Proceedings of the 50th Hawaii International Conference on System Sciences | 2017
2017 50th Hawaii International Conference on System Science

Learning to Shift Thermostatically Controlled Loads

Antoine Lesage-Landry and Joshua A. Taylor
The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
University of Toronto
Toronto, Ontario, Canada
{alandry@ece., josh.taylor@}utoronto.ca

Abstract

Demand response is a key mechanism for accommodating renewable power in the electric grid. Models of loads in demand response programs are typically assumed to be known a priori, leaving the load aggregator the task of choosing the best command. However, accurate load models are often hard to obtain. To address this problem, we propose an online learning algorithm that performs demand response while learning the model of an aggregation of thermostatically controlled loads. Specifically, we combine an adversarial multi-armed bandit framework with a standard formulation of load-shifting. We develop an Exp3-like algorithm to solve the learning problems. Numerical examples based on Ontario load data confirm that the algorithm achieves sub-linear regret and performs within 1% of the ideal case when the load is perfectly known.

1. Introduction

The random nature of wind and solar power necessitates new sources of flexibility in power systems. Demand response, paying loads to modify their consumption to benefit the power system, can help accommodate renewables, improve power system efficiency, and ensure that supply and demand balance at all times [1]–[3]. Therefore, efficient and easy-to-use demand response (DR) models are a key development to average the demand and, thus, add flexibility to the power grid. In this work, we use thermostatically controlled loads (TCLs) to flatten the total power demand over time [4]–[6].

However, *a fundamental challenge* arises in DR: one must precisely know the model of the loads. Characterizing loads is difficult for several reasons, such as high number of the load, remoteness and inability to perform pilot studies. To address this challenge, we propose an online learning algorithm [7], [8] that learns the parameters of TCLs while using them for DR. Our approach allows, therefore, the aggregator to avoid any on-site measurement which would require an important deployment of resources.

The proposed approach is based on the *multi-armed bandit* framework [9], [10]. More precisely, the adversarial version [11] of the multi-armed bandit framework is used to determine which model or combination of models best fits the load. In this setting, each *arm* represents a potential model (set of thermal parameters of the load). The aggregator's or player's task is to simultaneously shift load while determining which arm yields to the best performance. We quantify performance in terms of a loss function, defined as the deviation from the predicted total power consumption to the observed one. This is the only observation that the aggregator has access to. Consequently, the aggregator does not have access to the feedback for all his potential model (arms), but only for the selected one. This limited feedback corresponds to the bandit setting, as opposed to the full information *expert* framework [9].

The multi-armed bandit was first formulated in [12] and later solved by the same author in [13]. Since the first formulation, the original problem has been divided into several families of bandit problem [10] which all express the exploration-exploitation tradeoff according to a different type of arm. In the stochastic bandit, the arms are characterized by an unknown probability distribution function and the player is looking to maximize its expected gain. The player's best strategy is then to build a policy based on the experimental mean gain using to the UCB family of algorithms [14]. On the other hand, in the adversarial bandit the gain (or loss) is fixed (randomly or deterministically) by an opponent or Nature. In this case, the best strategy can be found using the Exp3 family of algorithms [11]. Here we employ this version of the bandit problem. Finally, the third family of bandit is the Markovian bandit. The classical Markovian bandit and its *restless* extension can be solved using index policies derived respectively in [15] and [16].

More recently, these theoretical frameworks have been applied to demand response. For example, in [17],

URI: http://hdl.handle.net/10125/41522
ISBN: 978-0-9981331-0-2

[18], Markovian bandits were used to obtain a policy for curtailing TCLs. Then, in [19], the stochastic bandit was used to learn the curtailment signal response by the consumer. Online convex optimization has also been used in the demand response literature [20]–[23]. The proposed online formulation differs from all mentioned work and uses the adversarial bandit to directly learn load parameters which are central to the DR problem.

The novel contributions of our approach are:

- We apply bandit learning to load-shifting with TCLs. Specifically, we learn their models while utilizing them for DR.
- We invoke theoretical regret bounds from the literature that guarantee the performance of our approach.
- Our approach can flexibly accommodate a variety of load models.

2. Background

2.1. Load parameters

A TCL is characterized by the following thermal parameters [4]–[6]:

(i.) R, the thermal resistance [°C/kW];
(ii.) C, the thermal capacitance [kWh/°C];
(iii.) η, the coefficient of performance (COP);
(iv.) P_r, the thermal transfer power rate [kW].

Let $y(t)$ denote the control variable of the cooling or heating system.

2.2. Load power consumption

Let N denote the number of TCLs and M be the number of time instants of length h in a day. Let $p_n = h\,P_{r_n}/\eta_n$ and $\mathbf{p} \in \mathbb{R}^N$, be the vector $\left[\frac{hP_{r_1}}{\eta_1}, \frac{hP_{r_2}}{\eta_2}, \ldots, \frac{hP_n}{\eta_n}\right]^{\mathrm{T}}$. Finally for the controls, let $y_n = [y_n(0)\ y_n(1)\ \ldots\ y_n(M-1)]^{\mathrm{T}}$ denote the control vector for the n-th TCL. Then, the set of controls for all TCLs is expressed by $\mathbf{Y} \in \mathbb{R}^{M\times N}$ and is given by,

$$\mathbf{Y} = \begin{bmatrix} | & | & \ldots & | \\ \mathbf{y}_1 & \mathbf{y}_2 & \ldots & \mathbf{y}_N \\ | & | & \ldots & | \end{bmatrix}. \tag{1}$$

The power consumption at time t of the n-th TCL is,

$$C_n(t) = y_n(t)p_n, \tag{2}$$

and the total power consumption in the grid at each time instance is given by the following equation.

$$C(t) = \sum_{n=1}^{N} y_n(t)p_n + b(t), \tag{3}$$

for all $t = 1, 2, \ldots M-1$ where $b(t)$ is the inflexible baseload at time t. Equivalently,

$$\mathbf{C} = \mathbf{Y}\mathbf{p} + \mathbf{b}, \tag{4}$$

where $\mathbf{b} \in \mathbb{R}^M$ is the base-load vector.

2.3. Load constraints

The temperature of a thermostatically controlled load like a house or commercial building is constrained by its occupants' comfort requirements. This is often represented as a dead-band around a nominal desired temperature [4], [5], [24]. Let θ_d be the desired temperature and Δ be the dead-band width. Let θ_- and θ_+ be respectively the lower and upper bounds of the dead band defined as,

$$\theta_- = \theta_d - \Delta\,, \qquad \theta_+ = \theta_d. \tag{5}$$

Next, the discrete-time model developed by [25]–[27] is used to model the temperature inside the TCLs subject to ambient temperature changes and to the operation of the cooling system. This model is given by

$$\begin{aligned}\theta_n(t+1) = {} & a\,\theta_n(t) + (1-a)[\theta_{amb_n}(t) \\ & - y(t)R_nP_{r_n}] + w(t),\end{aligned} \tag{6}$$

where $a = \mathrm{e}^{-\frac{h}{R_nC_n}}$, $w(t)$ represents system noise, θ_{amb} is the ambient (outside) temperature. Finally, $y(t) \in \{0,1\}$ is the control variable sent to the cooling system. For further comparison, let $y_{NoDR}(t)$ be the control variable when no DR is attempted. [25] defined this control variable as,

$$y_{NoDR}(t+1) = \begin{cases} 0, & \text{if } \theta(t+1) < \theta_- \\ y_{NoDR}(t), & \text{if } \theta(t+1) \in [\theta_-, \theta_+]. \\ 1, & \text{if } \theta(t+1) > \theta_+ \end{cases} \tag{7}$$

2.4. Optimal offline load-shifting

In this section, a DR model is presented. This DR model aims at flattening the load while ensuring that the temperature of each TCLs is at all time inside its dead-band. This model uses the aggregated power consumption and the base-load instead of listed prices as used in [4] to directly target load averaging instead of financial savings.

Given the base-load $\mathbf{b}$ of the time period and all thermal parameters of the TCLs, the set of controls $\mathbf{Y}$ that averages the power demand and ensures a temperature inside the dead-band is the optimum of the following problem:

$$\begin{aligned} \min_{\mathbf{Y}\in\mathbb{R}^{N\times M}} \quad & ||\,\mathbf{Y}\mathbf{p} + \mathbf{b}||^2 \\ \textit{subject to} \quad & \mathbf{0} \le \bar{\mathbf{Y}} \le \mathbf{1} \\ & \boldsymbol{\theta}_- \le \widehat{\boldsymbol{A}}\boldsymbol{\theta}(0) + \\ & \boldsymbol{\mathcal{X}}\bar{\boldsymbol{\theta}}_{amb} + \boldsymbol{\mathcal{A}}\bar{\mathbf{Y}} \le\ \boldsymbol{\theta}_+ \end{aligned} \tag{8}$$

where the over-line over a matrix is the unfolding operator (all columns of the matrix are stacked to form a vector).

In this optimization problem, the L^2-norm is used to discourage variation in the power consumption and hence fill valleys in the base-load. The second constraint is the vector version of (6) for all time-steps and all TCLs. Finally, the first constraint is a relaxed version of the previously stated definition of the set of controls. In this relaxation, the convex hull is considered and all values inside the 0-1 interval are permitted for the **Y** variable. Note that this relaxation is not required to fit the theoretical framework and that a non-convex programs could still be used in the bandit framework. This relaxation is used in our simulation to allow the optimization problem to be convex and hence to be efficiently solved numerically using cvx [28], [29] and MOSEK [30]. In the context of TCLs, this relaxation means that one can set the intensity of the cooling system rather than only turning it on or off.

3. Optimal online load-shifting

Our objective is to optimally flatten the load to the best of the aggregator's knowledge while, at each round, improving his knowledge of the load. Indeed, at each time step (round) a prediction of the actual parameters of the TCLs is made and then using the feedback from the load, the prediction is improved for the next round. Hence, to handle an uncharacterized load, an online learning algorithm can be deployed.

In the following sections, since the power consumption of a load can be easily accessed by the aggregator, the focus will be given on learning the thermal transfer rate P_r which is directly related to the TCL power consumption (cf. equation (2) and (3)).

For the present online model, the following assumptions are made,

Assumption 1. the aggregator has access to an accurate estimate of the next day base-load **b**;

Assumption 2. the aggregator has access to an accurate estimate of the next day ambient temperature θ_{amb};

Assumption 3. the thermal capacitance *C* and thermal resistance *R* of the TCLs are known and constant;

Assumption 4. the aggregator observes the aggregated power consumption of all TCLs.

Note that Assumption 3 could be dropped in future extensions where the aggregator has access to a feedback on the temperature. Alternatively, a learning algorithm could be applied to learn these two parameters as well.

Due to the non-convex loss function that will be given in the next section, an expert-like approach is used. Let K be the number of arms and κ be the set of arms. Then, each arm represents a potential model for the load. The algorithm must then choose which one yields to the minimum loss when playing it. Multi-armed bandit problems balance the tradeoff between exploration, in this case testing different arms, and exploitation, using the arms that appears best at present [10]. In this context, the aggregator has to look for the model that best represents the loads while trying to flatten the power usage. This is opposed to the *full information* settings where the loss for each model would be observed [9]. Indeed, since the only feedback is the power usage which is a function of the model, only the power consumption of the computed control with respect to the predicted model can be observed.

Hence, this problem can be modeled as an adversarial bandit where the adversary fixes the loss for each model at each time instant without knowledge of the player's strategy. This makes the process an oblivious game. Therefore, a natural choice of algorithm to shift load while learning the model of the load is based on the *Exponential weights for Exploration and Exploitation* algorithm (Exp3) [10], [11]. This algorithm enjoys sublinear regret bounds and uses randomization to deal with the exploration and exploitation tradeoff. The algorithm functions by evolving a probability distribution over the arms, and in each time period sampling an arm from the distribution.

Remark 1. (Time-scale) Note that our approach uses two different time-scales. The first one is the intra-day load-shifting time step and is represented by h. This time-scale is only used by the DR optimization problem and is used for load flattening. The second time-scale represents rounds t for the online learning algorithm and has a length of a day.

Remark 2. (State reset) To ensure that all the bandit framework's assumptions are respected, the state (temperature) is reset between each round to the initial temperature which corresponds to the dead-band upper bound. This mathematical assumption is made to make sure that each round is not a function of the previous ones and hence to ensure that the adversary is oblivious. Note that in the TCLs setting, the reset has only a very small influence since the final temperature should be approximatively given by the dead-band upper bound.

3.1. Regret

The performance of an online algorithm is defined by its cumulative regret. This regret represents the loss incurred by the player's choice compared to the minimal loss suffered if the best arm (model) was always picked. Let R_T be the cumulative regret at round T,

$$R_T = \sum_{t=1}^{T} \ell(I_t, Z_t) - \min_{i \in \kappa} \sum_{t=1}^{T} \ell(i, Z_t), \quad (9)$$

where ℓ is the loss function, I_t the choice of arm at time t and Z_t is the observation following Nature's choice of model. This choice corresponds to the actual load parameters and can be indirectly observed as the aggregated power used at round t. Taking into consideration the randomization of the player, the expected regret is,

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell(I_t, Z_t)\right] - \min_{i \in \kappa} \sum_{t=1}^{T} \ell(i, Z_t). \quad (10)$$

Observe that the adversary is oblivious because Nature always selects the observation Z_t using the true load model. Also, note that Nature's strategy is deterministic (for each round there exists a one-to-one mapping from the chosen arm to a unique loss value). For this reason, the expected regret can be expressed as (10). Hence only I_t is a random variable and the expected value is computed with respect to the randomized strategy. We seek an online learning algorithm that achieves sublinear regret, which implies that it improves with each time step.

3.2. Loss function

We quantify the performance of the algorithm in each time step with the loss function

$$\ell(i_t, Z_t) = 1 - \exp\left[\frac{-|\mathbf{1}^{\mathrm{T}}\mathbf{Y}(i_t)\mathbf{p}(i_t) - Z_t|}{\alpha}\right], \quad (11)$$

with $\alpha > 0$ and where $Z_t = \mathbf{1}^{\mathrm{T}}\mathbf{Y}(i_t)\mathbf{p}_{\mathrm{real}}$ represents the observed power consumption. i_t denotes the selected arm at time t and is an element of κ. This value will be discussed in the next section. α is a positive tuning factor for controlling the size of the loss function. Note that $\ell(i_t, Z_t) \in [0,1]\ \forall\ (i_t, Z_t)$. Then, the optimal load shifting strategy for the arm i_t is given by,

$$\mathbf{Y}(i_t) = \underset{\mathbf{Y} \in \mathcal{F}(i_t)}{\operatorname{argmin}} \, || \, \mathbf{Y}\mathbf{p}(i_t) + \mathbf{b}||^2, \quad (12)$$

where the feasible set is as discussed in Section 2:

$$\mathcal{F}(i_t) = \{\overline{\mathbf{Y}} \in \mathbb{R}^{M \times N} \mid \mathbf{0} \leq \overline{\mathbf{Y}} \leq \mathbf{1}, \\ \boldsymbol{\theta}_- \leq \widehat{\boldsymbol{A}}\boldsymbol{\theta}(0) + \boldsymbol{\mathcal{X}}\overline{\boldsymbol{\theta}}_{amb} + \boldsymbol{\mathcal{A}}(i_t)\overline{\mathbf{Y}} \leq \boldsymbol{\theta}_+\}. \quad (13)$$

Remark 3. (Choice of approach) The online problem as stated is not convex in $\mathbf{p}$ (the learned parameter) and hence other online approaches like online gradient descent [31] or online mirror descent [7] cannot be used. To overcome this problem, an expert-like or bandit algorithm is used.

We also make use of estimates of the loss function for unselected arms. We use the unbiased estimator proposed in [10]:

$$\tilde{\ell}(i, Z_t) = \frac{\ell(i, Z_t)}{q_i(t)} \mathbb{I}_{I_t, i}, \forall\ i = 1, 2, \ldots, K, \quad (14)$$

where $\mathbb{I}_{I_t,i}$ is an indicator function and $q_i(t)$ is the probability mass associated with the i-th model.

3.3. Models (a.k.a. arms)

Each arm is a candidate set of parameters that models the load aggregation. Here, each TCL has an unknown parameter P_r which lies in the interval $[P_{r_{min}}, P_{r_{max}}]$ [6], [32]. This approach is similar to [24], in which experts represent different models of TCL aggregations. Each of the K arms is given by

$$Arm_k = [u_1^k \ u_2^k \ \ldots \ u_N^k], \quad (15)$$

where $u_i^k \sim Uniform[P_{r_{min}}, P_{r_{max}}]$.

Note that more arms, i.e. a larger value of K, increases the chances that there is a better model in the set of arms, but also increases the time needed for the algorithm to converge to the best arm or combination of arms.

3.4. Proposed algorithm for DR

We now give the algorithm, Exp3 for DR, for learning while load-shifting. Then, theoretical bounds on the regret are given in Proposition 1 and in Proposition 2.

Remark 4. (Exp3 for DR is an Exp3 algorithm) The DR problem with a partially uncharacterized load described here fits the multi-armed bandit framework which can be solved, with sub-linear regret, using the Exp3 algorithm [9]–[11]. The application of Exp3 to the DR context respects all assumptions and hence is an Exp3 algorithm.

Proposition 1. (Bounded regret of Exp3 for DR) *Let K be the number of models, t the rounds and T the time horizon. If $\eta_t = \sqrt{\frac{\ln K}{tK}}$, the expected regret of* Exp3 for DR *is bounded by,*

$$\mathbb{E}[R_T] \leq 2\sqrt{TK \ln K}. \quad (16)$$

The proof is given in [10] for the Exp3 algorithm for the pseudo-regret $\bar{R}_T$. Then, [33] showed that $\mathbb{E}[R_T] = \bar{R}_T$ when the adversary is deterministic, yielding to the previous result.

Proposition 1 implies that the proposed algorithm asymptotically converges to the best probability distribution over the arms. This implies that the aggregator will asymptotically achieve optimal averaging with respect to the sampled models without any prior knowledge of the load power transfer rate.

Exp3 for DR is also subject to a lower bound in its regret. In other words, Exp3 for DR will always commit a certain error yielding to a regret always greater than a certain constant. This is a consequence of the randomization of the forecaster [10]. The result is given in Proposition 2.

Algorithm 1. Exp3 for DR

Parameters: Given R and C for all TCLs, the base-load $\mathbf{b}_t \; \forall \; t = 1, 2, \ldots$ and K the number of models.

Initialization: Sample the set of models κ, set $q_i(0) = \frac{1}{K} \; \forall \; i = 1, 2, \ldots, K$ and set the learning rate $\eta_1 = \sqrt{\frac{\ln K}{K}}$.

for $t = 1, 2, \ldots$ **do**

- Sample a model I_t according from the probability distribution $q_i(t)$;
- Solve the DR optimization problem with model I_t,
$$\mathbf{Y}(I_t) = \underset{\mathbf{V} \in \mathcal{F}(I_t)}{\operatorname{argmin}} \, || \, \mathbf{V}\mathbf{p}(I_t) + \mathbf{b}||^2$$
and send the control to the load.
- Observe the power usage of the aggregated load Z_t
- Compute the estimated loss of each model,
$$\tilde{\ell}(i, Z_t) = \frac{\ell(i, Z_t)}{q_i(t)} \mathbb{I}_{I_t, i}, \forall \; i = 1, 2, \ldots, K$$
with $\ell(i, Z_t) = 1 - \exp[-|\mathbf{1}^{\mathrm{T}}\mathbf{Y}(i)\mathbf{p}(i) - Z_t|/\alpha]$
- Update the cumulative estimated loss for all model i,
$$\tilde{L}_i(t) = \tilde{L}_i(t-1) + \tilde{\ell}(i, Z_t)$$
- Decrease the learning rate, $\eta_t = \sqrt{\frac{\ln K}{tK}}$
- Update the probability distribution over all models,
$$q_i(t+1) = \frac{\mathrm{e}^{-\eta_t \tilde{L}_i(t)}}{\sum_{j \in \kappa} \mathrm{e}^{-\eta_t \tilde{L}_j(t)}}$$

end

Proposition 2. (Minimax lower bound of Exp3 for DR) *Let K be the number of models and T the time horizon, then the expected cumulative regret is lower bounded by*,

$$\mathbb{E}[R_T] \geq \frac{1}{20}\sqrt{TK}. \tag{17}$$

The proof is given in [10] for Exp3.

Remark 5. (Dynamical Model) Due to the randomization, the exploration phase is always present in the player's strategy. This allows the online model to dynamically adapt its strategy if there is a change in the load parameters (e.g. due to seasonal change or to a broken cooling system).

4. Numerical example

We now present numerical results obtained with the proposed model. Ontario's base-load is used to simulate real values for **b**. Note that the base-load is scaled down by a factor of 2500 since only a few TCLs are used in this simulation. For the following simulation, K the number of models, is fixed to 20 and the number of TCLs is fixed to $N = 10$. The optimal load shifting algorithm is executed for each day using an arm selected by the learning algorithm. Each day corresponds to an iteration of the learning algorithm and the simulation is computed over a period of 730 days with a load-shifting time step $h = 5$ minutes.

We limit the population to 10 TCLs so that we can run the simulation for two years with a reasonable amount of computation time. In a real implementation, there would be one iteration per day, and thus we could accommodate a far larger population of TCLs.

For the TCL, the R, C and η values are fixed to 3°C/kW, 12 kWh/°C and 2.5 respectively for all units and $w(t)$ is omitted. The P_r are sampled randomly using the same prior distribution as the models with $P_{r_{min}} = 10$ kW and $P_{r_{max}} = 18$ kW. Lastly, we fix $\theta_d = 23$℃ and $\Delta = 1$℃ for all TCLs. All TCL parameters are fixed according to [32].

To represent the variation in temperature throughout the day, a simplified version of [34] is used for the ambient temperature. The simplified model is given by,

$$\theta_{amb}(t) = \theta_{max} \left| \sin \frac{2\pi t}{2M} \right| + \theta_{min}, \tag{18}$$

with $\theta_{max} = 10$℃, $\theta_{min} = 21$℃ and recall that M is the number of load-shifting time step in a day and is equal to 288. Finally, we fix the loss function tuning parameter $\alpha = 4$.

4.1. Regret analysis

We plot the estimated cumulative regret of the proposed model is first presented in Figure 1 with its lower and upper bounds. Figure 1 shows that the cumulative regret is sub-linear and, therefore, as stated in Section 3.4, will converge to the best sampled model.

4.2. Demand response performance analysis

We now compare the performance of the learning algorithm to the case where the true parameters are known by the optimal load-shifting routine. A metric is

defined to allow this comparison. The relative demand flattening ratio, Δ, is given by,

$$\Delta = \frac{\left\|\mathbf{Y}_{comp}\mathbf{p} + \mathbf{b}\right\|^2 - \left\|\mathbf{Y}_{Exp3\ for\ DR}\mathbf{p} + \mathbf{b}\right\|^2}{\left\|\mathbf{Y}_{comp}\mathbf{p} + \mathbf{b}\right\|^2}, \quad (19)$$

where the subscript *comp* stands for the case to which the algorithm is compared to. Table 1 compares the performance of the algorithm with this indicator for two cases. First against the ideal case where the real TCLs parameters are known and second against the case where no DR is attempted.

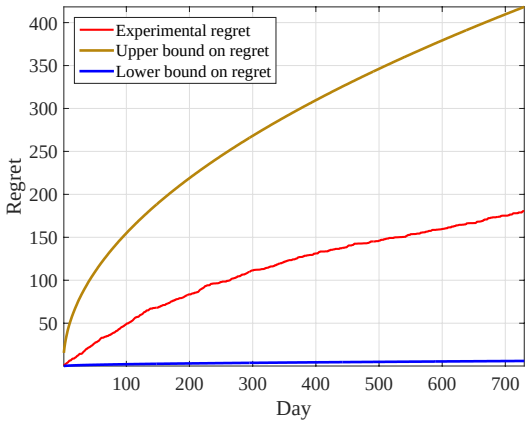

Figure 1. Estimated cumulative regret of Exp3 for DR

Table 1. DR performance for the proposed learning algorithm

Comparison	Δ
vs. ideal case	0.35%
vs. No DR	-11.50%

The ideal case comparison in Table 1 shows that the performance of the algorithm is similar to the DR problem in which all parameters are known. This motivates thus the use of such an algorithm for demand response instead of other algorithms that required a significant amount of on-site measurement. Using a better prior when sampling the models could improve performance further. Nevertheless, a non-zero deviation is unavoidable since at some point the algorithm will, in its *exploration* phase, test high-loss models. Therefore, asymptotically, one will perform as well as the best arm or combination of arms available from the sampling step.

The averaging performance indicator with respect to the no DR case is high which shows the averaging ability of the approach. However, note that (19) is a function of the base-load which has been scaled down to better represent the ratio of power used by TCLs on the base-load. Therefore, such an indicator will be a function of the scaling factor and of the number of considered TCLs.

To illustrate the power usage and demand averaging in the grid, two figures are presented here. In Figure 2, the power usage of the TCLs is shown for the three stated scenarios: the proposed bandit-learning approach, the ideal case when all parameters are known exactly and finally for the scenario where no DR is attempted. These three scenarios are respectively labeled *bandit*, *ideal* and *NoDR*.

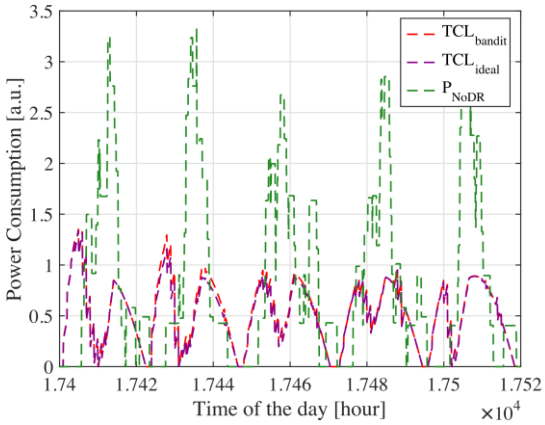

Figure 2. Power usage of the TCLs for the last five days of the simulation

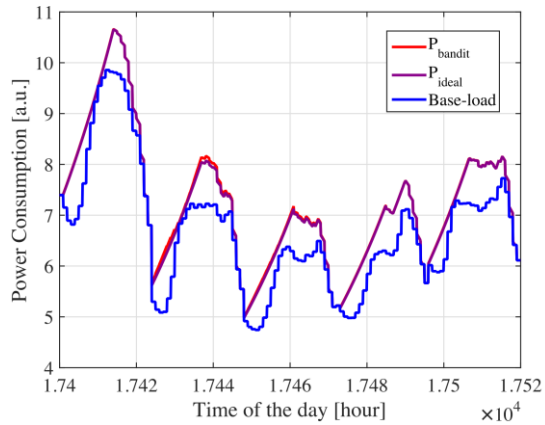

Figure 3. Total power usage (TCLs + base-load) for the five last days

Figure 2 shows that similarity of the *ideal* and *bandit* curves is high and that the approach avoids daily peaks encountered in the no-DR case. Figure 3 shows the averaging ability of the model. In both figures, the *ideal* and *bandit* curves are almost perfectly superimposed reflecting the learning performance. Note that the model

shifts daytime loads to the night time valleys in the base-load.

Lastly, Figures 4 and 5 show the exploration-exploitation process underlying the multi-armed bandit. We see that the 16th arm has the largest probability and has been selected the most often after 730 iterations of the learning algorithm. This is then illustrated by the high probability of the 16th arm presented in Figure 5.

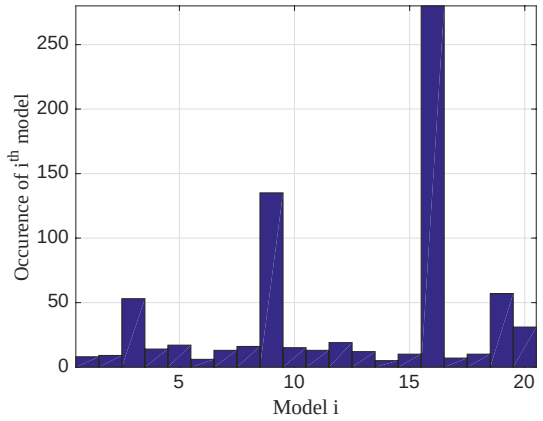

Figure 4. Distribution of the chosen models after 730 rounds (2 years)

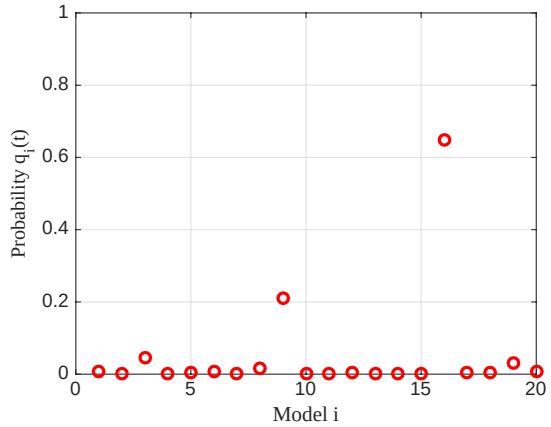

Figure 5. Probability mass distribution $q_i(t)$ of the models after 730 rounds (2 years)

5. Conclusion

In this work, we address load uncertainty in demand response. By using an adversarial multi-armed bandit framework, the aggregator can select the best model from an arbitrary set of candidate models. In the present case, the aggregator goal is to average the power demand. Numerical simulations showed the advantage of using this approach.

In future works, the online model will be extended to address the problem of a totally uncharacterized load by adding feedback on the temperature of the load and extending sampled models to other load parameters.

6. Acknowledgements

This work was funded by a CGS-M research fellowship and a Discovery grant both awarded by the NSERC.

7. References

[1] M. Alizadeh, X. Li, Z. Wang, A. Scaglione, and R. Melton, "Demand-side Management in the Smart Grid: Information processing for the power switch," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 55–67, 2012.

[2] P. Siano, "Demand response and smart grids - A survey," *Renew. Sustain. Energy Rev.*, vol. 30, pp. 461–478, 2014.

[3] R. Deng, Z. Yang, M.-Y. Chow, and J. Chen, "A Survey on Demand Response in Smart Grids: Mathematical Models and Approaches," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 1, pp. 152–178, 2015.

[4] M. Kamgarpour, C. Ellen, S. E. Z. Soudjani, S. Gerwinn, J. L. Mathieu, N. Mullner, A. Abate, D. S. Callaway, M. Franzle, and J. Lygeros, "Modeling Options for Demand Side Participation of Thermostatically Controlled Loads," in *Proceedings of IREP Symposium: Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid, IREP 2013*, 2013, pp. 1–15.

[5] J. L. Mathieu, M. Kamgarpour, J. Lygeros, G. Andersson, and D. S. Callaway, "Arbitraging Intraday Wholesale Energy Market Prices With Aggregations of Thermostatic Loads," *IEEE Trans. Power Syst.*, vol. 30, no. 2, pp. 763–772, 2015.

[6] D. S. Callaway, "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy," *Energy Convers. Manag.*, vol. 50, no. 5, pp. 1389–1400, 2009.

[7] S. Bubeck, "Introduction to Online Optimization," *Lect. Notes*, pp. 1–86, 2011.

[8] S. Shalev-Shwartz, "Online Learning and Online Convex Optimization," *Found. Trends Mach. Learn.*, vol. 4, no. 2, pp. 107–194, 2011.

[9] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. 2006.

[10] S. Bubeck and N. Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

[11] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The Nonstochastic Multiarmed Bandit Problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[12] H. Robbins, “Some Aspects of the Sequential Design of Experiments,” in *Herbert Robbins Selected Papers*, Springer, 1985, pp. 169–177.
[13] T. L. Lai and H. Robbins, “Asymptotically Efficient Adaptive Allocation Rules,” *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.
[14] P. Auer, N. Cesa-bianchi, and P. Fischer, “Finite-Time Analysis of the Multiarmed Bandit Problem,” *Mach. Learn.*, vol. 47, no. 2–3, pp. 235–256, 2002.
[15] J. C. Gittins, “Bandit Processes and Dynamic Allocation Indices,” *J. R. Stat. Soc. Ser. B*, vol. 45, no. 2, pp. 148–177, 1979.
[16] P. Whittle, “Restless Bandits: Activity Allocation in a Changing World,” *J. Appl. Probab.*, vol. 25, no. 1988, pp. 287–298, 1988.
[17] J. A. Taylor and J. L. Mathieu, “Index Policy for Demand Response,” *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1287–1295, 2014.
[18] Q. Wang, M. Liu, and J. L. Mathieu, “Adaptive demand response: Online learning of restless and controlled bandits,” *2014 IEEE Int. Conf. Smart Grid Commun. SmartGridComm 2014*, pp. 752–757, 2015.
[19] D. Kalathil and R. Rajagopal, “Online Learning for Demand Response,” *Fifty-third Annual Allerton Conference*. pp. 218–222, 2015.
[20] N. Y. Soltani, S. J. Kim, and G. B. Giannakis, “Real-Time Load Elasticity Tracking and Pricing for Electric Vehicle Charging,” *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1303–1313, 2015.
[21] S.-J. Kim and G. Giannakis, “An Online Convex Optimization Approach to Real-Time Energy Pricing for Demand Response,” *IEEE Trans. Smart Grid*, pp. 1–10, 2016.
[22] G. S. Ledva, L. Balzano, and J. L. Mathieu, “Inferring the Behavior of Distributed Energy Resources with Online Learning,” pp. 187–194, 2015.
[23] W. Ma, V. Gupta, and U. Topcu, “Distributed Charging Control of Electric Vehicles Using Regret Minimization,” *arXiv ID 1507.07123*, pp. 1–10, 2014.
[24] J. L. Mathieu, M. Kamgarpour, J. Lygeros, and D. S. Callaway, “Energy Arbitrage with Thermostatically Controlled Loads,” in *European Control Conference*, 2013, pp. 2519–2526.
[25] R. E. Mortensen and K. P. Haggerty, “A Stochastique Computer Model for Heating and Cooling Loads,” *IEEE Trans. Power Syst.*, vol. 3, no. 3, pp. 1213–1219, 1988.
[26] R. Malhame and Chee-Yee Chong, “Electric Load Model Synthesis by Diffusion Approximation of a High-Order Hybrid-State Stochastic System,” *IEEE Trans. Automat. Contr.*, vol. 30, no. 9, pp. 854–860, 1985.
[27] C. Uçak and R. Çaglar, “The Effects of Load Parameter Dispersion and Direct Load Control Actions on Aggregated Load,” *1998 International Conference on Power System Technology Proceedings. POWERCON’98*, vol. 1. pp. 280–284, 1998.
[28] M. Grant and S. Boyd, “CVX: Matlab Software for Disciplined Convex Programming, version 2.1.” Mar-2014.
[29] M. Grant and S. Boyd, “Graph implementations for nonsmooth convex programs,” in *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.
[30] MOSEK ApS, “The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28).” 2015.
[31] M. Zinkevich, “Online Convex Programming and Generalized Infinitesimal Gradient Ascent,” *Sch. Comput. Sci. Carnegie Mellon Univ.*, 2003.
[32] J. L. Mathieu and D. S. Callaway, “State estimation and control of heterogeneous thermostatically controlled loads for load following,” in *45th Hawaii International Conference on System Science (HICSS), 2012*, 2012, pp. 2002–2011.
[33] J. Audibert, “Regret Bounds and Minimax Policies under Partial Monitoring,” *J. Mach. Learn. Res.*, vol. 11, pp. 2785–2836, 2010.
[34] W. J. Parton and J. A. Logan, “A Model for Diurnal Variation in Soil and Air Temperature,” *Agric. Meteorol.*, vol. 23, pp. 205–216, 1981.