Energy Management Strategy for Smart Meter Privacy and Cost Saving

Yang You, *Student Member, IEEE,* Zuxing Li, *Member, IEEE,* and Tobias J. Oechtering, *Senior Member, IEEE*

***Abstract*—We design optimal privacy-enhancing and cost-efficient energy management strategies for consumers that are equipped with a rechargeable energy storage. The Kullback-Leibler divergence rate is used as privacy measure and the expected cost-saving rate is used as utility measure. The corresponding energy management strategy is designed by optimizing a weighted sum of both privacy and cost measures over a finite time horizon, which is achieved by formulating our problem into a belief-state Markov decision process problem. A computationally efficient approximated Q-learning method is proposed as a generalization to high-dimensional problems over an infinite time horizon. At last, we explicitly characterize a stationary policy that achieves the steady belief state over an infinite time horizon, which greatly simplifies the design of the privacy-preserving energy management strategy. The performance of the practical design approaches are finally illustrated in numerical experiments.**

***Index Terms*—Smart meter privacy, privacy-utility trade-off, Kullback-Leibler divergence, MDP, Q-learning.**

I. Introduction

In future smart grids, smart meters are essential components to deliver information about consumers' energy demand to the energy provider (EP). This information can help the EP to improve the prediction on the future energy demands and therefore to increase the efficiency of the whole smart grid [1]. However, this benefit is at a cost of privacy of consumers, since an adversary (this could be a legitimate receiver of the data, e.g., the energy grid operator) can use standard energy load disaggregation algorithms, e.g., non-intrusive load monitoring algorithms [2]–[6] to learn the consumers' private activities.

Regarding this issue, different approaches have been proposed previously. One approach is to modify the smart metering data before it is sent to the EP by using of-the-shelf methods, such as obfuscation [7], anonymization [8], and data aggregation [9]. The major limitation of these methods is that they hide the real energy flow in the grid so that these methods fail if the legitimate receiver of the data requires exact measurements. Moreover, the adversary (even a compromised EP) may decide to install a sensor for directly monitoring the energy request of a household or a business. In the United States, in the case of Naperville Smart Meter Awareness v. City of Naperville [10], the court has decided that the Fourth Amendment [11] protects the energy consumption data collected by smart meters. The EU General Data Protection Regulation (GDPR) [12] calls for an authorized data recipient to hold and process only the data absolutely necessary for the completion of its duties as well as limiting the access to personal data to those needed to act out the processing [12]. To achieve this, GDPR advocates for innovative privacy-by-design approaches as considered here. Using an energy storage such as rechargeable battery [13]–[17], or an alternative energy supply such as renewable energy source [18]–[20], the actual consumer profile can be modified by a privacy-enhancing energy management strategy.

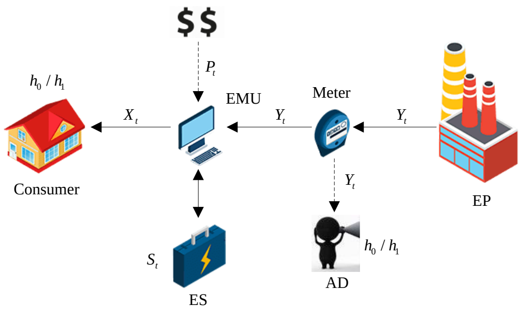


Fig. 1. Smart metering system with rechargeable energy storage (ES) and privacy-cost-aware energy management unit (EMU) that modifies energy consumption profile to protect against unauthorized hypothesis testing (h_0/h_1) of adversary (AD) taking dynamic energy prices into account.

A. Related Works

Privacy measures. Different privacy measures have been considered for such privacy-by-design approaches. In [21] and [22], differential privacy and in [16], the variance of random energy supplies from the EP have been used as privacy measure. In [13], [14], [18], [19], [23], the privacy leakage is measured by different information theoretic metrics such as mutual information or conditional entropy rates. Another important approach is to consider a privacy-preserving objective derived from a specific adversarial hypothesis test scenario, e.g., [24], [25]. Compared to the aforementioned approaches, the hypothesis testing privacy measure has a clear operational meaning, but it is also limited by the specific assumptions of the considered scenario. In more detail, in [24], the privacy-preserving problem is evaluated under Bayesian detection setting. The work has been extended in [25] considering the trade-off between privacy and energy storage loss. Recently, the Kullback-Leibler (KL) divergence is used to measure privacy leakage in [26] and [27]. The KL divergence characterizes the asymptotic Neyman-Pearson hypothesis testing performance of the adversary with independent and identically distributed (i.i.d.) observations. In this work, we also adopt

Y. You, Z. Li and T. J. Oechtering are with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: youy@kth.se; zuxing@kth.se; oech@kth.se).

the KL-divergence as the privacy measure. Correspondingly, the objective is to minimizing the KL-divergence between the distributions of energy request considering a binary hypothesis test on the consumers' behavior.

Design approaches. Different approaches have been proposed for the design of privacy-preserving mechanisms. This includes heuristic approaches, such as the best-effort water-filling algorithm in [28] that aims to keep the output load at its most recent value. A battery-based noise adding mechanism is designed in [21] and [22] to achieve differential privacy. In [29] and [30], control optimization methods such as model-distribution predictive control and cloud-based control have been applied to enhance the privacy. Likewise, a stochastic control model is considered in [13], [14], [18], [19], where the privacy-preserving energy management design problem is transferred into an optimal control strategy design problem in the Markov decision process (MDP) framework. In particular, [16], [17], [31] proposed different online control algorithms based on the dynamic programming framework considering the realistic case where the consumers' energy consumption profile can only be known casually.

Privacy-cost tradeoffs. In fact, in most cases, the privacy enhancement will lead to an increased energy cost, which may often violate the original cost-saving motivation of the energy storage investment for consumers. In such cases, the design of a privacy-enhancing and cost-efficient energy management strategy becomes even more important. While most of the aforementioned papers focus on how to preserve the privacy, only [14], [16], [17], and [19] have taken the consumers' cost for purchasing the energy into consideration. [14] and [19] assume in their design that the statistics of the energy profile is known. [16] provides an online strategy and [17] studies the case where the realization of the energy profile is non-causally known as well as an online strategy. In more details, in [16] the online dynamic programming problem is relaxed to a Lyapunov optimization problem which jointly optimize the privacy and the energy cost. In [17], first an offline convex optimization problem for the privacy-cost trade-off is solved, then a low-complexity heuristic online control algorithm is proposed as an alternative solution to the original online dynamic programming problem. Along a different line, the authors in [14] and [19] formulate the privacy-cost trade-off problem into the offline stochastic control problem under the MDP framework. However, characterizing the optimal strategy is computationally challenging due to the continuous state-action space. Thus approximate solutions under specific cases, or upper and lower bounds have been derived and proposed.

B. Contributions

In this paper, we consider a smart metering system with a rechargeable energy storage at the consumer using KL-divergence rate as the privacy leakage measure assuming that the statistics of the energy profile is known. For the cost measure, we use the expected cost-saving rate as defined in [14]. In order to design policies that optimally trade-off between privacy and cost, a weighted sum of them is considered as overall objective function. We formulate the energy management design problem as an equivalent average reward MDP problem so that the optimal solution is given by a Bellman dynamic program. The main purpose and contribution of this paper is then to develop computationally complexity efficient design approaches to circumvent the computational complexity that come with solving infinite horizon belief state MDP problems. In more detail, we utilize the techniques of reinforcement learning to propose a sub-optimal but computationally efficient approximated Q-learning method. As an alternative approach, we explicitly characterize a stationary policy that achieves the steady belief state over an infinite time horizon by assuming a simplified setting with i.i.d. energy demands. As a consequence, the original privacy measurement, i.e., n-letter KL-divergence can be simplified to a single-letter expression, which allows the derivation of a sufficient condition for perfect privacy.

Accordingly, the contribution of this paper can be accordingly summarized as follows: (i) We provide a problem formulation for the design of an energy management strategy that aims for privacy enhancement and energy cost-saving, and we specifically use KL-divergence as a novel privacy measurement; (ii) We provide the MDP reformulation of our original optimization problem such that the MDP framework can be utilized for the explicit design of optimal strategies; (iii) A more computational efficient approximated Q-learning approach is proposed as a generalization to the high-dimensional problem under infinite time horizon; (iv) A stationary energy management strategy is provided under the special case of i.i.d. energy demand without a cost-saving concern. The results are developed and motivated by the smart meter privacy problem, but can be transferred to other settings where a demand profile should be protected and one has the opportunity to modify the instantaneous demand within same range.

Notation: In the following, we denote a random variable by the capital letter, its realization by the corresponding lowercase letter, and its alphabet by the corresponding calligraphic letter. We further denote a random sequence $(X_t, ..., X_{t+k})$ and its realization $(x_t, ..., x_{t+k})$ by X_t^{t+k} and x_t^{t+k} respectively. In particular, X^t stands for X_1^t. For probability mass function $P_X(x)$ of random variable X, we write it as $P(x)$ if it is clear from the context. We use $P_{A|h_i}$ and $P_{A|B,h_i}$ as the short notations for the (conditional) distributions when the hypothesis h_i holds. Also, let $D(\cdot||\cdot)$ denote the KL-divergence.

II. Privacy-Cost trade-off under Markovian energy demand and price

A. System Model

Consider a smart metering system as shown in Fig. 1. The consumer's privacy-sensitive behavior over a certain time period T is modeled by a binary hypothesis h_0 or h_1, e.g., cooking during Ramadan, working on Shabbat or using a health equipment that reveals a disease. Under each hypothesis, the consumer will have a certain energy consumption profile. Time and energy levels are assumed to be discretized. At time step t, denote consumers' energy demand by $x_t \in \mathcal{X} = \{0, 1, ..., x_{max}\}$, energy supply from the EP by

$y_t \in \mathcal{Y} = \{0, 1, ..., y_{max}\}$, instantaneous price by $p_t \in \mathcal{P} = \{1, ..., p_{max}\}$. The energy storage (ES), e.g., a rechargeable battery, has a finite capacity with its instantaneous storage level denoted by $s_t \in \mathcal{S} = \{0, 1, ..., s_{max}\}$. The instantaneous energy consumption x_t should always be satisfied by supplies from either EP or ES without wasting energy. Then, the ES level evolves as

$$s_{t+1} = s_t + y_t - x_t. \tag{1}$$

In addition, to guarantee $0 \leq s_t \leq s_{max}$, the energy supply y_t should be chosen within the following feasible set:

$$\begin{aligned} &\overline{\mathcal{Y}}(\, x_t, s_t) \\ &= \{y_t \in \mathcal{Y} : \max\{0, x_t - s_t\} \leq y_t \leq s_{max} + x_t - s_t\}, \end{aligned} \tag{2}$$

where the lower bound $x_t - s_t$ ensures that the energy supply y_t provides at least the rest energy when ES level s_t cannot solely satisfy the consumer demand; and the lower bound 0 is because no energy can be sold back to the grid; the upper bound is due to the constraints of finite maximum ES capacity and that no energy should be wasted.

For the design of the policies, we employ a probablistic approach. We assume that the consumer energy demand X_t and the dynamic price P_t follow first order Markov processes with time-invariant transition probabilities $P_{X_{t+1}|X_t,h_i}, i \in \{0, 1\}$, and $P_{P_{t+1}|P_t}$. Over a T-time horizon, the energy management unit (EMU) requests energy supply Y_t from the EP based on an energy management strategy $f = \{f_t\}_{t=1}^{T} \in \mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times .... \times \mathcal{F}_T$, with $f_t \in \mathcal{F}_t$. The set $\mathcal{F}_t$ of the possible strategies is given by the set of conditional probability mass functions (pmfs):

$$\mathcal{F}_t = \{P_{Y_t|X^t,S^t,P^t,Y^{t-1}} : \sum_{\substack{y_t \in \overline{\mathcal{Y}}_t(x_t,s_t) \\ \forall x_t \in \mathcal{X}, s_t \in \mathcal{S}}} P(y_t|x^t, s^t, p^t, y^{t-1}) = 1\}. \tag{3}$$

Thus, strategy f_t decides on the random amount of energy to request from the EP at time t, given the demands x^t, ES levels s^t, prices p^t and supplies y^{t-1}.

For $i \in \{0, 1\}$, after initializing the joint pmf of X_1, S_1 and P_1 as $P_{X_1,S_1,P_1|h_i}$, over a finite horizon with length T, the joint conditional pmf of (X^T, S^T, Y^T, P^T) induced by f can be written as:

$$\begin{aligned} &P^f_{X^T,S^T,Y^T,P^T|h_i}(x^T, s^T, y^T, p^T) \\ &= \underbrace{P_{X_1,S_1,P_1|h_i}(x_1, s_1, p_1)}_{Initialization} \times \underbrace{P(y_1|x_1, s_1, p_1)}_{f_1(y_1|x_1,s_1,p_1)} \\ &\prod_{t=1}^{T-1} [\underbrace{P(p_{t+1}|p_t)}_{Price\ evolution} \times \underbrace{P(x_{t+1}|x_t, h_i)}_{Demand\ evolution} \\ &\times \quad \underbrace{\mathcal{I}_{s_{t+1}}(y_t + s_t - x_t)}_{Energy\ storage\ level\ evolution} \\ &\times \underbrace{P(y_{t+1}|x^{t+1}, s^{t+1}, p^{t+1}, y^t)}_{f_{t+1}(y_{t+1}|x^{t+1},s^{t+1},p^{t+1},y^t)}], \end{aligned} \tag{4}$$

where $\mathcal{I}$ is the indicator function, i.e., $\mathcal{I}_{s_{t+1}}(y_t + s_t - x_t) = 1$ if $s_{t+1} = y_t + s_t - x_t$, $\mathcal{I}_{s_{t+1}}(y_t + s_t - x_t) = 0$, otherwise.

B. Privacy-by-Design Approach

We assume that an adversary (AD) has access to the smart metering data sequence y^T, price sequence p^T, and is fully informed about the statistics of the system, i.e., $P_{Y^T,P^T|h_0}$ and $P_{Y^T,P^T|h_1}$, and infers on the consumers' privacy-sensitive consumption behavior using statistical inference methods. In our problem, due to the uncertainty about the inference behavior of the AD, we use the KL-divergence rate as privacy leakage measure, since the KL-divergence measures the similarity between two distributions. Over a finite time horizon T, given a strategy $f \in \mathcal{F}$, the privacy leakage is measured by the following KL-divergence rate between joint pmfs of (Y^T, P^T) conditioned on hypotheses h_0 and h_1:

$$L_T(f) = \frac{1}{T} D(P^f_{Y^T,P^T|h_0} \| P^f_{Y^T,P^T|h_1}), \tag{5}$$

where $P^f_{Y^T,P^T|h_i}$, for $i = 0, 1$, denotes the joint distribution of (Y^T, P^T) conditioned on h_i induced by f:

$$P^f_{Y^T,P^T|h_i}(y^T, p^T) = \sum_{x^T,s^T} P^f_{X^T,S^T,Y^T,P^T|h_i}(x^T, s^T, y^T, p^T). \tag{6}$$

Besides the privacy enhancement objective, we are looking for a policy f that is also cost-efficient. We define the cost-saving at time t as $\Delta V_t = (X_t - Y_t)P_t$. The expected cost-saving rate induced by f over a finite horizon T can then be written as:

$$V_T(f) = \frac{1}{T} \sum_{t=1}^{T} (\mathbb{E}[\Delta V_t|h_0]P(h_0) + \mathbb{E}[\Delta V_t|h_1]P(h_1)), \tag{7}$$

where the expectation is taken with respect to the joint conditional distribution $P^f_{X_t,Y_t,P_t|h_i}$, for $i = \{0, 1\}$, induced by f.

To study the trade-off between privacy enhancement and cost-saving, the overall objective is to choose a strategy $f \in \mathcal{F}$ that minimizes the following weighted sum objective:

$$C_T(f, \lambda) = \lambda L_T(f) - (1 - \lambda)V_T(f), \tag{8}$$

where $\lambda \in [0, 1]$. In more detail, the trade-off between privacy and cost is realized by choosing different values of λ, e.g., $\lambda = 1$ leads to finding the optimal privacy-enhancing strategy, while $\lambda = 0$ leads to the objective function of finding the optimal cost-saving strategy. Then, the optimal strategy is

$$f^* = \arg\min_{f \in \mathcal{F}} C_T(f, \lambda). \tag{9}$$

C. MDP Formulation

By iteratively applying the chain rule for KL-divergence, $L_T(f)$ can be written in the following form:

$$\begin{aligned} L_T(f) &= \frac{1}{T} \sum_{t=1}^{T} D(P^f_{Y_t,P_t|h_0,Y^{t-1},P^{t-1}} \| P^f_{Y_t,P_t|h_1,Y^{t-1},P^{t-1}}) \\ &= \frac{1}{T} \sum_{t=1}^{T} \sum_{y^t} \sum_{p^t} P^f(y^t, p^t|h_0) \log \frac{P^f(y_t, p_t|h_0, y^{t-1}, p^{t-1})}{P^f(y_t, p_t|h_1, y^{t-1}, p^{t-1})}. \end{aligned} \tag{10}$$

Meanwhile, the $V_T(f)$ can be expressed by:

$$\begin{aligned}V_T(f) =& \frac{1}{T}\sum_{t=1}^{T}\sum_{x_t,p_t,y_t}(P(h_0)P^f(x_t,p_t,y_t|h_0)\\ &+P(h_1)P^f(x_t,p_t,y_t|h_1))(x_tp_t - y_tp_t).\end{aligned} \tag{11}$$

As observed from (10) and (11), at each time step t, the current choice of the energy management strategy will affect the future statistics of the smart metering system as well as the choices of the future energy management strategies. In this case, successive independent optimization of the energy management strategy at each time step will not necessarily lead to an optimal solution over the whole time horizon. Instead, our optimization problem needs to be formulated into a sequential decision making problem, e.g., Markov decision process.

We first identify a structural simplification for the energy management strategy f. Define a new strategy $f^{'} = \{f_t^{'}\}_{t=1}^{T} \in \mathcal{F}^{'} = \mathcal{F}_1^{'} \times \mathcal{F}_2^{'} \times .... \times \mathcal{F}_T^{'}$, with $f_t^{'} \in \mathcal{F}_t^{'}$. And $\mathcal{F}_t^{'}$ is the set of pmfs that:

$$\mathcal{F}_t^{'} = \{P_{Y_t|X_t,S_t,P^t,Y^{t-1}} : \sum_{\substack{y_t\in\overline{\mathcal{Y}}_t(x_t,s_t)\\ \forall x_t\in\mathcal{X},s_t\in\mathcal{S}}} P(y_t|x_t,s_t,p^t,y^{t-1}) = 1\}. \tag{12}$$

Thus, $f_t^{'}$ is the energy management strategy that decides on the energy supply based on the current consumer demand x_t, current ES level s_t, prices p^t, and supplies y^{t-1}.

Proposition 1. *For the optimization problem proposed in (9), there is no loss of optimality by focusing only on strategies in $\mathcal{F}^{'}$. And the equivalent optimization problem is:*

$$f^{'*} = \arg\min_{f^{'}\in\mathcal{F}^{'}} C_T(f^{'},\lambda). \tag{13}$$

Proof: For any realization (x^t, s^t, y^t, p^t), construct $f_t^{'}$ by the following equation:

$$P^{f_t^{'}}(y_t|x_t,s_t,p^t,y^{t-1}) = P^{f_t}(y_t|x_t,s_t,p^t,y^{t-1}). \tag{14}$$

Our aim is to show, under this transformation, the probabilities terms in our objective function are identical when they are induced by either $f^{'}$ or f. On observing the following relationships between $P_{X_t,S_t,Y^t,P^t|h_i}$, $P_{Y_t,P_t|h_i,Y^{t-1},P^{t-1}}$, $P_{Y^t,P^t|h_i}$ and $P_{X_t,P_t,Y_t|h_i}$:

$$\begin{aligned}P_{Y^t,P^t|h_i} &= \sum_{X_t,S_t} P_{X_t,S_t,Y^t,P^t|h_i},\\ P_{Y_t,P_t|h_i,Y^{t-1},P^{t-1}} &= \frac{P_{Y^t,P^t|h_i}}{P_{Y^{t-1},P^{t-1}|h_i}},\\ P_{X_t,P_t,Y_t|h_i} &= \sum_{S_t,Y^{t-1},P^{t-1}} P_{X_t,S_t,Y^t,P^t|h_i},\end{aligned} \tag{15}$$

in order to prove the above proposition, it is then sufficient to show that:

$$P^f_{X_t,S_t,Y^t,P^t|h_i} = P^{f^{'}}_{X_t,S_t,Y^t,P^t|h_i}. \tag{16}$$

Next we use the induction method for the proof. For $t = 1$, according to the construction in (14), there is $f_1^{'}(y_1|x_1,s_1,p_1) = f_1(y_t|x_1,s_1,p_1)$. Additionally, the initial state is generated regardless of the strategy, i.e., $P^f_{X_1,S_1,P_1|h_i} = P^{f^{'}}_{X_1,S_1,P_1|h_i}$. Therefore, we have $P^f_{X_1,S_1,P_1,Y_1|h_i} = P^{f^{'}}_{X_1,S_1,P_1,Y_1|h_i}$.

Assume $P^f_{X_t,S_t,Y^t,P^t|h_i} = P^{f^{'}}_{X_t,S_t,Y^t,P^t|h_i}$, we will show that $P^f_{X_{t+1},S_{t+1},Y^{t+1},P^{t+1}|h_i} = P^{f^{'}}_{X_{t+1},S_{t+1},Y^{t+1},P^{t+1}|h_i}$ also holds. According to the following expression we have:

$$\begin{aligned}&P^f(x_{t+1},s_{t+1},y^{t+1},p^{t+1}|h_i)\\ &\quad= \sum_{x_t,s_t} P^f(x_{t+1},x_t,s_{t+1},s_t,y^{t+1},p^{t+1}|h_i)\\ &\quad= \sum_{x_t,s_t}\big(P(p_{t+1}|p_t)P(x_{t+1}|x_t,h_i)\mathcal{I}_{s_{t+1}}(y_t+s_t-x_t)\\ &\qquad P^f(y_{t+1}|x_{t+1},s_{t+1},p^{t+1},y^t)\big)\times P^f(x_t,s_t,y^t,p^t|h_i)\\ &\quad\overset{(a)}{=} \sum_{x_t,s_t}\big(P(p_{t+1}|p_t)P(x_{t+1}|x_t,h_i)\mathcal{I}_{s_{t+1}}(y_t+s_t-x_t)\\ &\qquad f_{t+1}^{'}(y_{t+1}|x_{t+1},s_{t+1},p^{t+1},y^t)\big)\times P^{f_t^{'}}(x_t,s_t,y^t,p^t|h_i)\\ &\quad= P^{f^{'}}(x_{t+1},s_{t+1},y^{t+1},p^{t+1}|h_i),\end{aligned} \tag{17}$$

where (a) holds due to the construction in (14) and the induction hypothesis. Thus (16) holds for all t according to the principle of induction. □

To transform the optimization problem (13) into an MDP, we need to identify the states, the transition between different states, the control actions, and the per-step cost which can be expressed in terms of the current state and the corresponding control action. Thus, to end up with such a per-step cost formulation, we will follow the standard approach of introducing *control actions*, which jointly with the corresponding *policies* describe the energy management strategy.

At time step t, let the control action be $a_t \in \mathcal{A}$ which is the condition pmf $P_{Y_t|X_t,S_t,P_t}$ taken from the following set:

$$\mathcal{A} = \{P_{Y|X,S,P} : \sum_{y\in\overline{\mathcal{Y}}(x,s)} P(y|x,s,p) = 1, \forall x\in\mathcal{X}, s\in\mathcal{S}\} \tag{18}$$

Thus, the control action a_t randomly decides on the energy supply Y_t according to the conditional pmf $P_{Y_t|X_t,S_t,P_t}$. Note that in this section we assume that the EMU is unaware of the hypothesis h_0 or h_1, i.e., the consumers behavior. Thus, the EMU will choose an action according to a hypothesis independent policy $\pi_t \in \mathit{\Pi}_t$, where $\mathit{\Pi}_t$ denotes the set of deterministic mappings from the historical observations (y^{t-1}, p^{t-1}) to a corresponding action a_t, i.e., $a_t = \pi_t(y^{t-1}, p^{t-1})$. Thus, the policy over a T-time horizon is $\pi = \{\pi_t\}_{t=1}^{T} \in \mathit{\Pi} = \mathit{\Pi}_1 \times \mathit{\Pi}_2 \times .... \times \mathit{\Pi}_T$. At each time step, any energy management strategy $f_t^{'}$ can be equivalently represented by a combination of a policy π_t and a control action a_t. For each $f_t^{'}$, we apply the policy π_t to establish a mapping from the historical sequence (y^{t-1}, p^{t-1}) to $a_t = P^{f_t^{'}}_{Y_t|X_t,S_t,P_t,Y^{t-1}=y^{t-1},P^{t-1}=p^{t-1}}$. This control action a_t is then used to obtain Y_t according to $P_{Y_t|X_t,S_t,P_t}$. Thereby, we end up with control action that itself does not depend on the historical sequence (y^{t-1}, p^{t-1}), while the history decides which control action is applied at time t.

Let $Q_t = (Y^t, P^t, A^t)$, for the purpose of formulating the dynamic programming as well as expanding the per-step cost in terms of state-action pairs. It is more convenient to write the policy π_t as $a_t = \pi_t(q_{t-1})$. This can be seen from the following: Since (y^{t-1}, p^{t-1}) determines $a_t = \pi_t(y^{t-1}, p^{t-1})$ and contains the information that determines a^{t-1}, the expression $a_t = \pi_t(q_{t-1})$ is equivalent to $a_t = \pi_t(y^{t-1}, p^{t-1})$.

With the above definition and conclusion, we obtain an equivalent reformulation of the previous problem (13) as stated in the following proposition.

Proposition 2. *The optimization problem in Proposition 1 is equivalent to finding a policy $\pi \in \Pi$ that minimizes the following weighted sum objective:*

$$C_T(\pi, \lambda) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[C_t(\pi_t, \lambda, Q_{t-1})], \tag{19}$$

where the per-step expected cost conditioned on each possible historical sequence q_{t-1} can be specified as

$$\begin{aligned} &C_t\ (\pi_t, \lambda, q_{t-1}) \\ &\quad = \lambda \sum_{p_t, y_t} P^{\pi_t}(y_t, p_t|q_{t-1}, h_0) \log \frac{P^{\pi_t}(y_t, p_t|q_{t-1}, h_0)}{P^{\pi_t}(y_t, p_t|q_{t-1}, h_1)} \\ &\quad - (1-\lambda) \sum_{x_t, p_t, y_t} (P(h_0) P^{\pi_t}(x_t, p_t, y_t|q_{t-1}, h_0) \\ &\quad + P(h_1) P^{\pi_t}(x_t, p_t, y_t|q_{t-1}, h_1))(x_t p_t - y_t p_t). \end{aligned} \tag{20}$$

And the optimal policy is given by $\pi^* = \arg\min_{\pi \in \Pi} C_T(\pi, \lambda)$.

Proof: To prove this proposition, we need to show $C_T(\pi, \lambda)$ is equal to $C_T(f^{'}, \lambda)$ under transition from strategy $f^{'}$ to policy π. Thus, we need to further show that the probability terms in these two objective functions are equal. Since the proofs for the other probability terms are similar, we only prove $P^{f^{'}}_{Y^T, P^T|h_i} = P^{\pi}_{Y^T, P^T|h_i}$ here.

After expanding $P^{\pi}_{Y^T, P^T|h_i}$ according to the policy π, we obtain

$$\begin{aligned} &P^{\pi}{}_{Y^T, P^T|h_i}(y^T, p^T) \\ &\quad = \sum_{x^T, s^T} P_{X_1, S_1, P_1|h_i}(x_1, s_1, p_1) \times a_1(y_1|x_1, s_1, p_1, h_i) \\ &\quad \prod_{t=1}^{T-1} [P(p_{t+1}|p_t) P(x_{t+1}|x_t, h_i) \mathcal{I}_{s_{t+1}}(y_t + s_t - x_t) \\ &\qquad \times a_{t+1}(y_{t+1}|x_{t+1}, s_{t+1}, p_{t+1})]. \end{aligned} \tag{21}$$

By applying the equivalence between $f_t^{'}$ and (π_t, a_t), which is derived above, we get $P^{f_t^{'}}(y_{t+1}|x_{t+1}, s_{t+1}, p^{t+1}, y^t) = a_t(y_{t+1}|x_{t+1}, s_{t+1}, p_{t+1})$. Also, we have $P^{f_t^{'}}(y_1|x_1, s_1, p_1) = a_1(y_1|x_1, s_1, p_1)$. Thus, $P^{f^{'}}_{Y^T, P^T|h_i} = P^{\pi}_{Y^T, P^T|h_i}$ holds. □

The model described above can be regarded as a partially observed MDP: At each time step, the EMU observes the historical sequence q_{t-1} and makes a guessing on the current system state (X_t, S_t, P_t), i.e., the posterior distribution of (X_t, S_t, P_t) given realization q_{t-1}. Based on this guessing, the EMU will further choose a control action according to a specific policy. In contrast to a standard MDP problem, as shown in (20), the per-step conditional expected cost will depend on not only the current state and control action but also the historical sequence q_{t-1}. In order to formulate it into a standard MDP, we introduce belief states which will be used to replace the growing historical sequences by identifying an update rule. To this end, we define a belief state $\theta_{q_{t-1}} = (\theta^0_{q_{t-1}}, \theta^1_{q_{t-1}})$ as the posterior distributions of (X_t, S_t, P_t) conditioned on the realization q_{t-1} under the corresponding hypotheses as:

$$\theta^i_{q_{t-1}} = P_{X_t, S_t, P_t|q_{t-1}, h_i}, i \in \{0, 1\}. \tag{22}$$

At time step t, for $i \in \{0, 1\}$, given any $\theta^i_{q_{t-1}}$, observation y_t, and action a_t, the updating of belief state is given by $\theta^i_{q_t} = \varphi(\theta^i_{q_{t-1}}, a_t, y_t)$ in (23). And this evolution function ϕ can be derived by first applying Bayes rule and then substituting the corresponding terms by a_t and $\theta^i_{q_{t-1}}$.

C_t in (20) is indeed the per-step expected cost of the belief state $\theta_{q_{t-1}}$ when taking action a_t, since each realization q_{t-1} decides on a unique belief state. Thus, $C_t(\pi_t, \lambda, q_{t-1})$ can be expressed in terms of the corresponding belief state-action pairs, i.e., we have

$$\begin{aligned} &P^{\pi_t}(x_t, p_t, y_t|q_{t-1}, h_i) \\ &\quad = \sum_{s_t} \theta^i_{q_{t-1}}(x_t, s_t, p_t) a_t(y_t|x_t, s_t, p_t). \end{aligned} \tag{24}$$

With the above formulation, according to the definition of belief-state MDP in [32, pp.150-151], we have the following Theorem.

Theorem 1. *The original optimization problem in (13) can be modeled as a belief-state MDP problem such that: (i) the state at time t is given by (22) and evolves according to (23); (ii) the control action at time t is specified by $a_t(y_t|x_t, s_t, p_t)$; (iii) the per-step expected cost corresponding to a state-action pair is given by (20); (iv) the optimal policy π can be derived by using Bellman dynamic programming.*

Remark 1. *In the reformulated belief-state MDP problem, at each time step, the decision maker observes the historical sequence q_{t-1} and identifies a unique belief state $\theta_{q_{t-1}}$. Based on this belief state, the decision maker will further decide on an action according to the optimal strategy π_t derived from the Bellman dynamic programming.*

D. Backward Dynamic Programming over Finite Time Horizon

Lemma 1. *For any action a_t, according to [32, pp. 152-153], the modified Bellman operator B_{a_t} for our belief-state MDP problem can be written as:*

$$\begin{aligned} &(B_{a_t} V)(\theta_{q_{t-1}}) = C_t(\pi_t, \lambda, q_{t-1}) + \\ &\quad \sum_{y_t} [(\sum_{i=0,1} \sum_{x_t, s_t, p_t} P(h_i) \theta^i_{q_{t-1}}(x_t, s_t, p_t) \\ &\qquad a_t(y_t|x_t, s_t, p_t)) V(\varphi(\theta_{q_{t-1}}, a_t, y_t))], \end{aligned} \tag{25}$$

where V denotes the value function. The first term is the per-step expected cost for any given belief state $\theta_{q_{t-1}}$ and the second part denotes the corresponding expected cost-to-go.

$$\varphi(\theta^i_{q_{t-1}}, a_t, y_t) = \frac{\sum\limits_{x_t,s_t,p_t} \theta^i_{q_{t-1}}(x_t,s_t,p_t)a_t(y_t|x_t,s_t,p_t)P(p_{t+1}|p_t)P(x_{t+1}|x_t,h_i)\mathcal{I}_{s_{t+1}}(y_t+s_t-x_t)}{\sum\limits_{x_t,s_t,p_t} \theta^i_{q_{t-1}}(x_t,s_t,p_t)a_t(y_t|x_t,s_t,p_t)} \tag{23}$$

Proof: For the traditional belief state MDP with one belief state variable, which is described in [32], with an abuse of notation, the Bellman operator can be written as:

$$(B_aV)(b) = r(b,a) + \sum_{y\in\mathcal{Y}} P(y|b,a)V(\varphi(b,a,y)), \tag{26}$$

where b is the current belief state, a is the corresponding action, $r(b,a)$ denotes the per-step reward, and $\varphi(b,a,y)$ denotes the evolution of the belief state given a specific action a and a specific observation y. Most importantly, the term $P(y|b,a)$ denotes the probability of observing y at belief state b given a specific action a, i.e., the transition probability between belief state b and belief state $\varphi(b,a,y)$ given any action a. For our problem, the belief state $\theta_{q_{t-1}} = (\theta^0_{q_{t-1}}, \theta^1_{q_{t-1}})$ is a vector that contains two beliefs conditioned either on hypothesis h_0 or h_1. In this case, with the given prior of the hypotheses $P(h_0)$ and $P(h_1)$, the probability for observing y_t at belief state $\theta_{q_{t-1}}$ given action a_t can be then calculated by:

$$\begin{aligned} P(y_t|a_t,&\theta_{q_{t-1}}) \\ &= P(h_0)\sum_{x_t,s_t,p_t}(\theta^0_{q_{t-1}}(x_t,s_t,p_t)a_t(y_t|x_t,s_t,p_t)) \\ &+ P(h_1)\sum_{x_t,s_t,p_t}(\theta^1_{q_{t-1}}(x_t,s_t,p_t)a_t(y_t|x_t,s_t,p_t)), \end{aligned} \tag{27}$$

Plugging the above equation into (27) will lead to our modified Bellman operator as defined in (25). □

In this case, the value function is updated according to:

$$V(\theta_{q_{t-1}}) = \min_{a_t\in\mathcal{A}_t}(B_{a_t}V)(\theta_{q_{t-1}}). \tag{28}$$

Thus, the optimal policy $\pi^*_t(\theta_{q_{t-1}})$ and the corresponding optimal control action $a^*_t = \pi^*_t(\theta_{q_{t-1}})$ is given by the optimizer of (28). Let θ_1 denote the initial joint distributions of (X_1, S_1, P_1) conditioned on h_0 and h_1, then the minimum value of average expected cost C_T is given by $V(\theta_1)/T$.

E. Optimization over Infinite Time Horizon

In this section, we consider the case of infinite time horizon, i.e., $T \to \infty$. For the reason of simplicity, we use $(s,a) \in \mathcal{S}\times\mathcal{A}$ to denote the belief state and action pair, $c(s,a)$ and $P(s'|s,a)$ as the cost and transition probability to state s' from the corresponding state-action pair (s,a). Under the infinite time horizon, the optimal Bellman equation of our average expected cost MDP is given by:

$$h^*(s) = \min_{a\in\mathcal{A}}[c(s,a) - \rho^* + \sum_{s'} P(s'|s,a)h^*(s')], \tag{29}$$

where ρ^* denotes the optimal average expected cost of the optimal policy, i.e., the optimal gain of total expect cost in steady state. In the following, we assume the optimal stationary policy exists for our infinite horizon average reward MDP problem.[1]

Let $h(s)$ be the relative value function, i.e., the asymptotic difference between the total expected cost that starting from state s and the total expected cost that would be incurred if the per-step cost is equal to ρ^* for all states. The optimal relative value function $h^*(s)$ is given by the $h(s)$ that satisfies the above optimal Bellman equation, which represents the minimum asymptotic difference. The optimal policy for the above problem $\pi^* = (\pi, \pi, ...)$ is stationary and is defined by:

$$\pi(s) = \arg\min_{a\in\mathcal{A}}[c(s,a) + \sum_{s'} P(s'|s,a)h^*(s')], \quad \forall s. \tag{30}$$

Given state and action sets are with small cardinalities, the relative value iteration (RVI) method [34] can guarantee a fast convergence of the above operator if an optimal stationary solution exists. For our problem, both of the action space and the belief state space are continuous, i.e., with an infinite number of states and actions. Thus, the continuous space needs to be quantized to a finite set with a relatively low cardinality so that the RVI algorithm can work efficiently. However, the solution derived under the quantized spaces can only be regarded as a sub-optimal solution to the original problem. A better solution can be found by increasing the resolution of the quantization, or even considering the whole space without quantization. However, when the cardinality of the set increases, the system dynamics become impractical to characterize, i.e., cost function and transition behavior corresponding to each state-action cannot be fully characterized. Thus, the exact solution methods such as RVI will be inapplicable.

To address this issue, we propose two alternative methods. We first propose to use the Q-learning algorithm, since the reinforcement learning algorithms could help to solve the MDP problem without the knowledge of the cost function and the transition probabilities. Since it is infeasible to explicitly represent the Q-function over the continuous belief state space[2], a general function approximator is used to approximate the Q-function given each possible state. In more detail, we provide the framework of relative Q-learning with linear function approximation as a sub-optimal but computationally more efficient solution to our original optimization problem. On the other hand, in Section III, under the special case of

[1] To guarantee the existence of stationary optimal policy, we need to have some restrictions on the underlying Markov chains. For instance, the Markov chain induced by any policy should be unichain. However, the problem of checking such a unichain condition is NP-hard [33]. We thus assume that there exists an optimal stationary policy for our infinite horizon average MDP problem. For our numerical experiments, we can see that our relative iteration algorithm convergences under our discretized state-action space settings, which indicates that the optimal stationary policy exists under this specific setting.

[2] For the reason of simplicity, we restrict our problem to the case with infinite state space but a finite action space, i.e., the action takes values from a finite subset of the continuous action set $\mathcal{A}$.

i.i.d. energy demand, we study a time-invariant strategy that preserves the stationarity of each belief state, which can avoid the curse of the dimensions and the growth of the complex system dynamics.

F. Q-learning Based Stationary Energy Management Strategy Design

In this section, we first provide a brief outline of the relative Q-learning method. More details on this method can be found in [35]. The main contribution here is the linear functional approximation and the corresponding feature selection that results in a good performance of our approach.

Given the optimal relative value function $h^*(s)$ in (26), we define the optimal Q-function $Q^*(s,a)$ as the minimum asymptotic difference between total expected cost starting from state s with action a and the optimal total expected cost:

$$Q^*(s,a) = c(s,a) - \rho^* + \sum_{s'} P(s'|s,a)h^*(s'). \tag{31}$$

Since $h^*(s) = \min_{a\in\mathcal{A}} Q^*(s,a)$, we have:

$$Q^*(s,a) = c(s,a) - \rho^* + \sum_{s'} P(s'|s,a) \min_{b\in\mathcal{A}} Q^*(s',b). \tag{32}$$

Further we define an operator H as:

$$(HQ)(s,a) = c(s,a) - \rho^* + \sum_{s'} P(s'|s,a) \min_{b\in\mathcal{A}} Q(s',b), \tag{33}$$

the optimal Q-function then becomes a fixed point of operator H. According to the Robbins-Monro algorithm [36], the optimal Q-function can be learned by utilizing the temporal difference between the new estimate and the old estimate, which is given by the following:

$$\begin{aligned} &Q_{n+1}(s,a) \\ &\quad = Q_n(s,a) + \alpha[c(s,a) - \rho^* + \min_{b\in\mathcal{A}} Q_n(s',b) - Q_n(s,a)] \\ &\quad = (1-\alpha)Q_n(s,a) + \alpha[c(s,a) - \rho^* + \min_{b\in\mathcal{A}} Q_n(s',b)], \end{aligned} \tag{34}$$

where $\alpha \in (0,1]$ denotes the learning rate, which can be kept as constant during the learning process. The new estimate term $c(s,a) - \rho^* + \min_{b\in\mathcal{A}} Q_n(s',b)$ is sampled in the system by executing an action a selected by ϵ-greedy policy[3] which results in state s'. Note that the optimal gain ρ^* is unknown in advance. Thus, we introduce the following relative Q-function iteration to overcome this problem.

First, we select an arbitrary state-action pair $(\hat{s},\hat{a})$ before the algorithm starts, this state-action pair is fixed and acts as the reference state-action pair in each iteration until the algorithm ends. The Q-function corresponding to each possible state action pair (s,a) can be updated by:

$$\begin{aligned} &Q_{n+1}(s,a) \\ &\quad = (1-\alpha)Q_n(s,a) + \alpha[c(s,a) - Q_n(\hat{s},\hat{a}) + \min_{b\in\mathcal{B}} Q_n(s',b)]. \end{aligned} \tag{35}$$

It has been shown in [34] that as $n \to \infty$, the sequence $\big(Q_n(\hat{s},\hat{a})\big)_n$ will converge to ρ^*. As a result, this algorithm will converge to the fixed point of (32).

Remark 2. *The computational complexity of the previous RVI algorithm is $\mathcal{O}(|S|^2|A|)$ per iteration and $\mathcal{O}(T|S|^2|A|)$ overall [35], where T denotes the number of iterations to converge, $|S|$ and $|A|$ denote the cardinalities of state and action spaces. Meanwhile, in the above relative Q-learning algorithm, the computational complexity is only $\mathcal{O}(|A|)$ per iteration and $\mathcal{O}(N|A|)$ overall, where N is the number of iterations to converge. The Q-learning algorithms usually need a higher number of iterations to converge than RVI, i.e., $N > T$. However, when we have large state or action spaces, i.e., large $|S|$ or $|A|$, the Q-learning algorithms will lead a significant reduction of the computational complexity .*

1) Linear function approximation: Since our belief state space is continuous, the number of states to learn is infinite so that an explicit characterization of each $Q(s,a)$ is infeasible. For this reason, we propose to use the function approximation method to avoid the explicit characterization of the Q-function. In more detail, we use the following linear function $\hat{Q}(s,a)$ to approximate the Q-function $Q(s,a)$, since it is simple for mathematical analysis and it can inherit the useful convergence results from different kinds of learning systems [35]. Assume we have a finite action set $\mathcal{A}' \subset \mathcal{A}$ with a small cardinality, it is then practical to represent $\hat{Q}(s,a)$ by the following weighted sum of different features of state s :

$$\hat{Q}(s,a) = \sum_{i=1}^{N} w_i(a) f_i(s), \tag{36}$$

where $f_i(s)$ for $i = 1,2,...,N$ are N feature functions corresponding to each possible belief state s; and $w_i(a)$ for $i = 1,2,...,N$ are weights for different features given each possible action a. More details on how to select the features will be discussed later. Let M denote the cardinality of the finite action set $\mathcal{A}$. We thus transfer our original task of learning $Q(s,a)$ for an infinite number of state action pairs into the task of learning M different weight vectors, where each vector is of length N, i.e., $w(a) = [w_1(a), w_2(a), ..., w_N(a)]$ denotes the weight vector corresponding to action a.

Using the update rule of the Q-function $Q_n(s,a)$ in (35), we define the temporal difference between new estimate and old estimate as follows:

$$\Delta Q_n(s,a) = c(s,a) - Q_n(\hat{s},\hat{a}) + \min_{b\in\mathcal{A}} Q_n(s',b) - Q_n(s,a). \tag{37}$$

Equation (35) will converge to the optimal Q-function, when we reduce the magnitude of $\Delta Q_n(s,a)$, i.e., $\big(Q_n(\hat{s},\hat{a})\big)_n \to \rho^*$ as $\Delta Q_n(s,a) \to 0$, when $n \to \infty$. By applying the same underlying idea and substituting the Q-function with its linear function approximation (36), it has been shown in [37] that the optimal linear approximator, which satisfies (32), can be obtained by solving:

$$\min_{w(a),\forall i} \mathbb{E}(\Delta\hat{Q}_n(s,a))^2, \tag{38}$$

[3] In reinforcement learning scenario, under an ϵ-greedy policy, the agent choose the best action with probability $1-\epsilon$ and randomly choose an action with probability ϵ.

where $\Delta \hat{Q}_n(s,a)$ denotes the the temporal difference between new estimate and old estimate when the Q-function is approximated by $\hat{Q}$:

$$\Delta \hat{Q}_n(s,a) = c(s,a) - \hat{Q}_n(\hat{s},\hat{a}) + \min_{b \in \mathcal{A}} \hat{Q}_n(s',b) - \hat{Q}_n(s,a). \tag{39}$$

To find the optimal solutions to (38),[4] the stochastic gradient descent method [35] is used to update the weights $w(a)$ for all actions $a \in \mathcal{A}$. The updating rule is then given by:

$$w_i(a) = w_i(a) + \alpha \Delta \hat{Q}_n(s,a) f_i(s), \forall i \in \{1,2,...N\}, \tag{40}$$

where $\alpha \in (0,1]$ is the diminishing learning rate.

Remark 3. *Given the convergence of the above algorithm, the optimal strategy can be obtained by:*

$$\pi^*(s) = \arg\min_{a \in \mathcal{A}} \hat{Q}^*(s,a), \tag{41}$$

where $\hat{Q}^*$ is the optimal approximated Q-function.

2) Feature selection: In order to have a linear function $\hat{Q}(s,a)$ that can well approximate $Q(s,a)$, it is important to choose appropriate features $f_i(s)$ for the linear approximator. Given the Q-function defined by (31), we propose the following heuristic approach where we select features that describe well the per-step cost function $c(s,a)$ in the linear form (36). We first expand the per-step cost function (20) as follows:

$$\begin{aligned}
&C_t(\pi_t, \lambda, q_{t-1}) \\
&= \lambda \sum_{p_t,y_t} P^{\pi_t}(y_t,p_t|q_{t-1},h_0) \log P^{\pi_t}(y_t,p_t|q_{t-1},h_0) \\
&- \lambda \sum_{p_t,y_t} P^{\pi_t}(y_t,p_t|q_{t-1},h_0) \log P^{\pi_t}(y_t,p_t|q_{t-1},h_1) \\
&- (1-\lambda) \sum_{x_t,p_t,y_t} P(h_0) P^{\pi_t}(x_t,p_t,y_t|q_{t-1},h_0)(x_t p_t - y_t p_t) \\
&- (1-\lambda) \sum_{x_t,p_t,y_t} P(h_1) P^{\pi_t}(x_t,p_t,y_t|q_{t-1},h_1)(x_t p_t - y_t p_t).
\end{aligned} \tag{42}$$

Let the cardinality of the energy supply set $\mathcal{Y}$ be K, and the cardinality of price set $\mathcal{P}$ be L. Also let the operator $|\cdot|$ denotes the L_2 norm. For any belief state (θ^0,θ^1)[5] and price p_i, define θ_i^0 as the vector which contains elements $[\theta^0(x,s,p_i)]_{(x,s)\in\mathcal{X}\times\mathcal{Y}}$ and θ_i^1 as the vector with elements $[\theta^1(x,s,p_i)]_{(x,s)\in\mathcal{X}\times\mathcal{S}}$. Further, let a_i^j be the vector with elements $[a_t(y_j|x,s,p_i)]_{(x,s)\in\mathcal{X}\times\mathcal{S}}$. Let ϕ_i^j be the angle between vector a_i^j and θ_i^0, and ψ_i^j be the angle between vector a_i^j and θ_i^1. By finding a feature-action representation for each term of (42), the features given any action are characterized in the following proposition.

Proposition 3. *For any belief state (θ^0,θ^1) and action a, by doing a decomposition of (42), the corresponding heuristic feature selection is characterized as follows:*

$$\begin{aligned}
&f_1((\theta^0,\theta^1)) = 1, \\
&f_2((\theta^0,\theta^1)) = |\theta^0|, \\
&f_3((\theta^0,\theta^1)) = |\theta^1|, \\
&f_4((\theta^0,\theta^1)) = \sqrt{\sum_i^L (|\theta_i^0| \log |\theta_i^0|)^2}, \\
&f_5((\theta^0,\theta^1)) = \sqrt{\sum_i^L (|\theta_i^0| \log |\theta_i^1|)^2}.
\end{aligned} \tag{43}$$

Proof: For reason of simplicity, we provide the derivation for the features-action representation of the first term in (42), and the others can be derived in a similar way.

Ignoring λ, the first term in (42) can be further decomposed into the following:

$$\begin{aligned}
&\sum_{p_t,y_t} P^{\pi_t}(y_t,p_t|q_{t-1},h_0) \log P^{\pi_t}(y_t,p_t|q_{t-1},h_0) \\
&= \sum_{p_t,y_t} \Big(\sum_{x_t,s_t} \theta^0(x_t,s_t,p_t) a_t(y_t|x_t,s_t,p_t) \\
&\times \log \sum_{x_t,s_t} \theta^0(x_t,s_t,p_t) a_t(y_t|x_t,s_t,p_t) \Big) \\
&\overset{(a)}{=} \sum_i^L \sum_j^K |a_i^j||\theta_i^0| \cos\phi_i^j (\log|a_i^j| + \log|\theta_i^0| + \log\cos\phi_i^j) \\
&= \sum_i^L |\theta_i^0| \sum_j^K |a_i^j| \cos\phi_i^j (\log|a_i^j| + \log\cos\phi_i^j) \\
&+ \sum_i^L |\theta_i^0| \log|\theta_i^0| \sum_j^K |a_i^j| \cos\phi_i^j \\
&\overset{(b)}{=} \vec{A}_1 \cdot \vec{B}_1 + \vec{A}_2 \cdot \vec{B}_2 \\
&= |\vec{A}_1||\vec{B}_1| \cos\left\langle \vec{A}_1, \vec{B}_1 \right\rangle + |\vec{A}_2||\vec{B}_2| \cos\left\langle \vec{A}_2, \vec{B}_2 \right\rangle,
\end{aligned} \tag{44}$$

where $\cdot$ denotes the inner product between two vectors in the Euclidean space. The equality (a) in (44) follows from considering sum over all possible (y_t,p_t) and the inner product between vector θ_i^0 and a_i^j. Likewise, the equality (b) is an inner product where the i-th element of $\vec{A}_1, \vec{A}_2, \vec{B}_1, \vec{B}_2$ are defined by the following:

$$\begin{aligned}
A_{1i} &= \sum_j^K |a_i^j| \cos\phi_i^j (\log|a_i^j| + \log\cos\phi_i^j), \quad B_{1i} = |\theta_i^0|, \\
A_{2i} &= \sum_j^K |a_i^j| \cos\phi_i^j, \quad B_{2i} = |\theta_i^0| \log|\theta_i^0|.
\end{aligned} \tag{45}$$

[4] For the discounted expected reward MDP, there exist some conditions to guarantee the convergence of Q-learning combined with linear function approximation, e.g., see [37]. However, to the best of our knowledge, there exist no such theoretical convergence guarantees for the average reward MDP.

[5] For reason of simplicity, we use (θ^0,θ^1) as short notation for $(\theta^0_{q_{t-1}}, \theta^1_{q_{t-1}})$.

Similarly, the second term in (42) can be decomposed by:

$$\begin{aligned}\sum_{p_t,y_t} & P^{\pi_t}(y_t,p_t|q_{t-1},h_0)\log P^{\pi_t}(y_t,p_t|q_{t-1},h_1)\\ &= \vec{A}_3\cdot\vec{B}_1+\vec{A}_2\cdot\vec{B}_3\\ &= |\vec{A}_3||\vec{B}_1|\cos\left\langle\vec{A}_3,\vec{B}_1\right\rangle+|\vec{A}_2||\vec{B}_3|\cos\left\langle\vec{A}_2,\vec{B}_3\right\rangle,\end{aligned} \tag{46}$$

where the i-th element of $\vec{A}_3,\vec{B}_3$ are defined by the following:

$$A_{3i}=\sum_j^K|a_i^j|\cos\phi_i^j(\log|a_i^j|+\log\cos\psi_i^j),\ B_{3i}=|\theta_i^0|\log|\theta_i^1|. \tag{47}$$

Thus, the decomposition of the first two terms identifies the features $f_2((\theta^0,\theta^1))=|\vec{B}_1|=|\theta^0|$, $f_4((\theta^0,\theta^1))=|\vec{B}_2|=\sqrt{\sum_i^L(|\theta_i^0|\log|\theta_i^0|)^2}$, and $f_5((\theta^0,\theta^1))=|\vec{B}_3|=\sqrt{\sum_i^L(|\theta_i^0|\log|\theta_i^1|)^2}$, and we can conclude these features will be highly relevant to the value of KL-divergence term.

Meanwhile, by applying the same method to decompose last two terms in (42), i.e., the cost-saving term, $f_2((\theta^0,\theta^1))=|\theta^0|$ and $f_3((\theta^0,\theta^1))=|\theta^1|$ are identified as the features that will be highly relevant to the value of the cost-saving term. Besides, feature "1" is selected to add an offset to the approximated function and compensate the errors. In this case, summarizing the above analysis leads to the feature selection scheme in Proposition 3. □

Remark 4. *The intuition behind the above derivation is to decompose the cost function $C_t(\pi_t,\lambda,q_{t-1})$ to identify the features that is only related to the belief state, and the result above shows the cost function can be written as a linear combination of these features. With this result, one can conclude that these features will be highly relevant to the value of the cost function.*

Remark 5. *As we can see from (44) amd (46), the correlation between different actions and states cannot be eliminated during the derivation due to the existence of the angles between state and action vectors, which means the linear combination of our selected feature cannot exactly represent the cost function and the Q-function. In this case, there will be a performance gap between our proposed algorithm and the optimal RVI algorithm. And this is reason why the linear approximated Q-learning algorithm requires to solve the optimization problem (38), where the proper weights that minimize the mean square approximation error over all possible states can be identified. Also, an extra feature "1" is added here to compensate the approximation error, which will further lead to a smaller performance gap.*

3) Approximated relative Q-learning: Based on the above discussion, we summarize our proposed linear function approximated relative Q-learning in Algorithm 1.

Algorithm 1: Approximated relative Q-learning

```
Input: Belief states and actions
Output: The optimal strategy
Initialization: Initialize $\alpha$, $\epsilon$, $w(a), \forall a \in \mathcal{A}'$, the reference
  state-action pair $(\hat{s},\hat{a})$, the initial belief state $s_1$.
Calculate the corresponding features of $s_1$.
for $n$=1:$T$ do
    Select action $a_n$ using $\epsilon$-greedy policy;
    Execute action $a_n$;
    Observe a new state $s_{n+1}$ and the per-step cost
      $c(s_n,a_n)$;
    Calculate the corresponding features of $s_{n+1}$;
    Update $w(a_n)$ by using (40);
end
Find the optimal strategy by using (41).
```

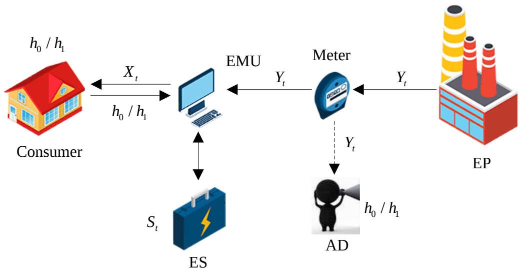


Fig. 2. Smart metering system with rechargeable energy storage, authorized adversary, and privacy-aware energy management unit that knows the consumer's energy consumption behavior.

III. Privacy-Preserving Under I.I.D Energy Demand

In the previous section, we studied the problem of designing the privacy-enhancing and cost-efficient energy management strategy under the MDP framework. However, with the increasing dimension and time horizon length, it becomes more difficult or even infeasible to find an analytical solution. Thus, we propose the following study of privacy-preserving stationary strategy design under the special case of i.i.d. energy demand. Although the real energy demand will not be i.i.d. the designed energy management strategy under this simplifying assumption can be still practically implemented. If perfect privacy is not achieved, then the adversary will be always able to learn the hyphothesis in the infinite time horizon [27]. Thus, in this section we study under the special case of i.i.d. energy demand without cost concern, where we can derive the sufficient condition to achieve the perfect privacy.

A. System Model

In this section, we consider privacy-preserving problem in the system shown in Fig. 2. Assume the energy demand $x_t\in\mathcal{X}=\{0,1,...,x_{max}\}$ is i.i.d. with distribution $P_X^0(x)$ under hypothesis h_0, and $P_X^1(x)$ under h_1 respectively. The energy supply $y_t\in\mathcal{Y}=\{0,1,...,y_{max}\}$ and battery level $s_t\in\mathcal{S}=\{0,1,...,s_{max}\}$ also satisfy the physical constraint described in (1) and (2). We further assume that the EMU knows the consumer's behavior, and design the corresponding energy management strategies under different energy consumption behaviors: $f^{(i)} = \{f_t^{(i)}\}_{t=1}^T \in \mathcal{F}^{(i)} = \mathcal{F}_1^{(i)} \times \mathcal{F}_2^{(i)} \times .... \times \mathcal{F}_T^{(i)}$, with $f_t^{(i)} \in \mathcal{F}_t^{(i)}$. $\mathcal{F}_t^{(i)}$ denotes the set of pmfs that:

$$\mathcal{F}_t^{(i)} = \\ \{P_{Y_t|X_t,S_t,Y^{t-1},h_i} : \sum_{\substack{y_t \in \overline{\mathcal{Y}}_t(x_t,s_t) \\ \forall x_t \in \mathcal{X}, s_t \in \mathcal{S}}} P(y_t|x_t, s_t, y^{t-1}, h_i) = 1, i \in \{0,1\}\}, \tag{48}$$

where $f^{(i)}$ denote the specific energy management strategy under consumer's behavior hypothesis h_i. Also, let the control action be $a_t \in \mathcal{A}_t$, which is the conditional pmf $P_{Y_t|X_t,S_t}$ taken from the following set:

$$\mathcal{A} = \{P_{Y|X,S} : \sum_{y \in \overline{\mathcal{Y}}(x,s)} P(y|x,s) = 1, \forall x \in \mathcal{X}, s \in \mathcal{S}\}. \tag{49}$$

To make it more clear, we denote the action chosen under hypothesis h_i as $a^{(i)}$. Different from the previous definition of policy π in Section II-C, at each time step t, the EMU will choose the action according to the policy $\pi_t^{(i)} \in \mathit{\Pi}_t^{(i)}$ under consumer's behavior hypothesis h_i. Define the set of binary consumer's behavior hypotheses as $\mathcal{H}$. Then $\mathit{\Pi}_t^{(i)}$ denotes the set of deterministic mappings from the historical observations (y^{t-1}, h_i) to a corresponding action $a_t^{(i)}$, i.e., $\mathit{\Pi}_t^{(i)} : \mathcal{Y}^{t-1} \times \mathcal{H} \to \mathcal{A}_t$ with $a_t^{(i)} = \pi_t^{(i)}(y^{t-1}, h_i)$. Thus, the policy under consumer's behavior hypothesis h_i over a T-time horizon is $\pi^{(i)} = \{\pi_t^{(i)}\}_{t=1}^T \in \mathit{\Pi} = \mathit{\Pi}_1^{(i)} \times \mathit{\Pi}_2^{(i)} \times .... \times \mathit{\Pi}_T^{(i)}$.

B. Design of Memory-Less Stationary Energy Management Strategy

With the above definition, we consider the problem of designing the memory-less stationary privacy-preserving strategies[6] $\pi_t^{(i)}$ depending on different h_i, with the following objective function:

$$\begin{aligned} L_T(\pi^{(0)}, \pi^{(1)}) &= \frac{1}{T} D(P^{\pi^{(0)}}_{Y^T|h_0} \| P^{\pi^{(1)}}_{Y^T|h_1}) \\ &= \frac{1}{T} \sum_{t=1}^T D(P^{\pi_t^{(0)}}_{Y_t|h_0,Y^{t-1}} \| P^{\pi_t^{(1)}}_{Y_t|h_1,Y^{t-1}}) \\ &= \frac{1}{T} \sum_{t=1}^T \sum_{y^t} P^{\pi^{(0)}}(y^t|h_0) \times \log \frac{P^{\pi_t^{(0)}}(y_t|y^{t-1}, h_0)}{P^{\pi_t^{(1)}}(y_t|y^{t-1}, h_1)}. \end{aligned} \tag{50}$$

Before designing the stationary energy management strategy for our proposed model, we first propose a structural simplification on the states and actions by introducing two new auxiliary random variables: $W_t \in \mathcal{W} = \{s_t - x_t : s_t \in \mathcal{S}, x_t \in \mathcal{X}\}$. Under policy $\pi^{(i)}$, define the posterior distributions of (X_t, S_t), W_t and S_t conditioned on the realization y^{t-1} as: $\theta_t^{(i)} = P_{X_t,S_t|y^{t-1},h_i}$, $\xi_t^{(i)} = P_{W_t|y^{t-1},h_i}$ and $\gamma_t^{(i)} = P_{S_t|y^{t-1},h_i}$. In particular, there is:

$$\theta_t^{(i)}(x_t, s_t) = P_X^i(x_t)\gamma_t^i(s_t), \tag{51}$$

$$\begin{aligned} \gamma_t^{(i)}(s_t) &= P^{\pi^{(i)}}(S_t = s_t|Y^{t-1} = y^{t-1}, h_i), \\ \xi_t^{(i)}(w_t) &= P^{\pi^{(i)}}(W_t = w_t|Y^{t-1} = y^{t-1}, h_i). \end{aligned} \tag{52}$$

Since the following derivation works for both hypotheses h_0 and h_1, we only discuss the case for h_0 and denote $\theta_t^{(0)}, \gamma_t^{(0)}, \xi_t^{(0)}$, $a_t^{(0)}$ and $\pi_t^{(0)}$ by $\theta_t, \gamma_t, \xi_t$, a_t and π_t.

Define $\mathcal{D}(w_t) = |(x_t, s_t) \in \mathcal{X}_t \times \mathcal{S}_t : \ s_t - x_t = w_t|$, there is $\xi_t(w_t) = \sum_{(x_t,s_t) \in \mathcal{D}(w_t)} \theta_t(x_t, s_t)$. At time t, define a new action $b_t \in \mathcal{B}_t$ which is the condition pmf $P_{Y_t|W_t}$ taken from the set $\mathcal{B}_t = \{P_{Y|W} : \sum_{y \in \overline{\mathcal{Y}}(w)} P(y|w) = 1, \ \forall w \in \mathcal{W}\}$, where $\overline{\mathcal{Y}}(w)$ is defined by replacing $s_t - x_t$ with w in (2). Thus, action b_t can be expressed in terms of original belief state θ_t and action a_t,

$$\begin{aligned} b_t(y_t|w_t) &= \frac{P^\pi(Y_t = y_t, W_t = w_t|Y^{t-1} = y^{t-1}, h_0)}{P^\pi(W_t = w_t|Y^{t-1} = y^{t-1}, h_0)} \\ &= \frac{\sum_{(x_t,s_t) \in \mathcal{D}(w_t)} a_t(y_t|x_t, s_t)\theta_t(x_t, s_t)}{\xi_t(w_t)}. \end{aligned} \tag{53}$$

Similarly, we can define policy $\hat{\pi}_t$ as the deterministic mappings from the historical observations (y^{t-1}, h_0) to a corresponding action b_t, i.e., $b_t = \hat{\pi}_t(y^{t-1}, h_0)$. We further define the following distributions:

$$\begin{aligned} \gamma_t'(s_t) &= P^{\hat{\pi}}(S_t = s_t|Y^{t-1} = y^{t-1}, h_0) \\ \xi_t'(w_t) &= P^{\hat{\pi}}(W_t = w_t|Y^{t-1} = y^{t-1}, h_0). \end{aligned} \tag{54}$$

Thus at time step t, for any realization of y_t and b_t, the evolution of ξ_t can be expressed in terms of b_t as follows:

$$\xi_{t+1}'(w_{t+1}) = \varphi'(\xi_t', y_t, b_t) = \\ \frac{\sum_{(x_{t+1},s_{t+1}) \in \mathcal{D}(w_{t+1})} \sum_{w_t} \xi_t'(w_t) b_t(y_t|w_t) P_X^0(x_{t+1}) \mathcal{I}_{s_{t+1}}\{y_t + w_t\}}{\sum_{w_t} b_t(y_t|w_t)\xi_t'(w_t)} \tag{55}$$

Lemma 2. *Given historical observations (y^{t-1}, h_0), the posterior distributions of W_t and S_t induced by policy π and $\hat{\pi}$ are the same, i.e.,*

$$\xi_t(w_t) = \xi_t'(w_t), \quad \gamma_t(s_t) = \gamma_t'(s_t), \ \forall s_t \in \mathcal{S}, \ w_t \in \mathcal{W}. \tag{56}$$

Proof: The proof is provided in Appendix A. □

With the above derivation, we have the following proposition.

Proposition 4. *Under the above assumption and transition, there is no loss of optimality in focusing on action $b_t \in \mathcal{B}_t$ and new belief state ξ_t instead of (a_t, θ_t). The per-step cost defined in (50) can be equivalently described by a function of state-action pair (b_t, ξ_t').*

Proof: To prove this proposition, we need to show that the conditional probability $P(y_t|y^{t-1}, a^{t-1}, h_0)$ remains identical

[6] Memoryless control strategies have been previously shown to be optimal for hypothesis privacy in some other contexts, e.g., Linear Quadratic Gaussian (LQG) system [38], [39].

under transition from $(a_t, \theta_t) \to (b_t, \xi_t)$. Note that the same arguments can be applied to h_1.

$$\begin{aligned} P^{\pi}(y_t|y^{t-1}, h_0) &= \sum_{w_t} P^{\pi}(Y_t = y_t, W_t = w_t|y^{t-1}, h_0) \\ &= \sum_{w_t} \sum_{(x_t,s_t)\in\mathcal{D}(w)} a_t(y_t|x_t, s_t)\theta_t(x_t, s_t) \\ &= \sum_{w_t} b_t(y_t|w_t)\xi_t(w_t) \\ &= \sum_{w_t} b_t(y_t|w_t)\xi'_t(w_t) \\ &= \sum_{w_t} P^{\hat{\pi}}(Y_t = y_t, W_t = w_t|y^{t-1}, h_0) \\ &= P^{\hat{\pi}}(y_t|y^{t-1}, h_0) \quad \square \end{aligned} \tag{57}$$

Theorem 2. *Given $y \in \mathcal{Y}$, $w \in \mathcal{W}$, and any possible distributions γ'_t, ξ'_t, a time-invariant policy $\hat{f}$, with $\hat{f}(\xi_t) = \hat{b}_t$, leads to steady states, i.e., $\xi'_t = \xi'_1$ and $\gamma'_t = \gamma'_1$, if and only if $\hat{b}_t$ satisfies the following structure:*

$$\hat{b}_t(y|w) = \begin{cases} Q_Y(y)\frac{\gamma'_t(y+w)}{\xi'_t(w)}, \ y \in \overline{\mathcal{Y}_t}(w), \\ 0, \ otherwise, \end{cases} \tag{58}$$

where $Q_Y(y)$ is an arbitrary probability distribution over all feasible $y \in [0, \min\{y_{max}, s_{max} + x_{max}\}]$ [7], and the same $Q_Y(y)$ is applied for the design of $\hat{b}_t$ at different time steps.

Proof: The proof is provided in Appendix B. □

Moreover, an important conclusion drawn from the proof of Theorem 2 is summarized into the following corollary.

Corollary 1. *The distribution $Q_Y(y)$ that is used for designing the structured action $\hat{b}_t$ in Theorem 2 should satisfy the following equation:*

$$Q_Y \overset{(\Delta)}{=} P_{Y_t|Y^{t-1}=y^{t-1}}, \ \forall y^{t-1}. \tag{59}$$

Thus, the marginal distributions of Y_t at each time step are identical, i.e.,

$$Q_Y \overset{(\Delta)}{=} P_{Y_t}, \ \forall t. \tag{60}$$

C. Privacy-Preserving under Steady-State Strategy

For $i \in \{0, 1\}$, let $\hat{f}^{(i)}$ be the time-invariant policy under hypothesis h_i which leads to the steady state. Also define $\hat{b}_t^{(i)}$ as the action decided by $\hat{f}^{(i)}$ at time t under hypothesis h_i, which satisfies the structure in Theorem 2. Further denote $Q_Y^0(y)$ and $Q_Y^1(y)$ as the distributions of Y used for constructing action $\hat{b}_t^{(0)}$ and $\hat{b}_t^{(1)}$, respectively. By combining the results in Theorem 2 and Corollary 1, we have the following corollary.

Corollary 2. *The objective function in (50) can be simplified to the single-letter expression by the following:*

$$\begin{aligned} &L_T(\hat{f}^{(0)}, \hat{f}^{(1)}) \\ &= \frac{1}{T} D(P^{\hat{f}^{(0)}}_{Y^T|h_0} \| P^{\hat{f}^{(1)}}_{Y^T|h_1}) \\ &= \frac{1}{T}\Big(D(P^{\hat{f}^{(0)}}_{Y^{T-1}|h_0} \| P^{\hat{f}^{(1)}}_{Y^{T-1}|h_1}) + \sum_{y^{T-1}} P^{\hat{f}^{(0)}}(y^{T-1}|h_0) \\ &\qquad \sum_{y_T} P^{\hat{f}^{(0)}}(y_T|y^{T-1}, h_0) \log \frac{P^{\hat{f}^{(0)}}(y_T|y^{T-1}, h_0)}{P^{\hat{f}^{(1)}}(y_T|y^{T-1}, h_1)}\Big) \\ &\overset{(a)}{=} \frac{1}{T}\Big(D(P^{\hat{f}^{(0)}}_{Y^{T-1}|h_0} \| P^{\hat{f}^{(1)}}_{Y^{T-1}|h_1}) + \sum_{y} Q_Y^0(y) \log \frac{Q_Y^0(y)}{Q_Y^1(y)}\Big) \\ &\overset{(b)}{=} \frac{1}{T}\Big(T \times \sum_{y} Q_Y^0(y) \log \frac{Q_Y^0(y)}{Q_Y^1(y)}\Big) \\ &= D(Q_Y^0(y) \| Q_Y^1(y)), \end{aligned} \tag{61}$$

where (a) holds due to the fact $Q_Y^i(y) \overset{(\Delta)}{=} P^{\hat{f}^{(i)}}_{Y_t|Y^{t-1}=y^{t-1}, h_i}$, $\forall y^{t-1}$, and (b) follows from iteratively applying the chain rule of KL-divergence.

On observing the single-letter expression (61) given in Corollary 3, the following corollary is proposed as a consequence.

Corollary 3. *The time-invariant policies $\hat{f}^{(i)}$ will achieve the zero-lower bound of Kullback-Leibler divergence, i.e., perfect privacy, if and only if the distributions of Y used for constructing $\hat{b}_t^{(i)}$ (decided by $\hat{f}_t^{(i)}$) are equal, i.e., $Q_Y^0(y) = Q_Y^1(y)$, $\forall y$.*

Remark 6. *Due to the physical constraints of the system we may not find the feasible Q_Y^0 and Q_Y^1 that satisfy the condition in the above corollary.*

Next, we present a case in which the perfect privacy cannot be achieved. It follows from (1) that the per-step expected energy amount constraint should hold as:

$$\mathbb{E}_{P_X}[X_t] + \mathbb{E}_{\gamma_t(s)}[S_t] - \mathbb{E}_{\gamma_{t-1}(s)}[S_{t-1}] = \mathbb{E}_{Q_Y}[Y_t], \tag{62}$$

which means, at each time step, the average expected amount of energy requested from the grid should be equal to sum of the average expected amount of energy consumed by the consumer and the average expected amount of energy stored into the ES. Since $\gamma_t = \gamma_{t-1}$ under a stationary strategy, equation (62) then further reduces to $\mathbb{E}_{P_X}[X_t] = \mathbb{E}_{Q_Y}[Y_t]$. In this case, we cannot have arbitrary Q_Y^i for constructing $\hat{b}_t^{(i)}$. Instead, Q_Y^i needs to satisfy the constraint $\mathbb{E}_{P_X}[X_t] = \mathbb{E}_{Q_Y}[Y_t]$ for both hypotheses. For the perfect privacy case, in order to satisfy the condition $Q_Y^0(y) = Q_Y^1(y)$, we should at least have $\mathbb{E}_{Q_Y^0}[Y] = \mathbb{E}_{Q_Y^1}[Y]$, which will conflict with the practical case of $\mathbb{E}_{P_X^0}[X] \neq \mathbb{E}_{P_X^1}[x]$. However, only considering the constraint on expected energy amount is not enough, since several different distributions of Y might lead to the same expectation and there might be some other constraints which requires $Q_Y^0(y) \neq Q_Y^1(y)$. Thus, the condition $Q_Y^0(y) = Q_Y^1(y)$ may still not be satisfied even if we have $\mathbb{E}_{Q_Y^0}[Y] = \mathbb{E}_{Q_Y^1}[Y]$. The above analysis can be summarized in the following proposition.

[7] The energy supply y should lie in this feasible set due to the constraint of ES capacity.

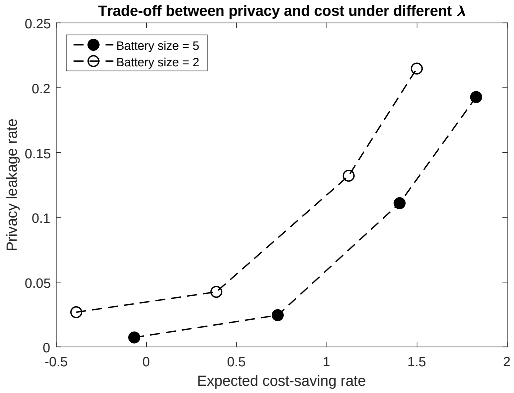


Fig. 3. Trade-off between privacy leakage and cost-saving for different ES sizes. From left to right the data points represent: $\lambda = 1, 0.8, 0.2, 0$.

TABLE I
VALUE COMPARE BETWEEN DIFFEREN λ FOR BATTERY SIZE 5 IN FIG. 3

λ	Privacy leakage rate	Expected cost-saving rate
0	0.1928	1.8276
0.2	0.1109	1.4034
0.8	0.0245	0.7277
1	0.0073	-0.0675

TABLE II
VALUE COMPARE BETWEEN DIFFEREN λ FOR BATTERY SIZE 2 IN FIG. 3

λ	Privacy leakage rate	Expected cost-saving rate
0	0.2148	1.4987
0.2	0.1321	1.1211
0.8	0.0425	0.3875
1	0.0268	-0.389

Proposition 5. *If the consumer's expected demand of energy under h_0 and h_1 are different, i.e., $\mathbb{E}_{P_X^0}[X] \neq \mathbb{E}_{P_X^1}[X]$, the perfect privacy cannot be achieved, since we cannot find Q_Y^0 and Q_Y^1 satisfying the conditions in Corollary 3.*

IV. NUMERICAL RESULTS

A. Finite Horizon Dynamic Programming

For simplicity we do not include units in the following. We consider a finite horizon with length $T = 10$. The energy demand, supply and price alphabets are set as $\mathcal{X} = \{0, 1, 2, 3, 4, 5\}$, $\mathcal{Y} = \{0, 1, 3, 4, 5\}$, $\mathcal{P} = \{5, 10\}$, and the ES capacity can be $s_{max} = 2$ or $s_{max} = 5$. The transition probabilities of x_t under both hypotheses, the transition probability of p_t and the initial belief state are set as following:

$$\begin{gathered} P(h_0) = P(h_1) = 0.5, \\ P_{X_{t+1}|h_0,X_t}(x_{t+1}|h_0, x_t) = \frac{1}{6}, \forall x_t, x_{t+1} \in \{0, 1, 2, 3, 4, 5\}, \\ P_{X_{t+1}|h_1,X_t}(x_{t+1}|h_1, x_t) = 0.4, \forall x_{t+1} = x_t \\ P_{X_{t+1}|h_1,X_t}(x_{t+1}|h_1, x_t) = 0.12, \forall x_{t+1} \neq x_t \\ P_{P_{t+1}|P_t}(p_{t+1}|p_t) = 0.5, \forall p_{t+1} \in \{5, 10\}, p_t \in \{5, 10\}, \\ \theta(x_1, s_1, p_1|h_0) = \theta(x_1, s_1, p_1|h_1) = \frac{1}{12 \times (s_{max} + 1)} \\ \forall x_1 \in \{0, 1, 2, 3, 4, 5\}, s_1 \in \{0, 1, ...s_{max}\}, p_1 \in \{5, 10\}. \end{gathered} \tag{63}$$

For the reason of simplification, the continuous belief state space is discretized into 36 different distributions (for ES capacity $s_{max} = 2$) and 100 belief sates (for ES capacity $s_{max} = 5$), i.e., 36^2 or 100^2 belief state vectors in total. Also, we have a finite action set $\mathcal{A}'$ including 20 different actions. At first, for different battery capacities, we investigate the trade-off between privacy enhancement and cost-saving by setting $\lambda = 1, 0.8, 0.2, 0$. The variation of privacy leakage rate against expected cost-saving rate with respect to λ is shown in Fig. 3. As λ increases, both the corresponding privacy leakage rate and expected cost-saving rate increase, which confirms the intuition that more cost-saving can be achieved at a cost of larger privacy leakage. We can also see from the figure that the performance will improve when the ES capacity gets larger.

B. Experiments for Solutions over Infinite Time Horizon

In this section, we compare the optimal and sub-optimal solutions to our belief-state MDP problem over the infinite time horizon, where the optimal solution is derived by the relative value iteration (RVI) and the sub-optimal one is derived by the approximated linear function approximated relative Q-learning (LARQL).

For this part, we have the ES capacity as $s_{max} = 5$ and the finite action set $\mathcal{A}'$ includes 20 different actions. The continuous belief state space is first discritized into 36^2 belief states. We compare the optimal and sub-optimal overall objective function (weighted sum between privacy leakage and cost savings) derived by RVI and LARQL approach. As we can see from Fig. 4, the performance of our proposed LARQL algorithm is close to be optimal. We can also see in the figure that the gap between LARQL and the optimal solution when $\lambda = 1$ is larger than the gap for $\lambda = 0$, which means that our linear function can approximate the Q-function when only induced by expected cost-saving ($\lambda = 0$) better than when only induced by the KL-divergence term ($\lambda = 1$).

Next, as shown in Fig. 5, with the increased number of belief states to be 4368, i.e., 4368^2 possible belief state vectors in total, the sub-optimal results using LARQL are shown in the figure, while the RVI method is too time-consuming and it is impractical to get an exact solution.

Based on the result, we notice that the performance of L-ARQL improves a lot with larger amount of belief states. The gap when $\lambda = 0$ is larger is again because our linear approximator can approximate Q-function induced only by expected cost-savings better than the Q-function only induced by KL-divergence. Besides, the performance of L-ARQL with 4368^2 is better than the optimal solutions with 36^2 belief states, this is due to the larger amount of belief states offers the system more freedom.

C. Experiments on real data

In this section, to better demonstrate the working mechanism of our proposed algorithm, we present our simulation results using real data from the reference data set REDD [40]. We consider a kitchen with a dishwasher which has

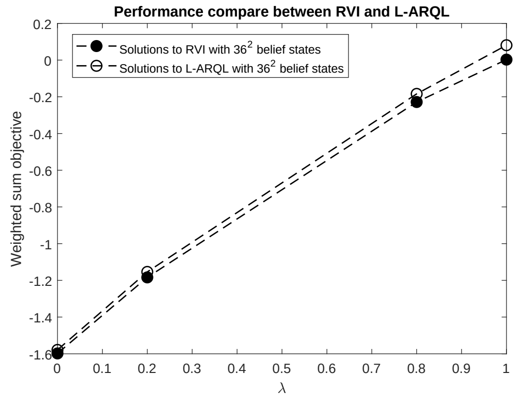


Fig. 4. Performance compare between LARQL and RVI with 36^2 belief states.

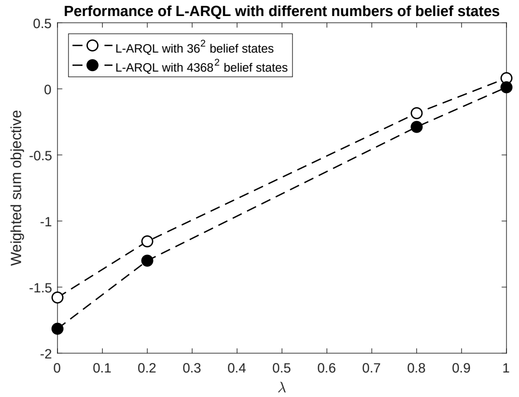


Fig. 5. Performance compare between LARQL with 36^2 and 4368^2 belief states.

two different operation modes: types A (hypothesis h_0) and B (hypothesis h_1). Over a time horizon with 1000 sampling instances, the load signatures of the two different operation types are illustrated in the upper two figures of Fig. 6. Both operation modes involve three different states $x \in [10, 200, 1100]$. The duration of staying in a state and the transition probabilities between different states depend on the operation mode. Under each operation mode, we learn the different initial state distributions and the state transition probabilities by calculating the empirical distributions utilizing the corresponding training dataset. In this case, given a specific operation mode (either A or B), i.e., under hypothesis h_i, the probability that the dishwasher is initially at operating state k is approximated by:

$$P(x_1 = k|h_i) = \frac{\# : \mathit{Appearance\ of\ initial\ state\ k\ under\ hypothesis\ h_i}}{\# : \mathit{Running\ times\ under\ hypothesis\ h_i}}. \tag{64}$$

And the transition probability from operating state k at time step t to operating state j at time step $t+1$ is approximated by:

$$P(x_{t+1} = j|x_t = k, h_i) = \frac{\# : \mathit{Transition\ from\ state\ k\ to\ j\ under\ hypothesis\ h_i}}{\# : \mathit{All\ transitions\ from\ state\ k\ under\ hypothesis\ h_i}}. \tag{65}$$

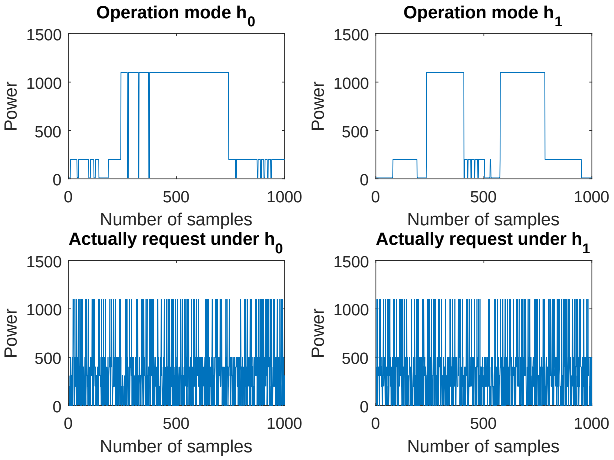


Fig. 6. Upper figures show dishwasher signatures of two different operation modes (hypothesis). Lower figures show realizations of the requested energy profiles for both hypotheses when privacy-preserving policy is used. Operation modes are hard to differentiate due to randomness in profiles.

We restrict the energy supply to take values within the set $\mathcal{Y} = [0, 10, 200, 310, 400, 500, 1100]$. The battery level is quantized into $[0, 200, 800]$ and we further assume the initial battery level is uniformly distributed within this set, i.e., $P(s_1 = 0|h_i) = P(s_1 = 200|h_i) = P(s_1 = 800|h_i) = \frac{1}{3}$. In this case, the initial belief state can be calculated by $\theta(x_1, s_1|h_i) = P(x_1|h_i)P(s_1|h_i)$[8], and the belief state will evolve according to different actions (MDP solution), corresponding observations as well as the transition probabilities we learned above. For a time-horizon of 1000 samples, we implement and simulate our privacy-preserving energy management strategy derived from our finite horizon belief-state MDP design. Two realizations of the requested energy profiles for both operation modes, i.e., the random output of our energy management strategy, are presented in the lower two figures of Fig. 6. From the visual comparison of the profiles, we can see that it becomes very difficult for the AD to identify the hypothesis from the energy supply data.

V. Conclusion

In this work, we have shown that an energy storage can be used for both privacy enhancing and cost saving. Using the belief-state MDP framework, an energy management strategy that optimally trade-offs KL-divergence and expected cost-saving rates can be derived using the Bellman dynamic programming. The complexity of the optimal design problems

[8] In this experiment, under each hypothesis, the belief state space is again discretized into 36 different distributions. In this case, the approximated initial belief state derived above should be further quantized to its nearest neighbor among those 36 distributions.

$$\xi_{t+1}(w_{t+1}) = \frac{\sum\limits_{(x_{t+1},s_{t+1})\in\mathcal{D}(w_{t+1})}\sum\limits_{x_t,s_t}\theta_t(x_t,s_t)a_t(y_t|x_t,s_t)P_X^0(x_{t+1})\mathcal{I}_{s_{t+1}}\{y_t+s_t-x_t\}}{\sum\limits_{x_t,s_t}a_t(y_t|x_t,s_t)\theta_t(x_t,s_t)}, \tag{66}$$

$$\xi'_{t+1}(w_{t+1}) = \frac{\sum\limits_{(x_{t+1},s_{t+1})\in\mathcal{D}(w_{t+1})}\sum\limits_{w_t}\sum\limits_{(x_t,s_t)\in\mathcal{D}(w_t)}\theta_t(x_t,s_t)a_t(y_t|x_t,s_t)P_X^0(x_{t+1})\mathcal{I}_{s_{t+1}}\{y_t+s_t-x_t\}}{\sum\limits_{w_t}\sum\limits_{(x_t,s_t)\in\mathcal{D}(w_t)}a_t(y_t|x_t,s_t)\theta_t(x_t,s_t)}. \tag{67}$$

grows quickly, which calls for computationally efficient solutions. Our proposed sub-optimal linear function approximated relative Q-learning approach is computationally efficient and also works for an infinite time horizon. With the identified feature vector, the linear function approximated Q-function can be efficiently learned and therefore leads to a practical online energy management design approach. Another approach to reduce the strategy design complexity is to assume an i.i.d. energy demand, which allows further analysis, in particular the derivation of a steady state strategy. Moreover, we provide sufficient conditions to achieve perfect privacy. Our numerical experiments show that the framework leads to energy management strategies that optimally trade-offs privacy enhancing and cost saving. They also show that our proposed LARQL method is close to optimal performance but is significantly computationally more efficient.

As future extensions of the current work, a potential direction would be studying the privacy-preserving problem under the multiple hypothesis testing scenario and designing the corresponding privacy-preserving energy management strategy. Also. in this work, we adopted the natural choice of Q-learning as algorithm that following the value function based approach under the proposed MDP framework. Since there also exists other efficient reinforcement learning techniques to deal with the continuous state-action space MDP problem, e.g., policy gradient algorithms, another interesting extension would be to implement such algorithms and assess which approach can provide a better solution to our proposed privacy-cost trade-off problem.

Appendix A
Proof of Lemma 2

Since $\xi_t(w_t)$ and $\gamma_t(s_t)$ are linearly related to each other by $\xi_t(w_t) = \sum\limits_{(x_t,s_t)\in\mathcal{D}(w_t)} P_X(x_t)\gamma_t(s_t)$, it is then sufficient to show $\xi_t(w_t) = \xi'_t(w_t),\ \forall w_t \in \mathcal{W}$ at each time step.

In the following, we use the induction method to prove that $\xi_t(w_t) = \xi'_t(w_t),\ \forall w_t \in \mathcal{W}$ at each time step. For $t = 1$, the initial distributions $\xi_1(w)$ and $\xi'_1(w)$ are identical since they do not depend on the actions a_t or b_t. Then, for any $t > 1$, given $\xi_t(w_t) = \xi'_t(w_t),\ \forall w_t \in \mathcal{W}$, we need to show $\xi_{t+1}(w_{t+1}) = \xi'_{t+1}(w_{t+1}),\ \forall w_{t+1} \in \mathcal{W}$ holds.

Knowing that $\xi_{t+1}(w_{t+1}) = \sum\limits_{(x_{t+1},s_{t+1})\in\mathcal{D}(w_{t+1})} \theta_t(x_t,s_t)$, we can derive the expression for $\xi_{t+1}(w_{t+1})$ as shown in (64). Meanwhile, $\xi'_{t+1}(w+1)$ can be expressed by Equation (55). According to Equation (53), we have $b_t(y_t|w_t)\xi_t(w_t) = \sum\limits_{(x_t,s_t)\in\mathcal{D}(w_t)} a_t(y_t|x_t,s_t)\theta_t(x_t,s_t)$. Since $\xi_t(w_t) = \xi'_t(w_t),\ \forall w_t \in \mathcal{W}$ holds according to our assumption, we can substitute $b_t(y_t|w_t)\xi'_t(w_t)$ by $\sum\limits_{(x_t,s_t)\in\mathcal{D}(w_t)} a_t(y_t|x_t,s_t)\theta_t(x_t,s_t)$ in Equation (55), which then leads to the expression as shown in (65). On noticing that the operator $\sum\limits_{w_t}\sum\limits_{(x_t,s_t)\in\mathcal{D}(w_t)}$ is equivalent to $\sum\limits_{(x_t,s_t)}$, we can conclude $\xi_{t+1}(w_{t+1}) = \xi'_{t+1}(w_{t+1})$. Thus, according to the principle of induction, there is $\xi_t(w_t) = \xi'_t(w_t),\ \forall w_t \in \mathcal{W}$ at each time step.

Appendix B
Proof of Theorem 2

At first, $\hat{b}_t$ can be easily verified to be a feasible action which belongs to the set $\mathcal{B}_t$ by the following:

$$\begin{aligned} P_Y(y)\times&\gamma'_t(y+w) \\ &= P_Y(y)\times\gamma_t(y+w) \\ &= P^\pi(S_t = y+w, Y_t = y|Y^{t-1} = y^{t-1}, h_0) \\ &= P^\pi(W_t = w, Y_t = y|Y^{t-1} = y^{t-1}, h_0). \end{aligned} \tag{68}$$

We next show the sufficiency. Since γ'_t and ξ'_t are easily shown to be equivalent, it is sufficient to check if $\gamma'_t = \gamma'_1$ for all t. For a time invariant policy, it is then sufficient to show $\gamma'_2 = \gamma'_1$. Consider a realization s of S_2, y of Y_1, and $w = s - y$. If $y \in \overline{\mathcal{Y}}(w)$ holds, we have:

$$\begin{aligned} P^{\hat{f}}(S_2 = s, Y_1 = y) &= P^{\hat{f}}(W_1 = s-y, Y_1 = y) \\ &= \xi'_1(s-y)\hat{b}_1(y|s-y) \\ &= P_Y(y)\gamma'_1(s). \end{aligned} \tag{69}$$

Marginalize over all possible s, we can get $P_{Y_1}(y) = Q_Y(y)$. Divide both sides by $Q_Y(y)$, it follows that $\gamma'_2(s) = P^{\hat{f}}(S_2 = s|Y_1 = y) = \gamma'_1(s)$. Regarding the proof of necessity, we divide it into two parts, where we first show that under the time-invariant policy which leads to the steady state, $P_{Y_t|Y^{t-1}=y^{t-1}},\ \forall y^{t-1}$ remains identical.The joint distribution $P^{\hat{f}}_{W_t,Y_t|Y^{t-1},h_0}$ can be decomposed by the following:

$$\begin{aligned} P^{\hat{f}}_{W_t,Y_t|Y^{t-1},h_0} &= P^{\hat{f}}_{W_t|Y^{t-1},h_0} \times P^{\hat{f}}_{Y_t|W_t,Y^{t-1},h_0} \\ &= \xi'_t(w)b_t(y|w) \\ &\overset{(a)}{=} \xi'_1(w)b_1(y|w) \\ &= P^{\hat{f}}_{W_1,Y_1|h_0}, \end{aligned} \tag{70}$$

where (a) holds due to the stationarity of states ξ'_t. Marginalizing over W, we can get $P^{\hat{f}}_{Y_t|Y^{t-1},h_0} = P^{\hat{f}}_{Y_1|h_0},\ \forall y^{t-1}$, which proves the above lemma.

Given the conclusion $P^{\hat{f}}_{Y_t|Y^{t-1},h_0} = P^{\hat{f}}_{Y_1|h_0}$, $\forall y^{t-1}$ remains indentical, we further show that the structure of the strategy in (58) should always be satisfied with $Q_Y \overset{(\Delta)}{=} P_{Y_t|Y^{t-1}=y^{t-1}}$, $\forall y^{t-1}$. Considering a realization s of S_2, y of Y_1, and $w = s - y$. If $y \in \overline{\mathcal{Y}}(w)$ holds, we have,

$$\begin{aligned} \mathbb{P}^{\hat{f}}(S_2 = s, Y_1 = y) &= \mathbb{P}^{\hat{f}}(W_1 = s - y, Y_1 = y) \\ &= \xi_1'(s-y) b_1(y|s-y) \end{aligned} \tag{71}$$

Since $\gamma_2'(s) = \gamma_1'(s)$ holds due to the stationarity of the states, divide both sides by $P^{\hat{f}}_{Y_1}(y)$, we can get,

$$\begin{aligned} &\mathbb{P}^{\hat{f}}(S_2 = s | Y_1 = y) = \gamma_1'(s) = \frac{\xi_1'(s-y) b_1(y|s-y)}{P^{\hat{f}}_{Y_1}(y)} \\ &\implies b_1(y|w) = P^{\hat{f}}_{Y_1}(y) \frac{\gamma_1'(y+w)}{\xi_1'(w)} \end{aligned} \tag{72}$$

Since γ_t' and ξ_t' remain identical over whole time horizon, and there is $P^{\hat{f}}_{Y_t|Y^{t-1}} = P^{\hat{f}}_{Y_1}$, the above equation implies that the action b_t satisfies the structure in (58) with $Q_Y \overset{(\Delta)}{=} P_{Y_t|Y^{t-1}=y^{t-1}}$, $\forall y^{t-1}$ over whole time horizon.

References

[1] Y. Mo, T. H. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber-physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, Jan 2012.

[2] G. W. Hart, "Residential energy monitoring and computerized surveillance via utility power flows," *IEEE Technology and Society Magazine*, vol. 8, no. 2, pp. 12–16, June 1989.

[3] J. Kolter and T. Jaakkola, "Approximate inference in additive factorial hmms with application to energy disaggregation," in *Artificial intelligence and statistics*, 2012, pp. 1472–1482.

[4] M. Zeifman, "Disaggregation of home energy display data using probabilistic approach," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 1, pp. 23–31, 2012.

[5] Q. Liu, K. M. Kamoto, X. Liu, M. Sun, and N. Linge, "Low-complexity non-intrusive load monitoring using unsupervised learning and generalized appliance models," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 1, pp. 28–37, 2019.

[6] M. Figueiredo, B. Ribeiro, and A. de Almeida, "Electrical signal source separation via nonnegative tensor factorization using on site measurements in a smart home," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 2, pp. 364–373, 2014.

[7] Y. Kim, E. C. H. Ngai, and M. B. Srivastava, "Cooperative state estimation for preserving privacy of user behaviors in smart grid," in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct 2011, pp. 178–183.

[8] C. Efthymiou and G. Kalogridis, "Smart grid privacy via anonymization of smart metering data," in *2010 First IEEE International Conference on Smart Grid Communications*, Oct 2010, pp. 238–243.

[9] J. M. Bohli, C. Sorge, and O. Ugus, "A privacy model for smart metering," in *2010 IEEE International Conference on Communications Workshops*, May 2010, pp. 1–5.

[10] United States Court of Appeals for the Seventh Circuit, "Naperville smart meter awareness v. city of naperville," Tech. Rep., 2018.

[11] U.S. Constitution, "Fourth amendment-searches and seizures," Tech. Rep.

[12] "The EU General Data Protection Regulation," Available online: https://eugdpr.org/.

[13] S. Li, A. Khisti, and A. Mahajan, "Privacy-optimal strategies for smart metering systems with a rechargeable battery," in *2016 American Control Conference (ACC)*, July 2016, pp. 2080–2085.

[14] J. Yao and P. Venkitasubramaniam, "On the privacy-cost tradeoff of an in-home power storage mechanism," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2013, pp. 115–122.

[15] Y. You, Z. Li, and T. J. Oechtering, "Optimal privacy-enhancing and cost-efficient energy management strategies for smart grid consumers," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*, 2018, pp. 826–830.

[16] L. Yang, X. Chen, J. Zhang, and H. V. Poor, "Cost-effective and privacy-preserving energy management for smart meters," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 486–495, Jan 2015.

[17] O. Tan, J. Gómez-Vilardebó, and D. Gündüz, "Privacy-cost trade-offs in demand-side management with storage," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1458–1469, 2017.

[18] G. Giaconi and D. Gündüz, "Smart meter privacy with renewable energy and a finite capacity battery," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–5.

[19] E. Erdemir, P. L. Dragotti, and D. Gündüz, "Privacy-cost trade-off in a smart meter system with a renewable energy source and a rechargeable battery," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2687–2691.

[20] O. Tan, D. Gündüz, and H. V. Poor, "Increasing smart meter privacy through energy harvesting and storage devices," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, pp. 1331–1341, 2013.

[21] Z. Zhang, Z. Qin, L. Zhu, W. Jiang, C. Xu, and K. Ren, "Toward practical differential privacy in smart grid with capacity-limited rechargeable batteries," 2015.

[22] M. Backes and S. Meiser, "Differentially private smart metering with battery recharging," in *Data Privacy Management and Autonomous Spontaneous Security*, pp. 194–212. Springer, 2013.

[23] L. Sankar, S. R. Rajagopalan, S. Mohajer, and H. V. Poor, "Smart meter privacy: A theoretical framework," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 837–846, June 2013.

[24] Z. Li, T. J. Oechtering, and M. Skoglund, "Privacy-preserving energy flow control in smart grids," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2194–2198.

[25] R. R. Avula, T. J. Oechtering, J. Chin, and G. Hug, "Smart meter privacy control strategy including energy storage degradation," in *2019 IEEE Milan PowerTech*, June 2019, pp. 1–6.

[26] Z. Li, T. J. Oechtering, and D. Gündüz, "Smart meter privacy based on adversarial hypothesis testing," in *IEEE International Symposium on Information Theory (ISIT) 2017*, 2017, pp. 774–778.

[27] Z. Li, T. J. Oechtering, and D. Gündüz, "Privacy against a hypothesis testing adversary," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1567–1581, June 2019.

[28] G. Kalogridis, Z. Fan, and S. Basutkar, "Affordable privacy for home smart meters," in *2011 IEEE Ninth International Symposium on Parallel and Distributed Processing with Applications Workshops*, May 2011, pp. 77–84.

[29] J. X. Chin, T. Tinoco De Rubira, and G. Hug, "Privacy-protecting energy management unit through model-distribution predictive control," *IEEE Transactions on Smart Grid*, vol. PP, no. 99, pp. 1–1, 2017.

[30] F. Farokhi and H. Sandberg, "Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries," *IEEE Transactions on Smart Grid*, 2017.

[31] J. Koo, X. Lin, and S. Bagchi, "PRIVATUS: Wallet-friendly privacy protection for smart meters," in *Computer Security – ESORICS 2012*, Sara Foresti, Moti Yung, and Fabio Martinelli, Eds., Berlin, Heidelberg, 2012, pp. 343–360, Springer Berlin Heidelberg.

[32] V. Krishnamurthy, *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*, Cambridge University Press, 2016.

[33] J. Tsitsiklis, "NP-hardness of checking the unichain condition in average cost mdps," *Operations research letters*, vol. 35, no. 3, pp. 319–323, 2007.

[34] A. Gosavi, *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*, Springer US, 2015.

[35] R. Sutton and A. Barto, *Introduction to reinforcement learning*, vol. 2, MIT press Cambridge, 1998.

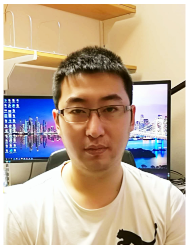
[36] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[37] F. Melo and M. Ribeiro, "Q-learning with linear function approximation," in *International Conference on Computational Learning Theory*. Springer, 2007, pp. 308–322.

[38] R. Zhang and P. Venkitasubramaniam, "Stealthy control signal attacks in linear quadratic Gaussian control systems: Detectability reward trade-off," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1555–1570, 2017.

[39] C. Bai, V. Gupta, and F. Pasqualetti, "On Kalman filtering with compromised sensors: Attack stealthiness and performance bounds," *IEEE Transactions on Automatic Control*, vol. 62, no. 12, pp. 6641–6648, 2017.

[40] J Zico Kolter and Matthew J Johnson, "REDD: A public data set for energy disaggregation research," in *Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA*, 2011, vol. 25, pp. 59–62.



Yang You (S'18) received the B.Eng. degree in information engineering from Beijing Institute of Technology, China, in 2014, the M.Sc. degree in electrical engineering from the KTH Royal Institute of Technology, Sweden, in 2017. He is currently working toward the Ph.D. degree with the division of Information Science and Engineering at the KTH Royal Institute of Technology. His research interests include statistical signal processing, sequential decision making, and information security and privacy.

Zuxing Li (S'14-18'M) received the B.Eng. degree in information security from Shanghai Jiao Tong University, China, in 2009, the M.Sc. degree in electrical engineering from the Technical University of Catalonia, Spain, and the KTH Royal Institute of Technology, Sweden, in 2013, and the Ph.D. degree in electrical engineering from the KTH Royal Institute of Technology in 2017. He was a postdoctoral researcher with CentraleSupélec, France, in 2018-2019, and with the KTH Royal Institute of Technology in 2019-2020. He has been an Assistant Professor with School of Electronics and Information Engineering, Tongji University, China, since September 2020. His research interests include statistical inference, information theory, reinforcement learning, and information security and privacy.

Tobias J. Oechtering (S'01-M'08-SM'12) received his Dipl-Ing degree in Electrical Engineering and Information Technology in 2002 from RWTH Aachen University, Germany, his Dr-Ing degree in Electrical Engineering in 2007 from the Technische Universität Berlin, Germany. In 2008 he joined KTH Royal Institute of Technology, Stockholm, Sweden and has been a Professor since 2018. In 2009, he received the "Förderpreis 200" from the Vodafone Foundation.

Dr. Oechtering is currently Senior Editor of IEEE Transactions on Information Forensic and Security since May 2020 and served previously as Associate Editor for the same journal since June 2016, and IEEE Communications Letters during 2012-2015. He has served on numerous technical program committees for IEEE sponsored conferences, and he was general co-chair for IEEE ITW 2019. His research interests include physical layer privacy and security, statistical signal processing, communication and information theory, as well as communication for networked control.