2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Online Learning for Incentive-Based Demand Response

Deepan Muthirayan, and Pramod P. Khargonekar

Abstract— **In this paper, we consider the problem of learning online to manage Demand Response (DR) resources. A typical DR mechanism requires the DR manager to assign a baseline to the participating consumer, where the baseline is an estimate of the counterfactual consumption of the consumer had it not been called to provide the DR service. A challenge of estimating the baseline is the incentive the consumers have to inflate the baseline. We consider the problem of learning online to estimate the baseline and to optimize the operating costs over a period of time under such incentives. We propose an online learning scheme that employs least-squares for estimation with a perturbation to the reward price (for the DR services or load curtailment) that is designed to balance the exploration and exploitation trade-off that arises with online learning. We show that, our proposed scheme is able to achieve a very low regret of $\mathcal{O}\left((\log T)^2\right)$ with respect to the optimal operating cost over T days of the DR program with full knowledge of the baseline, and is individually rational for the consumers to participate. Our scheme is significantly better than the averaging type approach, which only fetches $\mathcal{O}(T^{1/3})$ regret.**

I. Introduction

Demand Response (DR) programs [1] are potentially powerful tools to modulate the demand for electricity in a wide variety of situations. For example, at certain times such as mid-afternoons on hot summer days, the supply of additional electric power is scarce and expensive. At these times, it is more cost-effective to reduce demand than to increase supply to maintain power balance. Another scenario is a grid with high renewable penetration. Here, DR promises to be a better alternative compared to other expensive and polluting reserves to balance the variability in renewable generation. Realizing its potential, the 2005 Energy Policy Act provided the Congressional mandate to promote DR in organized wholesale electricity markets. The FERC order 745 [2] met this mandate by prescribing that demand response resource owners should be allowed to offer their demand reduction as if it were a supply resource rather than a bid to reduce demand so that the market operates fairly.

Dynamic pricing based DR programs can ideally achieve market efficiency, but they require complex metering and communication infrastructure to achieve this, which raises their implementation costs [3]. Furthermore, consumers may not be responsive to dynamic pricing [4]. Alternatively, consumers could be signaled to reduce consumption and paid for their load reductions. Such schemes are referred to as Incentive-based DR programs or Demand Reduction programs. There are two key components of any incentive-based DR program: (a) a baseline against which demand reduction is measured, (b) a payment scheme for agents who reduce their consumption from the baseline.

Thus, incentive-based DR programs require an established baseline against which consumer's load reduction is measured. The baseline is an estimate of the consumption when the consumer is not participating in the DR program. There are several ways to approach the estimation of baseline. One could, for example, use data to estimate the baseline. For example, the California Independent System Operator (CAISO) uses the average of the consumption on the ten most recent non-event days as the baseline estimate [5]. Data driven approaches can be broadly classified as (a) averaging, (b) regression, and (c) control group methods. Typically, these methods are prone to baseline manipulation [6]. There have have been reported cases in the past where the participants have artificially inflated their baseline for increasing payments [7]. Another class of approaches are based on *mechanism design*, where the consumers are elicited to report their baselines [8]–[10]. These approaches rely on suitably designed payment schemes to ensure that the manipulation or gaming of the reporting is minimal.

While many data-driven approaches for estimating baseline have been proposed in the literature [11]–[15], [15]–[20], they typically consider the offline setting where sufficient data is available for estimation prior to the start of the DR program. The limitation is that these approaches cannot be used when the data is limited or when the underlying conditions can change. The mechanism design approaches can avoid the need for learning, but have limitations because the consumers can be unwilling to reveal their baselines or can even be unaware of their baselines. These considerations necessitate approaches that can learn online while running the DR program without needing the consumers to report any information.

Contribution: In this work, we consider the problem of managing DR resources where a participating consumer's baseline is to be estimated online, i.e., while running the DR program. We consider the setting where the DR program can only learn from the consumption data that it gathers over the course of time. The unique challenge of our setting is that the DR program manager has to simultaneously learn the consumers' baselines and optimize its operating costs with the information it gathers along the way. This makes this problem a typical online learning problem. Therefore, the exploration-exploitation trade-off in any online learning problem applies to our problem as well. The added complexity in a setting like ours is the incentive the consumers have to interfere with the baseline estimation. Our formulation of the online learning DR problem incorporates all of these aspects. We propose an online learning DR scheme for this problem and show that our method achieves $\mathcal{O}((\log T)^2)$ regret with

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

respect to the optimal operating cost over T days of the DR program and is individually rational for the consumers to participate with an upfront payment for participation. Our main contribution is an online learning scheme for (incentive-based) DR that converges to the optimal cost with a very low regret and is at the same time individually rational. Ours is also the first work to formally study incentive-based DR as an online learning problem and present algorithms and regret guarantees.

A. Related Works

There exists substantial literature on baseline estimation methods [8], [9], [11]–[15], [15]–[20]. These can be broadly classified into four classes: (a) averaging, (b) regression, (c) control group methods and (d) baseline reporting approaches.

Averaging methods determine baselines by averaging the consumption on past days that are similar (*e.g.,* in weather conditions) to the event day. A detailed comparison of different averaging methods is offered in [11], [12], [14]. Averaging methods are simple but they suffer from estimation biases [14]–[16]. *Regression methods* estimate a load prediction model based on historical data which is then used to predict the baseline [13], [18]. They can potentially overcome biases incurred by averaging methods [15]. *Control group methods* have been suggested to have better accuracy than averaging or regression type methods and do not require large amounts of historical data [19]. However, these methods require the SO to recruit an additional set of consumers and install additional metering infrastructure. In addition, prior data based analysis might be required to identify the most appropriate control group depending on the control group method deployed. This can raise the costs of implementation [19].

Baseline reporting approaches were proposed in [9], [10], [21] as an alternative baseline estimation method. These approaches employ the framework of mechanism design to design payment and selection schemes to ensure that the consumers report the correct baseline values. While these methods can provably reduce the baseline error from that of averaging type methods [8], they violate privacy and are infeasible when the consumers can be unaware of their baselines.

In contrast to the above approaches, we propose an online approach that does not require large quantities of historical data, or a control group, or reporting private information.

Notation: We denote the expectation over a probability distribution by $\mathbb{E}[\cdot]$. We use $\mathcal{O}(\cdot)$ for the standard big-O notation while $\widetilde{\mathcal{O}}(\cdot)$ denotes the big-O notation neglecting the poly-log terms. We denote the sequence $(x_{m_1}, x_{m_1+1}, \dots, x_{m_2})$ compactly by $x_{m_1:m_2}$ and the sequence $(x^{m_1}, x^{m_1+1}, \dots, x^{m_2})$ compactly by $x^{m_1:m_2}$.

II. Demand Response Formulation

We consider the problem of managing Demand Response (DR) in an online setting, where the consumers' utility functions are unknown to the System Operator (SO) and the SO has to learn the necessary consumer relevant parameters online. In a typical demand response program, the SO recruits consumers for demand response and calls them to provide load curtailment on certain days. To incentivize the consumers to curtail, the SO typically pays a certain price (reward/kWh) for the load curtailment the consumers provide. Therefore, the consumer's response depends on the incentive or the reward to reduce, which is the price set by the SO. In addition to the payment for the DR services, the SO incurs an additional cost for serving the final consumption after load curtailment. Thus, the total cost for the SO depends on the payments for the DR services and the cost to serve the final load after the curtailment.

Typically, the SO can only observe the final consumption and not the load curtailment. Therefore, in addition to the price (reward/kWh), the SO needs to specify a baseline consumption to quantify the load curtailment. Baselines are estimates of the power that would be consumed had the consumer not been called to provide load curtailment. The SO, typically, announces the mechanism to assign the consumer's baseline to the consumers participating in the DR program. Thus, the SO's DR policy is the procedure to set the price (reward/kWh) and the baseline. The objective of the System Operator is to choose a DR policy that minimizes its overall cost.

We note that it is impossible for the SO to avoid under payment or over payment for the load curtailment without the knowledge of the consumer's correct baseline, which the SO need not know apriori. Here, we consider the setting, where the SO learns to set the correct baseline during the course of the DR program. The price and the baseline that the SO sets can vary from one day to the other and can be adapted depending on the response that the SO observes over its operation. The SO on any day has the following information: (i) the price for DR on all the previous days (ii) the baseline set for all the previous days and (iii) the final consumption of the consumers on all the previous days. The SO can use this information to set the price and the baseline consumption. Since the SO has to learn with the observations made on the fly and there is a cost that the SO incurs every day, this problem in effect is an online learning problem.

Like in a typical online learning problem, the SO has to balance the exploration and exploitation trade-off. Specifically, in this context, the SO cannot afford to set a constant price throughout to learn the correct baseline. This is because the total payment is a function of the assigned baseline, which creates an incentive for consumers to modify their consumption so as to inflate their future baselines and thence their future payments. Therefore, the SO has to be strategic in how the prices are set initially so as to infer the correct baseline over time. This is the exploration part. The exploration thus has to be balanced against deviating from the optimal outcome so as to ensure that on average the SO does not deviate from the best outcome. We characterize the effectiveness of the SO's DR policy, as is done typically in online learning, through a metric called *regret*, which in our case is the difference between the cumulative cost over a set of days and the achievable optimal cost with the full baseline information over the same set of days. Our objective is to develop an online learning scheme that achieves sublinear regret and thereby achieve an outcome that on average converges to the best DR outcome.

A. DR Setting

We index the days by t. We denote the number of consumers participating in the DR program by N. The SO, before any given day, decides whether to call a DR event or not. If it decides to call a DR event, it assigns a baseline $\hat{b}_t^i$ to the ith consumer participating in the DR program and the price for DR p_t (reward/kWh) prior to day t. The consumers are paid at the price p_t for the reduction of consumption from the assigned baseline. The price p_t is a reward or incentive for the consumer to reduce its consumption. The price and the baseline is set by the SO using the following information: (i) the price for DR on all the previous days (ii) the baseline set for all the previous days and (iii) the final consumption of the consumers on all the previous days. Thus, the SO can adapt its price and the baselines online as it makes newer observations. As in any DR program, the SO announces the procedure for setting the price for DR and the baseline prior to the first day, which is the SO's DR policy.

B. Consumer Model

We denote the electricity consumption of a consumer i on day t by q_t^i. Then, the payment P to consumer i for curtailing from $\hat{b}_t^i$ is given by

$$P_t^i = p_t(\hat{b}_t^i - q_t^i).$$

We denote the utility that the consumer derives from the electricity consumption q_t^i by

$$u_t^i = u^i(q_t^i, \epsilon_t^i) = \left(a^i + \epsilon_t^i\right) q_t^i - \frac{d^i (q_t^i)^2}{2},$$

where ϵ_t^i is a zero mean random variable and models the unpredictability or the uncertainty in the consumer's behavior. The assumption is that, by day t, the consumers observe their respective ϵ_t^is and that this information is private to them.

As in a typical power market, the consumers pay a retail price to the electricity provider for their daily consumption. We denote the retail price that the consumers pay by p_0. Therefore, the net utility to consumer i on day t is given by

$$U_t^i(q_t^i) = u_t^i - p_0 q_t^i + P_t^i.$$

The correct average baseline $\tilde{b}^i$ for a consumer i can be derived from the consumer's utility function. Following the definition that the correct baseline is the optimal consumption when the consumer is not called to provide DR, the correct average baseline for a consumer i is given by

$$\tilde{b}^i = \mathbb{E}_{\epsilon_t^i} b_t^i = \frac{a^i - p_0}{d^i}, \quad \text{where } b_t^i = \frac{a^i + \epsilon_t^i - p_0}{d^i}.$$

The optimal consumption in the hypothetical case when the set baselines are fixed to the correct values and do not depend on the past consumption can be derived by minimizing U_t^i individually. Therefore, the consumption for this hypothetical case is given by

$$s_t^i(p_t) = \arg\min_{q_t^i} U_t^i(q_t^i) = \frac{a^i + \epsilon_t^i - p_0 - p_t}{d^i}. \qquad (1)$$

Consumer's Optimal Decision: In a DR setting, since a consumer's current consumption determines the future baseline and payments, the consumer typically has an incentive to modify its consumption to influence the future baselines and the DR payments. To model this effect, we consider the setting, where a consumer's current decision is also determined by its effect on the outcome of the next m days. In this setting, the optimal consumer response on a day t is given by

$$q_t^{*i} = \arg\max_{q_{t:t+m}^i} \mathbb{E} \sum_{s=t}^{t+m} U_s^i(q_s^i), \qquad (2)$$

where expectation is over all randomness in $\epsilon_s^i, p_s, \hat{b}_s^i$ for all $s > t$.

C. System Operator's Objective

The system operator's decision variables on a day t are the price for DR and the baselines, which we collectively denote by $(p_t, \hat{b}_t^{1:N})$. We denote the aggregate of the assigned baselines on a day t by

$$\hat{b}_t = \sum_{i=1}^{N} \hat{b}_t^i.$$

Similarly, we denote the aggregate consumption on a day t by

$$q_t = \sum_{i=1}^{N} q_t^i.$$

Typically, the SO has to procure power from an external market to serve the demand of the consumers. Therefore, the SO incurs a cost for procuring the power consumed by the consumers. We denote the cost of procuring an unit of power by c. Therefore, the total cost that is incurred by the SO on a day t is the sum of the purchase cost and the cost for DR:

$$C_t(p_t, \hat{b}_t) = c q_t + p_t(\hat{b}_t - q_t).$$

Therefore, the SO's expected cost on day t conditioned on the set baseline and the price is given by

$$\widetilde{C}_t(p_t, \hat{b}_t) = \mathbb{E}[C_t(p_t, \hat{b}_t) | p_t, \hat{b}_t] = c\tilde{q}_t + p_t(\hat{b}_t - \tilde{q}_t),$$

where $\tilde{q}_t = \mathbb{E} q_t$. In the analysis, we assume that consumer chooses q_t according to Eq. (2).

SO's Objective: Let the price that minimizes SO's expected cost when the baselines are set to the correct values be denoted by p^*. Then, the consumption under this price, with the baselines set to the correct values, is given by $s_t^{*i} = s_t^i(p^*)$ for all i. Therefore, the optimal expected cost for the SO, when the baselines are set to the correct values, is given by

$$\widetilde{C}_t^* = c\tilde{s}_t^* + p^*(\tilde{b} - \tilde{s}_t^*),$$

where $\tilde{s}^*_t = \mathbb{E}_{\epsilon_t} s^*_t$ and $s^*_t = \sum_{i=1}^N s^{*i}_t$, $\tilde{b} = \sum_{i=1}^N \tilde{b}^i$. Since the primary objective of the SO is not to inflate the baseline and over pay, it is reasonable to define the regret with respect to the total optimal cost when the baselines are set to the correct values. Therefore, we define the SO's expected regret over a time period T, under a DR policy, as

$$R_T = \sum_{t=1}^{T} \left(\mathbb{E}[\widetilde{C}_t(p_t, \hat{b}_t)] - \widetilde{C}^*_t \right).$$

The SO's objective is to prescribe a DR policy such that

$$\lim_{T \to \infty} \frac{R_T}{T} = 0 \quad \text{(No Regret)}.$$

The SO has to achieve zero regret on average while ensuring that the consumer's individual rationality is satisfied on average, i.e.,

$$\lim_{T \to \infty} \frac{\mathbb{E}[\sum_{t=1}^T U^i_t(q^i_t)] - TU^{*i}}{T} \geq 0 \ \forall \ i,$$
$$\text{(Individual Rationality)},$$
$$\text{where } U^{*i} = \mathbb{E}[u^i(s^i_t(0), \epsilon^i_t) - p_0 s^i_t(0)]. \tag{3}$$

Remark 1 (Individual Rationality). The individual rationality condition in Eq. (3) is essential, since, otherwise the SO can set a very low baseline and under pay the consumers for the DR services. Thus, this condition is essential in the formulation. Moreover, the consumer will not participate in the DR program if the consumer does not receive a benefit that is on average at least as much as the benefit when not participating in the DR program. Therefore, to enforce this constraint, we set U^{*i} in the individual rationality condition as the optimal expected utility for a consumer when not participating in the DR program.

Remark 2 (Regret Definition). The question is whether $\widetilde{C}^*_t$ is appropriate as the cost to be compared with in the regret. It can be shown that if the SO inflates the baseline by a certain quantity $\Delta b > 0$ then the optimal expected cost necessarily increases till the incentives for the consumers to participate in the DR program are positive. Therefore, given that the primary objective is to mitigate over payment while ensuring individual rationality, it follows that $\widetilde{C}^*_t$ is the right candidate for the cost to be compared with.

III. Online Learning DR Mechanism

In this section, we discuss our algorithm and present the properties of our algorithm formally.

We recall that $q_1, q_2, q_3, ...$ denote the sequence of consumption by the consumer and $p_1, p_2, ...$ denote the sequence of price set by the SO for DR. The SO, at the end of a day t, calculates a $\hat{b}^i_{1,t+1}$ and $\hat{b}^i_{t+1}$ for each consumer i by

$$\begin{bmatrix} \hat{b}^{e,i}_{1,t+1} \\ \hat{b}^{e,i}_{t+1} \end{bmatrix} = \arg\min_{\hat{b}, \hat{b}_1} \sum_{k=1}^{t} (q^i_k - (\hat{b} - \hat{b}_1 p_k))^2. \tag{4}$$

The SO then assigns $\hat{b}^i_{t+1} = \hat{b}^{e,i}_{t+1}$ as the baseline for day $t+1$ to the ith consumer and calls all the consumers to provide DR service. The price p_t for DR is given by

$$p_t = p^* + \delta p e^{-t}, \tag{5}$$

where δp is a constant. In addition, the SO also pays a payment P_o to the consumer upfront. This payment is needed for meeting the individual rationality condition. Later, we give the specific form of this payment.

Algorithm 1 Online Learning DR Mechanism (OL-DRM)

```
Input: $N, P^i_o$ for each $i \in [1, N]$
Make the payment $P^i_o$ to each $i \in [1, N]$.
Announce the price sequence for the DR program as given
  by Eq. (5) and the process of baseline estimation.
Initialize $\hat{b}^i_1$ for each $i \in [1, N]$ arbitrarily.
for $t = [1, T]$ do
    Assign $\hat{b}^i_t$ as the baseline for each $i \in [1, N]$.
    Set $p_t$ as the price for DR.
    Receive the demand request $q^i_t$ from each consumer $i \in$
      $[1, N]$.
    Serve $q^i_t$ to each consumer $i \in [1, N]$.
    Incur the purchase cost $cq_t$ and the DR cost $p_t(\hat{b}_t - q_t)$.
    Update $\hat{b}^i_t \to \hat{b}^i_{t+1}$ for each $i \in [1, N]$ according to Eq.
      (4).
end
```

Remark 3 (Optimal Price for DR). For the consumer utility model considered here (1), it is straightforward to show that the optimal price p^*, given by the condition $\frac{d\widetilde{C}_t(p_t, \tilde{b})}{dp_t} = 0$, is $p^* = c/2$.

Remark 4 (Exploration Strategy). The prescribed policy for the SO explores by perturbing the price from the optimal price p^*. These perturbations cannot be persistent and therefore the prescribed policy reduces the perturbations with time. The decreasing of the perturbation is the balancing part of the exploration necessary to achieve sub-linear regret or "No Regret".

Definition 1.

$$\widetilde{\Delta}^i_t = \frac{1}{d^i} \sum_{j=1}^{m} \frac{p_{t+j} \left(-\sum_{s=1}^{t+j-1} p_s p_t + \sum_{s=1}^{t+j-1} p_s^2 \right)}{(t+j-1) \sum_{s=1}^{t+j} \left(p_s - \frac{1}{t+j-1} \sum_{l=1}^{t+j-1} p_l \right)^2}.$$

$$P^i_o = \sum_{t=1}^{T} p_t \left(\frac{\sum_{k=1}^{t} p^* \delta p e^{-k} \widetilde{\Delta}^i_k}{\sum_{k=1}^{t} \left(p_k - \frac{\sum_{l=1}^{t} p_l}{t} \right)^2} \right).$$

Next, we present the regret guarantee for our algorithm.

Theorem 1. *Under the Algorithm OL-DRM (Algorithm 1), with P^i_o given by Definition 1, for $\delta p = \mathcal{O}(1)$, and $T > 1$,*

$$R_T = \mathcal{O}((\log T)^2),$$

and individual rationality holds for each $i \in [1, N]$.

We give the detailed analysis in the next section. We observe that the regret guarantee for our approach leads to the desired “No Regret” and individually rational outcome.

Remark 5 (Approach). Our approach is the online equivalent of the regression approach to estimate baseline. An alternate approach is to not call the consumers for a certain number of days and use the consumption on these days to set the baseline for the future. This is the online equivalent of the averaging approach to estimate baseline. It can be shown that this approach leads to $\mathcal{O}(T^{1/3})$ regret. Our result, therefore, provides theoretical justification that regression type approaches can be superior to averaging type approaches.

Remark 6 (Extensions). Our setting assumes that the consumer utility model is quadratic, and the consumer's decision depends only on a finite horizon m. Our approach can be trivially extended to the infinite horizon case, where the future benefits are discounted by a discounting factor. For this case, the algorithm requires no change except for the definition of P^i_o. We can use the same proof technique to analyse this case to show that the same regret is achievable. The extension to a general utility model is a subject of future work.

IV. Regret Analysis

First, we derive an expression for $\hat{b}^i_t$.

Lemma 1. *Under the baseline estimation procedure of OL-DRM Algorithm 1, for any $t > 1$,*

$$\hat{b}^i_{t+1} = \frac{-\sum_{k=1}^t p_k \sum_{k=1}^t p_k q^i_k + \sum_{k=1}^t p_k^2 \sum_{k=1}^t q^i_k}{t\sum_{k=1}^t \left(p_k - \frac{\sum_{l=1}^t p_l}{t}\right)^2}.$$

Proof: Let

$$\Phi = \begin{bmatrix} \sum_{k=1}^t p_k^2 & -\sum_{k=1}^t p_k \\ -\sum_{k=1}^t p_k & t \end{bmatrix}.$$

The determinant of matrix Φ is given by

$$\text{Det}(\Phi) = t\sum_{k=1}^t \left(p_k - \frac{\sum_{l=1}^t p_l}{t}\right)^2.$$

Now, given how p_t is defined (Eq. (5)), $\left(p_k - \frac{\sum_{l=1}^t p_l}{t}\right)^2 > 0$ for $k = 1$ and any $t > 1$. Therefore, Φ is invertible for any $t > 1$. Therefore, from standard least-squares estimation solution, it follows that

$$\begin{bmatrix} \hat{b}^{e,i}_{1,t+1} \\ \hat{b}^{e,i}_{t+1} \end{bmatrix} = \Phi^{-1} \begin{bmatrix} -\sum_{k=1}^t p_k q^i_k \\ \sum_{k=1}^t q^i_k \end{bmatrix}.$$

By using the standard formula for the inverse of a matrix, which for a given matrix A is $\text{Adj}(A)^\top/\text{Det}(A)$, we get that, for any $t > 1$,

$$\begin{bmatrix} \hat{b}^{e,i}_{1,t+1} \\ \hat{b}^{e,i}_{t+1} \end{bmatrix} = \frac{\begin{bmatrix} t & \sum_{k=1}^t p_k \\ \sum_{k=1}^t p_k & \sum_{k=1}^t p_k^2 \end{bmatrix}}{t\sum_{k=1}^t \left(p_k - \frac{\sum_{l=1}^t p_l}{t}\right)^2} \begin{bmatrix} -\sum_{k=1}^t p_k q^i_k \\ \sum_{k=1}^t q^i_k \end{bmatrix},$$

i.e., $\hat{b}^{e,i}_{t+1} = \dfrac{-\sum_{k=1}^t p_k \sum_{k=1}^t p_k q^i_k + \sum_{k=1}^t p_k^2 \sum_{k=1}^t q^i_k}{t\sum_{k=1}^t \left(p_k - \frac{\sum_{l=1}^t p_l}{t}\right)^2}.$

Next, we derive an expression for the optimal consumer response given by Eq. (2).

Lemma 2. *The optimal consumer response under OL-DRM Algorithm 1 is given by*

$$q^{*i}_t = b^i_t - \frac{p_t}{d^i} + \widetilde{\Delta}^i_t, \quad \widetilde{\Delta}^i_t = \frac{1}{d^i}\sum_{j=1}^m p_{t+j}\frac{\partial \hat{b}^i_{t+j}}{\partial q^i_t},$$

$$\frac{\partial \hat{b}^i_{t+j}}{\partial q^i_t} = \frac{-\sum_{s=1}^{t+j-1} p_s p_t + \sum_{s=1}^{t+j-1} p_s^2}{(t+j-1)\sum_{s=1}^{t+j}\left(p_s - \frac{1}{t+j-1}\sum_{l=1}^{t+j-1} p_l\right)^2}.$$

Proof: Recall that the optimal consumer decision is given by

$$q^{*i}_t = \arg\max_{q^i_{t:t+m}} \mathbb{E}_{\epsilon^i_{t+1:t+m}}[J(q^i_{t:t+m})],$$

$$J(q^i_{t:t+m}) = \left(\sum_{j=1}^m U^i_{t+j}(q^i_{t+j})\right).$$

The general expectation in Eq. (2) reduces to the specific expectation in the equation above because (i) the price sequence for DR is set deterministically, and (ii) the fact that, at time t, the only randomness in the baselines to be assigned in the future, $\hat{b}^i_{t+j}$ for $j \in [1, m]$, is in the q^i_ks for $k \in [t+1, t+m]$; refer to Lemma 2 for the expression for $\hat{b}^i_t$.

Next, we observe that the only term in $U^i_{t+j}(q^i_{t+j})$ that is a function of q^i_t for all $j \in [1, m]$ is the assigned baseline $\hat{b}^i_{t+j}$; see Lemma 2 for the dependence of $\hat{b}^i_{t+j}$ on q^i_t. Given that $J(q^i_{t:t+m})$ is concave in q^i_{t+j}s and q^i_t, it follows that $\mathbb{E}_{\epsilon^i_{t+1:t+m}}[J(q^i_{t:t+m})]$ is also concave in q^i_t. Therefore, given the concavity, by applying first order condition for optimality, we get that the optimal q^{i*}_t satisfies

$$a^i + \epsilon^i_t - d^i q^{*i}_t - (p_0 + p_t) + d^i\widetilde{\Delta}^i_t = 0.$$

The final expression follows from here.

Remark 7. The optimal consumer response has the following terms: (i) b^i_t, the standard response when not participating in DR, (ii) the second term is the reduction incentivized by the reward/kWh p_t, and (iii) the third term is the inflation in consumption that arises from the incentive to inflate future baseline assignments.

In the next lemma, we derive an upper bound for the regret in terms of the cumulative error in the baseline estimation.

Lemma 3. *The regret under the OL-DRM Algorithm 1 is*

$$R_T = \sum_{t=1}^T (c - p_t)\widetilde{\Delta}_t + \sum_{t=1}^T \frac{(\delta p)^2 e^{-2t}}{d}$$
$$+ \sum_{t=1}^T p_t\mathbb{E}[\hat{b}_t - \tilde{b}] + P_o,$$

where $1/d = \sum_{i=1}^{N} 1/d^i, \quad \widetilde{\Delta}_t = \sum_{i=1}^{N} \widetilde{\Delta}_t^i, P_o = \sum_i P_o^i.$

Please see [22] for the proof.

Remark 8. The regret has the following terms: (i) the first term reflects the increase in the power purchase cost from consumption inflation and the decrease in DR payments from consumption inflation, (ii) the second term reflects the exploration cost that arises from the deviation from the optimal price, (iii) the third term reflects the increase in DR payments from baseline inflation, and (iv) the final term is the total payment made to the consumers upfront.

Next, we bound a key term that contributes to the consumption inflation term.

$$\Delta_{t,k} := \frac{p_{t+1}}{d}\left[\frac{-\sum_{s=1}^{t} p_s p_{t-k} + \sum_{s=1}^{t} p_s^2}{t\sum_{s=1}^{t}(p_s - \frac{\sum_{l=1}^{t} p_l}{t})^2}\right]. \qquad (6)$$

Lemma 4. *Under the OL-DRM Algorithm 1, for any* $t > 1$

$$\Delta_{t,k} = \mathcal{O}\left((p^* + \delta p)t^{-1}\right).$$

Please see [22] for the proof. The bound on the consumption inflation term $\widetilde{\Delta}_t$ follows from adding m terms of the type bounded in Lemma 4; see Lemma 2. In the next lemma, we derive an upper bound for the cumulative error in the baseline estimation and the payment P_o.

Lemma 5. *Under the OL-DRM Algorithm 1, for* $T > 1$*,*

$$\sum_{t=1}^{T} p_t \mathbb{E}[\hat{b}_t - \tilde{b}] + P_o = \mathcal{O}\left((\log T)^2\right).$$

Please see [22] for the proof. Then, combining Lemma 3 and Lemma 5, we get the final regret result. Also, see [22] for the proof for *individual rationality*.

V. Conclusion

In this work, we study the DR problem where the participating consumers' baselines have to be estimated online. The online nature of our baseline learning problem results in an exploration-exploitation trade-off between learning the baseline and optimizing the overall operating cost simultaneously, with an added complexity that the consumers can have incentives to inflate the baseline estimate. We propose a novel, online learning DR scheme for this problem and show that our approach achieves a low regret of $\mathcal{O}((\log T)^2)$ over T days of the DR program with respect to the best DR outcome with full information of the baselines and ensures that participating is individually rational for each consumer. The utility of our approach lies in the fact all prior approaches either require large quantity data or the consumers to report their baselines, both of which could be infeasible. Our contribution is a low regret online learning DR scheme.

References

[1] M. H. Albadi and E. El-Saadany, "A summary of demand response in electricity markets," *Electric power systems research*, vol. 78, no. 11, pp. 1989–1996, 2008.

[2] F. E. R. Commission, "Demand response compensation in organized wholesale energy markets," *Final Rule Report*, 2011.

[3] J. L. Mathieu, T. Haring, J. O. Ledyard, and G. Andersson, "Residential demand response program design: Engineering and economic perspectives," in *European Energy Market (EEM), 2013 10th International Conference on the*. IEEE, 2013, pp. 1–8.

[4] A. Faruqui and J. Palmer, "Dynamic pricing and its discontents," *Regulation*, vol. 34, no. 3, pp. 16 – 22, 2011.

[5] CAISO, *Demand Response User Guide. Version 4.3*, California ISO, May 2017.

[6] J. Vuelvas and F. Ruiz, "Rational consumer decisions in a peak time rebate program," *Electric Power Systems Research*, vol. 143, pp. 533–543, 2017.

[7] J. Pierobon. (2013) Two FERC settlements illustrate attempts to 'game' demand response programs. Last accessed 2019-3-30. [Online]. Available: https://www.energycentral.com/c/ec/ferc-settlements-illustrate-attempts-game-demand-response-programs

[8] D. Muthirayan, D. Kalathil, K. Poolla, and P. Varaiya, "Mechanism design for demand response programs," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 61–73, 2019.

[9] D. Muthirayan, E. Baeyens, P. Chakraborty, K. Poolla, and P. P. Khargonekar, "A minimal incentive-based demand response program with self reported baseline mechanism," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2195–2207, 2019.

[10] B. Satchidanandan, M. Roozbehani, and M. A. Dahleh, "A two-stage mechanism for demand response markets," *arXiv preprint arXiv:2205.12236*, 2022.

[11] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, "Estimating demand response load impacts: Evaluation of baseline load models for non-residential buildings in california," *Lawrence Berkeley National Laboratory*, 2008.

[12] C. Grimm, "Evaluating baselines for demand response programs," in *AEIC Load Research Workshop*, 2008.

[13] J. L. Mathieu, P. N. Price, S. Kiliccote, and M. A. Piette, "Quantifying changes in building electricity use, with application to demand response," *IEEE Transactions on Smart Grid*, vol. 2, no. 3, pp. 507–518, 2011.

[14] T. K. Wijaya, M. Vasirani, and K. Aberer, "When bias matters: An economic assessment of demand response baselines for residential customers," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1755–1763, 2014.

[15] S. Nolan and M. O'Malley, "Challenges and barriers to demand response deployment and evaluation," *Applied Energy*, vol. 152, pp. 1–10, 2015.

[16] Y. Weng and R. Rajagopal, "Probabilistic baseline estimation via gaussian process," in *IEEE Power & Energy Society General Meeting*, 2015.

[17] Y. Zhang, W. Chen, R. Xu, and J. Black, "A cluster-based method for calculating baselines for residential loads," *IEEE Transactions on smart grid*, vol. 7, no. 5, pp. 2368–2377, 2016.

[18] X. Zhou, N. Yu, W. Yao, and R. Johnson, "Forecast load impact from demand response resources," in *IEEE Power and Energy Society General Meeting*, 2016.

[19] L. Hatton, P. Charpentier, and E. Matzner-Løber, "Statistical estimation of the residential baseline," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1752–1759, 2016.

[20] F. Wang, K. Li, C. Liu, Z. Mi, M. Shafie-Khah, and J. P. Catalão, "Synchronous pattern matching principle-based residential demand response baseline estimation: Mechanism analysis and approach description," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6972–6985, 2018.

[21] D. Muthirayan, D. Kalathil, K. Poolla, and P. Varaiya, "Mechanism design for self-reporting baselines in demand response," in *2016 American Control Conference (ACC)*. IEEE, 2016, pp. 1446–1451.

[22] D. Muthirayan and P. P. Khargonekar, "Online learning for incentive-based demand response," *arXiv preprint arXiv:2303.15617*, 2023.