Real-time Retail Electricity Pricing Using Offline Reinforcement Learning

Sharath Ram Kumar, Arvind Easwaran, Benoit Delinchant, Rémy Rigo-Mariani

Real-time Retail Electricity Pricing Using Offline Reinforcement Learning

Sharath Ram Kumar
Univ. Grenoble Alpes, CNRS, Grenoble INP, G2Elab
Grenoble, France
Nanyang Technological University
Singapore
CNRS@CREATE LTD, 1 Create Way, 08-01 CREATE Tower
Singapore
sharath.ramkumar@cnrsatcreate.sg

Arvind Easwaran
Nanyang Technological University
Singapore
CNRS@CREATE LTD, 1 Create Way, 08-01 CREATE Tower
Singapore
arvinde@ntu.edu.sg

Benoit Delinchant
Univ. Grenoble Alpes, CNRS, Grenoble INP, G2Elab
Grenoble, France
CNRS@CREATE LTD, 1 Create Way, 08-01 CREATE Tower
Singapore
benoit.delinchant@g2elab.grenoble-inp.fr

Rémy Rigo-Mariani
Univ. Grenoble Alpes, CNRS, Grenoble INP, G2Elab
Grenoble, France
CNRS@CREATE LTD, 1 Create Way, 08-01 CREATE Tower
Singapore
remy.rigo-mariani@g2elab.grenoble-inp.fr

ABSTRACT

Real-time electricity pricing has the potential to provide incentives for retail consumers to offer flexibility services by altering their consumption patterns. However, such incentive schemes have met with limited success in the real world due to problems such as low consumer interest and the creation of rebound peaks after periods of high pricing. In this paper, a model of an individual consumer's response to real-time prices, which captures these effects, is presented. A contract between a retail service provider and a consumer is proposed, and a method for personalized real-time price generation based on smart meter data, using reinforcement learning, is implemented. Initial results suggest that the approach can be used to achieve grid-level objectives while rewarding consumer flexibility.

CCS CONCEPTS

• **Computing methodologies** → *Intelligent agents.*

KEYWORDS

Reinforcement Learning

ACM Reference Format:
Sharath Ram Kumar, Arvind Easwaran, Benoit Delinchant, and Rémy Rigo-Mariani. 2024. Real-time Retail Electricity Pricing Using Offline Reinforcement Learning. In *The 15th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '24), June 04–07, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3632775.3661964

E-Energy '24, June 04–07, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0480-2/24/06
https://doi.org/10.1145/3632775.3661964

1 INTRODUCTION

One of the major issues faced by modern power system operators is the need to cater to short-term peaks in consumer energy demand. Traditionally, these are met by the activation of peaking power plants such as diesel generators and gas plants, which can be operated on demand. However, such solutions are capital intensive and cause significant emissions, and are only operated for brief periods during the day. In recent years, with the increasing penetration of clean energy sources such as solar and wind power, there has also been growing interest in the use of Energy Storage Systems (ESS) to address the issue. These are promising alternatives, but are hindered by the high costs for large-scale ESSs and the intermittency in renewable generation. [12]

In this context, there is significant interest in the use of dynamic electricity pricing as an incentive for consumers to shift their consumption from periods of high demand to those of low demand. Theoretically, such approaches are considered more market efficient, with the promise of financial benefits both for consumers as well as for the utility company [2]. Various tariff structures, such as time-of-use (TOU), peak-load pricing and day-ahead dynamic pricing, have been implemented across the world to mixed success, often due to low consumer awareness and the formation of rebound peaks [4].

This paper considers the most extreme form of dynamic electricity pricing, real-time pricing (RTP), from the perspective of a utility provider. In this scheme, each consumer is offered a new price for each time slot in a day based on real-time demand. Compared to existing works[5, 6, 11], the key novelties here are:

- A price-response model of a consumer that captures real-world aggregated effects such as rebound peaks.

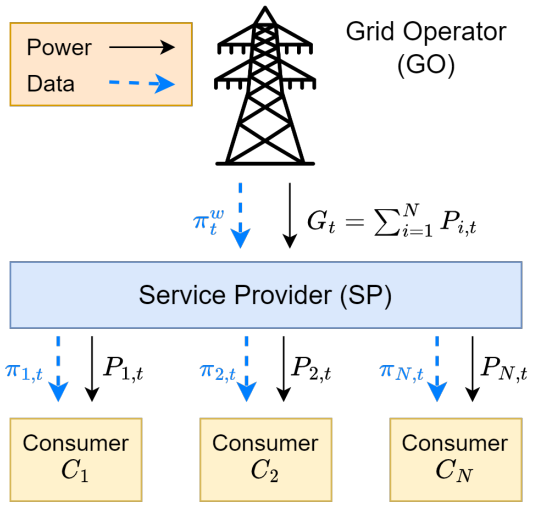

Figure 1: System Model

- A contract between the service provider and the consumer that does not rely on behind-the-meter information.
- A scalable method for price generation based on offline reinforcement learning.

2 SYSTEM MODEL

The system consists of a grid operator (GO), a service provider (SP) and a set of consumers (C) as shown in Fig 1. Each day, there are T time slots for which the SP provides a real-time price $\pi_{i,t}$ ($t \in \{1, ..., T\}$) to the consumer C_i based on the information available at time $t-1$. At each step, the GO sends the SP the spot wholesale electricity price π_t^w. At the end of the time slot, the consumer's consumption $P_{i,t}$ is measured using smart meter readings. It is notable that the SP does not have access to any behind-the-meter measurements or scheduled demand from each consumer, thus protecting consumer privacy.

Section 2.1 discusses the interaction between the SP and a single consumer - as such, the index i is dropped for clarity.

2.1 Consumer Model

The consumer model is based on the concept of price elasticity [2], which measures the change in electricity consumption due to a change in the price. It is similar to the models presented in [5], [6] and [11], with the important change that the price at the current time, π_t, can also directly affect the future demand.

At the start of the day, the consumer C has a baseline demand for each time slot, $\mathbf{D} = \{d_0, d_1, ..., d_T\}$ which represents their consumption at some static price $\pi^b \in \mathbb{R}^+$. The deviation from $\mathbf{D}$ due to π_t is determined by a price-dependent function $\varepsilon(\pi) \in (0, 1)$, where higher values indicate a consumer that is more responsive to price changes.

During high-price periods ($\pi_t > \pi^b$), the consumer reduces their consumption and incurs an immediate backlog b_t given by Eq 1, where $\mathbb{I}_{\pi > \pi^b}$ is the indicator function:

$$b_t = \mathbb{I}_{\pi > \pi^b}(\pi_t) \times \varepsilon(\pi_t) \times d_t \tag{1}$$

The cumulative backlog B_t is tracked as in Eq 2, where $B_0 = 0$. The condition $B_{T+1} = 0$ is enforced, so that the total energy consumption over the day is conserved regardless of the price sequence.

$$B_t = b_{t-1} + B_{t-1} \tag{2}$$

Similarly, when $\pi_t < \pi^b$, the consumer advances their planned consumption to take advantage of the lower prices, as modelled by Eqs 3 - 6. In this formulation, the consumer prefers to shift the load from future periods where they are planning to consume more energy to this time slot.

$$w_{t,k} = \frac{d_{t+k}}{\max(\mathbf{D}_{t+1:T})} \tag{3}$$

$$a_{t,k} = \mathbb{I}_{\pi < \pi^b}(\pi_t) \times w_{0,k} \varepsilon(\pi_t)^{(2 - w_{t,k})} \times d_{t+k} \tag{4}$$

$$A_t = \sum_{k=1}^{T-t} a_{t,k} \tag{5}$$

$$d_{t+k} \leftarrow d_{t+k} - a_{t,k} \forall k \in \{1, 2, ..., T-t\} \tag{6}$$

Here, $w_{t,k}$ is a weight assigned by the consumer to the demand d_{t+k}, $a_{t,k}$ is the actual shifted part of d_{t+k}, and A_t is the total load shifted from future time steps to this step.

The final power consumption at time t is then given by Eq 7.

$$P_t = d_t + \alpha_t B_t - b_t + A_t \tag{7}$$

Here, $\alpha_t \in (0, 1)$ is a consumer-independent parameter which represents the fraction of the cumulative backlog that will be consumed in this time step. Its value linearly increases from 0.25 to 1.0 from 5 AM to 10 PM, and is constant outside these periods at the corresponding value.

The cumulative backlog B_{t+1}, initially calculated using Eq 2, is then updated as shown in Eq 8.

$$B_{t+1} \leftarrow B_{t+1} - \alpha_t B_t \tag{8}$$

Fig 2 shows the final consumption calculated by the model, under a typical time-of-use (TOU) pricing scheme, for the same $\mathbf{D}$ for different ε. Here, $\varepsilon(\pi)$ linearly maps π from 0 to ε^{max}.

2.2 SP Objectives and Contract

In this work, the objective of the SP is to maximize its profits while reducing the peak-to-average ratio (PAR) of the aggregated consumption curve of the consumers. The aggregated consumption for a cluster of N consumers is denoted as $G_t = \sum_{i=1}^{N} P_{i,t}$. Then, the PAR of the power profile $\mathbf{G} = \{G_1, G_2, ..., G_T\}$ is described in Eq 9. The SP achieves its objective by supplying each consumer with a new price $\pi_{i,t}$ at every time slot in a sequential manner.

$$\text{PAR}(\mathbf{G}) = \frac{\max(\mathbf{G})}{\bar{\mathbf{G}}} \tag{9}$$

The proposed contract imposes constraints on the prices offered by the SP as shown in Eqs 10 and 11. By linking the prices to a baseline price, π^b, the consumer is guaranteed to receive periods of higher and lower prices through the day.

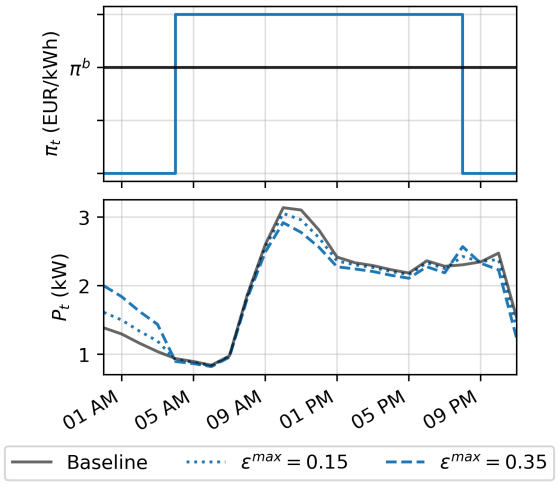

Figure 2: Consumer Model Operation

$$\frac{|(\frac{1}{T}\sum_{\tau=1}^{T}\pi_{i,\tau}) - \pi^b|}{\pi^b} < 3\% \tag{10}$$

$$|\pi_{i,t} - \pi^b| <= \delta\pi, \forall t \in [1,T], \forall i \in [1,N], \delta\pi \in \mathbb{R}^+ \tag{11}$$

Under the contract, the i-th consumer's electricity bill, c_i^c (in €), is given by Eq 12. This formulation is typical in dynamic pricing contracts, and implicitly rewards those consumers who alter their consumption to align to the pricing strategy.

$$c_i^c = \sum_{\tau=1}^{T} \pi_{i,\tau} P_{i,\tau} \Delta t \tag{12}$$

The GO sells electricity to the SP at the wholesale price π_t^w, with an additional surcharge, $c^{sur}(G_t)^2$, if $G_t > G^{peak}$. Here, G^{peak} is a threshold power (in kW) agreed upon by the SP and the GO, and c^{sur} is a scaling factor (in €/kW^2). Thus, the profit achieved by the SP for trading with this consumer cluster, c_t^{sp}, is calculated as shown in Eq 13.

$$c_t^{sp} = [(\sum_{i=1}^{N} \pi_{i,t} P_{i,t}) - \pi_t^w G_t]\Delta t - (\mathbb{I}_{G>G^{peak}}(G_t) \times c^{sur}(G_t - G^{peak})^2) \tag{13}$$

The SP has historical data of $\mathbf{P}_{i,t^-}$ and Π_{i,t^-} for each consumer, using which it is able to extract the historical demand $\mathbf{D}_{i,t^-}$. This work also assumes that the SP knows the consumer's response function $\varepsilon_i(\pi)$ perfectly - however, it does not know the consumer's intended consumption ($\mathbf{D}_i$) over the next day.

3 METHODS

3.1 Reinforcement Learning

Reinforcement Learning (RL) is a branch of machine learning focused on training agents, through interaction, to make sequential decisions in an environment to maximize cumulative rewards [9]. The goal of the agent is to find the optimal mapping (*policy*) from a given state (s_t) to an action (a_t). In this work, the implementation of the Proximal-Policy Optimization (PPO) RL algorithm available in the Stable-Baselines3 [8] python library was used, along with a custom Gymnasium [10] environment for training and testing. The default hyperparameters were used for all runs, with one exception - the discount factor, γ, was set to 0.958 to reflect the time horizon $T = 24$[1].

An RL approach is chosen in this context for its model-free nature and its ability to generalize, given sufficient training resources [8]. Compared to model-based approaches such as Model Predictive Control (MPC), it also eliminates the need to have a forecast of the demand. The PPO algorithm is used to train a neural network which accepts the local state of a given consumer $s_{i,t}$ and outputs a price $\pi_{i,t}$. This is done in an offline simulation with artificial data - as such, the network is never trained by direct interaction with the consumers.

3.2 Generation of Demand Profiles

An approach based on Kernel Density Estimation (KDE) is used [1] to generate new power profiles based on the historical demand data, $\mathbf{D}_{i,t^-}$. First, a set of KDEs $\{K_\tau\}$ are fit to the data $\{\mathbf{D}_{i,\tau}\}$ - ie, a separate kernel is created for each time slot τ. Then, by sampling iteratively from the kernels as described in Eq 14, new sequences $\hat{\mathbf{D}} = \{\hat{d}_1, \hat{d}_2, ..., \hat{d}_T\}$, which mimic the consumption patterns embdedded in $\mathbf{D}_{i,t^-}$, are generated.

$$\hat{d}_t = \lambda k_t + (1-\lambda)\hat{d}_{t-1} \tag{14}$$

Here, $k_t \sim K_t$ and $\lambda \in (0,1)$ is a mixing parameter to capture the temporal correlation of electrical demand. In this study, $\lambda = 0.4$.

3.3 Offline RL Training Environment

The SP creates an offline training environment for the RL agent using artificial data generated as described in Section 3.2, and the consumer model described in Section 2.1. Under the terms of the contract described in Section 2.2, the SP's pricing strategy should closely follow the consumer's expected consumption patterns. As such, it first groups together consumers who have similar consumption patterns and price-response behaviours. For each group, a dedicated training environment is created to train a corresponding RL agent. The process is outlined in Fig 3.

3.4 Price Generation

To solve the sequential price generation problem described in Section 2.2, an offline RL approach is proposed here. The state of each consumer C_i at time t is denoted by $s_{i,t} = \{t, \bar{\pi}_{i,1:t-1}, P_{i,t-1:t-2}, \bar{P}_{i,1:t-1}, \max(P_{i,1:t-1}), \pi_t^w\}$ [2], giving $s_t \in \mathbb{R}^7$. The action taken by the RL agent, $a_{i,t} = \pi_{i,t}$, and $a_{i,t} \in [\pi^b - \delta\pi, \pi^b + \delta\pi]$. A training environment is created to generate artificial values for $s_{i,t}$ as described in Section 3.3, and an RL agent is trained here in an offline manner using the PPO algorithm. The reward function, used to calculate the agent's reward $r_{i,t}$, is described in Eqs 15 - 16 [3].

[1] The approximation $T \sim \frac{1}{1-\gamma}$ is used here

[2] The notation $\bar{\mathbf{X}}$ represents the mean value of a vector $\mathbf{X}$

[3] Γ is a large positive number

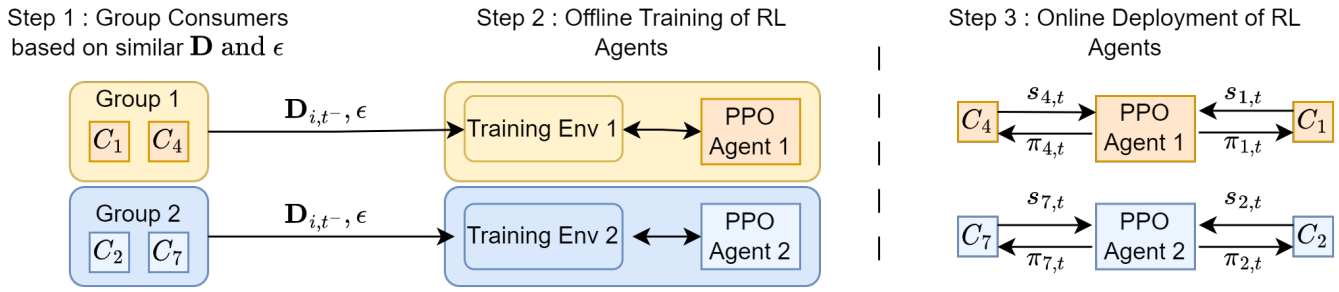

Figure 3: RL Training and Deployment Methodology

$$v_{i,t} = (\pi_{i,t} - \pi_t^w)P_{i,t} \tag{15}$$

$$r_{i,t} = \begin{cases} v_{i,t} - \Gamma, & \text{if } t = T \text{ and Eq 10 is violated} \\ v_{i,t} + \Gamma, & \text{if } t = T \text{ and PAR}(\mathbf{P}) < \text{PAR}(\mathbf{D}) \\ v_{i,t}, & \text{otherwise} \end{cases} \tag{16}$$

Here, the agent aims to maximize the expected income obtained from each consumer ($v_{i,t}$) while following the constraints embedded in the contract. This formulation is suitable for this problem since dedicated agents are required for consumers with different consumption patterns. As such, each agent tends to offer higher prices at times of higher expected consumption for its cluster, which implicitly encourages consumers to shift their consumption to lower price periods.

As more data becomes available, the steps in Fig 3 are repeated periodically with new agents and groups. It must be noted that the reward $r_{i,t}$ does not have any significance in the deployment phase (ie, Step 3 in Fig 3). The fact that Eq 16 is calculated using PAR($\mathbf{D}$) also implies that it cannot be calculated outside a simulation.

4 RESULTS AND DISCUSSION

To test the performance of the approach, a cluster of $N = 100$ consumers with similar consumption patterns was chosen. Each consumer was assigned a linear price response function $\epsilon(\pi)$, with a maximum value ϵ^{max} of either 0.15 or 0.4 with equal probability. The SP is required to generate hourly prices, ie, $T = 24$. Two PPO agents were trained over a period of 5000 simulated days using artificially generated data, with one agent for each ϵ^{max}. The other parameters were set as $\pi^b = 0.25$ €/kWh, $\delta\pi = 0.05$ €/kWh, $c^{sur} = 0.02$ €/kW^2 and $G^{peak} = 230$ kW. The consumer consumption profiles are based on [7], and the wholesale spot electricity prices are obtained from [3]. The train and test data for the study were generated as described in Section 3.2, using consumption data for July and August 2018 respectively.

The key metrics are reported in Table 1 - for RL, the numbers are averaged over 10 runs with different random seeds. Fig 4 shows the aggregated power profile for the cluster under different pricing schemes.

4.1 SP Profits and Consumer Bills

Under the proposed scheme, the results suggest that the SP profits would increase by around 7%, while the consumer's electricity bill would remain unchanged compared to the static pricing case. Such an outcome is disadvantageous for the consumer - as such, it is important to offer additional financial incentives, such as a profit-sharing mechanism, to compensate the consumer for the flexibility provided.

In this study, the average consumer bill would reduce by 0.7% for every 1% of SP profit shared equally across the cluster. As such, the SP would be able to achieve a 5% reduction in the average bill while maintaining the same profit as in the static case.

Pricing	PAR	SP Profit c^{sp} (€)	Mean c^c (€)
Static (π^b)	1.46	775.05	10.75
TOU	1.46	798.15	10.73
RL RTP	**1.29 ± 0.05**	**828.74 ± 8.33**	10.72 ± 0.07

Table 1: Results

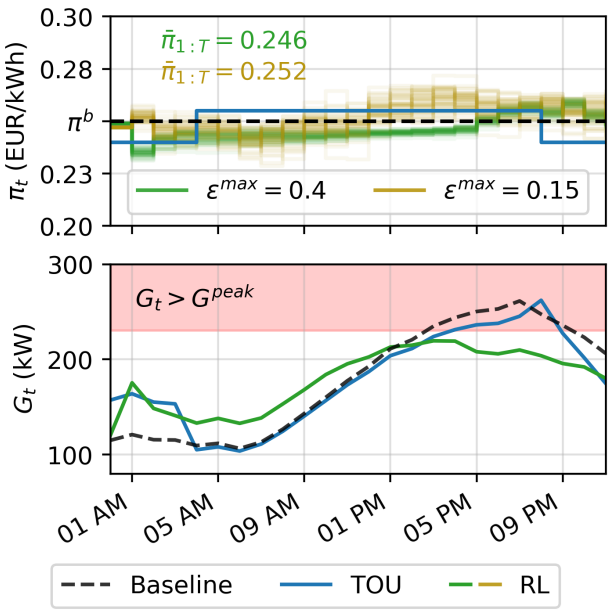

Figure 4: Prices and Aggregated Grid Profiles

4.2 Reduction in PAR

It is evident from Table 1 and Fig 4 that the proposed pricing strategy can achieve a reduction in the PAR of **G**. By supplying gradual and personalized price variations across the day, the agent provides a real grid benefit without causing significant rebound peaks, which is a major improvement compared to TOU pricing.

On the test day, the RL pricing strategy achieved a PAR reduction of 11.64% compared to the other strategies. This is a notable reduction especially considering the fact that the SP does not have direct control over any flexibility resources in the cluster.

5 CONCLUSION AND FUTURE WORK

A mathematical model of a consumer's response to real-time pricing is presented, and a novel contract between a service provider (such as a utility company) and the consumer is proposed. An offline reinforcement learning approach is used to develop agents that generate real-time personalized prices under the conditions in the contract. The resulting prices are able to simultaneously increase the profit of the SP, reduce the bills of the consumer and reduce the PAR of the aggregated load when compared to traditional pricing strategies.

A limitation of the consumer model here is that it only captures load shifting effects, and cannot model flexibility offered by curtailment or sobriety. This work also makes two strong assumptions about the data available to the SP - first, that it is possible to extract the historical demand D_{i,t^-} from consumption and pricing data, and second, that it has perfect knowledge of the consumer's price response function $\varepsilon_i(\pi)$. Future work should focus on the case where this is not true, to provide a more realistic evaluation of the method.

ACKNOWLEDGMENTS

This research is part of the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

REFERENCES

[1] Yen-Chi Chen. 2017. A Tutorial on Kernel Density Estimation and Recent Advances. (4 2017).

[2] Goutam Dutta and Krishnendranath Mitra. 2017. A literature review on dynamic pricing of electricity. , 1131-1145 pages. Issue 10. https://doi.org/10.1057/s41274-016-0149-4

[3] Matt Ewan. 2024. European wholesale electricity price data. https://ember-climate.org/data-catalogue/european-wholesale-electricity-price-data/

[4] Natalia Fabra, David Rapson, Mar Reguant, and Jingyuan Wang. 2021. Estimating the Elasticity to Real-Time Pricing: Evidence from the Spanish Electricity Market. *AEA Papers and Proceedings* 111 (5 2021), 425–429. https://doi.org/10.1257/pandp.20211007

[5] Byung-Gook Kim, Yu Zhang, Mihaela van der Schaar, and Jang-Won Lee. 2016. Dynamic Pricing and Energy Consumption Scheduling With Reinforcement Learning. *IEEE Transactions on Smart Grid* 7 (9 2016), 2187–2198. Issue 5. https://doi.org/10.1109/TSG.2015.2495145

[6] Renzhi Lu and Seung Ho Hong. 2019. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. *Applied Energy* 236 (2 2019), 937–949. https://doi.org/10.1016/j.apenergy.2018.12.061

[7] Oliver Parson, Grant Fisher, April Hersey, Nipun Batra, Jack Kelly, Amarjeet Singh, William Knottenbelt, and Alex Rogers. 2015. Dataport and NILMTK: A building data set designed for non-intrusive load monitoring. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 210–214. https://doi.org/10.1109/GlobalSIP.2015.7418187

[8] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. http://jmlr.org/papers/v22/20-1364.html

[9] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press. http://incompleteideas.net/book/the-book-2nd.html

[10] Mark Towers, Jordan K Terry, Ariel Kwiatkowski, John U. Balis, Gianluca Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Jin Shen Tan, and Omar G. Younis. 2023. Gymnasium. https://doi.org/10.5281/ZENODO.8127025

[11] Georgios Tsaousoglou, Nikolaos Efthymiopoulos, Prodrommos Makris, and Emmanouel Varvarigos. 2019. Personalized real time pricing for efficient and fair demand response in energy cooperatives and highly competitive flexibility markets. *Journal of Modern Power Systems and Clean Energy* 7 (1 2019), 151–162. Issue 1. https://doi.org/10.1007/s40565-018-0426-0

[12] Moslem Uddin, Mohd Fakhizan Romlie, Mohd Faris Abdullah, Syahirah Abd Halim, Ab Halim Abu Bakar, and Tan Chia Kwang. 2018. A review on peak load shaving strategies. *Renewable and Sustainable Energy Reviews* 82 (2 2018), 3323–3332. https://doi.org/10.1016/j.rser.2017.10.056