Optimal Pricing for Demand Response

***Abstract*—**

I. Introduction

A. Motivations and Techniques

Industrial and economic developments during last decades has led a significant growth in energy demand. Consequently, energy systems reliability and performance has been compromised due to this unprecedented increase (e.g. transmission line congestion, system security issues and etc.). However, one of the main solutions to deal with these issues is implementing demand response programs in energy systems. Demand response is an efficient form of load reduction to keep system from extreme changes which could jeopardize system performance requirements. This programs mainly provide monetary incentives or penalties to end-users to encourage them to change their level of demand according to supply side availability. As an implementation approach for demand response programs, load serving entity (LSE) uses pricing models (e.g. real time pricing (RTP), time of use pricing (TOUP), critical peak pricing (CPP) and etc.) to coordinate demand responses to the improvement of system reliability and performance. This pricing scheme allows the dynamic adjustment of electrical elastic load to adapt their consumption level according to energy generation costs. For example, LSE announces higher price rates at peak load due to higher cost of energy procurement. Then, consumers reduce their demand to avoid excessive payments.

In this paper, we consider an optimal pricing scheme to the demand response program from LSE point of view. LSE tends to supply the demand side by setting the best price rates to meet her own financial and technical objectives. For example, having a flat load profile could let her to have more robust supply even at times of sudden changes happening in the system. Moreover, lower generation level could lead to lower generation costs which in turn can increase the net profit of LSE. Besides, another important objective which should be taken into account is minimizing the incentive payments to participants in demand response program. It should be noted that the demand side responses to the announced price rates could have a significant impact on optimizing the LSE's objectives. However, it is very difficult to have a proper estimation of consumers' response function due to divergent consumption habituates of demand side. Hence, using learning algorithms to learn volatile behavior of demand side could pave the way to acquire an optimal pricing scheme according to the LSE's objectives.

B. Related Works

The authors in [**?**], [**?**] have considered reinforcement learning (RL) algorithm to solve the problem from user point of view.

II. Optimal Pricing for Demand Response

A. System Model

In this paper, we consider a discrete-time system, where in each time slot t, the LSE wishes to procure a total amount $d(t)$ of load reduction from a set of users $\mathcal{N}$. To procure this load reduction, at time slot t, the utility announces a price $p(t)$ and pays user $n \in \mathcal{N}$ the amount $p(t)s_n(t)$ when user n reduces consumption by $s_n(t) \geq 0$. The LSE's task is to choose a price $p(t)$ from a pricing plan $\mathcal{P}$ so that the LSE achieves the desired amount of curtailment.

Note that the reduction $s_n(t)$ depends on user n's reaction to the price. Let $r_n(\cdot)$ denote user n's response function to the price, which may depend on the energy usage state $e_n(t)$ of user n and other parameters, e.g., weather conditions. For example, given the same price, a user may reduce less power for heating in a cold day, compared with a warm day. In practice, these parameters of the response function are not available at the LSE and may differ across the users and change over the time. Therefore, given the price $p(t)$, the reduction $s_n(t)$ is a random quantity, i.e., $s_n(t) = r_n(p(t), e_n(t), x_n(t))$, where $x_n(t)$ denotes the unknown parameter to the LSE and is assumed to be independent and identically distributed.

Due to the uncertainty of users' responses, this curtailment may not match the demand response target $d(t)$. Any deviation from the target may incur some penalty to the LSE. Let $h(\cdot)$ denote the penalty function to capture the penalty of deviation from the target, which is assumed to be convex, non-negative, and has a global minimum $h(0) = 0$ (e.g., $h(x) = x^2$ is often considered in the literature [**?**], [**?**]). The penalty can be represented as $h(d(t) - \sum_n r_n(p(t), e_n(t), x_n(t)))$. Therefore, the LSE's overall system cost U at time slot t is equal to the sum of the penalty of deviation from the demand response target $d(t)$ and the total incentives paid to users for reducing their consumptions, i.e.,

$$\begin{aligned} U(p(t), d(t), \mathbf{e}(t), \mathbf{x}(t)) = & h(d(t) - \textstyle\sum_n r_n(p(t), e_n(t), x_n(t))) \\ & + p(t) \textstyle\sum_n r_n(p(t), e_n(t), x_n(t)), \end{aligned} \tag{1}$$

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

B. Problem Formulation

The objective of the LSE is to minimize the total expected system cost over time, as the LSE's overall system cost (**??**) is a random variable due to $\mathbf{x}(t)$. Therefore, the demand response problem can be formulated as the following stochastic optimization problem:

$$\begin{aligned} &\text{minimize} \ \lim_{t\to\infty} \tfrac{1}{t} \sum_{\tau=1}^{t} \mathbb{E}[U(p(\tau), d(\tau), \mathbf{e}(\tau), \mathbf{x}(\tau))] \\ &\text{subject to} \ \ p(\tau) \in \mathcal{P}, \forall \tau. \end{aligned} \tag{2}$$

Solving the proposed demand response problem (**??**) requires to find a pricing strategy π that maps each state $\{\mathbf{e}(t), d(t)\}$ to a price $p(t)$. One key challenge of deriving the optimal pricing strategy is that $r_n(\cdot)$ may not be available at the LSE, i.e., the LSE is not aware of users's response strategies, which is the major difference between this work and the existing studies, where users' response functions are often assumed to be availabile at the LSE or can be predicted accurately. Note that when $r_n(\cdot)$ is a linear function, it can be predicted from historical consumption data (see e.g., [**?**], [**?**], [**?**], [**?**]). However, such predictions are error prone and the prediction errors can be large when users' response functions are complicated. To tackle this challenge, we propose a reinforcement learning based approach such that the LSE can learn to adjust its pricing strategy adaptively based on users' responses.

C. Reinforcement Learning based Pricing Strategy

Specifically, the LSE determines the price based on the perception values. Each perception value corresponds to the LSE's current perception of the expected system cost when choosing a price at a given system state. Here the system state $S = \{d, \mathbf{e}\} \in \mathcal{S}$ includes the target $d \in \mathcal{D}$ and the energy usages of users $\mathbf{e} \in \mathcal{E}$, where $\mathcal{S}$ denotes the set of all possible states. Let $V_S^p(t)$ denote the perception of the expected system cost at time slot t when a price p is announced in state S. When a price p is announced, the LSE updates only the corresponding perception based on the perceived system cost in current state $S(t)$ from (**??**) as follows:

$$V_S^p(t) = \begin{cases} (1-\alpha_t)V_S^p(t-1) + \alpha_t U(t), & \text{if } p(t) = p, S(t) = S \\ V_S^p(t-1) & \text{otherwise} \end{cases} \tag{3}$$

where α_t is a learning rate parameter satisfying that $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$. $U(t)$ is the system cost at time slot t, which can be computed based on users' responses $s_n(t)$ at time slot t. Simply put, based on (**??**), the LSE updates only the perception of system cost in the current state under price p and keeps the perceptions in other states unchanged.

Then, the price for the next time slot is announced based on the pricing strategy $\pi_S(t) = \{\pi_S^p(t)\}_{p\in\mathcal{P}}$, where $\pi_S^p(t)$ denotes the probability of announcing the price p at state S. Using the Boltzmann distribution, the pricing strategy is given as follows:

$$\pi_S^p(t) = \frac{\exp(-\beta V_S^p(t))}{\sum_{p'\in\mathcal{P}} \exp(-\beta V_S^{p'}(t))}, \tag{4}$$

Algorithm 1 Reinforcement learning based pricing algorithm

Initialization: Given the set of states $\mathcal{S}$ and the pricing plan $\mathcal{P}$, set the initial perception values $V_S^p(0) = 1$.
For each time slot t
1) Select a price $p(t)$ according to (**??**).
2) Compute the perceived system cost according to (**??**) using users' responses $s_n(t)$.
3) Update the perception values $V_S^p(t)$ according to (**??**).

where β is the parameter that makes the exploration versus exploitation tradeoff. When β is large, the utility company will choose these prices with currently lower perceptions with higher probabilities. When β is small, the utility company will explore these prices with currently higher perceptions. Intuitively, the choice of β would play a very important role in finding the optimal pricing strategy for the LSE. And, a moderately small β is needed to increase the randomness of the pricing strategy, in order to ensure sufficient exploration over the possible prices to guarantee the convergence of the algorithm as well as small performance loss. The proposed reinforcement learning based pricing algorithm is summarized in Algorithm **??**.

D. Convergence of Reinforcement Learning based Pricing Strategy

We define the mapping function from the perception $\mathbf{V}(t)$ to the conditional expected cost value of the utility given state S and price p, as

$$R_S^p(\mathbf{V}(t)) \triangleq E\left[U(t)|\mathbf{V}(t), S(t) = S, P(t) = p\right] \tag{5}$$

where $E[\cdot]$ is taken with respect to the random variable $x(t)$.

Lemma 1: if the parameter β follows condition bellow,

$$\beta < \frac{1}{max_{U^p(t)\in\mathcal{U}}\{|U^p(t)|\}} \tag{6}$$

where,

$$\begin{aligned} U^p(t) = h\left(d(t) - \sum_n r_n(p, e_n(t), x_n(t))\right) \\ + p\sum_n r_n(p, e_n(t), x_n(t)),\ U^p(t) \in \mathcal{U} \end{aligned}$$

Then, the mapping function $\mathbf{R}_S(V(t)) \triangleq (R_S^p(\mathbf{V}(t)), \forall S \in \mathcal{S}, p \in \mathcal{P})$ forms a maximum norm contraction mapping, i.e.

$$\|\mathbf{R}_S(\mathbf{V}) - \mathbf{R}_S(\hat{\mathbf{V}})\|_\infty \le \epsilon \|\mathbf{V} - \hat{\mathbf{V}}\|_\infty$$

where $\epsilon \triangleq \beta max_{U^p(t)\in\mathcal{U}}\{|U^p(t)|\}$, $0 < \epsilon < 1$ and $\mathcal{U}$ is finite action-state space,.

Proof of Lemma 1:

Let define $\pi_S(t)$ as pricing strategy that, given the expected system cost perception is $\mathbf{V}(t)$, the state of system is S and

the announced price is p as follow,

$$\pi_S^p(t) = Pr\{p(t) = p|\mathbf{V}(t), S(t) = S\} \tag{7}$$

According to **??** we have that

$$R_S^p(\mathbf{V}(t)) = E\left[h\left(d(t) - \sum_n r_n(p, e_n(t), x_n(t))\right)\right.$$
$$\left.+p\sum_n r_n(p, e_n(t), x_n(t))\right] \times Pr\{p(t) = p|\mathbf{V}(t), S(t) = S\}$$

We then form $|R_S^p(\mathbf{V}(t)) - R_S^p\hat{\mathbf{V}}(t)|$ given two arbitrary perceptions $V(t)$ and $\hat{V}(t)$ as

$$|R_S^p(\mathbf{V}(t)) - R_S^p(\hat{\mathbf{V}}(t))| = \left|E\left[h\left(d(t) - \sum_n r_n(p, e_n(t), x_n(t))\right)\right.\right.$$
$$\left.\left.+p\sum_n r_n(p, e_n(t), x_n(t))\right] \times (\pi_S^p(t) - \hat{\pi}_S^p(t))\right|$$
$$= |E[U^p(t)] \times (\pi_S^p(t) - \hat{\pi}_S^p(t))| \tag{8}$$

Define a function $f(\mathbf{V}(t)) \triangleq \pi_S^p(t) = \frac{\exp(-\beta V_S^p(t))}{\sum_{p'\in\mathcal{P}}\exp(-\beta V_S^{p'}(t))}$ where $V_S^p(t)$ is perception value by being in state S and taking action p. Since the $f(V(t))$ is continuously differentiable, we know by Mean Value Theorem that there exist $f(\tilde{V}(t)) = \frac{f(\mathbf{V}(t)) - f\hat{\mathbf{V}}(t)}{\mathbf{V}(t) - \hat{\mathbf{V}}(t)}$ then,

$$\pi_S^p(t) - \hat{\pi}_S^p(t)$$
$$= \frac{\exp(-\beta V_S^p(t))}{\sum_{p'\in\mathcal{P}}\exp(-\beta V_S^{p'}(t))} - \frac{\exp(-\beta \hat{V}_S^p(t))}{\sum_{p'\in\mathcal{P}}\exp(-\beta \hat{V}_S^{p'}(t))} =$$
$$-\beta\frac{\exp\left(-\beta\tilde{V}_S^p\right)\sum_{p'\in\mathcal{P}}\exp\left(-\beta\tilde{V}_S^{p'}(t)\right) - \left(\exp -2\beta\tilde{V}_S^p(t)\right)}{\left(\sum_{p'\in\mathcal{P}}\exp(-\beta\tilde{V}_S^{p'}(t)\right)^2}$$
$$\times (V_S^p(t) - \hat{V}_S^p(t))$$
$$+\beta\sum_{p'\in\mathcal{P},p'\neq p}\frac{\exp\left(-\beta\tilde{V}_S^{p'}(t)\right)\exp\left(-\beta\tilde{V}_S^p(t)\right)}{\left(\sum_{p'\in\mathcal{P}}\exp(-\beta\tilde{V}_S^{p'}(t))\right)^2}$$
$$\times (V_S^p(t) - \hat{V}_S^p(t))$$

Let $C_p = \frac{\exp\left(-\beta\tilde{V}_S^p\right)\sum_{p'\in\mathcal{P}}\exp\left(-\beta\tilde{V}_S^{p'}(t)\right) - \exp\left(-2\beta\tilde{V}_S^p(t)\right)}{\left(\sum_{p'\in\mathcal{P}}\exp(-\beta\tilde{V}_S^{p'}(t)\right)^2}$ and $C_{p'} = \sum_{p'\in\mathcal{P},p'\neq p}\frac{\exp\left(-\beta\tilde{V}_S^{p'}(t)\right)\exp\left(-\beta\tilde{V}_S^p(t)\right)}{\left(\sum_{p'\in\mathcal{P}}\exp(-\beta\tilde{V}_S^{p'}(t))\right)^2}$. We can show that $C_p = \sum_{p'\in\mathcal{P},p'\neq p} C_{p'}$ and $2C_p \leq 1$. Thus, by using triangle inequality it can be obtained that,

$$|\pi_S^p(t) - \hat{\pi}_S^p(t)| \leq \beta C_p|V_S^p(t) - \hat{V}_S^p(t)|$$
$$+\sum_{p'\in\mathcal{P},p'\neq p} C_{p'}\beta C_{p'}|V_S^p(t) - \hat{V}_S^p(t)|$$
$$\leq \beta\left(C_p + \sum_{p'\in\mathcal{P},p'\neq p} C_{p'}\right)\|\mathbf{V}(t) - \hat{\mathbf{V}}(t)\|_\infty$$
$$\leq \beta\|\mathbf{V}(t) - \hat{\mathbf{V}}(t)\|_\infty \tag{9}$$

We can show by combining (**??**) and (**??**) that,

$$|R_S^p(\mathbf{V}(t)) - R_S^p(\hat{\mathbf{V}}(t))| \leq \beta E[|U^p(t)|]\|\mathbf{V}(t) - \hat{\mathbf{V}}(t)\|_\infty$$

thus,

$$\|\mathbf{R}_S(\mathbf{V}) - \mathbf{R}_S(\hat{\mathbf{V}})\|_\infty \leq \beta E[|U^p(t)|]\|\mathbf{V} - \hat{\mathbf{V}}\|_\infty \tag{10}$$

Then, if $\beta < \frac{1}{max_{U^p(t)\in\mathcal{U}}\{|U^p(t)|\}}$, $\mathbf{R}(\mathbf{V})$ follows a maximum contraction mapping, i.e., $\|R_S^p(\mathbf{V}(t)) - R_S^p(\hat{\mathbf{V}}(t)\|_\infty \leq \epsilon\|\mathbf{V}(t) - \hat{\mathbf{V}}(t)\|_\infty$ with $0 < \epsilon < 1$.

Now, we can show by using the property of contraction mapping that the $\{\mathbf{V}(t), \forall t \geq 0\}$ is a sequence converging to a unique fixed point (i.e. optimal point) $\mathbf{V}^*$.

Theorem 1: For the presented reinforcement learning algorithm, if the parameter β follows (**??**), then the sequence $\{\mathbf{V}(t), \forall t \geq 0\}$ converges to an optimal point $\mathbf{V}^*$.

Proof of Theorem 1:

We can show that the sequence $\{\mathbf{V}(t)\}$ is a Cauchy sequence, hence it will converge to a limiting point based on the optimality of fixed point i.e. $\lim_{t\to\infty}\mathbf{V}(t) = \mathbf{V}^*$ [**?**]. The detailed optimality proof of fixed point can be followed in [**?**].

Now, we investigate the property of the fixed point $\mathbf{V}^*$ as an equilibrium of the reinforcement learning algorithm. theorem 1 implies that the utility can achieve an accurate estimation of objective value based on perception V_S^{p*} at the equilibrium. The following Theorem shows an important result for perception value at $\mathbf{V}^*$ based on pricing strategy $\pi_S(t)$.

Theorem 2: For the reinforcement learning algorithm, the pricing strategy $\pi_S^*(t) = \{\pi_S^{p*}\}_{p\in\mathcal{P}}$ at the equilibrium $\mathbf{V}^*$ approximately minimize the expected cost function, i.e.,

$$\sum_{p\in\mathcal{P}}\pi_S^{p*}R_S^p(\mathbf{V}^*) \leq \min_{\pi_S}\left\{\sum_{p\in\mathcal{P}}\pi_S^p R_S^p(\mathbf{V}^*)\right\} - \varphi$$

where the approximation gap φ is at most $\frac{1}{\beta}\ln\mathcal{P}$.

Proof of Theorem 2:

First, we need to form an optimization problem that balances between pricing strategy exploitation and exploration.

Hence, we consider the following problem:

$$\max_{\pi_S} \left(-\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) - \frac{1}{\beta}\sum_{p\in\mathcal{P}} \pi_S^p \ln\pi_S^p \right)$$
$$\text{subject to} \sum_{p\in\mathcal{P}} \pi_S^p = 1, \pi_S^p \geq 0, \forall p \in \mathcal{P} \tag{11}$$

the first term in (**??**) indicates to performance of pricing strategy (i.e. exploitation) while the second term in (**??**) represents the entropy which measures the randomness of the pricing strategy (i.e. exploration). In other words, the minimization problem (**??**) tries to find the best trade-off between the price exploitation and exploration. To solve (**??**), since the problem function is convex, using KTT conditions can result in the optimal solution as follows:

$$\tilde{\pi}_S^p = \frac{\exp(-\beta R_S^p(\mathbf{V}^*))}{\sum_{p'\in\mathcal{P}} \exp(-\beta R_S^{p'}(\mathbf{V}^*))}$$

we know by the concept of fixed point that at the point $\mathbf{V}^*$, $R_S^p(\mathbf{V}^*) = V_S^{p*}$. Hence, it can be shown that $\tilde{\pi}_S^p = \pi_S^{p*}$, i.e. the pricing strategy at $\mathbf{V}^*$ is the optimal solution to the problem (**??**). Also, we can show that,

$$\sum_{p\in\mathcal{P}} \pi_S^{p*} R_S^p(\mathbf{V}^*) = \max_{\pi_S} \left(-\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) - \frac{1}{\beta}\sum_{p\in\mathcal{P}} \pi_S^p \ln\pi_S^p \right) + \frac{1}{\beta}\sum_{p\in\mathcal{P}} \pi_S^{p*} \ln\pi_S^{p*} \tag{12}$$

Furthermore, it is easy to verify that

$$\max_{\pi_S} \left(-\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) - \frac{1}{\beta}\sum_{p\in\mathcal{P}} \pi_S^p \ln\pi_S^p \right) = \min_{\pi_S} \left(\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) + \frac{1}{\beta}\sum_{p\in\mathcal{P}} \pi_S^p \ln\pi_S^p \right) \leq \min_{\pi_S} \left(\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) \right) \tag{13}$$

Since the uniform distribution results in maximum entropy, we can say that

$$-\sum_{p\in\mathcal{P}} \pi_S^{p*} \ln\pi_{S*}^p \leq \ln\mathcal{P}$$

then, it can be shown from (**??**) and (**??**) that

$$\sum_{p\in\mathcal{P}} \pi_S^{p*} R_S^p(\mathbf{V}^*) \leq \min_{\pi_S}(\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*)) + \frac{1}{\beta}\sum_{p\in\mathcal{P}} \pi_S^{p*} \ln\pi_S^{p*} \leq \min_{\pi_S}(\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*)) - \frac{1}{\beta}\ln\mathcal{P}$$

Thus, the theorem follows.

load.png

Fig. 1: Aggregated load profile before and after pricing

III. Numerical Result

In this section, the performance of pricing model will be evaluated by considering a numerical example. Also, we will show the impact of trade-off variable, β, in convergence of reinforcement learning algorithm.

As an experimental setup, we assume that the price set has action space $\mathcal{P} = \{0.1 : 0.15 : 1\}(\$/KWh)$, and the demand response target is equal to the mean value of historical data for evaluating time window. It should be noted that demand response target could be considered time variant. However, for the sake of clarity, we consider a constant value for our target. Furthermore, we assume that consumers' behaviors in different time slots are independent. This assumption implies that there is no constraint on total demand of consumers in the evaluating time window and pricing strategy plays important role in steering up the consumers' behaviors.

We first show the performance of our pricing methodology by showing the aggregated load profile of the consumers before and after new price rates which is shown in Figure **??**. As expected, the pricing strategy announces new prices according to consumers' usage state. In other words, the announced price rate in peak occasions is higher in order to encourage the consumers to decrease their energy usage and in the same approach, it is lower for time slots that the demand is less than considered demand response target, $d(t)$. Figure **??** illustrates that how the reinforcement learning algorithm learns over the consumers' behaviors based on the new price rates. For example, since the consumer demand is in the minimum state at times 2 to 5, the new prices are in the minimum rate to urge the consumers to enhance their consumptions. Besides, based on the defined utility cost function, there is a huge

Fig. 2: New price rates in different time slots

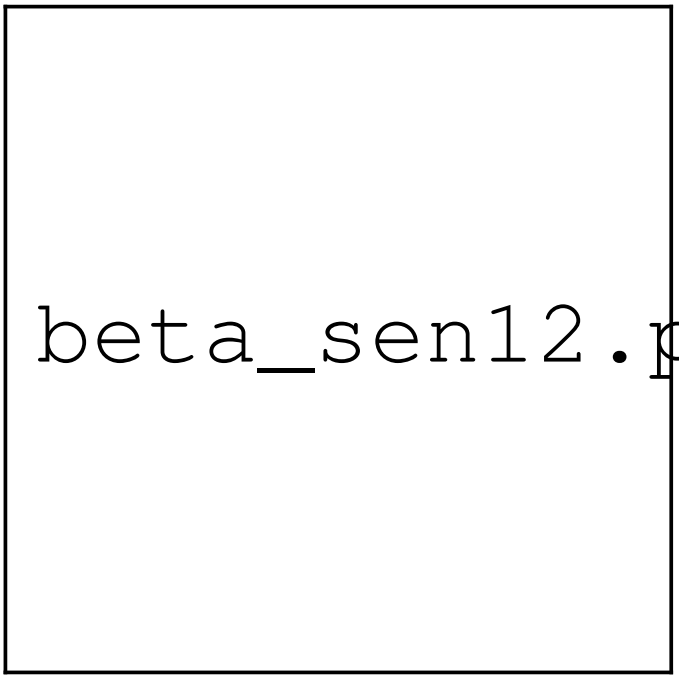

Fig. 3: System performance with different choices of β

gap between demand response target and energy usage state during these time slots. Hence, the learning algorithm takes into account both affecting values and announce the lowest price rate to minimize the utility cost function.

We then evaluate the performance of the reinforcement learning algorithm for different choices of β. Figure **??** shows the system performance based on the variation of β. It can be inferred that small β means the utility tends to choose price rates more randomly (i.e. more exploration) which results in a large performance gap according to Theorem 2. On the other hand, large β can keep the algorithm from converging due to over exploitation and also can affect the system performance gap negatively. In other words, our previously defined mapping distribution from perception to policy space (i.e. Boltzmann distribution) can be equalized as $\pi_S^{p*} = argmax_{p \in \mathcal{P}}\{\pi_S^p\}$ by assuming large β parameter which means the best policy in each iteration. Thus, there will be a proper β which would balance between exploration and exploitation and offers the best performance. In our example, $\beta = 0.5$ is achieved since it yields the best trade-off between exploration and exploitation and provides the best system performance.

IV. Conclusion