

Learning a reward function for user-preferred appliance scheduling

Čović, Nikolina; Cremer, Jochen L.; Pandžić, Hrvoje

DOI

[10.1016/j.epsr.2024.110667](https://doi.org/10.1016/j.epsr.2024.110667)

Publication date

2024

Document Version

Final published version

Published in

Electric Power Systems Research

Citation (APA)

Čović, N., Cremer, J. L., & Pandžić, H. (2024). Learning a reward function for user-preferred appliance scheduling. *Electric Power Systems Research*, 235, Article 110667. <https://doi.org/10.1016/j.epsr.2024.110667>

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



Learning a reward function for user-preferred appliance scheduling[☆]

Nikolina Čović^{a,*}, Jochen L. Cremer^b, Hrvoje Pandžić^a

^a University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia

^b Delft University of Technology, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft, The Netherlands

ARTICLE INFO

Keywords:

Demand response

User comfort

Inverse reinforcement learning

ABSTRACT

Accelerated development of demand response service provision by the residential sector is crucial for reducing carbon-emissions in the power sector. Along with the infrastructure advancement, encouraging the end users to participate is crucial. End users highly value their privacy and control, and want to be included in the service design and decision-making process when creating the daily appliance operation schedules. Furthermore, unless they are financially or environmentally motivated, they are generally not prepared to sacrifice their comfort to help balance the power system. In this paper, we present an inverse-reinforcement-learning-based model that helps create the end users' daily appliance schedules without asking them to explicitly state their needs and wishes. By using their past consumption data, the end consumers will implicitly participate in the creation of those decisions and will thus be motivated to continue participating in the provision of demand response services.

1. Introduction

To accommodate the growth of electricity demand within the existing power grid, demand and generation must be locally balanced as much as possible. Energy communities allow their participants to respond with demand and generation adjustments to maintain the local energy balance. For example, a community participant (user) with an appliance such as a washing machine may be willing to shift operating the machine from demand peaks to times of higher power generation or to times when their neighbor's battery is fully charged. However, participation in these community concepts often lacks the right incentives, which represents an obstacle to forming energy communities.

The reasons for the insufficient deployment of energy communities are structural and motivational [1]. Structurally, individuals may find technology, such as smart appliances, too expensive, insecure, or unaccommodating. The additional investments required to control their appliances may be presented as an unnecessary financial burden, while potentially insufficiently secure information sharing may feel threatening to some users. These new technologies may seem quite menacing since most users tend to appreciate their autonomy and ownership, which is why relinquishing control of appliances could prove to be an extremely difficult step. Therefore, they decide not to participate. Motivationally, possibly impaired comfort of the user is identified as one of the key reasons for not participating in the

provision of the service. Individuals may be hesitant to participate or lack persistence due to expected discomforts despite the community benefits to the environment and their economics. Furthermore, human decision-making hesitations favor the status quo of not participating. In past research, the structural-technological reasons have been studied significantly, demonstrating that the financial benefits scale (economy of scale) and control algorithms may exist while managing users' privacy and autonomy. However, the motivational reasons were not studied adequately in these algorithms. In response, this paper proposes a control algorithm that directly interacts with (and learns from) the user, resolving issues of hesitations and persistence by optimally balancing the expected discomforts with their motivations, whether they are environmental or economic.

Household device management usually defines the objective function as a weighted sum of the predefined energy cost and the user (dis)satisfaction functions [2–6]. Similarly, most research papers arbitrarily define weights and analyze the sensitivity of the model to these weights, for example, [7]. Accordingly, the authors in [8] present a multi-objective optimization model for the optimal operation scheduling of shiftable household appliances. By using an iterative learning technique, they are trying to determine the weight value that will enable an optimal trade-off between achieving the minimum energy cost and the minimum user dissatisfaction, which are often in conflict. As a

[☆] This work was supported by the Croatian Science Foundation under project Active Neighbourhoods energy Markets participation—ANIMATION (IP-2019-04-09164) and the WeForming project that has received funding from the European Union's Horizon Europe Programme under the Grant Agreement No. 101123556. Employment of Nikolina Čović is fully funded by the Croatian Science Foundation under project DOK-2020-01-3911.

* Corresponding author.

E-mail addresses: nikolina.covic@fer.hr (N. Čović), j.l.cremer@tudelft.nl (J.L. Cremer), hrvoje.pandzic@fer.hr (H. Pandžić).

<https://doi.org/10.1016/j.epsr.2024.110667>

Received 1 October 2023; Received in revised form 13 March 2024; Accepted 17 June 2024

Available online 29 June 2024

0378-7796/© 2024 Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

result, consumption costs are reduced, and the user's comfort remains intact. However, the presented model as the all available literature, to the best of our knowledge, assumes that the user's discomfort can be modeled by a predefined mathematical function. Due to each person having different preferences, a single discomfort function probably will not fit all users, so a proper control algorithm should be able to detect an appropriate discomfort function for each user. It might even be difficult to fit a single user throughout an indefinite time period. On shorter terms, people's preferences might change from morning until evening, just as they can change on a daily basis. In the longer course of time, a person may change their lifestyle and would need the control algorithm to recognize the change.

Another issue of presented methods that illustrates itself precisely in the model's inability to adapt to a changed environment stems from utilizing methods that require knowledge of all data beforehand (e.g. deterministic optimization). Reinforcement learning, on the other hand, in addition to its ability to learn offline, allows learning through interactions with the environment online, leaving room to improve the adaptability of the model [9]. This technique adds to the versatility of the model by making decisions, unlike the other mentioned approaches, solely based on the *interactions* with the environment, without actually requiring its exact mathematical model. This way, the developed algorithm can be used in completely different environments — in households with different appliances or the same appliances of different brands. Nevertheless, although reinforcement learning solves the issue of adaptability of the model to different households and to different/changed human habits, the available literature, e.g. [10–14], still uses fixed and predefined reward functions, which leaves the problem of unknown basic model of discomfort still unsolved.

A promising field of inverse reinforcement learning (IRL), whose foundations were laid by Ng and Russell in [15], gives a solution to the remaining matter by learning the reward function from data on previous user behavior. Particularly, allowing the user's data to speak for itself, while making no assumptions about the user's preferences beforehand, is the greatest benefit of this approach. Numerous works deal with learning the reward function in the field of robotics. [16] presents a method for learning the reward function based on the user's preferences expressed through their direct feedback and show successful results in the semi-autonomous driving example. Similarly, in [17] the authors propose reward learning from feedback that provides the most information to achieve the best solution. Once again, the authors simulate the theoretical considerations on the example of autonomous driving and two other robotic environments. Contrarily, no work has been found on reward function learning in power systems per our knowledge.

A household owning power-curtable and time-shiftable appliances provides demand response services through which it can obtain financial benefits, while also potentially impair its comfort. This paper attempts to find the underlying reward function that guides the user's daily decision-making process of household appliance scheduling solely based on the examples of the user's past demand response provisions using inverse reinforcement learning, without explicitly modeling the user's habits and wishes. It tackles the following issues present in the current demand response provision literature:

- (1) Unknown reward/objective function that guides the behavior;
- (2) Adaptability of the model to short-/long-term changes in human behavior;
- (3) Adaptability of the model to different users.

2. Demand response as a Markov decision process

The problem of demand response can be modeled in a simplified manner by a Markov Decision Process. Markov decision process is a stochastic model in which the current state depends only on the one that preceded it. Such a model is defined by a set of states S , a set of

actions A , the probability of transitioning from state s to state s' if an action a is taken ($P_{s,a}^{s'}$), the reward r in state s , and the discount factor γ . The agent in a certain state performs an action a , moves to state s' with probability $P_{s,a}^{s'}$ and receives a reward r for this action. A household providing demand response services can have their appliances clustered in three different categories:

1. time-shiftable (washing machine, dishwasher, dryer, electric vehicle, etc.)
2. power-curtable (AC, electric heater, etc.)
3. non-shiftable,

and based on their operation requirements, state and action spaces need to be determined. Time-shiftable devices can be shifted in time, however, once turned on, they operate at a constant power and cannot be turned off. On the other hand, power-curtable devices can change their power output at any given time. The last category, non-shiftable devices, should never have their power output or time schedule altered and should always have their demand fully satisfied. The state space of time-shiftable devices consists of time periods in which the operation of the appliance was requested but delayed for demand response purposes, while in the case of power-curtable and non-shiftable appliance it contains information about their demand at a given time period. The action space, on the other hand, is modeled as discrete reductions in *total consumption*¹ to multiples of a tenth of total household demand at a given time period, for example to 20% or 60% of total demand:

$$A = \{0, 0.1 D, 0.2 D, \dots, D\}.$$

Later on, after the action is decided upon and, therefore, the total consumption in the given time period is determined as well, an optimization problem is run to distribute the total consumption per individual appliance.

Reward function in a demand response problem is usually formulated in two parts — on one side financial earnings from providing the service, while on the other, comfort disruption cost. Financial earnings are calculated as the price offered for service provision (c) multiplied by the change in consumption a user made as compared to the assumed baseline consumption² (b) it should have had without providing demand response. If demand response is used for congestion management purposes, only the reduction of consumption should be financially rewarded, as given in (1). Comfort disruption, or discomfort (ψ), however, is commonly modeled by a quadratic function or an absolute value of the deviation from the previously set baseline value (e.g. desired room temperature). Nonetheless, since comfort experience can vary among different users, different times of the day or year, etc., it is not possible to define a single comfort function that works for all purposes.

$$r = c \cdot \max[0, (b - p)] - \psi \quad (1)$$

Thus, at each hour of the day, the agent receives data on the household demand and makes decisions (chooses actions) on the amount of flexibility provided in the form of reduced consumption in order to achieve the highest reward. It is necessary to find such a sequence of actions $\{a_1, \dots, a_k\}$ that the user's total reward is ultimately the highest possible. The function that maps state s to action a is called a policy π , and one wants to find the one that is optimal. Goodness of the policy is most often determined by the value function $V(s)$, which gives the expected value of the total reward if we follow policy π from state s , given by the Bellman equation (2).

$$V^\pi(s) = r(s) + \gamma \sum_{s'} P_{s,\pi(s)}(s') V^\pi(s') \quad (2)$$

¹ This paper assumes demand response for congestion management to simplify the problem and reduce the action state space.

² The baseline consumption of a single hour of the day is calculated as the average consumption in that hour for the last 10 days.

Furthermore, it is possible to define the value brought to the agent by doing action a in state s — $Q(s, a)$, as in (3).

$$Q^\pi(s, a) = r(s) + \gamma \sum_{s'} P_{s,a}(s') V^\pi(s') \quad (3)$$

By comparing the Q-values of different actions one can determine which action is better.

3. Learning user preference with IRL

The scientific community has been working on developing algorithms for determining optimal π , which directly depends on the definition of the reward function. However, it is often impossible, or at least very difficult, to mathematically precisely determine the reward function for which the optimal behavior of the agent will then be sought. Ng and Russell in [15] approach the problem from another angle. Assuming the knowledge of the optimal policy either through knowing the environment or the available trajectory samples, they formulated the problem of discovering the underlying reward function.

Starting from the fact that (2) can be rewritten in a vector notation as $\mathbf{V}^\pi = \mathbf{R} + \gamma \mathbf{P}_{a_1} \mathbf{V}^\pi$ and $\pi(s) \equiv a_1$, then

$$\mathbf{V}^\pi = (\mathbf{I} - \gamma \mathbf{P}_{a_1})^{-1} \mathbf{R} \quad (4)$$

and the reward must satisfy the following inequality:

$$(\mathbf{P}_{a_1} - \mathbf{P}_a) (\mathbf{I} - \gamma \mathbf{P}_{a_1})^{-1} \mathbf{R} \geq 0 \quad (5)$$

Many rewards can satisfy this formula (including the one equal to 0), which is why the authors introduce additional conditions which the reward must satisfy in order to be considered appropriate: the reward must be such that any deviation from the optimal policy (choosing a suboptimal action) brings a large reduction to the value function (penalty). Furthermore, the reward must be as small as possible since it is assumed that it is better to have as simple rewards as possible. Therefore, the Q-value of the optimal action must be greater than all the other actions. By stating that condition as the objective function (6)

$$\sum_{s \in S} \left(Q^\pi(s, a_1) - \max_{a \in \mathcal{A} \setminus a_1} Q^\pi(s, a) \right) \quad (6)$$

and obeying constraint (5), the optimization problem can be formulated:

$$\begin{aligned} & \max \sum_{i=1}^N \min_{a \in \mathcal{A} \setminus a_1} \left[(\mathbf{P}_{a_1}(i) - \mathbf{P}_a(i)) (\mathbf{I} - \gamma \mathbf{P}_{a_1})^{-1} \mathbf{R} \right] - \\ & \lambda \|\mathbf{R}\|_1 \\ & \text{s.t. } (\mathbf{P}_{a_1} - \mathbf{P}_a) (\mathbf{I} - \gamma \mathbf{P}_{a_1})^{-1} \mathbf{R} \geq 0, \quad \forall a \in \mathcal{A} \setminus \pi(s) \\ & |\mathbf{R}_i| \leq R_{\max}, \quad i = 1, \dots, N \end{aligned} \quad (7)$$

where i denotes one of the N states in state space S , λ represents an adjustable penalty that balances two parts of the objective function and R_{\max} stands for the theoretical maximum of the reward value.

In order to further limit the search space of the reward function, the following linear approximation of reward function is assumed:

$$R(s) = \alpha_1 \phi_1(s) + \alpha_2 \phi_2(s) + \dots + \alpha_d \phi_d(s) \quad (8)$$

Furthermore, most problems in the real world, including demand response, have unknown environments which is why it is only possible to access the state of the environment at a particular time period. For such problems, different state samples or trajectories are collected from which policies can be determined, which are then used to calculate the expected value of the value function as

$$\hat{V}^\pi(s_0) = \alpha_1 \hat{V}_1^\pi(s_0) + \dots + \alpha_d \hat{V}_d^\pi(s_0),$$

$$\text{where } \hat{V}_i^\pi(s_0) = \phi_i(s_0) + \gamma \phi_i(s_1) + \gamma^2 \phi_i(s_2) + \dots \quad (9)$$

Finally, the optimization problem transforms into:

$$\begin{aligned} & \max_{\alpha_i} \sum_{s_0 \in S_0} \sum_{j=1}^k \left(\hat{V}^{\pi^*}(s_0) - \hat{V}^{\pi_j}(s_0) \right) \\ & \text{s.t. } |\alpha_i| \leq 1, \quad i = 1, \dots, d \end{aligned} \quad (10)$$

and is solved according to Algorithm 1. The requirements of the algorithm are the optimal policy π^* , which is assumed to be known for the demand response case through the user's past consumption data, and a randomly chosen policy π_1 .³ For a more detailed mathematical formulation, please refer to [15].

Algorithm 1 IRL algorithm.

Require: $\Pi = \{\pi^*, \pi_1\}$, $V_i^{\pi_j}(s_0)$, $k = 2$
while $k < N$ **do**
 $\alpha_i \leftarrow$ optimization problem in (10) for policies in Π
 $\pi_k \leftarrow$ train the agent with α_i and Eq. (8)
 add π_k to Π
if performance is satisfying **then**
 break
end if
 $k \leftarrow k + 1$
end while

4. Input data and model setup

The Pecan Street data set, which consists of data on consumption per individual household appliance for 25 households, was used [18]. Out of time-shiftable appliances, the data set contains consumption data for an electric vehicle, a washing machine, a dishwasher and a dryer. Power-curtable representative is the air conditioning, while the rest of the data falls under the category of non-shiftable devices. Depending on the user, various combinations of household appliances per household are possible (some users do not own an electric vehicle, some do not have a dishwasher, etc.) Data is available from the beginning of April 2018 until beginning of November in the same year. The training set covers the periods from April 1st until July 1st and from August 2nd until November 1st. The period in the middle is used for testing purposes.

Still, the available data show household consumption without demand response service provision. Due to the lack of consumption data of households that actually provide demand response service, which is required as an input (serves as the desired behavior) for the IRL algorithm from Section 3, demand response provision is foremost simulated by a reinforcement learning (RL) model. So, for the algorithm verifying purposes, it is assumed that the household with the consumption from the available data set is providing demand response under the following reward function (11):

$$r = c \cdot \max[0, (b - p)] - w_{ac} \cdot (\Delta p_{ac})^2 - \sum_{TS} w_m \cdot (t_m^d)^2 \quad (11)$$

where the discomfort is assumed to follow the quadratic function of air conditioning consumption deviation (Δp_{ac}) and quadratic function of time delay for each time-shiftable device (t_m^d). Coefficients w_{ac} and w_m are measures of dissatisfaction — the higher the value, the higher the inconvenience when the appliance schedule is changed. With this approach, the optimal policy π^* under the reward with assumed discomfort modeling is obtained and is considered to be the historic behavior of the user who provided demand response services. Nonetheless, it is important to note that in the case of having a data set that contains consumption data of a household providing demand

³ An example of the policy π_1 is to choose a random action a at a state s .

response, the previously described approach of generating data should be omitted and the available data should be used directly.

The basis functions $\phi_i(s)$ for approximating the reward function in Eq. (8) are domain-specific and should be carefully designed. Therefore, six different basis functions were assumed in this problem, which differed in the way the profit from the demand response service is calculated and in the functions that model the discomfort. More specifically, profit could be calculated as $\max[0, (b - p)]$ or just $(b - p)$ multiplied by price c . Discomfort could be neglected, considered just as the deviation in the AC consumption and time delay or by taking the square of deviation and time delays. Combining two ways of calculating the financial part of the reward function and three ways of calculating the discomfort part, results in six total combinations. Then optimizing coefficients α_i should find the appropriate reward function approximation that matches the desired behavior.

The agent is trained on different number of episodes (1500, 2500 and 3500), depending on the case study, where one episode corresponds to a single day divided into 96 time steps of 15 min each. Deep Q Network (DQN) with 1 hidden layer of 32 units and a ReLU activation function after the first two layers is used. All three layers are linear. The input data to the network is the state of the household, which consists of the demand for power-curtable and non-shiftable devices and open delays for time-shiftable devices (total of 6 data points), the unit price of providing flexibility, and the baseline demand. The output of network is the action, that is, the reduction in the total household consumption. To select the agent's action during the training process, an ϵ -greedy strategy with a decreasing value of ϵ is implemented. The initial value of epsilon is 1.0, and with each episode, the value drops to 99.9% of the previous one. Furthermore, during training, a soft update of the target network is used with the value of the update parameter $\tau = 0.001$. Among other hyperparameters important for the training, the batch size of 32, the discount rate when updating DQN network parameters of 0.9, the learning rate of 0.001, and the buffer size of 10,000 input data were selected. The end of the IRL algorithm is set to be reached after 10 iterations.

5. Case studies

In this section, three case studies are presented to obtain answers to the following queries:

- (1) whether the developed model can even succeed in reproducing the desired user behavior;
- (2) whether it is possible to find an approximate reward function that gives good enough results for different days/consumptions in a particular household;
- (3) whether the model will be able to properly function in households guided by a different reward function;

All simulations were performed using Gurobi 9.1.2 and run on an AMD Ryzen 7 4700U CPU with 16 GB of RAM at 3733 MHz. The code used in this paper is available in [19] and the data can be downloaded from [18].

5.1. Single day

To determine whether the presented model is at all capable of adequately approximating the reward function and recreating the desired behavior of the end user that provides a demand response service (obtained by following policy π^* for the observed day, explained in Section 4), the model was trained on two arbitrarily chosen representative days from the training set: one with few active controllable devices (April 2nd; trained on 1500 episodes) and one in which all controllable devices are running (August 31st; trained on 2500 episodes). Fig. 1 shows the rewards and consumption changes made by the end user following the policy π^* (desired, *optimal* behavior; in blue) and the behavior the user learned guided by the approximate reward function

Table 1

The average differences in demand response provision obtained with the true reward and with the learned reward for two days with different usage of controllable devices (April 2nd = low usage; August 31st = high usage).

	April 2nd	August 31st
MAE	0.0	0.12
MSE	0.0	0.09
Pearson	1.0	0.77

(in green). In order to compare and validate the results, both policies (optimal and learned) were tested with the true reward function⁴ for both days. For both days, it can be observed that the reward converges to the optimal one after a few iterations. However, in the case of the day with lower usage of controllable devices, the agent's decisions obtained with the learned reward mirror the decisions obtained with the policy π^* , given in Fig. 1(a). Similarly, the corresponding reward reaches the exact same value in the final iteration as with the policy π^* , as seen in Fig. 1(b). Increasing the number of active devices in a day also makes it difficult to approximate the reward function, as seen in Figs. 1(c) and 1(d). Although the agent's decisions guided by the learned policy were similar to the optimal ones, the accuracy considerably dropped compared to the case of April 2. This can be read more clearly in Table 1 where, for the 31st August, the mean absolute error (MAE) was 0.12, versus 0.00 for the 2nd April. The same can be recognized with the remaining two metrics given in Table. Finally, the presented model can satisfactorily reproduce the desired behavior, but increasing the number of controllable devices makes the approximation of the reward more difficult.

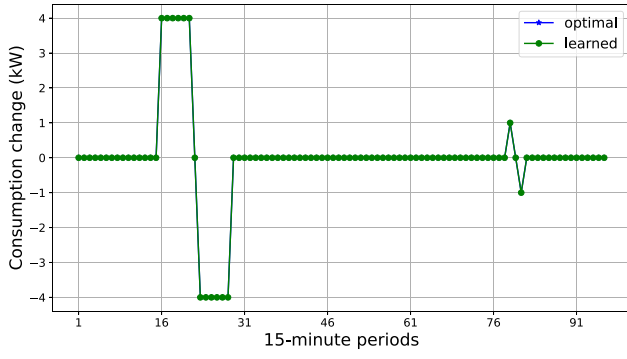
5.2. Adaptability to different days

The model is trained for 3500 episodes on the entire training set to determine the potential of the approach to approximate the reward on different consumption patterns. Testing is then done on 31 days from the test set and the results are presented in Table 2. The average (and median) MAE and mean squared error (MSE) on all tested days of 0.144 (0.159) and 0.065 (0.061) are comparable to the results for August 31 in Table 1, which leads to a conclusion of satisfactory model performance for different days. Among other things, dependency on the number of active controllable appliances on a certain day, resulted in the Pearson correlation coefficient for demand response provision based on the learned and optimal policy of 0.589 on the worst and 0.958 on the best tested day.

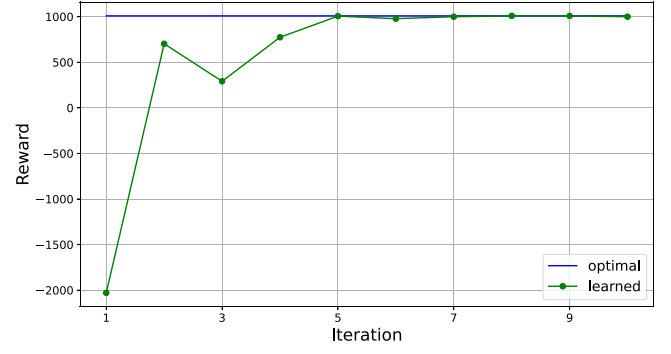
A more detailed view is provided in Fig. 2, where appliance schedules for the first twelve hours of August 8th under the optimal policy (Fig. 2(a)) and learned policy (Fig. 2(b)) are compared. The values in these Figs. are in the interval $[-1, 1]$ for all appliances aside from the AC, where 1 means that an incoming request to start the appliance operation was ignored, 0 that it was accepted, and -1 that the device was started even without a prior request. The AC can take values from completely reduced consumption (0) to no reduction at all (1). Mostly, the agent's actions with both policies are the same on the observed day. The difference can be seen, e.g. in the second 15-minute period (red at the bottom) when the EV is charged with the optimal policy and is not charged with the learned one. Furthermore, there should have been some reductions of the AC consumption (light blue shades) based on the optimal policy, which did not happen with the learned one. Although a satisfactory approximation of the reward function can be obtained for certain days or parts of the day, the insufficient complexity of the presented model makes it impossible to achieve extremely high accuracy (e.g. less than 5% of errors for all tested days).

⁴ This is only possible due to the lack of demand response consumption data and assumed reward function for generating policy π^* .

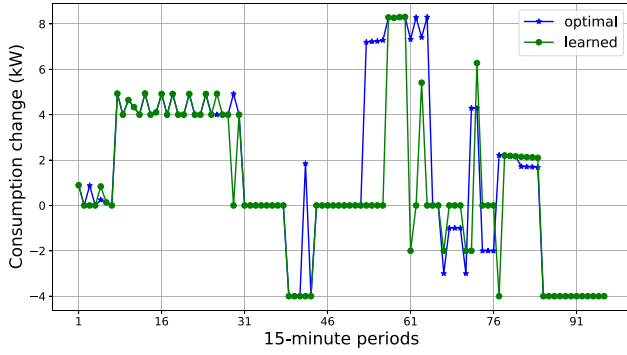
⁵ MAE and MSE throughout this paper are in the range of $[0, 1]$.



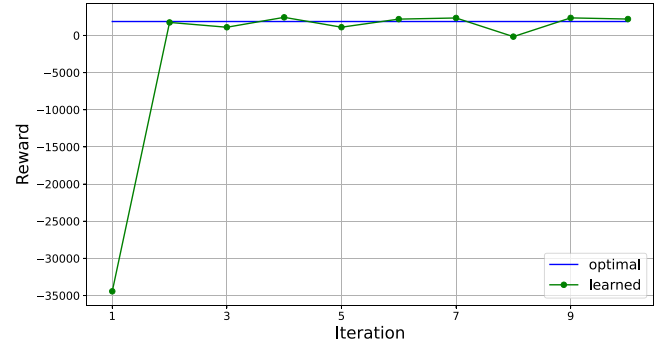
(a) Changes in consumption on April 2nd



(b) Obtained rewards on April 2nd

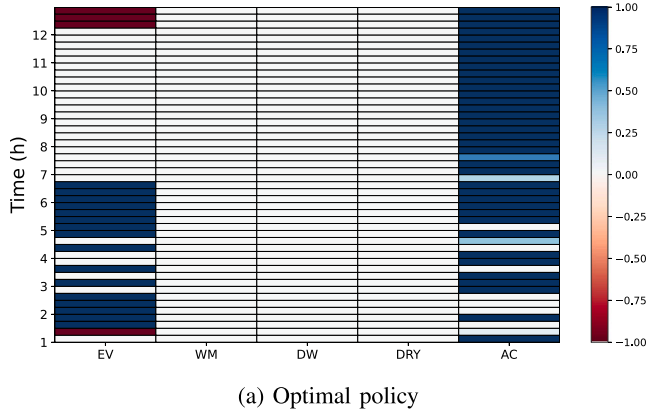


(c) Changes in consumption on August 31st

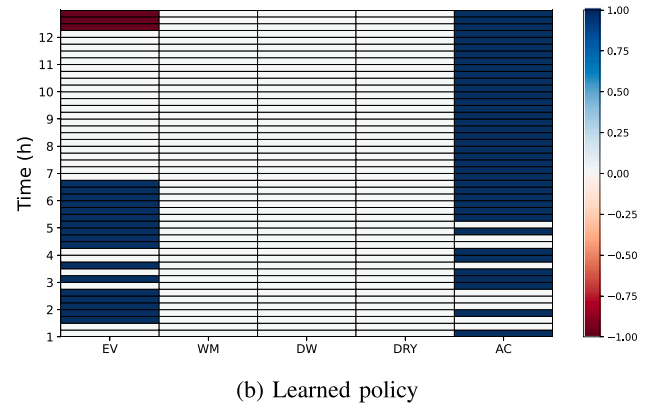


(d) Obtained rewards on August 31st

Fig. 1. Comparison of demand response provision obtained with true and learned reward evaluated with true reward.



(a) Optimal policy



(b) Learned policy

Fig. 2. Comparison of the devices' schedule with optimal and learned policy tested on August 8th.

Table 2

MAEs, MSEs and Pearson correlation between demand provision with optimal and learned behaviors over the whole test set.

	Average	Minimum	Maximum	Median
MAE	0.144	0.021	0.277	0.159
MSE	0.065	0.020	0.167	0.061
Pearson	—	0.589	0.958	0.843

5.3. Adaptability to different users

Concerning the adaptability to different households, five types of households with different underlying rewards were created that guide their scheduling decisions (actions). True rewards are assumed to vary in the way service provision revenue is modeled (either only the

reductions or changes in both directions are rewarded) and in the way discomfort is modeled (absolute value or quadratic function). For four of them, optimizing coefficients α_i could lead to the exact true underlying reward of a particular household, since these true rewards are modeled to be equal to basis functions used in the IRL problem.⁶ However, one reward (household 5 in Fig. 3) is modeled differently to investigate how the reward could adapt to a different environment. That reward is modeled as in (11), however, with different dissatisfaction (w_{ac}, w_m) coefficients. The resulting MAEs during the test period for all observed households are given in Fig. 3. It is observed that for all households, at least 75% of MAE falls below 0.20. Households 1 and 5 show a higher dispersity of errors (longer boxes) compared

⁶ One basis function is chosen for one household.

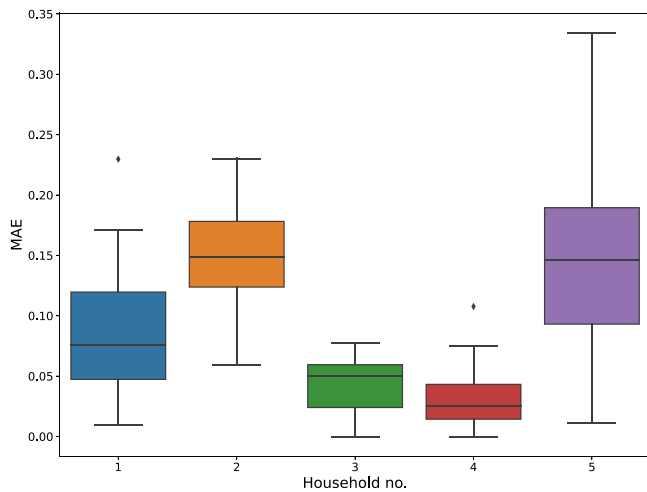


Fig. 3. Distribution of MAEs between optimal and learned demand response provision for 5 households over the test set.

to the remaining households. Furthermore, distributions of MAEs in households 1, 2 and 4 are slightly positively skewed (range of MAEs higher than the median is wider than of those with MAEs lower than the median), while in households 3 and 5 a negative skew can be seen (more dispersed lower range of MAEs). The smallest errors, which are also the least scattered, occur in households 3 and 4, which in the background have a reward with an absolute value for discomfort and payment for changes of consumption in both directions (house 3) and only for reductions (house 4). Household 5, with a different underlying reward function that generated demand response consumption data, generally shows comparable MAEs to the remaining houses, but the total size of the interval within which the errors are located is much larger (the errors of the best and worst day in terms of MAE differ by approximately 30%).

6. Conclusion

To facilitate the provision of residential demand response, it is necessary, in addition to development of the infrastructure, to motivate the end users to participate. Most of them will decide to participate solely if the need for their engagement in controlling the household appliances is not high and if their comfort remains (almost) intact. To make this possible, an inverse reinforcement learning model for learning the reward function of a household that provides demand response is presented in this paper. Through learning the reward function from the historic data, unnecessary user engagement in decision-making process is avoided and their comfort is kept unharmed.

The presented model showed satisfactory results of recreating the user-desired reward function. It is showed that the obtained reward function can generalize over different daily consumptions, however, the more complex the consumption (in terms of the number of active appliances) the harder it is to recreate the reward function. The model proved to be successful even when applied to different households with different consumption habits and desires.

Considering the complexity of the demand response problem, the presented model is too simple to significantly reduce errors in recreating the reward function. Assuming a family of functions for the reward function at the outset greatly reduces the solution space. Furthermore, the need to define the basis functions as a kind of input data also limits the solution space, and at the same time has a significant impact on the final result. In future work, it will be compared how more advanced techniques of inverse reinforcement learning, which do not restrict the reward function family of functions or learning the basis functions instead of assuming them known, may improve reward learning process

and demand response provision. Lastly, it is important to emphasize that the inherent problem of trying to learn the objective/reward function in the demand response problem in this way is the impossibility of objectively defining a satisfactory result. For a certain household, the fulfillment of only half of its wishes may be enough, while on the other hand, another household could expect a perfect performance in order to find the model's work satisfactory. In future work, efforts will be made to define metrics in order to enable the most objective evaluation of the model's performance in this research area.

CRedit authorship contribution statement

Nikolina Čović: Conceptualization, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jochen L. Cremer:** Supervision, Writing – review & editing. **Mrvoje Pandžić:** Funding acquisition, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

- [1] Bo de Wildt, Ross Quirke, Jos Sijm, Barriers To Demand Response, TNO, 2022.
- [2] Yi Liu, Yun Zhang, Kairui Chen, Si Zhe Chen, Bin Tang, Equivalence of multi-time scale optimization for home energy management considering user discomfort preference, *IEEE Trans. Smart Grid* 8 (4) (2017) 1876–1887.
- [3] Sajjad Ali, Ateeq Ur Rehman, Zahid Wadud, Imran Khan, Sadia Murawwat, Ghulam Hafeez, Fahad R. Albogamy, Sheraz Khan, Omaji Samuel, Demand response program for efficient demand-side management in smart grid considering renewable energy sources, *IEEE Access* 10 (2022) 53832–53853.
- [4] Motahhar Tehrani Nowbandegani, Mehrdad Setayesh Nazar, Miadreza Shafie-khah, João P.S. Catalão, Demand response program integrated with electrical energy storage systems for residential consumers, *IEEE Syst. J.* 16 (3) (2022) 4313–4324.
- [5] Ahmad Alzahrani, Ghulam Hafeez, Gul Rukh, Sadia Murawwat, Faiza Iftikhar, Sajjad Ali, Syed Irtaza Haider, Muhammad Iftikhar Khan, Azher M. Abed, Demand response for optimal power usage scheduling considering time and power flexibility of load in smart grid, *IEEE Access* 11 (2023) 33640–33651.
- [6] Jasper van Tilburg, Luciano C. Siebert, Jochen L. Cremer, MARL-iDR: Multi-agent reinforcement learning for incentive-based residential demand response, in: *Proceedings of the 2023 IEEE Belgrade PowerTech, 2023*, <http://dx.doi.org/10.1109/PowerTech55446.2023.10202941>.
- [7] Di Xie, Liang Yu, Tao Jiang, Yulong You, Distributed energy optimization for HVAC systems in university campus buildings, *IEEE Access* 6 (2018) 59141–59151, <http://dx.doi.org/10.1109/ACCESS.2018.2872589>.
- [8] Bo Chai, Zaiyue Yang, Kunlun Gao, Ting Zhao, Iterative learning for optimal residential load scheduling in smart grid, *Ad Hoc Netw.* 41 (2016) 99–111, <http://dx.doi.org/10.1016/j.adhoc.2016.01.005>.
- [9] Anusha Nagabandi, Clavera Ignasi, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, Chelsea Finn, Learning to adapt in dynamic, real-world environments through meta-reinforcement learning, in: *International Conference on Learning Representations*, 2018.
- [10] Liang Yu, Weiwei Xie, Di Xie, Yulong Zou, Dengyin Zhang, Zhixin Sun, Linghua Zhang, Yur Zhang, Tao Jiang, Deep reinforcement learning for smart home energy management, *IEEE Internet Things J.* 7 (2020) 2751–2762, <http://dx.doi.org/10.1109/JIOT.2019.2957289>.
- [11] Guanyu Gao, Jie Li, Yonggang Wen, DeepComfort: Energy-efficient thermal comfort control in buildings via reinforcement learning, *IEEE Internet Things J.* 7 (2020) 8472–8484, <http://dx.doi.org/10.1109/JIOT.2020.2992117>.
- [12] Liang Yu, Yi Sun, Zhanbo Xu, Chao Shen, Dong Yue, T. Jiang, X. Guan, Multi-agent deep reinforcement learning for hvac control in commercial buildings, *IEEE Trans. Smart Grid* 12 (1) (2021) 407–419.
- [13] Zhengbo Zou, Xinran Yu, Semiha Ergun, Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network, *Build. Environ.* 168 (2020).
- [14] Anchal Gupta, Youakim Badr, Ashkan Negahban, Robin G. Qiu, Energy-efficient heating control for smart buildings with deep reinforcement learning, *J. Build. Eng.* 34 (2021).

- [15] Andrew Y. Ng, Stuart J. Russell, Algorithms for inverse reinforcement learning, in: In Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00, 2000.
- [16] Dorsa Sadigh, Anca Dragan, Shankar Sastry, Sanjit Seshia, Active preference-based learning of reward functions, *Robotic Sci. Syst.* (2017) <http://dx.doi.org/10.15607/RSS.2017.XIII.053>.
- [17] Erdem Biyik, Malayandi Palan, Nicholas C. Landolfi, Dylan P. Losey, Dorsa Sadigh, Asking easy questions: A user-friendly approach to active reward learning, in: *Conference on Robot Learning*, 2019.
- [18] Pecan Street Dataport, Available: <https://www.pecanstreet.org/dataport/>.
- [19] Nikolina Čović, Jochen L. Cremer, Hrvoje Pandžić, Learning-reward-function-demand-response, 2023, GitHub repository, <https://github.com/nikskiks/learning-reward-function-demand-response>.