users in the DR problem. Besides, the incentive-based DR framework in the presence of the privacy-preserving users remains an uncharted territory. In this paper, we aim to tackle these challenges by developing a model-free incentive-based DR program that accounts for the privacy protection behaviors of users.

The rest of this paper is organized as follows. Section II presents the system model and the problem formulation. In Section III, we establish a learning-based approach to solve the problem of incentive-based DR program for privacy-preserving users and characterize its ϵ-optimal performance. Section IV presents numerical simulation results. The paper is concluded in Section V.

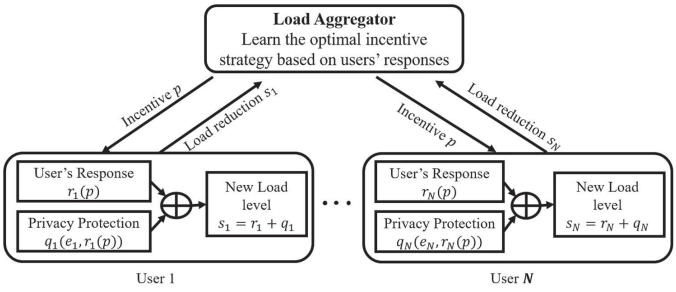

Fig. 1. Incentive scheme for privacy-preserving DR.

TABLE I
DEFINITION OF VARIABLES AND PARAMETERS

Variables and parameters	Definitions
$p_{inc}(t)$	Incentive rates in the retail market in $\$/kWh$
$p_{RR}(t)$	Retail rate of energy in the retail market in $\$/kWh$
$p_{sup}(t)$	Supply price in the wholesale market in $\$/kWh$
$s_n(t)$	User n's demand reduction after incentives in kWh
$e_n(t)$	Energy usage state of user n before incentives
$x_n(t)$	Random unknown parameters for user n
$r_n(.)$	User n's response function in kWh
$q_n(.)$	Noise generated by user n in kWh
$d(t)$	Desired load reduction in kWh
$h(.)$	Penalty function for the aggregated load reduction
$U(.)$	LSE's cost function in $
α	Load variation economic weight
$\mathcal{N}$	Set of users
$\mathcal{P}$	Set of all available incentives
$\mathcal{S}$	Set of all possible states of the system at time t
$S(t)$	State of the system at time t
$\Delta\mathcal{P}$	Set of all strategies
ψ_p	Probability that incentive p is chosen by the LSE
π_S	Incentive strategy at state S
$C_S(\pi_S)$	Total expected utility cost at state S and under the strategy π_S
$U_S(p)$	Expected utility cost at state S when the incentive p is selected
β	Reinforcement learning temperature parameter
γ^t	Learning rate of the fast time scale reinforcement learning algorithm
λ^t	Learning rate of the slow time scale reinforcement learning algorithm
$Q_S^t(p)$	Perception value of the expected utility cost at state S when the incentive p is selected
$b_S^t(p)$	Smooth best response of the LSE at time t at state S when p is the incentive chosen
ε	Differential privacy parameter
f	Query function
$\mathcal{A}$	Query mechanism
$\mathcal{M}$	Data set
$V(f)$	Global sensitivity of function f
$\mathcal{L}(\cdot)$	Independent random variable generated by Laplace distribution

II. DR FOR PRIVACY-PRESERVING USERS

A. System Model

As illustrated in Fig. 1, we consider a discrete-time system, where the LSE (e.g., the load aggregator) aims to procure a total load reduction, L, by incentivizing the end users to adapt their consumption. Specifically, at time slot t, the LSE announces the incentive $p_{\rm inc}(t)$ and pays user n the amount $p_{\rm inc}(t)s_n$ when user n reduce its consumption by $s_n \geq 0$. The market design task is to design $p_{\rm inc}(t)$ such that the LSE achieves the desired amount of load reduction. All the variables and parameters definitions are shown in Table I.

1) User-Side Model: The reduction s_n consists of two parts: 1) user n's response function to the announced incentive, $r_n(p_{\rm inc}(t), e_n(t), x_n(t))$, which depends on the announced incentive $p_{\rm inc}(t)$, energy usage state $e_n(t)$ of user n before incentives, and other unknown parameters $x_n(t)$, e.g., weather conditions, which is assumed to be independent and identically distributed, and 2) user n's privacy protection mechanism, which perturbs its demand to preserve the privacy [16], [17], [31]. Let $q_n(e_n(t), r_n(p_{\rm inc}(t), e_n(t), x_n(t)))$ denotes the "noise" generated by user n to protect its own privacy, which depends on the energy usage state $e_n(t)$ of user n, user n's response function $r_n(\cdot)$, and other unknown parameters $x_n(t)$. Practically, these variables represent the specific behaviors of the users and may not be available at the LSE side. The privacy protection function $q_n(\cdot)$ depends on the specific privacy protection strategy of user n to ensure privacy. For example, differential privacy based privacy protection strategies have been designed by perturbing users' load with a random noise following the Laplace distribution [31], [32]. Clearly, the users need to strike a trade-off between their privacy and the cost of electricity, and more privacy protection would result in a higher electricity cost [16].

2) LSE-Side Model: Due to users' privacy protection behaviors, users' responses are uncertain, and the total load reduction that the LSE achieves with $p_{\rm inc}(t)$ is a random quantity $L = \sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t))$. Therefore, this curtailment may not exactly match the desired load reduction $d(t)$, which reflects the supply availability. Note that $\sum_n e_n(t)$ is the aggregated consumption level before incentives, which is required in order to obtain the demand reduction in the incentive-based DR programs [33]. Let $h(\cdot)$ denote the penalty function to capture the penalty for deviation from $d(t)$. In particular, the penalty is $h(d(t) - \sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t)))$, which is assumed to be a quadratic function [8], [34]. Besides, there is a revenue function for the LSE due to the users' participation in the DR program [5], [7], [8]. More specifically, the LSE's revenue function can be constructed as follows:

$$\begin{aligned}\text{LSER} = {} & p_{\rm RR}(t)\left(\sum_n e_n(t) - \sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t))\right) \\ & - p_{\rm sup}(t)\left(\sum_n e_n(t) - \sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t))\right) \\ & - p_{\rm inc}(t)\left(\sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t))\right) \end{aligned} \quad (1)$$

where $p_{\rm RR}(t)$ is the retail price of energy, and $p_{\rm sup}(t)$ is the supply price of energy from the wholesale market. Note that the revenue function for the LSE can be reduced to the form $(p_{\rm inc}(t) + p_{\rm sup}(t) - p_{\rm RR}(t))\sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t))$ since $p_{\rm RR}(t)$ and $p_{\rm sup}(t)$ are available. Therefore, the LSE's overall utility cost $U(p_{\rm inc}(t), d(t), \mathbf{e}(t), \mathbf{x}(t))$ at time slot t is equal to the sum of the penalty of deviation from the desired load reduction target $d(t)$ and the LSE's revenue loss due to incentive payments, i.e.,

$$\begin{aligned} & U(p_{\rm inc}(t), d(t), \mathbf{e}(t), \mathbf{x}(t)) \\ & = h(d(t) - \sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t))) \\ & \quad + (p_{\rm inc}(t) + p_{\rm sup}(t) - p_{\rm RR}(t)) \\ & \quad \sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t)) \end{aligned} \quad (2)$$

where $\mathbf{e}(t) = \{e_n(t)\}$ denotes the set of energy usage states before incentives observed by the LSE in which the energy usage state $e_n(t)$ of each user n can be measured by smart meter, and $\mathbf{x}(t) = \{x_n(t)\}$ are the set of random variables to the LSE.

From (2), the LSE's cost function depends on the incentive rate $p_{\rm inc}(t)$, the desired load reduction target $d(t)$, and the aggregated load reduction $\sum_n s_n(p_{\rm inc}(t), e_n(t), x_n(t))$. Given

users' strategies [i.e., $r_n(\cdot)$ and $q_n(\cdot)$], the real energy usage state in the next time slot will depend on incentive $p_{\text{inc}}(t)$. As both $r_n(\cdot)$ and $q_n(\cdot)$ may not be available at the LSE side, the real consumption level may be unknown to the LSE, which is a key challenge of deriving the optimal incentive strategy for the LSE. In other words, a random perturbation is augmented to the actual load profile of each user according to their privacy protection strategies. Moreover, note that although users' response functions $r_n(\cdot)$ are not available to the LSE, it can be predicted from historical consumption data, especially when $r_n(\cdot)$ is a linear function, which is often considered in the literature, e.g., [26], [34], and [35]. We should caution that such predictions are error prone, and the prediction errors can be large when users' response functions are complicated.

B. Problem Formulation

We employ a game-theoretical framework to characterize the interaction between the users and the LSE. At each time slot, the users and the LSE are playing their best response against each other sequentially. Specifically, the LSE acts first by announcing the incentive rate, and then the users make their decisions based on the announced incentive. This interaction is formulated as a Stackelberg game, in which the LSE is the leader and the users are the followers. Let $\mathcal{N} = \{1, 2, \ldots, n\}$ be a set of users and $\mathcal{P}$ be the action space of the LSE, which corresponds to the set of available incentives. The LSE's objective is to minimize the total expected system cost, as the LSE receives noisy observations from the users. Solving the proposed problem requires to find an incentive strategy π_S that maps each state $S(t) = \{\mathbf{e}(t), d(t)\} \in \mathcal{S}$ to an incentive $p_{\text{inc}}(t)$, where the incentive strategy $\pi_S = \{\psi_p\} \in \Delta\mathcal{P}$ is a probability distribution over $\mathcal{P}$, and $\Delta\mathcal{P}$ denotes the strategy space. ψ_p, $p \in \mathcal{P}$, denotes the probability that incentive p is chosen by the LSE. Thus, the expected utility cost $C_S(\pi_S)$ of the LSE at state $S(t) = S$ and under strategy π_S can be written as follows:

$$C_S(\pi_S) = \sum_{p\in\mathcal{P}} \mathbb{E}\left[U(p_{\text{inc}}(t), S(t), \mathbf{x}(t))\right]\psi_p = \sum_{p\in\mathcal{P}} \bar{U}_S(p)\psi_p \tag{3}$$

where $\bar{U}_S(p)$ is the expected utility cost at state S when incentive $p_{\text{inc}}(t) = p$ is selected.

Summarizing, the problem can be cast as a stochastic Stackelberg game, denoted by a 3-tuple $\Gamma = (\mathcal{N}, \Delta\mathcal{P}, C)$, in which the LSE needs to find the optimal incentive strategy to minimize its own cost (3) by observing the users' perturbed responses. Note that we consider the one-step incentive strategy (i.e., greedy policy), such that in each time slot, we compute the best incentive policy for the current game state only. For the problem under consideration, the one-step optimal solution can be an efficient approximation for the stochastic game approach from the LSE point of view, since the LSE has limited knowledge of the future events.

III. Learning-Based DR Scheme

The LSE (i.e., leader) aims to minimize its expected utility cost at each time slot by announcing the best incentive rate with no knowledge of users' (i.e., followers') cost functions. Specifically, users take the incentive rate in each slot and apply it to their cost functions to obtain their energy consumption strategies. However, because of divergent behaviors of the users and privacy protection mechanisms, users' response functions cannot be easily predicted by the LSE. To this end, we devise a learning-based incentive mechanism to learn the users' response functions. In particular, we leverage a smooth best response rule, based on which the LSE updates its own incentive strategy, and, correspondingly, determines the ϵ-optimal strategy of the game [20]. In order to deal with noise-corrupted observations, a two-timescale reinforcement learning algorithm is proposed to achieve ϵ-optimality.

A. Smooth Best Response

The LSE has no perfect knowledge about the users' actual behaviors (i.e., the LSE can observe only the perturbed responses) due to their privacy protection strategies and divergent consumption behaviors; this renders difficulties in designing the learning algorithm using the classic best response. That is, the classic best response would result in an unstable mixed strategy equilibrium for the stochastic learning game [19]. To tackle this challenge, we adopt the notation of the smooth best response as a tool to overcome the difficulties [20], [36]. The smooth best response is defined as follows.

Definition 1 (Smooth best response): For the LSE with the expected utility cost function $C_S(\pi)$, given the users' perturbed aggregated reduction $\mathcal{L}$, the smooth best response $b_S \in \Delta\mathcal{P}$ at state S is a strategy defined as follows:

$$b_S = \arg\min_{\pi\in\Delta\mathcal{P}} \left[C_S(\pi) - \frac{1}{\beta}\nu(\pi)\right] \tag{4}$$

where $\beta > 0$ is a temperature parameter, and the smooth function $\nu(\pi) : \Delta\mathcal{P} \to \mathbb{R}$ is strictly differentiable and concave such that as π approaches the boundary of $\Delta\mathcal{P}$, the slope of ν becomes infinite (e.g., an entropy function).

Note that when $\beta \to 0$, b_S becomes the uniform probability distribution over the strategy space $\Delta\mathcal{P}$, whereas when $\beta \to \infty$, it boils down to $b_S = \arg\min_{\pi\in\Delta\mathcal{P}} C_S(\pi)$, which turns out to be classical best response.

In this paper, we choose the following entropy function as smooth function:

$$\nu(\pi) = -\sum_{i=1}^{|\mathcal{P}|} \psi_i \ln(\psi_i). \tag{5}$$

As an outcome of incorporating the term $\nu(\pi)$, the solution obtained by b_S will have a gap from the optimal strategy due to the tradeoff between exploration and exploitation of the smooth best response. Accordingly, we define such a shifted strategy as ϵ-optimal strategy.

Definition 2 (ϵ-optimal strategy): A mixed strategy $\pi^* = \{\psi_p\} \in \Delta\mathcal{P}$ is an ϵ-optimal strategy of the Stackelberg game if the following hold:

$$\left|\min_{\pi\in\Delta\mathcal{P}} C_S(\pi) - C_S(\pi^*)\right| \le \epsilon \tag{6}$$

where $\min_{\pi\in\Delta\mathcal{P}} C_S(\pi)$ is the optimal cost.

Algorithm 1: Learning-Based DR for Privacy-Preserving Users.

Initialization: Given the set of states, $\mathcal{S}$, and the incentive plan $\mathcal{P}$, randomly choose $p \in \mathcal{P}$ and set the initial perception values for $Q_S^0(p)$ and ψ_p^0, $\forall p \in \mathcal{P}$. Set parameter β and specify the timescale parameters γ^t and λ^t for both processes.
end initialization
loop for each episode
"Expected utility cost learning":
1) Compute the perceived system cost $U_S^t(p)$ according to (2) using users' perturbed aggregated responses L.
2) Update the estimation of the utility cost $Q_S^t(p)$ according to (7) specified by γ^t as learning rate.
"Stackelberg game incentive strategy learning":
if $t = kT+1$ **then**
3) Compute the smooth best response according to (8) using the estimation $Q_S^t(p)$.
4) Update the incentive strategy π_S according to (9) for each episode specified by λ^t as learning rate.
end if
end loop

We can see from (6) that any deviation of optimal cost $\min_{\pi \in \Delta \mathcal{P}} C_S(\pi)$ from the ϵ-optimal strategy cost $C_S(\pi^*)$ is bounded by ϵ. It should be stated that the value of ϵ depends on β due to smooth best response properties.

B. Two-Timescale Reinforcement Learning Algorithm

In this section, we propose a reinforcement learning algorithm, as outlined in Algorithm 1, to determine the optimal incentive strategy for the LSE. Note that the dynamics of the proposed Stackelberg game can be modeled as a stochastic fictitious play (SFP) [19]. However, since the LSE cannot observe users' actual responses, we consider a variant of SFP, where the LSE indirectly builds its own belief based on the expected utility cost (3), which depends on users' actions. Furthermore, since the users perturb their responses, the LSE's observation of users' responses is noisy. To tackle this challenge, we propose a two-timescale reinforcement learning algorithm consisting of a fast and a slow timescale learning processes. In the fast timescale, the LSE estimates the expected utility cost. Specifically, the LSE with action $p \in \mathcal{P}$ iteratively learns the expected utility from the perturbed observations of $U_S^t(p)$. When the fast timescale process is completed, the perturbed estimation $Q_S^t(p)$ of the expected utility is calibrated at the slow timescale, in which the LSE updates its incentive strategy using the smooth best response. Specifically, the learning process is divided into "episodes," as illustrated in Fig. 2. Let T be the duration of each episode. During each episode, the fast timescale learns the expected utility with a learning rate of γ^t at each step t. When the episode ends, the slow timescale updates the incentive strategy with the learning rate of λ^t at step $t = kT + 1$ where $k = 0, 1, \ldots$

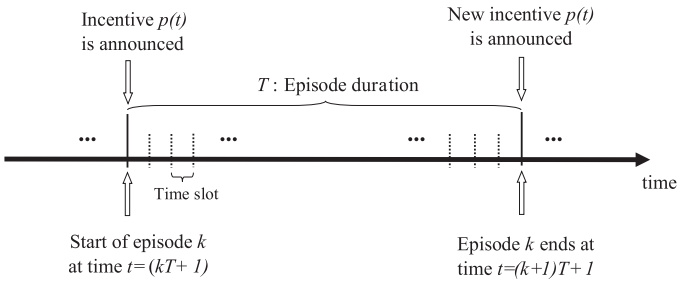

Fig. 2. Episode specified by fast and slow timescales.

1) Expected Utility Cost Learning: In the fast timescale, the LSE estimates the expected utility cost using the received users' responses. Specifically, the expected utility cost under each announced incentive $p \in \mathcal{P}$ and state S is estimated as follows:

$$Q_S^t(p) = Q_S^{t-1}(p) + \gamma^t I_{\{p_{\text{inc}}(t)=p\}} \left(U_S^t(p) - Q_S^{t-1}(p)\right) \quad (7)$$

where $Q_S^t(p)$ is the perception value of the expected utility cost at state S when p is the incentive chosen by the LSE at iteration t, and $I_{\{p_{\text{inc}}(t)=p\}}$ is the indicator function with $I_{\{X\}} = 1$ if the event X is true and $I_{\{X\}} = 0$ otherwise.

It is worth noting that at each iteration t, the LSE updates its instantaneous utility cost using (7). The term $U_S^t(p)$ is computed according to (2). Furthermore, it is shown in Theorem 1 that Algorithm 1 converges almost surely to an ϵ-optimal strategy.

2) Incentive Strategy Learning: In the slow timescale, the LSE updates the incentive strategy based on the estimated expected utility, which is learned in the fast timescale. Specifically, the LSE runs an SFP where its own strategy is updated based on smooth best response (4). The smooth best response of the LSE can be updated at time t according to Boltzmann distribution as follows:

$$b_S^t(p) = \frac{\exp(-\beta \cdot Q_S^t(p))}{\sum_{p' \in \mathcal{P}} \exp(-\beta \cdot Q_S^t(p'))} \quad \forall p \in \mathcal{P}. \quad (8)$$

Updating the incentive strategy assigns positive probabilities to all available incentives, which enables an exploration–exploitation tradeoff when searching for the best incentive strategy. Specifically, the incentive with a lower expected utility cost will be allocated with a larger probability, which represents the exploitation aspect of the tradeoff. Note that parameter β is used to control the tradeoff by tuning the probability assigned to each incentive. Intuitively, with a small β, the learning algorithm tends to explore the strategy space more to find the globally optimal solution. Then, the LSE updates its incentive strategy $\pi_S^t = \{\psi_p^t\}, p \in \mathcal{P}$, according to the computed smooth best response (8) at the end of each episode. For each $p \in \mathcal{P}$, we have the following:

$$\begin{cases} \psi_p^t = \psi_p^{t-1} + \lambda^t (b_S^t(p) - \psi_p^{t-1}), & \text{if } t = k \cdot T + 1 \\ \psi_p^t = \psi_p^{t-1}, & \text{otherwise} \end{cases} \quad (9)$$

where λ^t denotes the learning rate of the slow timescale. It is noteworthy that γ^t and λ^t should be calibrated correspondingly to ensure the convergence of Algorithm 1. To this end, we choose

γ^t and λ^t such that the following hold:

$$\lim_{t\to\infty} \frac{\lambda^t}{\gamma^t} = 0. \tag{10}$$

Based on (10), the first learning process would be processed faster than the second learning process. Thus, the current value of the slow timescale always can be adjusted with the outcome of the fast timescale.

C. Performance Analysis

In this section, we analyze the convergence of the proposed two-timescale reinforcement learning algorithm. As mentioned in Section III-B, the algorithm learns the expected utility of the LSE in the fast timescale and then updates its incentive strategy by the smooth best response in the slow timescale. Note that an ϵ-optimal strategy will be achieved asymptotically if both learning processes converge. Furthermore, the performance gap $\epsilon(\beta)$ between the expected utility cost at the ϵ-optimal point [i.e., $C_S(\pi^*)$] and the optimal expected utility [i.e., $\min_\pi C_S(\pi)$] is a function of parameter β. On the one hand, it is demonstrated in Theorem 2 that by choosing β as large enough, the performance gap will be reduced arbitrarily. On the other hand, for the proposed reinforcement learning algorithm, a small β is required to explore the strategy space to ensure the convergence of the algorithm to a globally optimal solution. The results for the performance behavior of Algorithm 1 are expressed in the following theorems. All proofs are relegated to the Appendix.

First, it is of paramount importance to show the general convergence property of the reinforcement learning algorithm where its strategy selection process is updated by Boltzmann distribution. When action p is selected during episode k, the LSE updates only the corresponding perception based on the perceived system cost in the current state $S(t) = S$ according to (7). Note that the LSE updates only the perception value in the current state under p and keeps the perceptions in other states unchanged. After learning the expected utility, the incentive in the next time slot is chosen based on the strategy $\pi_S^t = \{\psi_p^t(S)\}_{p\in\mathcal{P}}$, where $\psi_p^t(S)$ denotes the probability of choosing p at state S, and is updated based on (9). The smooth best response strategy in (9) is updated according to the Boltzmann distribution in (8). Note that the perception values would be different in different states, which, in turn, would result in different incentive strategies. However, they would converge to the expected utility value in all cases. We now characterize the optimality behavior of the proposed algorithm, which is presented in the following theorems.

Theorem 1: Algorithm 1 converges almost surely to an ϵ-optimal strategy in the game Γ if the following conditions hold:

C1: $\lim_{t\to\infty}\sum_{t\geq 0}\gamma^t = \infty, \lim_{t\to\infty}\sum_{t\geq 0}(\gamma^t)^2 < \infty$

C2: $\lim_{t\to\infty}\sum_{t\geq 0}\lambda^t = \infty, \lim_{t\to\infty}\sum_{t\geq 0}(\lambda^t)^2 < \infty$

C3: $\lim_{t\to\infty}\frac{\lambda^t}{\gamma^t} = 0$

Specifically, we have the following:

$$\lim_{t\to\infty} Q_S^t(p) = \bar{U}_S(p) \tag{11}$$

where $\bar{U}_S(p)$ is the expected utility cost of the LSE at state S and under its action p.

We use the convergence property of the martingale differences from stochastic approximation theory to show the convergence of the proposed two-timescale reinforcement learning algorithm. Moreover, a convex optimization problem is solved to show the optimality of the ϵ-optimal strategy in Theorem 2.

Theorem 2: For Algorithm 1, the incentive strategy of the LSE at the ϵ-optimal point minimizes the expected utility cost approximately, i.e.,

$$C_S(\pi^*) \leq \min_{\pi\in\Delta\mathcal{P}} C_S(\pi) + \epsilon(\beta) \tag{12}$$

where the approximation gap, $\epsilon(\beta)$, between the ϵ-optimal equilibrium and the minimum expected utility cost is at most $\frac{1}{\beta}\ln|\mathcal{P}|$.

Theorem 2 indicates that a large β is required to reduce the performance gap. However, increasing β would underrate the exploration aspect of the learning process, which may cause convergence to a suboptimal solution. In other words, a large β affects the system performance negatively by preventing the algorithm from finding the best incentive strategy due to overexploitation. Hence, a moderate β is required to strike a balance between optimality and convergence (i.e., exploration and exploitation) in order to achieve the best performance. It is noteworthy that the case $\epsilon = 0$ (i.e., no approximation gap) corresponds to the greedy mapping from perception to the policy space.

IV. Numerical Results

A. Data and Simulation Setting

In this section, the performance of the proposed algorithm is evaluated using 3000 independent load profiles generated by a domestic electricity demand model [37] in 1-min resolution. Note that each generated load profile is identical to different number of occupants, idiosyncratic behavior patterns, and running appliances on a weekday in April. A composite DR function [38] of linear, exponential, and logarithmic functions is considered to simulate the DR for each user, $s_n(t)$. Furthermore, the approximate range for the incentive rates is assumed to be $\mathcal{P} = \{0 : 0.02 : 0.2\}(\$/\text{kWh})$ with the flat retail rate of 0.15 ($\$/\text{kWh}$), and the wholesale price rates as shown in Fig. 4. Note that the LSE can estimate the supply price using historical and/or estimated demand and supply data [7]. Besides, the LSE can change the incentive ranges based on its requirements. The incentive rates are considered as discounts on the electricity retail rate for the participated users in the DR program. Since a flat load profile is a desirable output for the LSE considering technical requirements of the grid, the DR target [i.e., the desired load reduction $d(t)$ plus the actual aggregated load before incentives at each time step] is assumed to be 20% reduction during peak hours. The LSE can change the DR target based on its own technical requirements (e.g., operating reserve). It should be stated that at each episode, step size is 1 min, the number of steps is 60, and episode duration is 1 h. The learning rates, λ^t and γ_t, in Algorithm 1 are assumed to be $(t+1)^{-0.6}$ and t^{-1}, respectively so that conditions *C1–C3* are met. The LSE learns the expected utility cost during each episode using minute data and then up-

dates the incentive strategy at the end of the episode. Finally, it is noteworthy that the LSE has no knowledge of users' consumption behaviors and noise level. The users adjust the noise levels based on their own DR and privacy protection programs that are unavailable to the LSE.

We consider the use of differential privacy as a powerful tool to mask users' load profiles by adding noise into the real measurements [32]. In this paper, the adversary model follows a dishonest-but-nonintrusive model. The adversary aims to infer the detailed information about the occupants' activities (e.g., number of occupants and their consumption behaviors). In this sense, it needs to obtain the power consumption level of appliances, their periodicity, and the duration to extract complex usage patterns of households [31]. In what follows, we first provide a quick background of differential privacy and then investigate the proposed case studies under the assumption that differential privacy is the privacy protection mechanism adopted by the users.

Definition 3 (ε-differential privacy): Given datasets $\mathcal{M}_1$, $\mathcal{M}_2 \in \mathbb{M}$, a query $\mathcal{A}$ is ε-differential private if $\mathcal{M}_1$ and $\mathcal{M}_2$ differ in at most one element, and all subsets of possible answers follow $R \subseteq Range(\mathcal{A})$. We have the following:

$$P(\mathcal{A}(\mathcal{M}_1) \in R) \leq e^{\varepsilon} \cdot P(\mathcal{A}(\mathcal{M}_2) \in R) \tag{13}$$

where $P(\cdot)$ denotes the probability density for discrete random variables, and ε is a small value following $\ln(1+\varepsilon) \approx \varepsilon$.

This definition indicates that the results for any query over these two datasets differ up to a multiplicative factor e^{ε}. The parameter ε specifies the level of privacy. The lower value of ε represents stronger privacy. To implement the differential privacy on a dataset, we need to define the global sensitivity of a function. Global sensitivity is the maximum attainable change in the value of function f when its input differ only in one element.

Definition 4 (Global sensitivity): The global sensitivity of a function $f : \mathbb{M} \to \mathbb{R}^l$ is $V(f) = \max \|f(\mathcal{M}_1) - f(\mathcal{M}_2)\|_1$, where all pairs $\mathcal{M}_1, \mathcal{M}_2 \in \mathbb{M}$ differ in at most one element, $\|\cdot\|_1$ is the L_1 norm, and l denotes the number of independent Laplace variables.

It has been shown in [32] that to achieve differentially private output of function f, a random noise can be added to the value of f by calibrating the noise distribution to the global sensitivity of f. In our context, f is the set of the measurements from user n, and the sensitivity for the user is assumed to be its maximum consumption level. Note that the sensitivity can be defined with respect to any metric on the output space [32]. Simply put, user n can achieve ε-differential privacy by adding a random noise followed by Laplace distribution with scale parameter $V(f)/\varepsilon$, where $V(f)$ denotes the global sensitivity of f.

B. Case Studies

1) Load Profile Shaping: Fig. 3 shows the aggregated load profile before and after DR. The LSE learns the expected utility using high-resolution minute-by-minute data for each hour and then updates the incentive rate based on the perceived system cost for the next hour. In other words, learning the expected utility cost allows the LSE to remove the effects of the augmented noise generated by privacy-preserving mechanisms and announce the new incentive rates for the perturbed measurements. In this regard, considering the DR target, the LSE adjusts its incentive mechanism to achieve the desirable load profile. As illustrated in Fig. 4, if a higher reduction is required (e.g., $t = 19$), the LSE would announce a higher incentive to induce the customers' voluntary motives to curtail or shift their consumption level. Because of the penalty function in the LSE's cost function, LSE's incentive payments exceed electricity retail price in peak hours (e.g., $t = 19$) since it requires significant load reduction to flatten the aggregated load profile. However, lower incentive rates would be announced when a slight load reduction is required (e.g., $t = 9$ to 16). Hence, the LSE could use the incentive mechanism to smooth down the system's load profile, particularly in peak hours, which might cause technical issues (e.g., violating ramping constraints) at the supply side. Note that there are no incentives announced

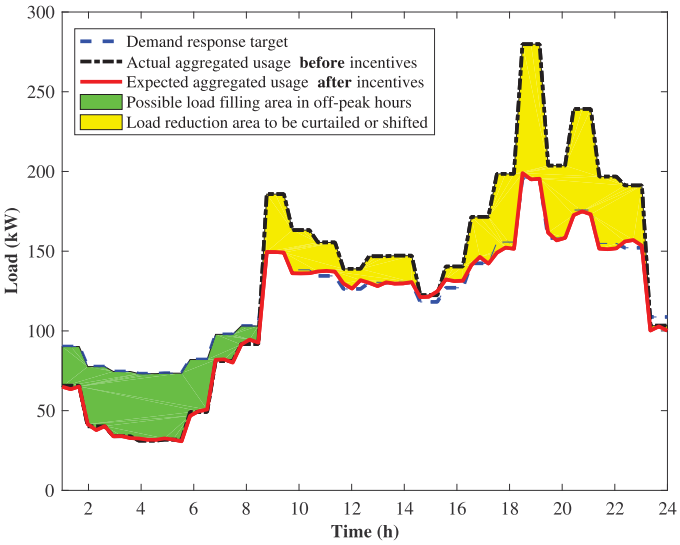

Fig. 3. Aggregated load profile before and after incentives.

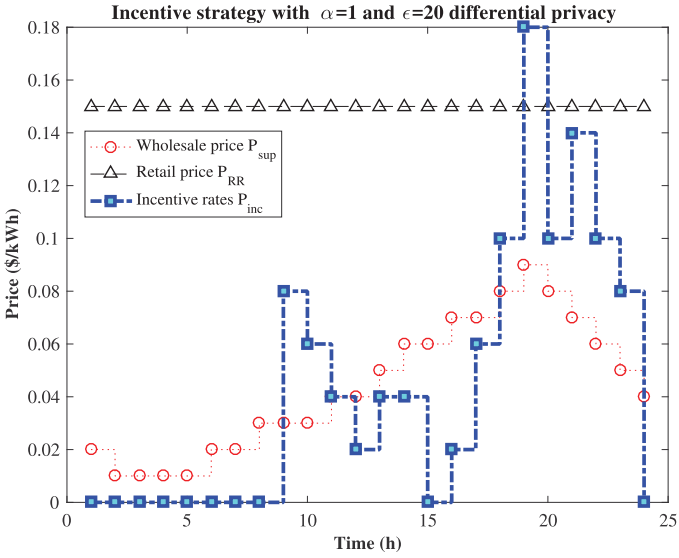

Fig. 4. Announced incentive rates in different time slots.

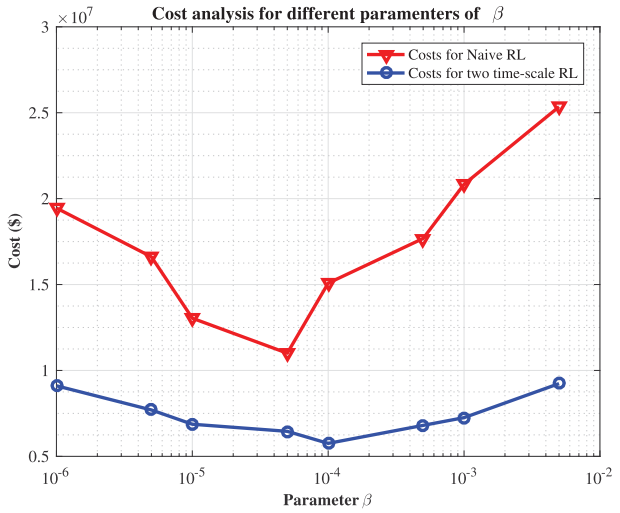

Fig. 5. System cost for different choices of β using the two-timescale algorithm and the naive reinforcement learning algorithm.

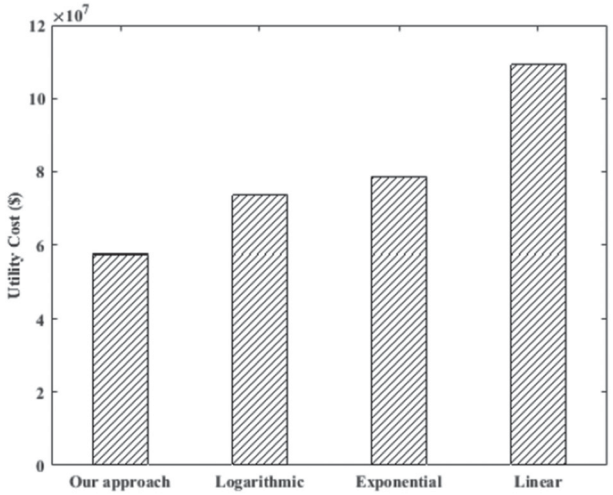

Fig. 6. Utility cost for different models of users responses.

during off-peak hours (e.g., $t = 1$ to 9) since no load reduction is favorable during these times. However, it can be implied that depending on the type of reduced loads (e.g., curtail-able or shift-able) during peak hours, it is possible that some of the shift-able loads can be shifted from peak hours (yellow shaded area in Fig. 3) to the off-peak hours (green shaded area in Fig. 3).

2) Cost Analysis: The performance of the proposed two-timescale reinforcement learning algorithm is evaluated by comparing the results with the naive reinforcement learning algorithm as proposed in [30]. The main difference between the two-timescale and naive reinforcement learning algorithms is that there is no expected utility cost learning step in the naive reinforcement learning algorithm. Simply put, the naive reinforcement learning algorithm is a one-timescale learning algorithm, which computes the system's perceived cost and then updates the incentive strategy using Boltzmann distribution at each time step. The main drawback for the naive algorithm is that if the load dynamics changes rapidly, it would be arduous for the naive algorithm to track down these variations and will result in a suboptimal incentive strategy.

The system costs using the two-timescale and the naive reinforcement learning algorithms are illustrated in Fig. 5 for different values of β. The results indicate that the obtained cost is decreased substantially by using the proposed two-timescale algorithm. It also implies that by calculating the expected utility cost, the proposed algorithm can track down the load variations efficiently and eliminate the effects of the aggregated noises to adjust the incentive strategy accordingly.

We also compare the results of different mathematical response functions as presented in [38]. The utility cost [i.e., (2)] using different response functions, i.e., our approach (composite function with random parameters), linear, logarithmic, and exponential, is illustrated in Fig. 6. The results show that the utility cost for composite function is less than the cost for other specified mathematical response functions. This implies that using specific mathematical response functions would result in a suboptimal solution and higher system cost due to mischaracterizing users' responses. Hence, it would be favorable to use the proposed reinforcement learning to learn the users' behaviors in order to reduce the system cost.

3) Impact of β: The performance of the two-timescale algorithm and the naive reinforcement learning algorithm using different β are illustrated in Fig. 5. The results indicate a tradeoff between the exploration (i.e., optimality) and exploitation (i.e., convergence) of the algorithms. In other words, on the one hand, more incentives are explored with a small β. which results in a large approximation gap according to Theorem 2. On the other hand, a large β may boil down the learning algorithm to a greedy paradigm, which can result in a suboptimal solution due to overexploitation. Simply put, the learning algorithm may fail to find the globally optimal solution because of the lack of exploration of the policy space. These intuitions stem from the mapping structure of the perception values to the policy space using Boltzmann distribution. Hence, a moderate β is required to achieve the best tradeoff between exploration and exploitation. As it is shown in Fig. 5, both algorithms achieve higher utility costs for small and large values of β. However, the utility cost improves as a moderate value of β is chosen. In this example, $\beta = 1e-4$ and $\beta = 5e-5$ are chosen for the two-timescale and the naive reinforcement learning algorithms, respectively, which strike a balance between the exploration and exploitation and yield the best performance. Note that the value of β can be calculated offline using historical data.

4) Impact of ε: It is of paramount importance to show the effect of parameter ε on the utility cost. Parameter ε allows the customers to control their privacy level. The smaller the value of ε, the more private the measurements are. However, a low ε may impose serious technical issues due to limitations in energy storage devices' capacity. In other words, if a low ε is chosen, the Laplace distribution's scale parameter would increase significantly. The noise augmentation in the case of a high scale parameter would be impractical because of storage capacity constraints [16]. Besides, it would be difficult for the LSE to learn the effect of the augmented noise due to higher variations. As it is shown in Fig. 7, the lower ε (i.e., the more private measurements) imposes slightly higher cost to the LSE.

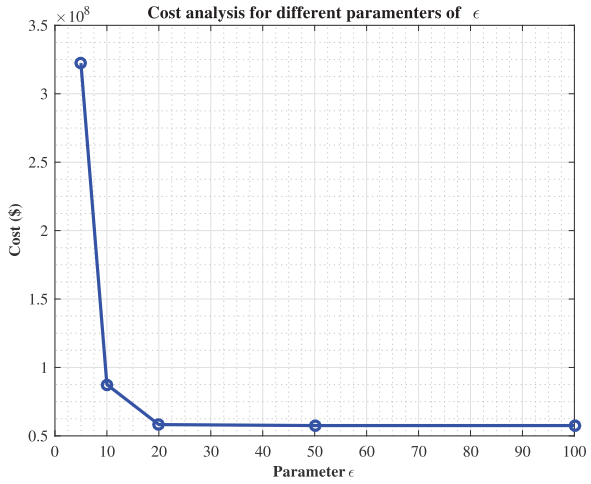

Fig. 7. System cost for different choices of ε using two-timescale reinforcement learning algorithm.

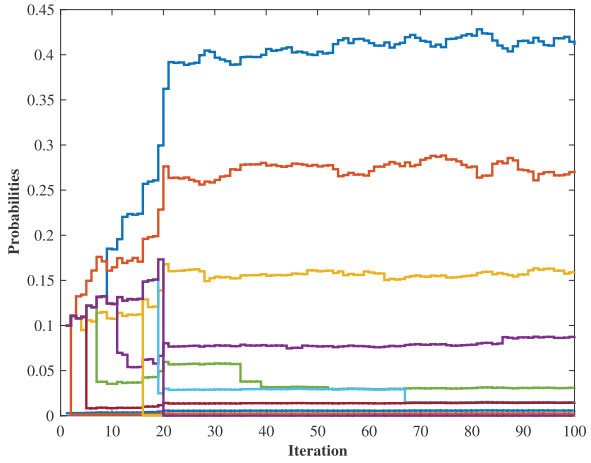

Fig. 8. Probabilities of incentives corresponding to the two-timescale reinforcement learning algorithm.

However, as the level of the privacy decreases, the utility cost degenerates to the nonprivacy preserved cost.

5) Convergence Speed: Historical data are used to obtain an ϵ-optimal strategy. Then, the available incentives in $\mathcal{P}$ are announced over time for each episode and then update the expected utility cost (7) and the corresponding incentive strategy (8). The incentive with a lower expected utility cost will gain a higher probability over time. The convergence speed of the proposed reinforcement learning algorithm is illustrated in Fig. 8. The probabilities of different incentives in the incentive set $\mathcal{P}$ are depicted for $\beta = 1e-4$ and $\varepsilon = 20$. It is shown that the probabilities are converged to an ϵ-optimal strategy in less than 20 iterations. The small variations in the probabilities after convergence are because of added perturbations. The perturbations change the value of the expected utility, which, in turn, result in modifications in the probabilities of incentives due to characteristics of the Boltzmann distribution, which tries to balance a tradeoff between exploration and exploitation. Furthermore, it is noteworthy that the converged probabilities are associated to a state. This means that the incentive probabilities would converge to different values in different states.

V. Conclusion

We have motivated and presented an optimal incentive scheme for privacy-preserved DR program from the LSE point of view. Specifically, we leveraged a two-timescale reinforcement learning algorithm to learn the perturbed behavior of customers to solve the optimal incentive strategy. The learning algorithm enables us to learn the users' privacy-preserving responses instead of using predefined response functions for all users. Besides, convergence and optimality analyses for the learning algorithm show that by choosing a moderate β, we can achieve the best performance from the perspective of cost. The efficacy of the proposed scheme is further evaluated using numerical case studies. The results showed that based on the defined objectives for the incentive scheme (i.e., desired load profile and minimum revenue loss for the LSE), the learning algorithm yields an ϵ-optimal incentive strategy in the presence of the perturbed measurements. The superior performance of the proposed algorithm was verified by comparing the results with those of other algorithms.

Appendix

Proof of Theorem 1: The proposed two-timescale reinforcement learning in Algorithm 1 can be expressed as a coupled stochastic approximation process with corresponding Lipshitz continuous functions F and martingale differences M as follows [39], [40]:

$$\begin{cases} Q_S^t(p) = Q_S^{t-1}(p) + \gamma^t\{F_Q(Q_S^{t-1}(p), \psi_p^{t-1}(S)) + M_Q^t\} \\ \psi_p^t(S) = \psi_p^{t-1}(S) + \lambda^t\{F_\psi(Q_S^{t-1}(p), \psi_p^{t-1}(S)) + M_\psi^t\} \end{cases} \tag{14}$$

where M_Q^t and M_ψ^t are the martingale differences, and F_Q and F_ψ are Lipshitz functions defined as below:

$$\begin{cases} F_Q(Q_S^{t-1}(p), \psi_p^{t-1}(S)) = \mathbb{E}\left[\dfrac{Q^t - Q^{t-1}}{\gamma^t} | Q^t, \psi^t\right] \\ F_\psi(Q_S^{t-1}(p), \psi_p^{t-1}(S)) = \mathbb{E}\left[\dfrac{\psi^t - \psi^{t-1}}{\lambda^t} | Q^t, \psi^t\right] \end{cases}. \tag{15}$$

We know from stochastic approximation theory that given the martingale differences, M_Q^t and M_ψ^t, and conditions, **C1** and **C2**, the sequences $\{\sum_{t=0}^k \gamma^t M_Q^t\}_k$ and $\{\sum_{t=0}^k \gamma^t M_\psi^t\}_k$ converge almost surely. Then, the discrete stochastic processes in (14) can be written as noisy discretization of the continuous ordinary differential equations (ODEs) according to stochastic approximation theory [39], [41] as follows:

$$\begin{cases} \dot{Q}_S^t(p) = F_{\dot{Q}}(Q^{t-1}, \psi^{t-1}) \\ \dot{\psi}_p^t(S) = F_\psi(Q^{t-1}, \psi^{t-1}) \end{cases}. \tag{16}$$

We first evaluate the fast learning process in (14). Assume that for each fixed strategy $\pi_S = (\psi_1, \ldots, \psi_{|\mathcal{P}|}) \in \Delta\mathcal{P}$, there exists a limiting expected utility cost, $Q_S(p)$, which is a unique globally

asymptotically stable point for the first ODE in (16). Then, according to (7), the first ODE can be written as follows:

$$\begin{aligned}\dot{Q}_S(p) &= \mathbb{E}_{\pi_S}\left[\frac{Q_S^t - Q_S^{t-1}}{\gamma^t}|Q^t, \psi^t\right] \\ &= \mathbb{E}_{\pi_S}\left[I_{\{p_{\text{inc}}(t)=p\}}\left(U_S^t(p) - Q_S^{t-1}(p)\right)\right] \\ &= \mathbb{E}_{\pi_S}\left[I_{\{p_{\text{inc}}(t)=p\}}\right]\left(\mathbb{E}_{\pi_S}[U_S^t(p)] - Q_S^{t-1}(p)\right) \\ &= \psi_p(S)\left(\bar{U}_S(p) - Q_S^{t'}(p)\right) \end{aligned} \tag{17}$$

where $t' \in (0, \infty)$ denotes a continuous time index. The solution of the derived ODE is as follows:

$$Q_S^{t'}(p) = \bar{U}_S(p) - \left(\bar{U}_S(p) - Q_S^0(p)\right) \cdot \exp(-\psi_p(S) \cdot t'). \tag{18}$$

Then, it follows that

$$\lim_{t' \to \infty} |Q_S^{t'}(p) - \bar{U}_S(p)| = 0. \tag{19}$$

The presented analysis indicates that the fast learning process (i.e., expected utility perception), $Q_S^t(p)$, will converge to a limiting expected cost $\bar{U}_S(p)$ at state S when incentive p is selected. In other words, the expected utility cost in (3) will be bounded to a limiting value of $C_S(\pi_S)$ with a fixed strategy π_S. Thereby, under condition **C3**, the convergence analysis of Algorithm 1 reduces to the convergence analysis of the second learning process (i.e., incentive strategy), which is a stochastic Stackelberg game between the LSE (i.e., leader) and the costumers (i.e., followers) by smooth best response.

The same analysis is performed for the slow learning process. The limiting behavior of the discrete stochastic process $\{\psi^t(S)\}$ in (14) turns out to be the same as the asymptotic behavior of the trajectories of the second ODE in (16), which describes the close form of the smooth best response dynamics [20] as follows:

$$\dot{\psi}_p^t(S) = b_S^t(p) - \psi_p^t(S). \tag{20}$$

The dynamic behavior of the smooth best response in the potential games has been studied in [36]. To show that the trajectories given by (20) converge to an approximate equilibrium in a potential game, we need to show the existence of a Lyapunov function for the cost function (2). Because of the fact that the considered action in the learning process is incentive, we can find a Lyapunov function for (2) by assuming integrability for $h(p)$ and $r_n(p)$, which indicates that the proposed game Γ is an exact potential game in long term, thereby establishing the convergence of Algorithm 1. Hence, the corresponding ODE for the smooth best response can be written as follows:

$$\dot{\psi}_p(S) = b_S(p) - \psi_p(S) \tag{21}$$

which converges to the associated zero point of the ODE. Then, it follows that

$$\psi_p(S) = b_S(p) \tag{22}$$

which proves that the convergence point is an ϵ-optimal strategy of game Γ. ■

Proof of Theorem 2: First, we need to form an optimization problem based on the properties of the smooth best response (4) that balances between incentive strategy exploitation and exploration. Thus, we consider the following problem:

$$\min_{\pi \in \Delta\mathcal{P}} \left[C_S(\pi) - \frac{1}{\beta}\nu_u(\pi)\right]. \tag{23}$$

The first term in (23) indicates the performance of the incentive strategy (i.e., exploitation) while the second term represents the entropy, which measures the randomness of the incentive strategy (i.e., exploration). In other words, the minimization problem (23) tries to find the best tradeoff between the incentive exploitation and exploration.

Given the smooth best response as follows:

$$\nu_u(\pi) = -\sum_{p \in \mathcal{P}} \psi_p \ln(\psi_p) \tag{24}$$

we have the following:

$$\frac{1}{\beta}\sum_{p \in \mathcal{P}} \psi_p \ln(\psi_p) \leq 0$$

since $\ln(\psi_p) \leq 0 \ \ \forall p \in \mathcal{P}$. Then, we have the following:

$$\min_{\pi \in \Delta\mathcal{P}} C_S(\pi) \geq \min_{\pi \in \Delta\mathcal{P}} \left(C_S(\pi) + \frac{1}{\beta}\sum_{p \in \mathcal{P}} \psi_p \ln(\psi_p)\right). \tag{25}$$

Since the uniform distribution results in the maximum entropy, the following can be shown:

$$\min_{\pi \in \Delta\mathcal{P}} C_S(\pi) \geq C_S(\pi^*) - \frac{1}{\beta}\ln|\mathcal{P}| \tag{26}$$

which directly leads to (12) with $\epsilon(\beta) = \frac{1}{\beta}\ln|\mathcal{P}|$. Then, the theorem follows. ■

REFERENCES

[1] B. P. Bhattarai *et al.*, "Design and cosimulation of hierarchical architecture for demand response control and coordination," *IEEE Trans. Ind. Inform.*, vol. 13, no. 4, pp. 1806–1816, Aug. 2017.

[2] F. Y. Xu and L. L. Lai, "Novel active time-based demand response for industrial consumers in smart grid," *IEEE Trans. Ind. Inform.*, vol. 11, no. 6, pp. 1564–1573, Dec. 2015.

[3] D. Li, W. 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.

[4] P. Centolella, "The integration of price responsive demand into regional transmission organization (RTO) wholesale power markets and system operations," *Energy*, vol. 35, no. 4, pp. 1568–1574, 2010.

[5] P. Faria and Z. Vale, "Demand response in electrical energy supply: An optimal real time pricing approach," *Energy*, vol. 36, no. 8, pp. 5374–5384, 2011.

[6] D. Fischer *et al.*, "Modeling the effects of variable tariffs on domestic electric load profiles by use of occupant behavior submodels," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2685–2693, Nov. 2017.

[7] H. Zhong, L. Xie, and Q. Xia, "Coupon incentive-based demand response: Theory and case study," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1266–1276, May 2013.

[8] A. Asadinejad and K. Tomsovic, "Optimal use of incentive and price based demand response to reduce costs and price volatility," *Elect. Power Syst. Res.*, vol. 144, pp. 215–223, 2017.

[9] A. Ehsanfar and B. Heydari, "An incentive-compatible scheme for electricity cooperatives: An axiomatic approach," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1416–1424, Mar. 2018.

[10] K. H. S. V. S. Nunna and S. Doolla, "Responsive end-user-based demand side management in multimicrogrid environment," *IEEE Trans. Ind. Inform.*, vol. 10, no. 2, pp. 1262–1272, May 2014.

[11] E. L. Quinn, "Smart metering and privacy: Existing laws and competing policies," *A Report for the Colorado Public Utilities Commission*, May 2009.

[12] G. Kalogridis, C. Efthymiou, S. Z. Denic, T. A. Lewis, and R. Cepeda, "Privacy for smart meters: Towards undetectable appliance load signatures," in *Proc. IEEE 1st Int. Conf. Smart Grid Commun.*, 2010, pp. 232–237.
[13] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," *IEEE Secur. Privacy*, vol. 8, no. 1, pp. 11–20, Jan.–Feb. 2010.
[14] S. McLaughlin, P. McDaniel, and W. Aiello, "Protecting consumer privacy from electric load monitoring," in *Proc. 18th ACM Conf. Comput. Commun. Secur.*, 2011, pp. 87–98.
[15] Z. Chen and L. Wu, "Residential appliance DR energy management with electric privacy protection by online stochastic optimization," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 1861–1869, Dec. 2013.
[16] L. Yang, X. Chen, J. Zhang, and H. V. Poor, "Cost-effective and privacy-preserving energy management for smart meters," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 486–495, Jan. 2015.
[17] J. Zhao, T. Jung, Y. Wang, and X. Li, "Achieving differential privacy of data disclosure in the smart grid," in *Proc. IEEE INFOCOM*, 2014, pp. 504–512.
[18] L. Yang, X. Chen, J. Zhang, and H. V. Poor, "Optimal privacy-preserving energy management for smart meters," in *Proc. IEEE INFOCOM*, 2014, pp. 513–521.
[19] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, vol. 2. Cambridge, MA, USA: MIT Press, 1998.
[20] D. S. Leslie *et al.*, "Convergent multiple-timescales reinforcement learning algorithms in normal form games," *Ann. Appl. Probability*, vol. 13, no. 4, pp. 1231–1251, 2003.
[21] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE Trans. Ind. Inform.*, vol. 7, no. 3, pp. 381–388, Aug. 2011.
[22] F. Li, B. Luo, and P. Liu, "Secure information aggregation for smart grids using homomorphic encryption," in *Proc. IEEE 1st Int. Conf. Smart Grid Commun.*, Oct. 2010, pp. 327–332.
[23] D. Seo, H. Lee, and A. Perrig, "Secure and efficient capability-based power management in the smart grid," in *Proc. IEEE 9th Int. Symp. Parallel Distrib. Process. Appl. Workshops*, May 2011, pp. 119–126.
[24] H. Li, X. Lin, H. Yang, X. Liang, R. Lu, and X. Shen, ''EPPDR: An efficient privacy-preserving demand response scheme with adaptive key evolution in smart grid," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 8, pp. 2053–2064, Aug. 2014.
[25] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "EPPA: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 9, pp. 1621–1631, Sep. 2012.
[26] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Basar, "Dependable demand response management in the smart grid: A Stackelberg game approach," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 120–132, Mar. 2013.
[27] D. O'Neill, M. Levorato, A. Goldsmith, and U. Mitra, "Residential demand response using reinforcement learning," in *Proc. IEEE 1st Int. Conf. Smart Grid Commun.*, Oct. 2010, pp. 409–414.
[28] Z. Wen, D. ONeill, and H. Maei, "Optimal demand response using device-based reinforcement learning," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2312–2324, Sep. 2015.
[29] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Appl. Energy*, vol. 220, pp. 220–230, 2018.
[30] A. Ghasemkhani and L. Yang, "Reinforcement learning based pricing for demand response," in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2018, pp. 1–6.
[31] G. Ács and C. Castelluccia, "I have a dream! (differentially private smart metering)." in *Information Hiding*, vol. 6958. New York, NY, USA: Springer-Verlag, 2011, pp. 118–132.
[32] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptographic Conference*, vol. 3876. New York, NY, USA: Springer-Verlag, 2006, pp. 265–284.
[33] H. Chao, "Demand response in wholesale electricity markets: The choice of customer baseline," *J. Regulatory Econ.*, vol. 39, no. 1, pp. 68–88, Feb. 2011.
[34] Z. Liu, I. Liu, S. Low, and A. Wierman, "Pricing data center demand response," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 42, no. 1, pp. 111–123, 2014.
[35] J. Yao, S. S. Oren, and I. Adler, "Two-settlement electricity markets with price caps and Cournot generation firms," *Eur. J. Oper. Res.*, vol. 181, no. 3, pp. 1279–1296, 2007.
[36] J. Hofbauer and E. Hopkins, "Learning in perturbed asymmetric games," *Games Econ. Behav.*, vol. 52, no. 1, pp. 133–152, 2005.
[37] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high-resolution energy demand model," *Energy Buildings*, vol. 42, no. 10, pp. 1878–1887, 2010.
[38] S. Yousefi, M. P. Moghaddam, and V. J. Majd, "Optimal real time pricing in an agent-based retail market using a comprehensive demand response model," *Energy*, vol. 36, no. 9, pp. 5716–5727, 2011.
[39] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. New York, NY, USA: Springer-Verlag, 2003.
[40] M. Zhang, L. Yang, D.-H. Shin, X. Gong, and J. Zhang, "Privacy-preserving database assisted spectrum access: A socially-aware distributed learning approach," in *Proc. IEEE GLOBECOM,*, 2015, pp. 1–6.
[41] M. Benaïm, "Dynamics of stochastic approximation algorithms," *Séminaire de Probabilités, XXXIII*, vol. 1709, pp. 1–68, 1999.

Amir Ghasemkhani (S'17) received the M.Sc. degree in electrical engineering (power systems) from the University of Tehran, Tehran, Iran, in 2014. He is currently working toward the Ph.D. degree in computer science and engineering at the University of Nevada, Reno, NV, USA.

His research interests include data analytics and stochastic optimization in power systems and smart grid.

Lei Yang (M'13) received the B.S. and M.S. degrees in electrical engineering from Southeast University, Nanjing, China, in 2005 and 2008, respectively, and the Ph.D. degree from the School of Electrical Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA, in 2012.

He was a Postdoctoral Scholar with Princeton University, Princeton, NJ, USA, and an Assistant Research Professor with the School of Electrical Computer and Energy Engineering, Arizona State University. He is currently an Assistant Professor with the Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA. His research interests include big data analytics, stochastic optimization and modeling in smart cities and cyber-physical systems, data privacy and security in crowdsensing, and optimization and control in mobile social networks.

Prof. Yang was the recipient of the Best Paper Award Runner-up at the IEEE INFOCOM 2014.

Junshan Zhang (S'98–M'00–SM'06–F'12) received the Ph.D. degree from the School of Electrical and Computer E, Purdue University, West Lafayette, IN, USA, in 2000.

He joined the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA, in August 2000, where he has been a Fulton Chair Professor since 2015. His research interests fall in the general field of information networks and data science, including communication networks, Internet of Things (IoT), fog computing, social networks, and smart grid. His current research focuses on fundamental problems in information networks and data science, including fog computing and its applications in IoT and 5G, IoT data privacy/security, optimization/control of mobile social networks, and stochastic modeling and optimization for smart grid.

Prof. Zhang was the recipient of the ONR Young Investigator Award in 2005 and the NSF CAREER Award in 2003. He was the recipient of the IEEE Wireless Communication Technical Committee Recognition Award in 2016. His papers have won a few awards, including the Best Student paper at WiOPT 2018, the Kenneth C. Sevcik Outstanding Student Paper Award of ACM SIGMETRICS/IFIP Performance 2016, the Best Paper Runner-up Award of IEEE INFOCOM 2009 and IEEE INFOCOM 2014, and the Best Paper Award at IEEE ICC 2008 and ICC 2017. Building on his research findings, he cofounded Smartiply, Inc., in 2015, a fog computing startup company delivering boosted network connectivity and embedded artificial intelligence. He was TPC Co-chair for a few major conferences in communication networks, including IEEE INFOCOM 2012 and ACM MOBIHOC 2015. He was General Chair for ACM/IEEE SEC 2017, WiOPT 2016, and IEEE Communication Theory Workshop 2007. He is currently serving as the Editor-in-Chief for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the Senior Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING.