A Multi-armed Bandit Approach for Electricity Reward of Smart Homes in a Smart Power Grid

Evangelos Spyrou and Vassilios Kappatos and Alkiviadis Tromaras and Afroditi Anagnostopoulou

***Abstract*— Smart homes play a crucial role in the smart grid infrastructure globally, offering substantial environmental and socioeconomic advantages. Often electricity comes from Hydroelectric power plants. Through the implementation of demand response programs by energy providers, these homes enable users to schedule household appliances strategically, leading to efficient energy management, cost savings, and improved reliability and efficiency of the power grid. The paper delves into the energy consumption dynamics of smart homes integrated with a smart grid. It encourages judicious electricity usage, rewarding proper practices with higher incentives and penalizing improper usage with lower rewards. The electricity usage reward, conveyed as a discount through the smart grid, is conceptualized as a multi-armed bandit problem and tackled with the simple epsilon-greedy approach. The findings reveal a sustained preference over the long term for the greater reward linked to a higher discount, leading to a consistent overall increase in rewards.**

I. Introduction

Renewable energy sources like solar, wind, and hydro power are pivotal in slashing carbon emissions. Smart grids, equipped with cutting-edge communication and control features, hold great promise for seamlessly incorporating renewable energy into the power grid. To grasp the workings of a hydroelectric power plant (HPP), it is helpful to first understand the broader principles of power generation. At its core, hydro power generation relies on the conservation of momentum. Here, the potential energy of water in a river dam is transformed into mechanical energy within the turbine rotor blades, which is then converted into electrical energy by the generator. [1]. HPPss can become key players in a power grid by being established as subsystems that connect to the smart grid and provide electricity to smart homes.

However, energy production from HPPs and decision making for the operation of an HPP is not always based in the most intelligent ways but rely on daily drinking water needs, irrigation needs in the area of the plant, cooling water needs for other power plants as well as minimum water flows downstream for ecological purposes. Intercommunication of the HPPs is also fragmented and in many cases plants for the same company are not connected or aware of energy demands in parts outside of their river complex. Building on this concept, digitisation of the hydropower sector is required. Digital tools of this kind aim to offer sustainable energy production by incorporating into decision making, operational and maintenance aspects, flow forecasting, environmental and biodiversity as well as socioeconomic factors.

Smart homes have evolved into integral components of the smart grid in numerous countries, owing to their substantial environmental and socioeconomic advantages. Through the facilitation of scheduling for home appliances based on demand response programs implemented by energy providers, smart homes empower users to optimize energy consumption, thereby reducing costs and bolstering the reliability and efficiency of the power grid. Additionally, smart homes play a pivotal role in curtailing the investments required for generation, transmission, and distribution to meet future electricity demands, primarily by advocating distributed energy generation [2].

The emergence of smart homes results from the convergence of state-of-the-art information and communication technologies, including smart sensors, advanced metering infrastructures, intelligent home appliances, and Internet-of-Things (IoT) devices [3]. This progressive trend has paved the way for the deployment of Home Energy Management Systems (HEMSs), charting the course toward the realization of future smart grids.

In recent years, HEMSs [4] have garnered global acceptance, emerging as indispensable tools for efficiently managing electricity demand within the smart grid. A growing body of global HEMS research focuses on enhancing energy efficiency, bolstering security, and minimizing electricity costs in both residential and commercial power systems. However, this study reveal persistent challenges in HEMSs, particularly concerning control and communication technologies, crucial components of their functionality.

Key issues include the effective integration of power electronic converters, renewable energy sources, and energy storage into HEMSs. Present HEMS research emphasizes theoretical design over implementation and operational concerns, creating an imbalance that warrants correction. Addressing this disparity is crucial, as real-world applications of HEMSs play a pivotal role in validating their designs and tackling deployment challenges.

The successful implementation of HEMSs hinges on the convergence of sensing, communication, and control technologies. These elements enable access to energy demand data and the dispatch of control strategies through the network promptly. Communication networks in smart grid applications are categorized based on coverage scale: Home Area Networks (HANs), Neighborhood Area Networks (NANs), and Wide Area Networks (WANs) [5]. A standard HAN encompasses a smart electricity meter interconnecting var-

Evangelos Spyrou, Vassilios Kappatos, Alkiviadis Tromaras and Afroditi Anagnostopoulou are with the Hellenic Institute of Transport, Centre for Research and Technology Hellas, Greece {espyrou, vkappatos, atromaras, a.anagnostopoulou }certh.gr

ious home devices, sensors, displays, gas and water meters, renewable energy sources, and electric vehicles.

A HEMS oversees and regulates the consumption, storage, and generation of power for various components, such as devices, sensors, displays, gas and water meters, renewable energy sources, and electric vehicles [6]. The central controller of the HAN is linked to the utility grid through its smart meter. Information from several HANs is consolidated and stored in a database, which then shapes either a Neighborhood Area Network (NAN) or a Wide Area Network (WAN), depending on the coverage scope. The compiled data from multiple NANs/WANs are transmitted to the utility administrator, aiding in decision-making regarding various system parameters, including pricing and anticipated load [7].

In this paper, we take the energy/electricity expenditure onboard of smart houses whereby the house is connected to a smart grid with electricity coming from HPPs. The appropriate use of the electricity coming from different parts of the smart house gets a high reward, while an improper use gets lower rewards. The reward of the proper use of electricity offers a discount by the smart grid. We formulate the problem as a simple multi-armed bandit problem and we solve it using the epsilon-greedy-approach. We show that in the long term, the higher reward that comes with the higher discount is selected and that the reward is increasing.

The remainder of this paper is as follows: section II provides the related work, section III gives the background of the multi-armed bandits framework, section IV provides information regarding the epsilon-greedy approach, section IV gives the system model, section VI gives the results, section VII presents the conclusions and section VIII gives the future work.

II. Related Work

There exist works that have been considering smart grid using reinforcement learning methods that appear in []. These include industrial scenarios [8] as well as residential settings [9]. A thorough survey on Deep Reinforcement Learning is given in [10] and a DRL with it individual components as approaches is given in [11].

In [12], the study introduces RSOTHA-QL, a Real-time Scheduling of Operational Time for Household Appliances using Quality Learning—a value iterative reinforcement learning method. The framework comprises Q learning agents engage with the smart home environment to schedule appliance operational times based on received rewards, ensuring minimal energy consumption. Dissatisfaction resulting from appliance scheduling is addressed by categorizing appliances as referable, non-referable, and controllable. Agents coordinate through shared memory synchronization. Simulations and experiments in a smart home setting illustrate that RSOTHA-QL, in comparison to previous research utilizing Least Slack Time and demand-response strategy scheduling, effectively minimizes both energy consumption and user dissatisfaction in scheduling household appliances' operational times.

The paper [13] introduces an innovative hierarchical deep reinforcement learning (DRL) approach designed to optimize energy consumption in smart homes, encompassing a variety of components such as appliances and distributed energy resources (DERs) like an energy storage system (ESS) and an electric vehicle (EV). Diverging from Q-learning algorithms that function in discrete action spaces, the proposed method schedules energy usage within a continuous action space, employing an actor–critic-based DRL framework. The two-tier DRL structure manages the scheduling of home appliances at the first level based on consumer preferences, while the second level computes ESS and EV charging/discharging schedules by considering optimal solutions from the first level and accounting for consumer environmental factors. Through simulation studies conducted in a single home setting, including appliances like an air conditioner, washing machine, solar photovoltaic system, ESS, and EV under time-of-use pricing, the paper establishes the effectiveness of the proposed method. Numerical examples, spanning various weather conditions, weekdays/weekends, and EV driving patterns, further validate its success in optimizing total electricity cost, ESS and EV state of energy, and meeting consumer preferences.

In [14], a robust residential energy management system for demand response, employing a synergy of Reinforcement Learning (RL) and Fuzzy Reasoning (FR) is presented. RL, functioning as a model-free control strategy, learns through environmental interactions by executing actions and evaluating outcomes. The proposed algorithm seamlessly integrates user preferences into its control logic using fuzzy reasoning as reward functions. Q-learning, an RL strategy, optimizes the scheduling of smart home appliances by shifting controllable devices from peak to off-peak hours, aligning with lower electricity prices and customer preferences. The method employs a single agent overseeing 14 household appliances, utilizing a streamlined set of state-action pairs and employing fuzzy logic for reward functions. Simulation results attest to the efficacy of the proposed scheduling method in smoothing power consumption profiles, reducing electricity costs, and accommodating user preferences. To illustrate the demand response scheme, a user-friendly interface for the Home Energy Management System (HEMS) is developed. This interface incorporates various features such as smart appliances, electricity pricing signals, smart meters, solar photovoltaic generation, battery energy storage, electric vehicles, and grid supply.

The paper [15] presents a novel data-driven approach that utilizes reinforcement learning to optimize energy consumption in a smart home featuring a rooftop solar photovoltaic system, an energy storage system, and smart appliances. Differing from conventional model-based optimization methods for home energy management, the proposed approach incorporates two distinct elements: the employment of a model-free Q-learning method to schedule energy consumption for individual controllable home appliances (such as air conditioners or washing machines) and manage the energy storage system's charging and discharging, and the integration of an

artificial neural network for indoor temperature prediction, enhancing the Q-learning algorithm's understanding of the relationship between indoor temperature and air conditioner energy consumption. The integrated Q-learning home energy management algorithm, along with the artificial neural network model, successfully reduces the consumer's electricity bill while adhering to preferred comfort levels, such as indoor temperature and appliance operation characteristics. Simulations focus on a single home with a solar photovoltaic system, air conditioner, washing machine, and energy storage system under time-of-use pricing, demonstrating a reduction in the relative electricity bill.

In [16], the article introduces the innovative Reward Shaping-based Actor–Critic Deep Reinforcement Learning (RS-ACDRL) algorithm, specifically designed for the effective management of residential energy consumption profiles when faced with limited information about uncertain factors. The intricate dynamics between the energy management center and residential loads are meticulously modeled as a Markov decision process, establishing a robust mathematical framework for decision-making in scenarios with partially random outcomes influenced by control signals. The RS-ACDRL algorithm, seamlessly integrating actor and critic networks along with a sophisticated reward shaping mechanism, is meticulously crafted to deduce optimal energy consumption schedules. Through real-world case studies that incorporate diverse datasets, the article validates the algorithm's supremacy over state-of-the-art RL methods, showcasing notable improvements in learning speed, solution optimality, and cost reduction.

III. Multi-armed Bandits

Multi-arm bandits [17] represent a core challenge in the realm of reinforcement learning. This problem entails managing multiple slot machines, or arms, each featuring distinct reward distributions. The goal is to identify the machine with the highest expected reward through sequential plays. This issue arises frequently in practical scenarios like advertising [18], healthcare [19], finance [20] among others [21], where decision-makers must select the optimal choice from a range of alternatives.

Two primary categories of multi-arm bandits exist: stochastic and adversarial. Stochastic multi-arm bandits involve arm rewards generated from a fixed, undisclosed probability distribution. Conversely, in adversarial multi-arm bandits, an adversary selects arm rewards, aiming to present a more challenging scenario for the agent. This adversary can adopt either a deterministic or randomized approach.

Numerous algorithms address the multi-arm bandit problem, each possessing distinct merits and drawbacks. Among the most renowned algorithms are epsilon-greedy, Upper Confidence Bound (UCB), and Thompson Sampling.

Epsilon-Greedy: This straightforward algorithm opts for the arm with the highest estimated reward with a probability of 1-epsilon, and it randomly selects an arm with a probability of epsilon. The choice of epsilon aims to strike a balance between exploration and exploitation. A setting of epsilon to zero leads the algorithm to consistently choose the arm with the highest estimated reward, potentially resulting in suboptimal solutions if the estimates are imprecise. Conversely, setting epsilon to one causes the algorithm to consistently choose a random arm, potentially resulting in inefficient exploration. This algorithm is going to be detailed further since it is going to be the one used.

UCB: The Upper Confidence Bound (UCB) algorithm seeks equilibrium between exploration and exploitation by selecting the arm with the highest UCB estimate. This estimate comprises two components: the estimated reward and the confidence interval. The confidence interval is proportional to the square root of the logarithm of the number of times the arm has been played. UCB has demonstrated effective performance in both stochastic and adversarial environments.

Thompson Sampling: Operating on Bayesian principles, Thompson Sampling updates its beliefs regarding the reward distribution of each arm after each play. Subsequently, the algorithm samples a reward from the updated distribution and chooses the arm with the highest sample. While Thompson Sampling has proven effective in stochastic environments, its performance in adversarial settings requires further understanding.

IV. Epsilon-Greedy Algorithm

Reading in [22], in essence, the epsilon-greedy agent combines features of both a fully exploratory agent and a completely greedy agent. In the context of the multi-armed bandit problem, a fully exploratory agent systematically samples all bandits at an equal rate, gradually accumulating knowledge about each bandit. However, this knowledge is not leveraged to make improved future decisions. Conversely, a completely greedy agent steadfastly selects a bandit and adheres to that choice indefinitely, neglecting the opportunity to test other bandits for potentially higher success rates and limiting its long-term rewards.

To strike a balance and create an agent with advantages from both approaches, the epsilon-greedy agent allocates an epsilon probability (e.g., 10%) for random exploration of bandits at any given state. For the remainder of the time, it acts greedily based on its current best estimate of bandit values. The rationale behind this approach is that the greedy mechanism enables the agent to concentrate on its presently most successful bandits, while the exploratory mechanism allows for the exploration of potentially superior bandits.

The remaining consideration is how to establish a concept of value for a bandit that allows the agent to make greedy choices. Drawing inspiration from reinforcement learning, we can introduce the action-value function $Q(s, a)$ to denote the anticipated long-term reward associated with taking action α from state s. In the context of the multi-armed bandit, where each action leads the agent to a terminal state and long-term rewards to immediate rewards, we streamline the notation for defining the action-value as

$$Q_k(\alpha) = \frac{1}{k}(r_1 + r_2 + ... + r_k) \tag{1}$$

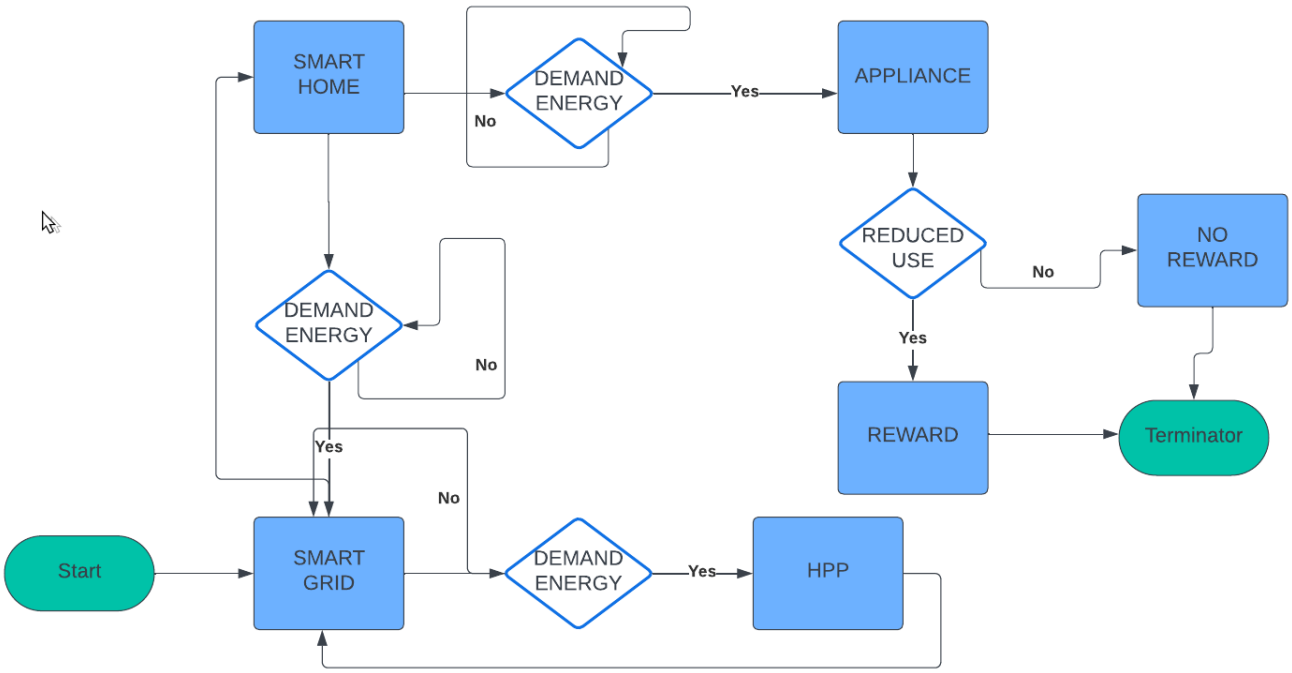

Fig. 1: Flowchart of Approach

where, k represents the count of how often action α (bandit) has been selected in the past, and r denotes the stochastic rewards for each instance the bandit was chosen. Through additional arithmetic manipulations, this definition can be reformulated in a recursive manner as

$$Q_{k+1}(\alpha) = Q_k(\alpha) + \frac{1}{k+1}(r_{k+1} - Q_k(\alpha)) \tag{2}$$

Given our initial lack of knowledge about the true values of $Q(\alpha)$, we can employ this recursive definition in an iterative fashion to approximate $Q(\alpha)$ at the conclusion of each episode.

To align the epsilon-greedy agent with our estimated action-values $Q(\alpha)$, we allow the epsilon-greedy agent to randomly select a bandit with an epsilon probability and, for the remaining instances, to choose an action greedily from our $Q(\alpha)$ estimates.

$$\alpha_{greedy} = \arg\max_{a} Q_k(\alpha) \tag{3}$$

V. Our System Model

We assume that we have a smart energy meter that is connected to the house and it measures the primary energy/electricity sources of the appliances the Heating, Ventilation and Air-Conditioning (HVAC), and the lights. Noteworthy, we are considering a smart house whereby the electricity sources are controlled and they can be tuned to spend more or less electricity. The house is connected to the utility provided via a smart grid. We can see a flowchart of our approach in Fig. 1.

The energy sources are denoted as P_a, P_{HVAC} and P_l respectively. The meter is connected to a smart grid that provides a reward when the energy sources are used in an appropriate manner by the residents. The reward is considered to be a rewards r_i of the proper use of the energy. The rewards are a discrete set of rewards $\{r_1, r_2, ..., r_i\}$ that the smart grid offers according to the use of energy. Each reward is essentially a percentage and comes with the respective discount coming from the smart grid.

The smart grid subsystems of the power grid are separated by the nature of the electricity source they use. In Fig. 2 we show different sources of renewable energy coming from different sources; however, we emphasise in the use of HPPs for the provision of electricity to the smart grid. In its turn the smart grid provides electricity to the smart house appliances as described earlier. To optimize both the electricity demand within the smart house and its supply, encompassing battery storage, self-generation, and engagement with the energy market, a comprehensive approach is imperative. This holistic management leads to the development of a smart grid solution tailored for industrial sites, with the energy coming from HPPs or renewable sources of energy.

We take a number of n predetermined bandits, each possessing distinct reward, bearing in mind that our agent remains unaware of these reward and can only discern them through individual bandit sampling. By inspection, we will be requiring our agent to pick the energy reward with the biggest value which corresponds to a specific energy plan.

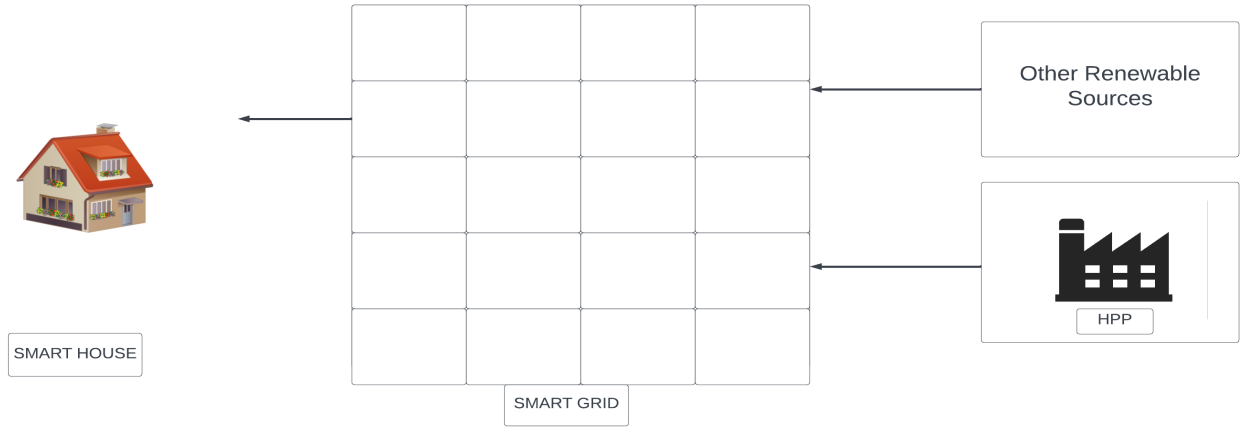

Fig. 2: HPP to smart grid to smart house

As such the smart grid will provide a service to the residents of the smart home by providing lesser expenditure in the bill.

Note that we can extend the energy sources with electric vehicle charging station and other sources, as well as have renewable sources of energy in our model. However, we will keep the work simple and leave the aforementioned for future work.

VI. Results

In a series of 1000 simulated experiments, the agent initiated its learning process with a 10% epsilon exploration probability, undergoing training for $10,000$ episodes per experiment. Here by arms, (we selected 8 arms as a sufficient number) we denote the rewards obtained by using the electricity in different manners. We provide the reader with the percentages in rewards in table I.

TABLE I: Arms and respective rewards

ARM	Reward
Arm 1	10 %
Arm 2	50 %
Arm 3	60 %
Arm 4	55 %
Arm 5	80 %
Arm 6	75 %
Arm 7	60 %
Arm 8	65 %

In Fig. 3, the initial bandit selection demonstrates a uniform distribution of approximately 10% across all bandits during the early training episodes (fewer than 10 episodes). This signifies the agent's exploratory phase, where it lacks knowledge about which bandits yield optimal rewards. As training advances, particularly beyond 100 episodes, a noticeable shift towards a greedy mechanism emerges, favoring Bandits 5, 6, and 8 based on sampled rewards. Towards the end of training, the agent consistently leans towards Bandit 5, while the remaining percentage is attributed to the fixed epsilon exploration parameter.

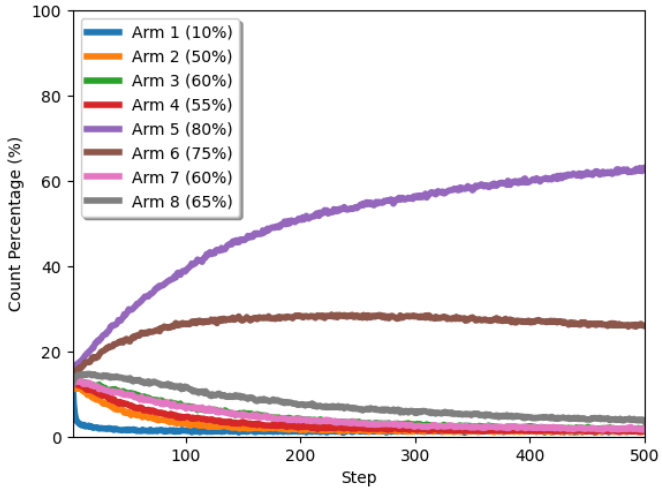

Fig. 3: Bandit actions taken over 500 steps

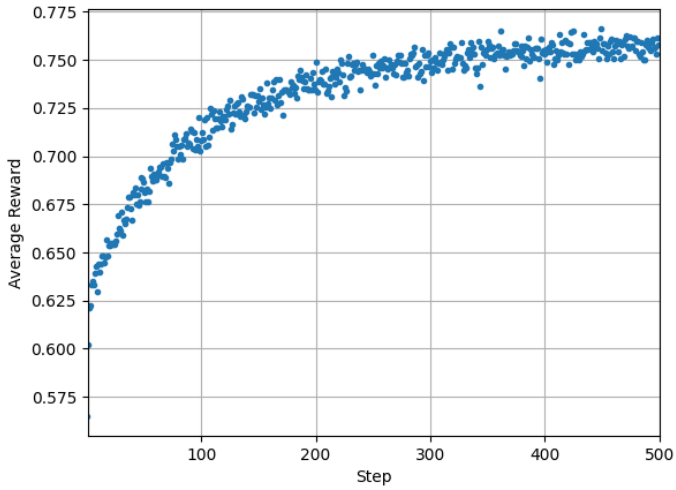

Fig. 4: Bandits average reward over 500 steps of our model

Moreover, Fig. 4 shows the average reward of the bandits for the given steps. As we can see the the average reward is

increasing.

While the optimal policy suggests selecting Bandit 5, it does not guarantee superiority in every trial due to stochastic rewards. However, over the long term, Bandit 5 tends to outperform others in average rewards. The agent's methodology represents just one among several approaches for adaptively maximizing long-term rewards. Different strategies, including fully exploratory or fully greedy agents, might surpass the epsilon=10%-greedy agent in specific scenarios. The deployment of such an agent proves appealing for automating the avoidance of re-selecting bandits that exhibited evidence of failure, potentially leading to time and resource savings in the quest to identify the best bandit.

VII. Conclusions

This paper explores the fact that ptimal utilization of electricity from various smart house components is rewarded with higher incentives, whereas inappropriate usage results in lower rewards. The correct electricity usage reward is provided in the form of a discount facilitated by the smart grid. The problem is presented as a multi-armed bandit problem and employ the epsilon-greedy approach to address it. Our findings demonstrate that over the long term, the preference shifts toward the greater reward associated with a higher discount, leading to an increasing overall reward by selecting the 80 % arm more frequently.

VIII. Future Work

For future work we aim to include renewable sources to the smart house to further optimise the utility and suggest a more sophisticated electricity expenditure plan. Moreover, we will add more resources of energy dissipation such as electric vehicles. Moreover, we will include thermal comfort as a variable of the use of energy to make sure that the residents get the reward not only from less energy use, but with ensuring that they feel comfortable as well [23]. Furthermore, we will use the work of [24] to further extend this research.

Acknowledgement

The results incorporated in this paper received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement N° 101021701, project title Di-Hydro: Digital maintenance for sustainable and flexible operation of HYDROpower plant.

References

[1] M. Sartor, L. Souza, A. Júnior, H. Rebelo, K. Cotta, L. Vianna, R. Pereira, and M. Morais, "Assets performance management systems for hydroelectric power plants—a survey," *Electric Power Systems Research*, vol. 228, p. 110080, 2024.

[2] A. Elrayyah and S. Bayhan, "Multi-channel-based microgrid for reliable operation and load sharing," *Energies*, vol. 12, no. 11, p. 2070, 2019.

[3] D. Mocrii, Y. Chen, and P. Musilek, "Iot-based smart homes: A review of system architecture, software, communications, privacy and security," *Internet of Things*, vol. 1, pp. 81–98, 2018.

[4] B. Zhou, W. Li, K. W. Chan, Y. Cao, Y. Kuang, X. Liu, and X. Wang, "Smart home energy management systems: Concept, configurations, and scheduling strategies," *Renewable and Sustainable Energy Reviews*, vol. 61, pp. 30–40, 2016.

[5] I. Serban, S. Cespedes, C. Marinescu, C. A. Azurdia-Meza, J. S. Gomez, and D. S. Hueichapan, "Communication requirements in microgrids: A practical survey," *IEEE Access*, vol. 8, pp. 47 694–47 712, 2020.

[6] T. Molla, B. Khan, B. Moges, H. H. Alhelou, R. Zamani, and P. Siano, "Integrated optimization of smart home appliances with cost-effective energy management system," *CSEE Journal of Power and Energy Systems*, vol. 5, no. 2, pp. 249–258, 2019.

[7] U. Zafar, S. Bayhan, and A. Sanfilippo, "Home energy management system concepts, configurations, and technologies for the smart grid," *IEEE access*, vol. 8, pp. 119 271–119 286, 2020.

[8] M. Roesch, C. Linder, R. Zimmermann, A. Rudolf, A. Hohmann, and G. Reinhart, "Smart grid for industry using multi-agent reinforcement learning," *Applied Sciences*, vol. 10, no. 19, p. 6900, 2020.

[9] T. Remani, E. Jasmin, and T. I. Ahamed, "Residential load scheduling with renewable generation in the smart grid: A reinforcement learning approach," *IEEE Systems Journal*, vol. 13, no. 3, pp. 3283–3294, 2018.

[10] Y. Li, C. Yu, M. Shahidehpour, T. Yang, Z. Zeng, and T. Chai, "Deep reinforcement learning for smart grid operations: Algorithms, applications, and prospects," *Proceedings of the IEEE*, 2023.

[11] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362–370, 2018.

[12] M. Khan, J. Seo, and D. Kim, "Real-time scheduling of operational time for smart home appliances based on reinforcement learning," *IEEE Access*, vol. 8, pp. 116 520–116 534, 2020.

[13] S. Lee and D.-H. Choi, "Energy management of smart home with home appliances, energy storage system and electric vehicle: A hierarchical deep reinforcement learning approach," *Sensors*, vol. 20, no. 7, p. 2157, 2020.

[14] F. Alfaverh, M. Denai, and Y. Sun, "Demand response strategy based on reinforcement learning and fuzzy reasoning for home energy management," *IEEE access*, vol. 8, pp. 39 310–39 321, 2020.

[15] S. Lee and D.-H. Choi, "Reinforcement learning-based energy management of smart home with rooftop solar photovoltaic system, energy storage system, and home appliances," *Sensors*, vol. 19, no. 18, p. 3937, 2019.

[16] R. Lu, Z. Jiang, H. Wu, Y. Ding, D. Wang, and H.-T. Zhang, "Reward shaping-based actor–critic deep reinforcement learning for residential energy management," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 3, pp. 2662–2673, 2022.

[17] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.

[18] V. Avadhanula, R. Colini Baldeschi, S. Leonardi, K. A. Sankararaman, and O. Schrijvers, "Stochastic bandits for multi-platform budget optimization in online advertising," in *Proceedings of the Web Conference 2021*, 2021, pp. 2805–2817.

[19] T. Zhou, Y. Wang, L. Yan, and Y. Tan, "Spoiled for choice? personalized recommendation for healthcare decisions: A multiarmed bandit approach," *Information Systems Research*, 2023.

[20] M. Zhu, X. Zheng, Y. Wang, Y. Li, and Q. Liang, "Adaptive portfolio by solving multi-armed bandit via thompson sampling," *arXiv preprint arXiv:1911.05309*, 2019.

[21] D. Bouneffouf, I. Rish, and C. Aggarwal, "Survey on applications of multi-armed and contextual bandits," in *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2020, pp. 1–8.

[22] A. Wong, "Solving the multi-armed bandit problem." [Online]. Available: https://towardsdatascience.com/solving-the-multi-armed-bandit-problem-b72de40db97c

[23] E. D. Spyrou and V. Kappatos, "Energy-efficient thermal comfort optimization game in office building networks," in *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 2023, pp. 1–5.

[24] T. T. Tsenis, G. Kapsimanis, and V. Kappatos, "Smartclima: reinforcement learning residential thermostat-less heating control system," in *2021 International conference on electrical, computer, communications and mechatronics engineering (ICECCME)*. IEEE, 2021, pp. 1–6.