Demand Response of a Heterogeneous Cluster of Electric Water Heaters Using Batch Reinforcement Learning

Frederik Ruelens*,Bert J. Claessens‡, Stijn Vandael*, Sandro Iacovella*, Pieter Vingerhoets* and Ronnie Belmans*
*Department of Electrical Engineering, Electa/EnergyVille
‡Flemish Institute for Technological Research VITO/EnergyVille
frederik.ruelens@esat.kuleuven.be

***Abstract*—A demand response aggregator, that manages a large cluster of heterogeneous flexibility carriers, faces a complex optimal control problem. Moreover, in most applications of demand response an exact description of the system dynamics and constraints is unavailable, and information comes mostly from observations of system trajectories. This paper presents a model-free approach for controlling a cluster of domestic electric water heaters. The objective is to schedule the cluster at minimum electricity cost by using the thermal storage of the water tanks. The control scheme applies a model-free batch reinforcement learning (batch RL) algorithm in combination with a market-based heuristic. The considered batch RL technique is tested in a stochastic setting, without prior information or model of the system dynamics of the cluster. The simulation results show that the batch RL technique is able to reduce the daily electricity cost within a reasonable learning period of 40-45 days, compared to a hysteresis controller.**

***Keywords*—Aggregator, demand response, batch reinforcement learning, electric water heater, fitted Q-iteration.**

I. Introduction

DUE to an increased awareness of the negative environmental impact of fossil fuels and its increasing prices, the share of renewable energy is rising in most industrialized countries. Simultaneously the electricity demand is likely to increase due to the electrification of transport and domestic heating. These two evolutions are expected to cause a paradigm shift in the way the grid is operated. Driven by advances in computing and communication technologies, demand response (DR) aggregators can play a central role in this shift. For instance, a DR aggregator can manage and control a cluster of flexibility carriers (e.g. heat pumps, electric water heaters or electric vehicles) in order to maximize an application dependent objective function, given certain grid and consumer constraints.

However as can be seen in the literature, scheduling such a cluster is a complex optimal control problem. In addition, in most smart grid applications an exact description of the the system dynamics and constraints is unavailable, and information comes mostly from observation of system trajectories. Several recent papers [1], [2], [3] have proposed the use of a simplified aggregated model in combination with a dispatching heuristic. They use model-based techniques, such as Model Predictive Control [4] (MPC) to solve the aggregated control problem. Once an aggregated solution has been found, a dispatching heuristic is used to distribute the aggregated solution over the cluster. Nevertheless, these model-based approaches rely on general system identification techniques to estimate the model parameters of the aggregated model before applying MPC. For example in [3], the authors represent the state of a cluster of thermostatically controlled loads by discrete temperature bins, through which the aggregated state probability mass of the population is moved. They show how the aggregated model parameters can be determined by observing the temperature dynamics of the loads using a Markov Chain technique. Similarly in [1], the authors use a lumped thermal first-order RC model to describe the dynamics of a heat pump portfolio.

In contrast to model-based techniques, learning-based techniques, such as (model-free) Reinforcement Learning [10] (RL), are model-free and do not rely on a priori information of the system dynamics. In the recent RL literature, batch Reinforcement Learning (batch RL) techniques have received a growing attention. Unlike classic Q-learning algorithms, batch RL techniques do not require many interactions until convergence to obtain reasonable policies [17], [5], [18], which makes them an attractive technique for real-world applications. An overview of real-world applications of batch RL techniques can be found in [18].

The objective of this paper is to apply a learning-based approach (model-free batch RL) to the control problem of a DR aggregator. This approach is inspired by the work of [5], [21] and extends on previous work in [6] and [12]. Similar as in [3], the approach is validated on a heterogeneous cluster of thermostatically controlled loads. The objective is to schedule a cluster of electric water heaters under a time-varying electricity price. Electric Water Heaters (EWHs) can modify their consumption profile by storing energy in their water tank, without violating the comfort of the consumers. However, in stead of using a generic first-order model to simulate the temperature dynamics of the population, this work uses a detailed stratified thermal tank model. This model allows us to calculate a temperature profile of each electrical water heater, which makes it possible to simulate the impact of the proposed batch RL approach on the comfort of the consumers.

The contributions of this paper are: (1) we presented a model-free batch RL technique in combination with a heuristic (2) the approach was tested in a stochastic setting (unknown tap water profile) using a stratified boiler model (3) we showed that batch RL can be seen as a valuable alternative to MPC methods; the presented control scheme converges within a relative limited number of training days (40-45 days).

The sequel of this paper is organized as follows. Section

Paper submitted to Power Systems Computation Conference, August 18-22, 2014, Wroclaw, Poland, organized by Power Systems Computation Conference and Wroclaw University of Technology.

II introduces the market-based heuristic (three-step approach), Section III gives the Markov decision process formulation and illustrates the considered batch RL technique (fitted Q-iteration). A stratified simulation model of an EWH is demonstrated in Section IV. Simulation results for a cluster of 100 EWHs are given in Section V, and finally Section VI summarizes the general conclusion of this work.

II. Three-step approach

Similar to [6] and [12], a three-step agent-based approach (TSA) for demand side management is used. This concept consists of three steps (1) combining data (2) solving an aggregated model that is reduced in size and complexity with respect to the original problem (3) using a heuristic to dispatch the aggregated solution over the cluster. These three steps are continuously repeated to adapt towards a dynamic and uncertain environment. Such aggregation-disaggregation schemes are widely used techniques in approximate dynamic programming to solve complex problems [15].

Three primary reasons motivate such an aggregation and disaggregation approach for demand response. Firstly the aggregation step results in a smaller and simpler problem (decrease in the number of decision variables), and hence makes the problem more tractable to solve. Secondly, the aggregated model makes it possible to rewrite the problem using a reduced state space, such that RL techniques can be used. Finally, the TSA provides a realistic decentralized solution with good scalability qualities [6], [7]. In addition, only limited intelligence for calculating the state of charge of an agent is needed at household level.

In the remainder of this paper a market-based heuristic is used in the first and third step. However it should be noted that, the presented approach can be used in combination with other dispatching methods [3], [1].

A. Step 1: Aggregation

In the first step, the cluster aggregator aggregates the power flexibility and state of charge information of each agent. The power flexibility is expressed through a bid function, which can be constructed model-free. The bid function of an EWH is expressed as follows

$$b^i(p) = \begin{cases} P_{rating} & if\ 0 < p \leq p_c^i, \\ 0 & if\ p > p_c^i \end{cases} \tag{1}$$

where p_c^i is the corner priority and the subscript i denotes the unit. The corner priority indicates the wish (priority) for consuming at a certain power rating P_{rating}. The closer the State of Charge (SoC) drops to zero, the more urgent its scheduling (high priority), the closer to 100%, the lower the scheduling priority. For an electric water heater the SoC is defined as the ratio of the energy content of the water to the reference energy content of a fully charged buffer. The corner priority of an EWH is calculated as follows

$$\forall j\,(j:0...n)\,, \forall k\,(k:0...n, T_k \geq T_{min})$$

$$p_c^i = 1 - \underbrace{\frac{\Sigma_k V_k\,(T_k - T_{min})}{\Sigma_j V_j\,(T_{max} - T_{min})}}_{\text{SoC}} \tag{2}$$

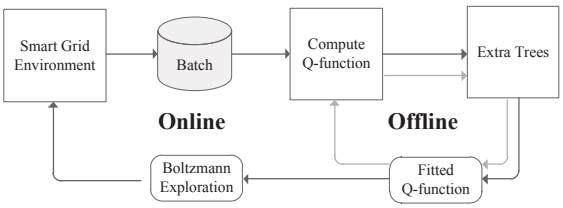

Fig. 1. This figure shows the online and offline part of the Fitted Q iteration algorithm [16]. The batch of data samples is increasing by means of Boltzmann exploration.

where n denotes the number of temperature measurements, V_k the volume of one layer, and T_{max} and T_{min} are the maximum and minimum allowed water temperature. Here, the calculation of the SoC is based on 8 temperature sensors, which are installed along the hull of the buffer tank [8]. Also, water layers where $T_k \leq T_{min}$ have no useful thermal energy and do not contribute to the SoC, since no water may leave the buffer at a temperature below T_{min}. Similar SoC formulations can be found in [8] and [11].

B. Step 2: Optimization

The goal of the second step is to determine an aggregated policy, which minimizes the costs for the cluster aggregator. The sole information the aggregator has for finding such a policy comes from observations of system trajectories, which contain limited state information, i.e. the reading of a temperature sensor. The next section describes how this problem can be formulated as a Markov Decision Process (MDP). By doing so we can use learning techniques from the reinforcement learning literature.

C. Step 3: Dispatch

Once an aggregated energy set point (u_t) is found, a clearing priority p^* is sent back to all flexibility carriers:

$$p^* = \underset{p}{\operatorname{argmin}} \left| \Sigma_{i=1}^{N} b^i(p) - u_t \right|, \tag{3}$$

here N represents the number of EWHs. Each EWH will locally match p^* in its own bid function $b^i(p)$ and start charging at a power corresponding to $b^i(p^*)$. Since an EWH is an ON/OFF appliance, agents with a corner priority lower than the clearing priority ($p^* > p_c$) switch off, while agents with a corner priority higher than the clearing priority ($p^* < p_c$) will switch on.

III. Batch mode reinforcement learning

This section describes how an approximate policy can be found using only a batch of observations of the system trajectories as input. These system trajectories contain state of charge information of the agents and are obtained by interaction with the smart grid environment.

A. Markov decision process

The control problem of finding a daily policy for the aggregated problem (step 2) is modeled as a stochastic Markov Decision Process (MDP). At every time step $t \in T_{96}$ the

aggregator takes a control action $u_t \in U$ and the state of the cluster $x_t \in X$ evolves according to the following stochastic transition function

$$x_{t+1} = f(x_t, u_t, w_t), \tag{4}$$

where $w_t \in W$ denotes a random disturbance. This disturbance depends on the stochastic user behavior of the consumers (unknown demand) and errors caused by the dispatching heuristic. After a transition to the next state x_{t+1}, an immediate cost is provided according to a problem specific objective function. The remainder of this paper considers a Time-of-Use (ToU) setting, with an hourly varying electricity price [19], and thus the associated cost is given by

$$c_t = \rho(x_t, u_t, w_t) = u_t^a \cdot \lambda_t, \tag{5}$$

where u_t is the desired energy set point of the aggregator, u_t^a is the actual consumed energy by the cluster and λ_t is the energy price during time step t. During the learning phase the desired energy set-point is not always achieved, since precautions have to be taken in order to guarantee the comfort and safety of the consumers. Each agent is equipped with a back-up controller, that can overrule the energy set points of the aggregator when comfort and safety constraints are being jeopardized.

The goal is to find an optimal policy $h^*(.) : X \rightarrow U$ that minimizes the expected cumulative return for any state in the state space. This accumulated cost over T stages for a given state x and policy h is defined as

$$J_T^h(x) = \underset{\substack{w_t \\ t=0\ldots T}}{E} \left[\Sigma_{t=0}^{T} \gamma^t \rho(x_t, h(x_t), w_t) | x_0 = x\right], \tag{6}$$

where γ is a discount factor and where the conditional mean is taken over all trajectories starting with the initial conditions $x_0 = x$. Since the policy, h, is time dependent with a finite horizon ($T = 96$), the discount factor γ is set to one, and thus is omitted in the remainder of this paper.

A convenient way to characterize a policy is by using a value function for each state-action pair (Q-function).

$$Q(x, u) = c_t + \underset{\substack{w_t \\ t=1\ldots T}}{E} \left[\Sigma_{t=1}^{T} \rho(x_t, h(x_t), w_t)\right] \tag{7}$$

This Q-value can be seen as the cumulative cost starting from state x and by taking action u and following h thereafter. Once the Q-function for every state-action pair is known, the policy can be calculated as follows

$$h(x) = \underset{u \in U}{\text{argmin}}\ Q(x, u). \tag{8}$$

The next section describes how this Q-function can be calculated using a batch of four tuples.

B. Fitted Q-iteration

Given full knowledge of the system dynamics and cost function an optimal policy can be found by solving (7) for every state-action pair [15]. However, in the remainder of this paper we assume that the transition function f is unknown (model-free learning). Thus the sole information available to solve the problem is the one obtained from daily observations of one step system transitions

$$\mathcal{F} = \{x_{t,d}, u_{t,d}, c_{t,d}, x_{t+1,d} | d = 1, ..., n_d, t = 1, ..., T-1\}, \tag{9}$$

where d denotes the day of the observation and each tuple is made up of the system state x_t, the control action u_t, the immediate cost c_t and the successor state of the system x_{t+1}.

Figure 1 outlines the batch RL learning framework as presented in [16], which consists of two interconnected loops. The offline loop applies the fitted Q-iteration algorithm (Algorithm 1) using the current batch of tuples as input and outputs a policy. The resulting policy is then used to generate new tuples by using a Boltzmann exploration strategy (online loop in Figure 1). At the end of the next day these new tuples are added to the old tuples.

Offline loop: The offline loop is repeated at the end of each day. At each iteration (k) the algorithm uses the current set of four-tuples with the Q-function gathered from the previous iteration $(k-1)$ to determine a new training set $\mathcal{TS}$ which is used by a regression method to compute the current Q-function.

Algorithm 1: Fitted Q-iteration

Data: set of four-tuples
$\mathcal{F} = \{x_i, u_i, c_i, x_{i+1}|, i = 1...\#\mathcal{F}\}$
initialize k = 0 and $\hat{Q}_0$ to zero;
repeat
  build up a training set
  $\mathcal{TS} = \{(in_i, out_i), i = 1, ..., \#\mathcal{F}\}$ where
  $in_i = (x_i, u_i)$
  $out_i = c_i + \min_{u'} \hat{Q}_{k-1}(x_i', u')$
  Use a regression algorithm to induce from $\mathcal{TS}$ the function $\hat{Q}_k$
  increment k
until $k = T$;
Result: $\hat{Q}_* = \hat{Q}_k$ and compute a policy according to (8)

In this work we follow [5] and use extremely randomized trees to approximate the Q-function. Each tree partitions the input space into a number of disjoint regions, and determines a constant prediction, by averaging the output values of the samples that belong to this region. The parameters of this method are M, the number of trees to create, k the dimension of the input space and n_{min} the minimal number of nodes in each region. The parameters M is set to 100, n_{min} to 5 and k to 4 (dimension of the Q-function). A more detailed analysis of the tree parameters can be found in [5].

Online loop: The objective of the online loop is to explore interesting tuples in order to improve the overall quality of the daily policies ("growing" BMRL). Here we use Boltzmann exploration to sample new tuples. With Boltzmann exploration new tuples are sampled with probability $P(u|x)$ given by

$$P(u|x) = \frac{e^{Q(x,u)/\tau_d}}{\Sigma e^{Q(x,u)/\tau_d}}, \tag{10}$$

where τ_d is called the Boltzmann temperature, which controls the exploration. If $\tau_d \rightarrow 0$ than the exploration decreases and the policy becomes more greedy. Thus by starting with a high

τ_d the exploration start completely random, however as τ_d decreases the exploration directs itself to the most interesting state action pairs. In future research, we will investigate other exploration strategies, such as active exploration by searching for trajectories that falsify the current policy [20]. By doing so, this strategy actively searches for valuable trajectories, by making use of a predictive model, that improve the policy as much as possible. Furthermore, it should be noted that each EWH is equipped with on internal controller, that guarantees the users' comfort settings.

IV. Electric water heater

In order to simulate the impact of the presented batch RL approach on the dynamics of the cluster of EWHs and on the comfort of the consumers a detailed thermally stratified model for each EWH is implemented. This model is based on existing literature [9] and was validated in a lab environment [8]. Table 1 shows the model parameters used in the following equations.

A. Simulation model

The considered buffer tank of an EWH contains exclusively water, where cold water enters the tank at the bottom and where hot water leaves the tank at the top. When hot water leaves the buffer tank it is replaces by cold water and heat can be injected by means of a resistance located at the bottom of the buffer. The water in the buffer tank is divided into n layers, with each layer a certain uniform temperature T_i. These layers can interact thermally through conduction and mixing, and can lose heat to the environment. The following equation is used to compute the thermal losses

$$Q_{loss,i} = Ah\left(T_{outside} - T_i\right), \tag{11}$$

where A is the total surface area of the buffer and h the heat loss coefficient of the buffer surface area. The conduction effects are computed as follows

$$\begin{array}{c} Q_{cond,i} = k_i A_i \frac{(T_i - T_{i+1})}{L_{cond,i}} \\ +k_i A_{i-1} \frac{(T_i - T_{i-1})}{L_{cond,i-1}}, \end{array} \tag{12}$$

where k indicates the thermal conductivity of water, A the conduction interface area and L the distance between the layers. The mixing effect are calculated as follows

$$\begin{array}{c} Q_{mix,i} = \dot{m}_i C_{p,i}\left(T_i - T_{i+1}\right) \\ +\dot{m}_{i-1} C_{p,i-1}\left(T_i - T_{i-1}\right), \end{array} \tag{13}$$

with $\dot{m}$ the flow rate, and C_p the specific heat of water. At each time step, $\Delta h = 10s$, the simulation model calculated the layer temperatures. However, when unstable layers are found (if the layer temperature is higher than the temperature of the layer above) an iteration is performed where the new temperature of the unstable layer is set to be equal to the average temperature of all unstable layers. This iteration process continues until convergence is reached.

B. Comfort settings

As mentioned before, each EWH is equipped with a backup controller, which can overrule the power set point of the aggregator.

$$P_{boiler} = \begin{cases} P_{rating} & if\ T_{n^*} \leq T_{min} \\ b(p_c) & if\ T_{n^*} > T_{min} \text{ and } T_1 < T_{max} \\ 0 & if\ T_1 \geq T_{max} \end{cases} \tag{14}$$

where T_1 is the temperature corresponding to the lowest layer and T_{n^*} corresponds with a temperature sensor near the top of the buffer. Here, the minimum temperature set point was set to $50°$C, as stated by the US Department of Energy [22], to avoid bacterial growth.

Model parameters	[/]
n_{layers}	50
Δh	10 s
h	0.8 $W/(m^2K)$
k	0.59 W/mK
μ_{coil}	0.75 W/mK
Simulation parameters	**[/]**
N_{boiler}	100
V_{boiler}	200-250 l
P_{rating}	2-2.5 kW
T_{min}	50 °C
T_{max}	70-80°C
T_{in}	10°C
$T_{outside}$	18°C

TABLE I. Model and simulation parmeters

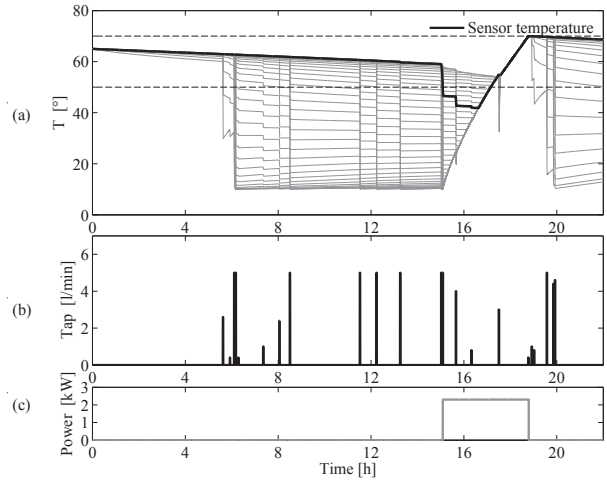

Fig. 2. This figure shows impact of the tap water demand on the temperature of the different layers.

C. Single unit simulations

Figure 2 illustrates the temperature profile and the corresponding tap water profile for one day, when no aggregated energy set point is provided.

V. Simulation Results

A. Problem description

The evaluation case concerns a DR aggregator with a portfolio of 100 EWHs (ON/OFF devices) in a Time-of-Use setting (ToU). This ToU price was obtained from [19] and consists of hourly-varying electricity prices. At the beginning of each day, the aggregator receives the ToU profile for the next day. The objective of the DR aggregator is to minimize its daily electricity cost, by using the thermal storage of the

water tanks. Each water heater has a buffer between 200 and 250 liter. It is important to note that, the aggregator has no information on the exact configuration of the cluster, i.e. buffer volumes, rated power and future tap water profiles. It should be clear that the RL-controller considers the cluster of EWHs as a black-box.

B. State variable

At every time step t, each EWH communicates its average temperature $\langle T_i \rangle$ to the aggregator agent. The aggregator uses these values to calculate a virtual temperature for the cluster $\langle T_{cluster} \rangle = \Sigma_{i=1}^{N} \langle T_i \rangle / N$. Also, at the beginning of each day, the aggregator makes a forecast of the aggregated tap water ($\hat{V}_{tap}$) the cluster will consume, e.g. the aggregator predicts the cluster will consume 8000 liters the next day. We assume that making such an aggregated forecast can be made based on historical data (weekend versus weekday or depending on the season)[1]. By doing so the regression algorithm can learn the relationship between the forecasted aggregated tap water and the cost, and thus improve the quality of the policy. The state variable is given by $\left(t, \langle T_{cluster} \rangle, \hat{V}_{tap}\right)$.

C. Simulation results

Figure 3 depicts the Q-values of the visited state-action pairs and the fitted Q-function (for a given $\hat{V}_{tap}$) during a moment in time when ToU prices were high. Firstly, it can be seen that for high average temperatures, the Q-values increase according to the power set point. Secondly, it is more valuable for the DR aggregator to be in high temperature states (low Q-value) than in low temperature states, since unwanted energy consumption occurs during low temperatures (back-up controller). This is because the reward in (5) is defined by the actual power consumption, and therefore the impact of the internal controller is captured in the shape of the Q-function. In our experience extremely randomized trees are very capable at capturing these sharp features in the shape of the Q-function.

Figure 4 compares the daily electricity cost of the no control option (backup controller) and the batch RL strategy. It can be seen that the daily electricity cost of the batch RL strategy decreases below the no control strategy after 40-45 days.

Figure 5 shows the average EWHs' temperatures, aggregated power set point, ToU profile and aggregated tap water consumption of the last 10 simulation days ($\tau_D \rightarrow 0$) for 100 EWHs. It shows that the policies succeeded at shifting the power consumption to low price periods, and they were able to minimize the activation of the internal controllers.

VI. Conclusions

Demand response aggregators face the difficult task of optimizing large clusters of flexibility carriers based on limited information. This information comes mainly from observations of system trajectories, which contain limited state information, e.g the output of a temperature sensor. With the presented work we showed how batch RL can be used in a realistic smart grid setting. In contrast to an MPC approach, the batch RL technique needs no system identification and can be applied directly to a batch of system trajectories. Our simulation results indicated that the proposed scheme can help reduce the electricity cost in a stochastic environments (unknown tap water profiles) of a cluster of 100 EWHs within a limited learning period of 40-45 days.

[1]These forecasts are considered deterministic, making such forecast is considered out of the scope of this paper.

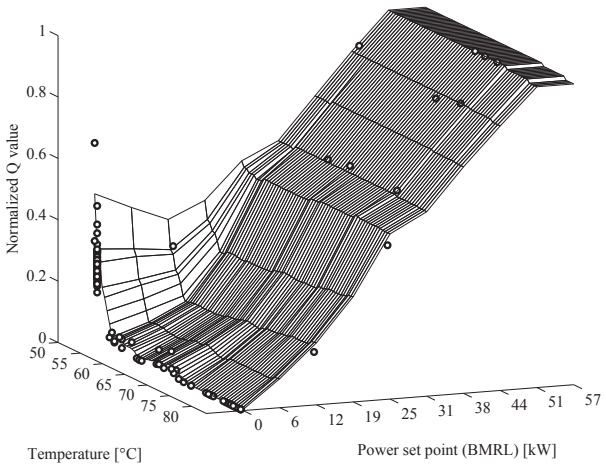

Fig. 3. This figure shows the effect of the backup-controller on the Q-function, high Q-values at low temperatures. As the temperature increases this effect disappears.

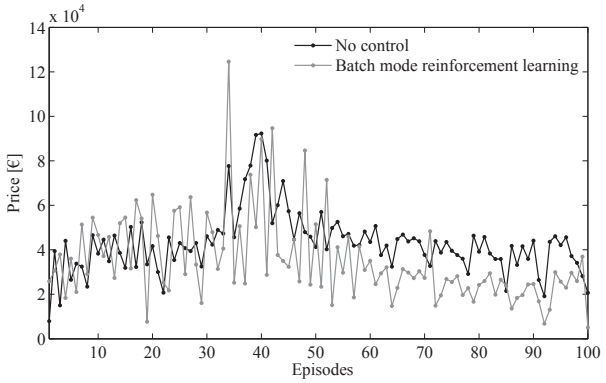

Fig. 4. The batch RL algorithm start performing better than the no control strategy after 40-45 days (episodes)

We think that in the context of smart grids where information on the state is limited and the system dynamics are complex (stochastic user behavior and unknown grid constraints) and time varying, "blind" batch RL techniques can be seen as a valuable alternative to an MPC approach. In our future research we will investigate alternative batch RL techniques, such as the synthesis of artificial trajectories [20]. Furthermore, we will start testing model-free batch RL techniques in the lab, applied to an electric water heater and a self-learning heat-pump thermostat. Also, other exploration strategies, that make use of predictive models, will be investigated, to lower the learning phase.

VII. Acknowledgements

We would like to thank Koen Vanthournout for his valuable work in the lab validating the stratified thermal buffer model. This work is supported by the Flemish Minister for Innovation

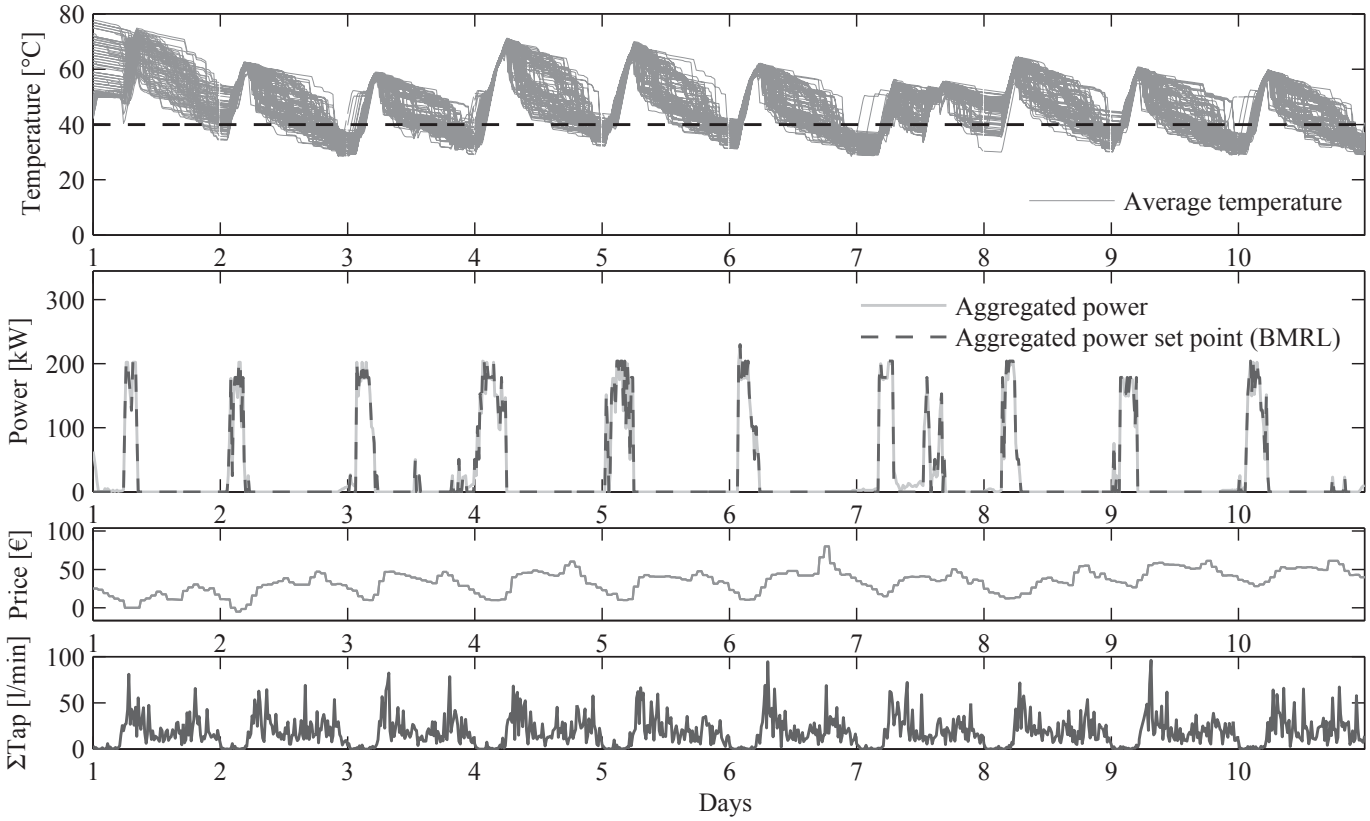

Fig. 5. This figure shows the simulation results of a cluster of 100 residential water heaters. (a) average temperature profile simulated using the individual models (Δh = 10s) (b) aggregated power consumption setpoint (BMRL) and actual aggregated power consumption (c) price profile (d) aggregated tap water consumption.

(Minister I. Lieten) via the project Linear organized by the Institute for Science and Technology (IWT).

REFERENCES

[1] Biegel, B.; Hansen, L.H.; Andersen, P.; Stoustrup, J., "Primary Control by ON/OFF Demand-Side Devices," Smart Grid, IEEE Transactions on , vol.PP, no.99, pp.1,11, 0 Camacho, Eduardo F., et al. Model predictive control. Vol. 2. London: Springer, 2004.

[2] Kara, E.C.; Berges, M.; Krogh, B.; Kar, S., "Using smart devices for system-level management and control in the smart grid: A reinforcement learning framework," Smart Grid Communications , 2012 IEEE Third International Conference on , vol., no., pp.85,90, 5-8 Nov. 2012

[3] Koch, Stephan, Johanna L. Mathieu, and Duncan S. Callaway. "Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services." Proc. PSCC. 2011.

[4] Camacho, Eduardo F., et al. Model predictive control. Vol. 2. London: Springer, 2004.

[5] D. Ernst, G. Pierre and L. Wehenkel. "Tree-based batch mode reinforcement learning.", Journal of Machine Learning Research. 2005.

[6] S. Vandael, B. Claessens, M. Hommelberg, T. Holvoet, G. Deconinck. "A scalable three-step approach for demand side management of plug-in hybrid vehicles.", *IEEE Transactions on Smart Grid.*

[7] Kok, J. K., C. J. Warmer, and I. G. Kamphuis. "PowerMatcher: multiagent control in the electricity infrastructure." Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems. ACM, 2005

[8] K. Vanthournout, R. D'hulst, D. Geysen, G. Jacobs. "A Smart Domestic Hot Water Buffer.", Smart Grid, IEEE Transactions on , vol.3, no.4, pp.2121,2127, 2012.

[9] TRNSYS [Online] http://www.trnsys.com/

[10] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. Vol. 1. No. 1. Cambridge: MIT press, 1998.

[11] Vrettos, E.; Koch, S.; Andersson, G., "Load frequency control by aggregations of thermally stratified electric water heaters," Innovative Smart Grid Technologies (ISGT Europe), vol., no., pp.1,8, 14-17 Oct. 2012

[12] Claessens, B.J.; Vandael, S.; Ruelens, F.; De Craemer K., Beusen B., "Peak shaving of a heterogeneous cluster of residential flexibility carriers using reinforcement learning," Innovative Smart Grid Technologies (ISGT Europe), 2013 4rd IEEE PES International Conference and Exhibition on , vol., no., pp.1,8, 6-9 Oct. 2013

[13] O'Neill, D.; Levorato, M.; Goldsmith, A.; Mitra, U., "Residential Demand Response Using Reinforcement Learning," Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on , vol., no., pp.409,414, 4-6 Oct. 2010

[14] Fonteneau, R., Murphy, S. A., Wehenkel, L., Ernst, D. (2012). Batch mode reinforcement learning based on the synthesis of artificial trajectories. Annals of Operations Research, 1-34.

[15] Bertsekas, Dimitri P. "Dynamic programming and optimal control 3rd edition, volume II." (2011).

[16] Gabel, Thomas, Christian Lutz, and Martin Riedmiller. "Improved neural fitted Q iteration applied to a novel computer gaming and learning benchmark." Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2011 IEEE Symposium on. IEEE, 2011.

[17] Ernst, Damien, et al. "Reinforcement learning versus model predictive control: a comparison on a power system problem." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 39.2 (2009): 517-529.

[18] Lange, Sascha, Thomas Gabel, and Martin Riedmiller. "Batch Reinforcement Learning." Reinforcement Learning. Springer Berlin Heidelberg, 2012. 45-73.

[19] Belpex - belgian power exchange, 2012. [Online]. Available: http://www.belpex.be/

[20] Fonteneau, R.; Murphy, S.A.; Wehenkel, L.; Ernst, D., "Active exploration by searching for experiments that falsify the computed control policy," Adaptive Dynamic Programming And Reinforcement Learning, 2011 IEEE Symposium on , vol., no., pp.40,47, 11-15 April 2011

[21] Busoniu, L., Babuska, R., De Schutter, B., Ernst, D. (2010). Reinforcement learning and dynamic programming using function approximators. CRC Press.

[22] U.S. Department of Energy (2014). [Online]. Available: http://energy.gov/energysaver/articles/tips-water-heating