Decentralized multi-agent multi-armed bandits for smart electric vehicles charging

Sharyal Zafar, Raphaël Féraud, Anne Blavette, Guy Camilleri, Hamid Ben Ahmed

Decentralized Multi-Agent Multi-Armed Bandits for Smart Electric Vehicles Charging

Abstract

Smart charging of electrical vehicles can help in avoiding congestion and peak load demands in an electrical distribution network. On the consumer's side, the advantage lies in minimizing the daily charging cost. He may also benefit from cheap photovoltaic electricity from local sources and therefore reduce his environmental impact. However, this cheaper electricity is variable and uncertain. In many research works, this has been formulated and solved as a centralized or hierarchical optimization problem. However, such systems may suffer from lack of scalability, single point of failures, and privacy breaches. We propose a fully decentralized and fair multi-agent system combined with reinforcement learning called "Decentralised multi-armed bandit (2-armed bandit) based on Thompson sampling"(D-MAB2AB-TS) to control the charging of electrical vehicles under uncertainties. The problem under consideration is formulated as a two-armed bandit (charging or not) for each instant. The proposed algorithm, based on Thompson Sampling, takes into account the uncertainties in the choice of arms combination of other players. The proposed algorithm finds the best combination of arms to play with a computational complexity $O(m)$ linear with the number of arms. The suggested system is also model-free, as it does not assume the model of the environment to be perfectly known, which is a common assumption in many of the existing centralized optimization strategies for smart charging.

Keywords:
Multi-Armed Bandits, Decentralized Smart Charging, Thompson sampling

1. Introduction

In recent years, a drastic increase in the adoption of electric vehicles (EVs) and photovoltaics (PVs) has been observed. This massive deployment of EVs and PVs can help to achieve targets in reducing greenhouse gas emissions. On the other hand, it may introduce new challenges such as network congestion and increased peak load demand at certain hours of the day on the existing electrical networks, which were not designed for this purpose. Also, the PV production is intermittent and uncertain [1]. These challenges could compromise the stability of the networks and the supply of electricity to the consumers. Grid reinforcement solutions can help tackle these challenges, but they may

come with higher network investment costs. A popular alternative is to control the charging of these EVs in real-time. This requires the synergy of different electricity market actors at different levels as the constraints of each actor must be satisfied. The distribution system operator (DSO) is responsible for the stability of the electrical distribution network, which is a global constraint. Also, each consumer would want its EV to be sufficiently charged at the time of its departure, which is a local constraint. Moreover, each of the consumers has an objective to minimize the total cost of charging the EV by exploiting dynamic electricity pricing during the day.

A wide range of solutions has been proposed in the literature to solve this optimization problem and can be classified as centralized, hierarchical, or decentralized [2]. The relatively simpler approach is to design a single node that gathers the information from all consumers, performs centralized optimization with this data, and sets the charging power of each consumer's EV according to the optimization results [3]. This approach has been used in [4, 5, 6] for smart charging of the EVs. Such systems are relatively easy to design, but this optimization approach is not scalable with large-scale electrical networks, not to mention their incompatibility with their real-time operation. Also, this approach raises concerns regarding privacy as a single operator may require to know private information such as arrival time, energy demand, etc. . These challenges are tackled in [7, 8, 9, 10, 11] by incorporating the philosophy of multi-agents in the system design to perform smart grid operations. These solutions are hierarchical in nature and tackle the challenge of scalability by dividing the optimization problem under study into smaller sub-optimization problems [12]. These systems may control a large-scale electrical network in real-time, but they can still suffer from other challenges such as a potential single point of failure in the system, and more importantly, concerns related to the data privacy of each consumer [13, 14]. Furthermore, in most such cases, an accurate model of the environment is assumed to be known, which is not always the case for electrical distribution networks. In summary, centralized and hierarchical approaches are unsuitable to solve large-scale problems such as on the one considered in this paper, and which focuses on the smart charging on large-scale EV fleets under grid constraints. However, decentralizing the system is a way to remove the mentioned drawbacks. A decentralized smart charging system can be scalable, and operate in real-time. It can also eliminate the possibility of a single point of failure, and may render privacy breaches more difficult [15].

In the past decades, a number of decentralized algorithms have emerged based on deep learning and reinforcement learning (RL). Although deep learning approaches usually proves superior performance for prediction tasks, it is notoriously less efficient than reinforcement learning for decision-making tasks under uncertainty [16, 17], especially when the situations faced by the system have not all been included in the training set. In the particular case of decision-making under uncertainty, reinforcement learning is therefore the tool of choice (potentially combined with deep learning to give deep reinforcement learning). In recent years, decentralized algorithms applied to the management of electrical distribution networks using reinforcement learning (RL) have been

proposed [18, 19, 20, 21, 22, 23]. In reinforcement learning, an agent interacts with an environment and learns to make decisions under uncertain scenarios. The agent makes an action according to the present state of the environment. This action changes the state of the environment, and the agent receives a reward from the environment. The goal of the agent is to maximize the running total of this observed reward [24].

Many research works have considered reinforcement learning for optimizing the energy management in power systems [25, 26, 27]. For instance, a decentralized multi-agent system is presented in [25] for the coordination of battery energy storage systems. The system in [25] uses Q-learning with neural networks as function approximators (Deep Q-Network (DQN) learning). While such RL algorithms may work well to optimize the studied cost function, the inclusion of multi-level (global and local) constraints can still be a difficult task, as unlike computer games, in most practical smart grid applications [28], there is no known oracle that can evaluate the performance of an action. In general, modeling a reward function that would direct the agent's policy towards a constraint-satisfying optimal policy is a difficult task, particularly in complex smart grid applications [28]. If the agent is learning through online interactions, then the total cost of the agent is directly linked to the number of samples required to find the constraint-satisfying optimal policy. Also, recent studies have shown that poor design choices can lead to non-optimal representation learning [29, 30]. Approaches such as Q-learning with function approximation can indeed suffer from over/under-estimation, instability, and even divergence due to the delusional bias in learning [31]. Surveys [26, 27] highlighted the vast number of studies performed on smart grid applications, and in particular EV charging, based on reinforcement learning (Q-learning, deep reinforcement learning, batched reinforcement learning, W-learning, State-Action-Reward-State-Action (SARSA), etc.). However, it is shown that multi-agent reinforcement learning methods based on Markov-decision processes (which are almost exclusively considered in the previously mentioned works) become intractable on large instances [27, 32]. Hence, these methods are not suitable for the energy management of large-scale EV fleets.

Unlike these previous approaches presenting obstacles rendering them unsuitable for solving our problem, there exists a subclass of reinforcement learning, called multi-armed bandits (MAB), where an agent tries to tackle the exploration-exploitation dilemma by finding the optimal arm (action offering the highest reward) through exploration [33] and that presents an interesting scalabilty potential [34]. Its key difference from standard RL is that the agent receives a reward based on the selected action only and not on the past sequence of actions. Also, there are no state transitions in multi-armed bandits problems. This simplicity allows multi-armed bandits to converge faster to the optimal solution (compared to DQN or Q-learning), when the considered problem can be formulated as a bandit problem. This constitutes a significant advantage in the context of pure online learning for smart grid applications, i.e. where an oracle that outputs the outcome of an action is not always available.

Bandits have been widely studied for communication management in mod-

ern communication networks, which presents similarities with the problem of energy management in smart grids [35, 36, 37, 38, 34]. Bandits have also found applications in the healthcare and e-commerce sectors [39, 40]. In [41, 42], smart charging strategies using multi-armed bandits have also been suggested. However, the architecture of both systems is not completely decentralized. In [35, 36], the authors have proposed multi-armed bandits in a fully decentralized approach for optimizing communication in Internet of Things (IoT) networks. A bandit algorithm is run on each agent for choosing the best channel. The authors report excellent results in practice [35, 36, 37, 38]. More specifically, the Thompson sampling algorithm has been shown to be asymptotically optimal under a stationary setting, and to work well in practice [43, 44].

In summary, multi-armed bandits based on Thompson sampling seem relevant for addressing the problem considered here. In this paper, we use and extend this approach in the context of smart charging [45]. Each EV is an agent that selfishly tries to optimize its charging during each day. In order to plan the charging during the day, the agent selects the time slots where it is going to charge. We model this selection of time slots as several independent bandit problems, with one bandit tackling a single problem associated to one time slot. Hence, for each day, a combination of arms is played. Each arm is selected based on its estimated reward and updated based on the reward actually obtained from the environment. The objective of each agent is to find the best combination of arms that maximizes its cumulative daily reward while satisfying the constraints.

1.1. Contributions

To the best of our knowledge, this paper is the first to propose a decentralized multi-agent smart charging system that uses the concepts of multi-agent multi-armed bandits based on Thompson sampling. The main contributions of this paper are summarized as follows:

1. A decentralized multi-agent system using 2-armed bandits with Thompson Sampling is proposed to control the charging of electrical vehicles and called "Decentralised multi-armed bandit (2-armed bandit) based on Thompson sampling" (D-MAB2AB-TS).
2. A comparative analysis with algorithms "Exponential Weights for Exploration and Exploitation" (EXP3) and "Upper Confidence Bound" (UCB) is performed.

The presented algorithm is decentralized as well as scalable. It is indeed shown that each agent can calculate the current best estimate of the combination of arms to play in a computational complexity $O(m)$, where m is the number of base arms. The proposed algorithm takes into account the uncertainties in the choice of super arms of other players and in local and cheap PV electricity production. Also, the system operates in real-time and ensures fairness among all participating agents. The proposed methodology is meant to be generic and so, the presented system is adaptable, i.e., it can also be used to control other elements in electrical distribution networks, such as household appliances.

It is shown through experimental evaluations that the proposed decentralized multi-agent system outperforms the basic charging strategy of electrical vehicles and produces near-optimal solutions under the mentioned uncertainties.

2. Preliminaries

Based on the observations detailed in the Introduction section, multi-armed bandits were selected for solving the problem presented in this paper. In particular, the Thompson sampling algorithm was selected. This section provides brief explanations of these concepts.

2.1. Multi-Armed Bandits

The combinatorial problem associated with smart charging consists in choosing K instants in a set of m instants. In this paper, we selected independent bandits for each instant to address this combinatorial problem, at a first stage, for the sake of computing speed. Each arm $i \in [m]$ is associated with a random variable $X_{i,t}$ for $1 \leq i \leq m$ and $t \geq 1$. Variable $X_{i,t}$ represents the random outcome of the i-th arm in its t-th trial. This random variable $X_{i,t}$ is independent and identically distributed according to some unknown distribution D_i with unknown expectation μ_i.

In our formulation of the problem, each EV agent selects a set of arms $S_t \in \mathcal{S}$, where $\mathcal{S}$ is the set of all possible arms, and which represents the charging instants. During each round t of the bandit problem, the agent selects a set of arms $S_t \in \mathcal{S}$. The environment provides feedback (rewards) to the agent based on the selected arms i in the arms combination S_t.

The feedback obtained by the agent from the environment can be of different types: *full information feedback*, when the agent observes the outcomes of all m arms; *bandit feedback*, when the agent observes an aggregated reward based on the played arms combination S_t; and *semi-bandit feedback*, when the player observes the rewards of the played $i \in S_t$ base arms only. In our proposed system, semi-bandit feedback is considered which is a realistic assumption: a consumer can only observe the reward of its played actions. Then, the t-th round observed reward $r(S_t)$ is a function of the played arms combination S_t and the observed feedback $X_t = \{X_{i,t} \mid i \in S_t\}$. This reward $r(S_t)$ can simply be the sum of the played base arms' feedback, $\sum_{i \in S_t} X_{i,t}$, or other non-linear reward functions can be modeled depending on the multi-armed bandits problem. The goal of the agent is to minimize its pseudo-regret over T rounds as shown in Equation (1):

$$R(T) = T\mathbb{E}[r(S^*)] - \sum_{t=1}^{T} \mathbb{E}\left[r(S_t)\right] \tag{1}$$

where the optimal arms combination is defined as $S^* = \arg\max_{S \in \mathcal{S}} \mathbb{E}[r(S)]$.

2.2. Thompson Sampling

Thompson Sampling (TS) is a natural randomized Bayesian algorithm to tackle the exploration-exploitation dilemma in multi-armed bandits problems. In each round, the algorithm selects an arm a out of a set of available arms $\mathcal{A}$ to play (based on its probability of being the optimal arm), and observes its corresponding reward r. A parametric likelihood function $P(r|a,\theta)$, parameterized by an unknown vector θ, is used to model the history $\mathcal{H}$ of past observations (a_i, r_i). The uncertainty about θ is expressed by a *prior* distribution $P(\theta)$. An initial belief is updated by calculating the *posterior* distribution using the Bayes rule, $P(\theta|\mathcal{H}) \propto P(\mathcal{H}|\theta)P(\theta)$.

In contrast to frequentist learning, in Thompson Sampling, the player forms its beliefs in each round by sampling $\tilde{\theta}$ from the posterior distribution $P(\theta|\mathcal{H})$ of the previous round and chooses an action $\tilde{a}$ that maximizes $\mathbb{E}\left[r|\tilde{a},\tilde{\theta}\right]$. The action $\tilde{a}$ is selected with the probability shown in Equation (2):

$$\int_{\theta} \mathbb{I}\left[\mathbb{E}[r|\tilde{a},\theta] = \max_{a'} \mathbb{E}[r|a',\theta]\right] P(\theta|\mathcal{H})\, d\theta \tag{2}$$

where $\mathbb{I}$ is the indicator function. A visual representation of Thompson sampling with Gaussian priors (parameterized by θ) is shown in Figure 1. The player has to find the best action among the 3 available actions, i.e., $\mathcal{A} = \{1,2,3\}$. Term n_i for $1 \leq i \leq 3$, represents the number of times i-th action has been selected. The player starts (round 0 in Figure 1) with a high variance in its beliefs (which motivates the agent to do exploration), as $n_i = 0 \;\forall\; 1 \leq i \leq 3$. This variance decreases as the number of observations increases for each action. Based on the observed reward r for the selected action a, the player forms the Gaussian posterior by Bayes rule shown in Equation (3) as:

$$P(\theta|r) \propto e^{-\frac{(n+2)}{2}(\theta-\hat{\mu}_{n+1})^2} \tag{3}$$

where n is the number of times an action has been played and $\mu_{n+1} = \frac{n\mu_n + r}{n+1}$ is the empirical average of $n+1$ samples. The right-hand side in Equation 3 is proportional to the probability density function of a normal distribution, $\mathcal{N}(\hat{\mu}_{n+1}, \frac{1}{n+2})$. After 100 observations in the given example, the expected return of each arm is better known (low variance in the estimations). Following this, the agent can exploit this information by picking the action with the highest expected reward value, to minimize its regret. Although Thompson sampling with Gaussian priors is described here, other types of priors can also be used depending on the studied problem [46].

In the multi-armed bandits setting with semi-bandit feedback, the agent holds a set of prior beliefs parameterized by θ for each base arm, $[P] = \{P(\theta_i)|i \in [m]\}$. The agent selects the arms combinations S_t that maximizes $\mathbb{E}[r(S_t)]$. The posterior for each played base arm is calculated using the Bayes rule, based on the observed feedback.

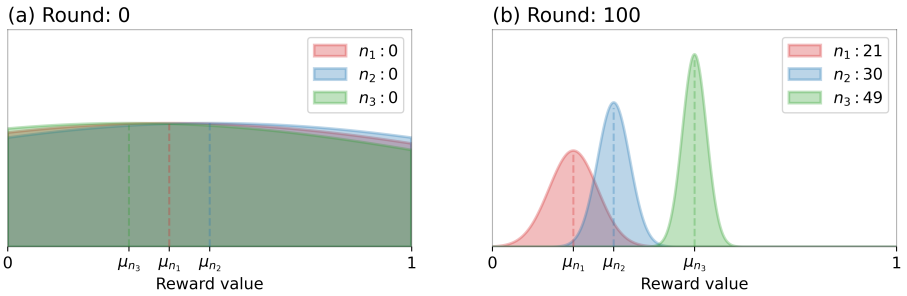

Figure 1: Thompson Sampling example with 3 base arms (a) Beliefs at the beginning of learning (b) Beliefs after 100 rounds of learning

3. Multi-armed Bandits for Smart Charging

3.1. Problem Description

The introduction of EVs in previously designed electrical distribution networks can cause congestion in the network during peak load hours as shown in Figure 2(a), for an example residential network. The electrical current drawn by the network goes beyond its rated value during later hours of the day when the majority of EV owners come back home and plug EVs for charging (basic charging), while the consumption of other loads (e.g. cooking appliances, etc.) is also at its peak. Electric vehicles owners can receive incentives through dynamic electricity pricing to avoid charging during peak load hours (dynamic price-based charging). But this could also lead to a severe congestion if all EVs would be charging during the lowest electricity price instants to minimize their cost of charging i.e., the avalanche effect [47], as shown in Figure 2(b). Furthermore, EVs may also benefit from cheaper or even free electricity generated from local sources (e.g. as in an energy community) such as PV, but this cheaper PV electricity production is variable and uncertain [1]. In this paper, we consider that free electricity may be obtained from local PV panels. The corresponding optimization problem is composed of an objective with a set of constraints:

- An objective to minimize the daily cost of charging, which can be done by learning the trend of free electricity from local PV in order to adjust the charging schedule to this production, and by completing the charge with electricity from the external network during cheap electricity price instants. Fairness among EV agents should be kept in mind as well.
- A set of constraints which includes the avoidance of congestion in the electrical distribution network (global constraint) and the satisfaction of mobility constraints, (i.e. EV battery sufficiently charged before departure), which represents a local constraint.

The graphs shown in Figure 2 correspond to the case study described later in in more details in the experimental evaluation section.

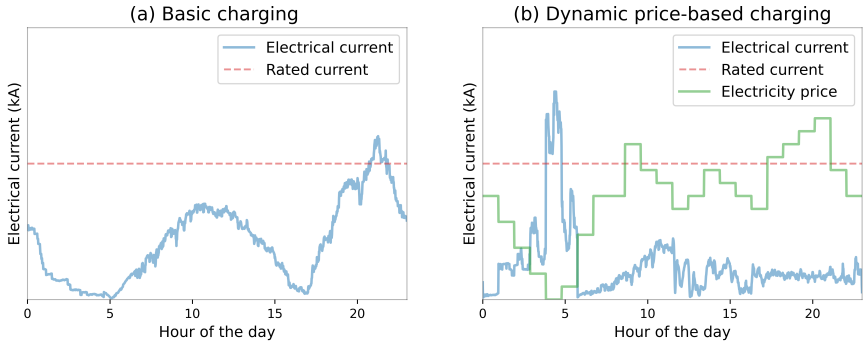

Figure 2: Thompson sampling example figure with 3 arms (before and after learning) indicating the total load current in the network transformer for (a) Basic charging (b) Dynamic price-based charging

3.2. Multi-agent 2-Armed Bandits Formulation

In the proposed approach, each EV is defined as an agent which includes m 2-armed bandits (one per instant of the day). For the sake of simplicity, no subscript will be used to characterize the bandits with respect to the EV to which they belong. However, it is important to emphasize that all EVs have independent bandits.

There is also a congestion management agent in the system. As a part of the environment, this latter agent rewards each of the EV agents based on the EV agent's action and the instantaneous electrical current drawn by the network. The learning objective of each EV agent is to minimize its daily charging cost while the goal of the congestion management agent is to make sure that the electrical distribution network is not congested.

Each day d is divided into $m \in [m] = \{1, 2, 3, ..., m\}$ equally spaced instants. Each instant $i \in [m]$ acts as a arm in this multi-armed bandits formulation and is associated with a normalized electricity cost $c(i) \in [0, 1]$. Let $S_d \in \{0, 1\}^m$ be the set of selected instants (i.e. the combination of arms) on day d by the EV agent, where $S_{i,d} = 1$ means that the instant $i \in [m]$ is selected by the EV agent for charging from the grid on day d and $S_{i,d} = 0$ means that the i-th instant on day d is not selected by the EV agent to charge from the grid. This formulation is referred to as a 2-armed bandits formulation because the action space consists of two actions $\mathcal{A} = \{0, 1\}$, and the EV agent makes a binary decision of selecting (or not selecting) each arm $i \in [m]$ to construct the combination of arms S_d on the day d. This corresponds to the EV agent deciding when to charge at its rated power (and when not to charge) from the grid, out of all available instances $[m]$ during the day d.

The daily reward of the EV agent is the sum of the observed rewards through semi-bandit feedback for each selected charging instant $i \in [m]$ during the day d.

The daily reward r_d is expressed as the sum of all individual rewards obtained by each arm during a given day d as shown in Equation (4):

$$r_d(S_d) = \sum_{i=1}^{M} k_i * Rew(i) \tag{4}$$

where k_i is a binary variable indicated whether arm k was selected or not to be played for day d, M is the number of arms (equal to the number of instants during the day) and $Rew(i)$ is the individual reward obtained by arm k if it is played.

The vector of estimated average rewards $\hat{\theta}$ for all arms is expressed as $\hat{\theta} = [\hat{\theta}_m]$, where $\hat{\theta}_m$ represents the estimated average reward for arm m. The actual selection of a given arm i depends on its estimated average reward $\hat{\theta}_i$. The goal of each EV agent is to learn the unknown vector θ. The agent improves its estimation $\hat{\theta}$ of θ after each interaction with the environment, as detailed in Algorithm 1.

3.2.1. Definition 1 (optimal policy π).

The optimal policy π is obtained by the agent when $\theta \in \mathbb{R}^m$ is known. The optimal policy π will play the combination of arms shown in Equation (5):

$$S^* = \arg \max_{S \in \{0,1\}^m} S_d^\top \theta \Rightarrow \forall m \in [m] \tag{5}$$

To put simply, the EV agent would select the combination of arms consisting of K ($0 \leq K \leq m$) number of arms (charging instants) during the day d, which would provide the maximum reward to the agent.

3.2.2. Definition 2 (pseudo-regret).

The pseudo-regret of the agent after D days of learning can be expressed as shown in Equation (6):

$$\begin{aligned} \mathbb{E}[R(D)] &= \sum_{d=1}^{D} \mathbb{E}[r(S^*)] - \sum_{d=1}^{D} \mathbb{E}[r_d(S_d)] \\ &= \sum_{d=1}^{D} {S^*}^\top \theta - \sum_{d=1}^{D} S_d^\top \hat{\theta}_d \end{aligned} \tag{6}$$

where $\mathbb{E}[r(.)]$ comes from Equation 4, $\hat{\theta}_d$ is the estimated vector on day d, and S^* is given by Equation 5.

The vector of unknowns θ is learned by the agent using Thompson sampling [48]. The agent holds a prior distribution belief (normal distribution) for each element of θ. The agent plays the combinations of arms S_d, consisting of $K \in [m]$ arms. After day d is finished, the beliefs of the played $K \in [m]$ arms are updated according to the observed set of rewards.

In the studied problem, there are two sources of uncertainty. As the proposed system is fully decentralized, from each EV agent's point of view, there is uncertainty in the choice of combination of arms of other agents. As other

agents are learning as well, this uncertainty is a non-stationary random variable. This uncertainty is addressed using Thompson Sampling [35, 36], which provides excellent experimental performances in practice, despite the fact that the original setting of Thompson sampling assumes stationary rewards while the learning of other agents makes the rewards non-stationary. This leads to the loss of theoretical guarantees as the original setting is not respected.

The other source of uncertainty is the "free" electricity production by local PV (e.g. if shared as part of the collective self-consumption operation). It is essential to learn this parameter because the number K of selected arms (EV charging instants from the grid) for the day d depends on this PV production during the day d. This information is learned using Bayesian learning. Bayesian learning is a simple tool compared to more sophisticated forecasting tools available nowadays. However, as the scope of this paper is on energy communities which may not have access to such sophisticated tools, more simple tools were considered to analyze the performance of our proposed algorithm. Let $\phi \in \mathbb{R}^m$ be the vector of length m representing the free PV electricity production value during each instant in $[m]$. In the beginning, the EV agent has no information regarding the magnitude and the trend followed by this free PV production during each day. It receives the instantaneous production information from the PV sensor and based on this information the agent updates its estimate of $\hat{\phi}_d$, using the Bayesian learning approach. Indeed through this approach, the agent learns the average value of the magnitude and the trend followed by the free PV production. As it will be shown later in the experimental evaluation section, this improves the performance of the agent significantly compared to the case when the agent has no estimate of the free PV electricity production. Furthermore, PV electricity production information without any uncertainty is required in the centralized approach to produce the theoretical optimal solution for each day. This would require a perfect forecaster for high-resolution PV forecasting, which is an extremely challenging and resource-intensive task because of the high variability and uncertainty in PV electricity production [49].

The total number of arms to be selected K, for day d can be calculated in Equation (7):

$$K = \left\lceil \frac{60 E_{bat}(SoC_f - SoC_s)}{\mathcal{D}([m]_i) P_{max} \eta} - \frac{\sum_{i=t_{start}}^{t_{depart}} \hat{\phi}_i}{P_{max} \eta} \right\rceil \tag{7}$$

where $\lceil . \rceil$ is the ceiling function, E_{bat} is the battery capacity of the EV, t_{start} is the arrival time of the EV, t_{depart} is the departure time of the EV, SoC_s is the state of charge of the EV's battery at the time of its arrival, SoC_f is the state of charge of the EV's battery at the time of its departure, P_{max} is the maximum charging power of the EV, η is the charging efficiency of the EV, and $\hat{\phi}_i$ is the estimated free PV electricity during the i-th instant. The idea is that the EV agent subtracts the estimated total free PV electricity production value during its connection time from the total energy it requires during the day to attain the desired final state-of-charge SoC_f. Based on this information the EV agent can select the arms $i \in [m] \ni t_{start} \leq i \leq t_{depart}$. To make the system adaptable

to changes in real-time, the EV agent is allowed to update its selection of arms combination S_d during the day as well. Let k_p be the total number of instants EV has already charged till the instant $u \ni 0 \leq u \leq m$ during the day and k_f is the number of instants EV still has to charge to attain SoC_f during this day. Then after every passing instant the EV agent can calculate the remaining number of instants k_f as shown in Equation (8):

$$k_f = K - k_p \tag{8}$$

Based on this information the EV agent can pick k_f number of arms (charging instants) from the remaining instants of the day, i.e., $||S_d||_1 = k_f$. Here, $||S_d||_1$ is the number of arms in the combination of arms S_d. The selected arms would be $i \in [m] \ni u \leq i \leq t_{depart}$.

3.2.3. Reward Function:

The feedback from the environment is a reward value. This value is determined by the congestion management agent in the environment and it is a function of the instantaneous electrical current drawn by the network and the instantaneous electricity price.

3.3. Congestion Model

The congestion model (working of the congestion management agent) is described in Algorithm 1. Let $[E] = \{1, 2, 3, ..., E\}$ be the set of E number of EV agents that have selected the i-th arm. Let $[X] \sim U(0, E)$ be the set of uniformly picked (without replacement) maximum number of EV agents from $[E]$ that can charge simultaneously at the i-th instant without causing congestion in the system. Then the reward obtained by each EV agent $e \in [E]$ is summarized in Equation (9):

$$Rew(i) = \begin{cases} 1 - c(i) & \text{if } I(i) < I_{rated} \\ 1 - c(i) & \text{if } I(i) \geq I_{rated} \ \& \ e \in [X] \sim U(0, E) \\ -1 & \text{if } I(i) \geq I_{rated} \ \& \ e \notin [X] \sim U(0, E) \end{cases} \tag{9}$$

where $I(i)$ is the instantaneous electrical current of the transformer, I_{rated} is its rated electrical current value and $c(i)$ is the electricity price at the i-th instant of the day. Fairness is ensured through uniform sampling (without replacement). In this setting, every agent $e \in [E]$ holds the probability $\frac{1}{E}$ of getting picked to receive the "good" reward. The agents receiving the "poor" reward would try to explore other arms to minimize their regrets.

It is important to note that congestion may occur, as the congestion penalty is included as a soft constraint in the reward, as usually done in many works on this topic. It is expected that congestions may occur frequently in the training phase, which can be performed through simulations. The deployment in real-life may be carried once the model is trained and avoids congestions. Alternatively, the number of EVs being trained simultaneously in real-life may be limited.

3.4. Multi-agent 2-Armed Bandits Algorithm

The functioning of the proposed decentralized multi-agent 2-armed bandits with Thompson Sampling (D-MAB2AB-TS) algorithm for each EV agent is described in Algorithm 2.

The agent plays the combination of arms S_d based on the estimated values of these parameters, and obtains instantaneous rewards. At the end of the day, the agent updates its estimates of the unknown parameter $\hat{\theta}_k$ and variance σ_k^2 based on the observed data. Photovoltaic power estimation is based on a relatively simple method (see Equations (16)-(18) of Algorithm 1). It is assumed to be representative of simple tools to be used in energy communities, while constituting a worst-case scenarios for other cases.

Algorithm 1 D-MAB2AB-TS (EV Agent)

Input: $\hat{\theta} := 0_m, \sigma^2 := 1_m, N_k(0) := 0, \hat{\phi}_m := 0, \sigma_\phi^2 := 1_m, N_{\phi,i}(0) := 0$

for $d = 1, 2, 3, ...$ **do**
  Sample estimated values for each intra-day instant t:
  $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, \sigma_t^2)$
  $\tilde{\phi}_t \sim \mathcal{N}(\hat{\phi}_t, \sigma_{\phi,t}^2)$
  Calculate K using Equation (7)
  for $i = 1, 2, 3, ..., m \ \forall \ t_{start} \leq i \leq t_{depart}$ **do**
    Calculate k_f using Equation (8)
    Play $S_d = \arg \max_{S \in \{0,1\}^m} S^\top \tilde{\theta}$ s.t. $\sum_{l>i,d} S_{l,d} = ||S_d||_1 = k_f$
    Rewards r_t are sampled according to Equation 9
    Update knowledge on best instants for selected instants k:
    $N_k(t) := N_k(t-1) + 1$
    $\hat{\theta}_k(t) := \frac{r_t + N_k(t-1)\hat{\theta}_k(t-1)}{N_k(t)}$
    $\sigma_k^2(t) := \frac{1}{N_k(t)}$
    Sampling instantaneous rewards $r_{\phi,i} \sim \mathcal{N}(\phi_a, \sigma_{\phi,a}^2)$
    Update knowledge on PV production for each instant i:
    $N_{\phi,i}(t_i) := N_\phi(t_i - 1) + 1$
    $\hat{\phi}_i(t_i) := \frac{r_{\phi,i} + N_{\phi,i}(t_i-1)\hat{\phi}_i(t_i-1)}{N_{\phi,i}(t_i)}$
    $\sigma_{\phi,i}^2(t_i) := \frac{1}{N_\phi(t_i)}$
  end for
end for

Remark 1. *The proposed D-MAB2AB-TS algorithm (Algorithm 2) is scalable for decentralized multi-agent applications in terms of each agent's computational time.*

As the selection of the best combination of arms depends only on the total number of arms m and not on the total number of agents in the decentralized system, this ensures that the computational time of each agent will remain the same (for a fixed m) irrespective of the number of agents in the system, and

follows a complexity of $O(m)$. Hence the system is scalable with the number of EV agents.

Remark 2. *The convergence time of the decentralized system, for the proposed D-MAB2AB-TS (Algorithm 2) algorithm, is linked to the congestion limit of the network.*

In smart charging (or similar smart grid applications), the scalability in terms of the convergence to a sufficiently high solution (in the mentioned reference time) is expected to be assisted by the fact that a larger electrical distribution network with a higher number of EV agents generally has a higher congestion limit as well. This can be observed later in the experimental evaluation section. The convergence of the large-scale electrical distribution network (10,175 agents) and that of the small-scale electrical distribution network (55 agents) took approximately the same amount of time (30 simulation days).

These remarks state some of the main contributions of the proposed system. Along with being decentralized, the system is expected to be scalable. Also, the local constraint of each EV agent is modeled inside the learning bandit algorithm (the agent must select exactly K arms during the day to achieve the desired state-of-charge at departure SoC_f). This eliminates the constraint-violating optimal policies with respect to the maximum allowed state-of-charge from the agent's search space. Hence, it can help the agent to converge faster compared to the case when the constraint is modeled inside the reward function, thus requiring the agent to explore many policies to find the optimal solution.

4. Centralized Optimization Formulation

Centralized optimization of the smart charging can provide the optimal solution of the complex smart charging problem [4]. This optimal solution can be used to evaluate the performance of the designed multi-agent system with reinforcement learning as a theoretical lower bound for small case studies, as the complexity class of this decision problem is nondeterministic polynomial time (NP)-hard [50]. Hence, it may become intractable when the number of agents is sufficiently large. The objective function with J number of total EV agents in the system is shown in Equation (10):

$$\begin{aligned} min \sum_{j=1}^{J} C_j(d) = min \sum_{j=1}^{J} \sum_{i=1}^{m} c(i) P_j(i) \Delta i \\ - \sum_{j'=1}^{J} \sum_{j=j'}^{J} \left| \frac{\sum_{i=1}^{m} c(i) P_{j'}(i) \Delta i}{\sum_{i=1}^{m} P_{j'}(i) \Delta i} - \frac{\sum_{i=1}^{m} c(i) P_j(i) \Delta i}{\sum_{i=1}^{m} P_j(i) \Delta i} \right| \end{aligned} \tag{10}$$

.

where $C_j(d)$ is the total charging cost of the EV agent j on day d, $P_j(i)$ is the j-th EV agent charging power at the i-th instant of the day, and $\Delta i = \mathcal{D}([m]_i)$ is the duration of each charging instant. Term $P_j(i) \in [0, P_{max}]$ is the decision

variable of this optimization problem. The first term in the objective function represents the daily charging cost of the EV agent j, and the second term takes into account the fairness constraint by making sure that the differences among per-unit charging costs (cost per energy unit) of all agents are minimized. This term is formulated as a soft constraint, and not as a hard constraint, because this per unit cost depends on the arrival and departure times of the EV. In case an EV is not present during the cheapest electricity price instants, the optimization problem with this fairness term as a hard constraint would not be able to converge. The hard constraints of this optimization problem include electrical network physical constraints, the network congestion constraint, and the charging constraints of each EV.

4.1. Network physical constraints:

These constraints make sure that the power flows in the electrical network are according to the physical constraints of the network. The formulation and linearization of these constraints are explained in [51].

4.2. Congestion constraint:

This is the global constraint of the optimization problem. The distribution system operator has to make sure that the instantaneous electrical current of the network remains below its rated limit (Equation (11)):

$$I(i) < I_{rated} \tag{11}$$

4.3. Electric vehicles constraints:

This is the set of local constraints for each of the EVs based on their state-of-charge (SoC). This set is shown in Equation (12) and (13):

$$SoC_{min} \leq SoC_j(i) \leq SoC_{max} \tag{12}$$

$$SoC_j(i) \geq SoC_f \ \forall \ i = t_{depart} \tag{13}$$

where SoC_{min} is the minimum allowed state-of-charge (SoC), SoC_{max} is the maximum allowed SoC, and $SoC_j(i)$ is the SoC of the j-th EV agent at instant i and SoC_f the desired state-of-charge before departure. The equation to calculate SoC has also been explained in [51].

5. Experimental Evaluation

5.1. Experimental Setting

Two simulation case studies were performed. First, on the small-scale (55 EV agents) IEEE low voltage test feeder (LVTF) distribution network [52]. Second, on the large-scale (10,175 EV agents) distribution network model consisting of 185 IEEE LVTF models as sub-districts. A real-life dataset is used to set the arrival and departure times of the EVs [53]. The daily PV irradiance data is

obtained from the national renewable energy laboratory (NREL) database [54]. The desired final state-of-charge SoC_f, the battery maximum power P_{max}, the battery capacity E_{bat}, and the battery efficiency ratio η are equal to 0.8, 7 kW, 52 kWh, and 0.95 respectively. Also, $m = 1440$, i.e., a day is divided into 1440 decision-making periods. The duration of each period is 1 minute. Communication between the EVs and a congested element in the network (here, the highest voltage transformer) is assumed to be done mostly through wired communication (e.g. Ethernet), thus being quicker than the 1 minute time step at which the agents make their decisions. The four following charging strategies are compared:

- **Basic (Naive) Charging Strategy:** In the basic charging strategy, the EV starts charging at its maximum charging power P_{max}, as soon as the EV owner returns home in the afternoon/evening (no charging during the day). This strategy is not optimal as it does not take the variable electricity prices into account. Also, no exploitation of free PV electricity production is done.
- **Centralized Optimization Charging Strategy:** In this strategy, all information is communicated to a single node, which performs the required centralized optimization to find the optimal solution (charging strategy of each EV). As this strategy uses a PV forecast, the error in this forecast is directly linked to the performance of this centralized optimization strategy. Nonetheless, an idealistic case when the daily PV forecast error is zero can give us the lower bound (optimal solution) for the studied smart charging problem.
- **Two-armed Bandits without PV Estimation Charging Strategy:** This is a variation of the proposed D-MAB2AB-TS algorithm in which the agent does not use any PV estimation, i.e., $\phi = \mathbf{0}$ in the proposed algorithm. This strategy would help in highlighting the improvements made by the proposed D-MAB2AB-TS algorithm which makes an estimation of the PV electricity production. In addition, two variants were also defined by replacing Thompson sampling by i) the Exponential Weights for Exploration and Exploitation algorithm (EXP3) [55], ii) the Upper Confidence Bound algorithm (UCB) [56]. A comparison in terms of cumulated average rewards was performed between these three variants (TS, EXP3, UCB).
- **Two-armed Bandits with PV Estimation (D-MAB2AB-TS) Charging Strategy:** This is the proposed decentralized multi-agent multi-armed bandits strategy. Each EV calculates its charging policy according to the algorithm described in Section 3.4. In addition, two variants were also defined by replacing Thompson sampling by i) the Exponential Weights for Exploration and Exploitation algorithm (EXP3), ii) the Upper Confidence Bound algorithm (UCB). A comparison in terms of cumulated average rewards was performed between these three variants (TS, EXP3, UCB).

The comparison is done regarding constraints satisfaction, total cost paid by the EV owners, and fairness in terms of the per-unit charging cost paid by each EV agent. The per-unit charging cost of the j-th agent is calculated as shown in Equation (14):

$$\frac{\sum_{i=1}^{m} c(i) P_j(i) \Delta i}{\sum_{i=1}^{m} P_j(i) \Delta i}. \tag{14}$$

Let $[D]$ be the set of per-unit costs of each EV agent, then the fairness value of this set is calculated as described in Equation (15):

$$\mathcal{F}(D) = \frac{1}{1 + \left(\frac{\sigma_D}{\overline{D}}\right)^2} \tag{15}$$

where σ_D is the standard deviation of the set D and $\overline{D}$ is the mean value of the set D [57]. The value of this index is in the range between 0 (infinite standard deviation i.e., completely unfair) and 1 (zero standard deviation i.e., completely fair).

The sensitivity of the proposed approach to variations in the PV resources was also investigated. A scaling factor was applied to the PV production time series, with values 0.2, 0.5 and 1. Unless specified, the scaling factor used by default is equal to one.

Simulations were performed using a PC with a 32-core AMD 3970X (3.69 GHz) processor and 128 GB of RAM running Windows 10.

5.2. Experimental Results

5.2.1. Small-Scale Study (55 Agents)

The mean reward of the electrical network (mean of all EVs' average rewards) for the proposed D-MAB2AB-TS algorithm and its variant, i.e., multi-agent 2-armed bandits without no PV estimation, are shown in Figures 3 and 4 respectively. The results for algorithms EXP3 and UCB are also presented.

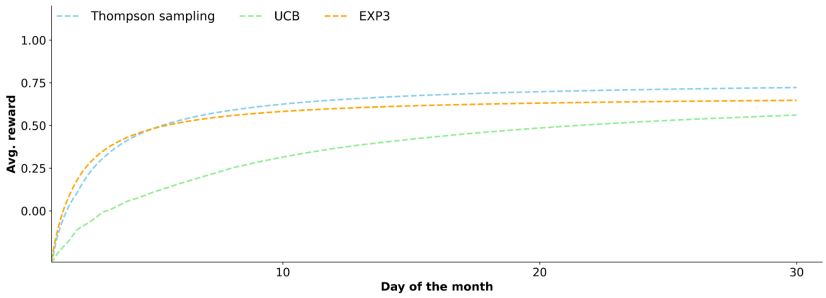

Figure 3: Average reward comparison between Thompson sampling, EXP3 and UCB (without PV prevision)

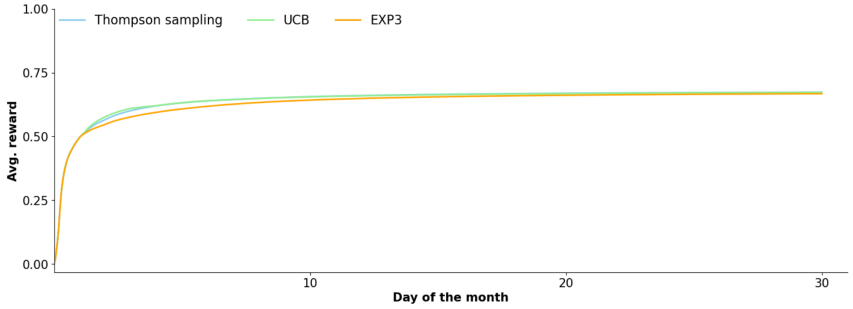

Figure 4: Average reward comparison between Thompson sampling, EXP3 and UCB (with PV prevision)

Analysis of Thompson sampling-based approaches. The analysis will focus at a first stage on the Thompson sampling-based approaches (D-MAB2AB-TS and its variant without PV prevision). The comparative analysis with algorithm EXP3 and UCB will be done in the next section.

In this case, the EV agents converge to their respective optimal policies within 30 days of learning. The next 30 days are used for performance evaluation.

To evaluate the cost of different strategies, the centralized optimization with no PV forecast error is taken as the lower bound. However, the cost of centralized optimization increases as the forecast error is increased, as shown in Figure 5(a). The average percentage increase in the cost, compared to this lower bound, is 138.0% for the basic charging strategy, 85.55% for the multi-agent 2-armed bandits with no PV estimation strategy, and only 10.4% for the proposed D-MAB2AB-TS strategy, as shown in Figure 5(b).

In terms of constraints, the local constraints are satisfied in all the studied charging strategies. The global congestion constraint results are shown in Figure 6. The congestion constraint is violated only in the case of the basic charging strategy for 4.1% of the total evaluation period.

The fairness comparison is presented in Figure 7. The basic charging strategy is not included in the comparison as it is not an optimization strategy. In all three optimization strategies, the local constraint of a desired state-of-charge before the departure of the EV is satisfied. The fairness index values $\mathcal{F}(D)$, calculated using Equation (15), are also presented. For the centralized optimization strategy, this number is 99.9%, while for the proposed D-MAB2AB-TS algorithm and its variant without PV estimation this number comes out to be 99.11% and 99.10% respectively. This confirms that the proposed system takes into account fairness as well.

The computing time was also investigated as shown in Figure 8 which presents the computing time as a function of the number of arms (in other words, the temporal granularity). The trend is clearly linear, as expected, with a standard

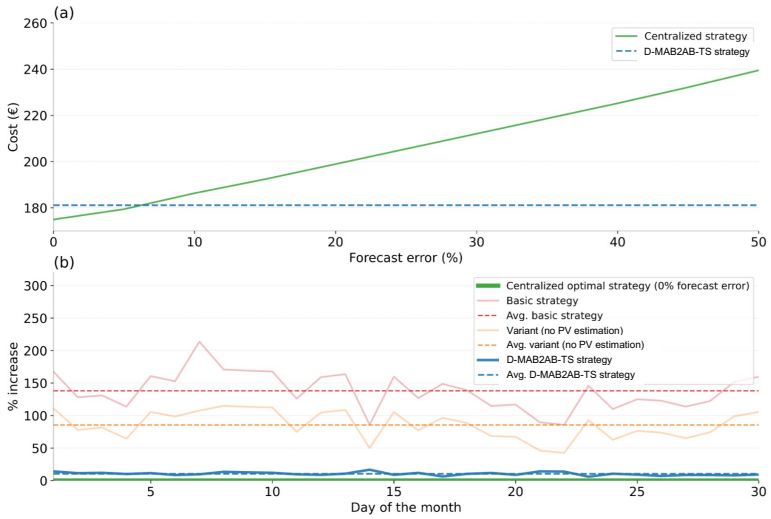

Figure 5: (a) Centralized strategy cost against PV forecast error (b) Percentage of cost increase of all strategies compared to the centralized lower bound.

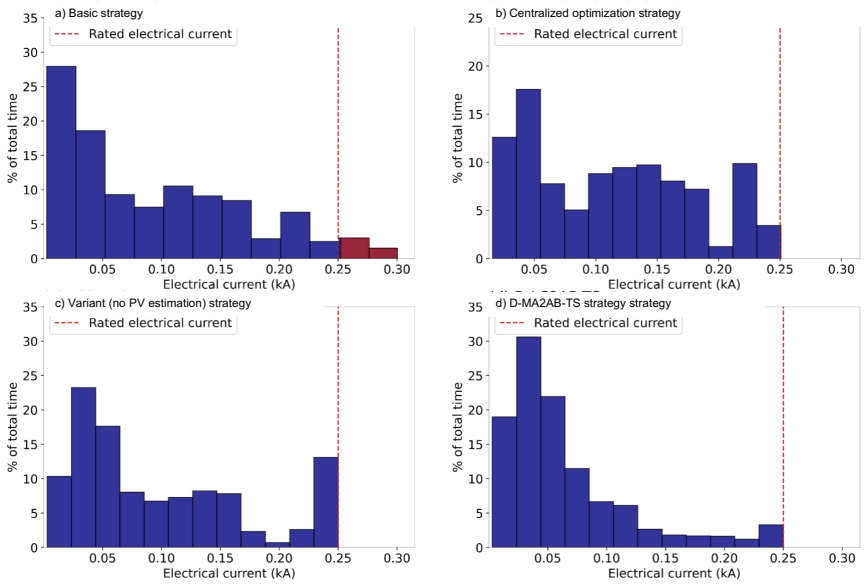

Figure 6: Frequency distribution of the electrical current at the network transformer for (a) Basic Charging (b) Centralized optimization (c) variant (no PV estimation) (d) D-MAB2AB-TS.

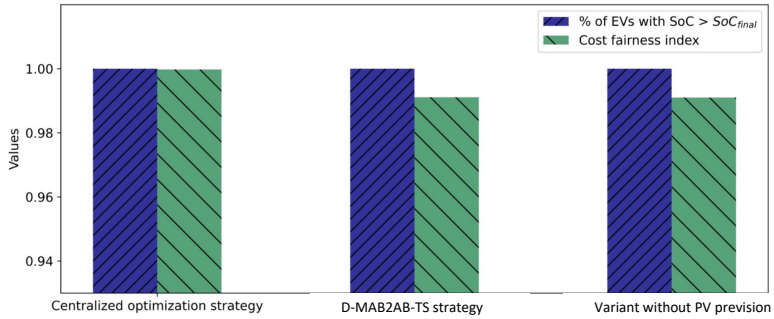

Figure 7: Fairness comparison between the centralized strategy, the proposed D-MAB2AB-TS algorithm and its variant without PV prevision for index $\mathcal{F}(D)$

deviation over 4 runs and normalized to the average ranging between 0.04% and 1.7%. The average computing time per iteration per EV agent is independent of the number of arms, thus proving the approach scalability potential, and is equal to 0.09s.The computing time in the case of a deployment in real-life would of course be different and dependent on the capacities of the decentralized computing devices (e.g. Raspberry pi, etc.).

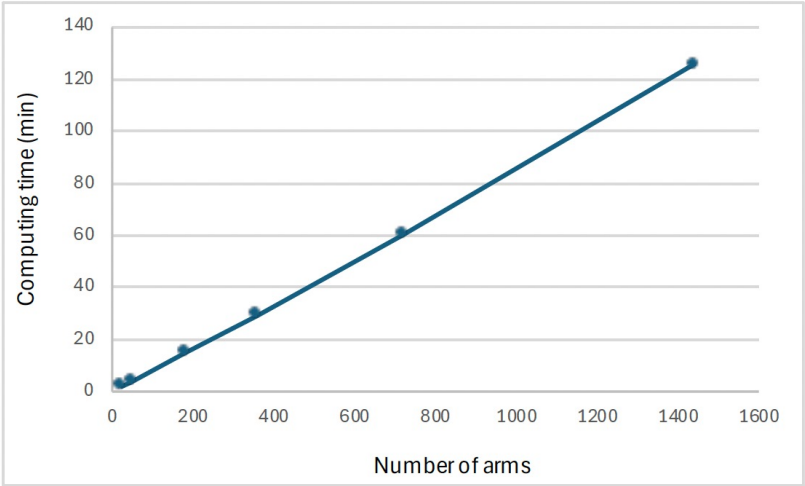

Figure 8: Computing time as a function of the number of arms (temporal granularity

The results also showed that close performance were obtained regardless of the number of arms, as shown in Figure 9, while still satisfying the grid constraints, as shown in Figure 10. It is interesting to note in Figure 9 that, with a lower granularity, a slightly higher average cumulated reward was achieved. This may be explained by the fact that exploration being a stochastic process,

increasing the search space by increasing the number of arms renders more difficult to reach the optimal solution. However, one advantage of increasing the number of arms is that the convergence speed is faster, as more exploration is possible within a single day. Another advantage is of course also the reactivity of the system to congestion that is managed in real-time. In summary, the choice of the number of arms should be result from an arbitrage between convergence speed, optimality and required reactivity to congestion.

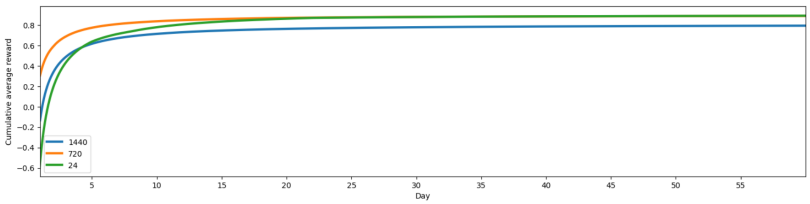

Figure 9: Average cumulated reward as a function of the number of arms

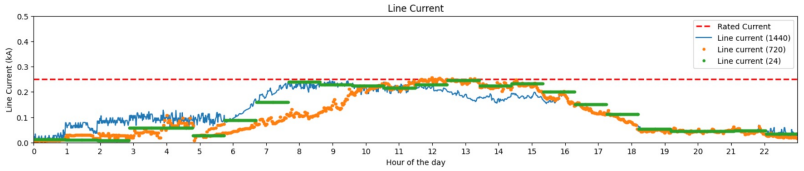

Figure 10: Current in the highest voltage transformer as a function of the number of arms

The sensitivity to PV production data was also investigated. Figure 11 presents the average cumulated reward when the NREL PV data being applied a scaling factor ranging between 0.2 and 1. The results show a relatively similar behaviour with increasing asymptotic values as a function of the scaling factor. This is expected as a higher amount of PV (i.e. higher scaling factor) considered as "free" in this paper decreases the EV owners' bill, thus increasing the average reward.

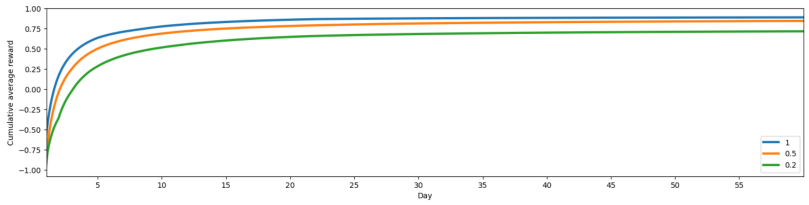

Figure 11: Average cumulated reward for different scaling factors of the PV production time series.

Comparative analysis of Thompson sampling-based approaches with EXP3 and UCB. Figure 3 presents the average reward for the variant of the D-MAB2AB-TS algorithm without PV estimation, as well the the same algorithm based on the EXP3 and the UCB algorithms. It can be seen in this figure that the EXP3 learning strategy is performing the best at the start of the learning phase. The adversarial EXP3 approach can be well suited to handle the non-stationarity in the choice of super arms of other players (EVs) from one player's (EV's) point of view. The mentioned non-stationarity is expected to be at its highest level at the beginning of the simulation-based experimentation. Thus, the EXP3 learning strategy shows this superior performance initially. However, as each agent learns its respective optimal policy, the non-stationarity in the choice of super arms of other EVs from one EV's perspective decreases. Hence, the Thompson sampling-based learning strategy starts performing better than the EXP3-based learning strategy. The UCB-based learning strategy shows inferior performance in this simulation-based experimentation comparatively, but it still converges to a near-optimal policy.

Figure 4 presents the average reward for our proposed algorithm D-MAB2AB-TS (including PV estimation), as well as its variants based on the EXP3 and UCB algorithms. In this case, it can be observed that when EV agents are also learning the trend of the daily freely available PV energy production, all learning strategies converge to the same average reward value. This is because EVs also utilize the available PV energy production during the day. Hence, the competition during the low electricity price instants is reduced. This reduced competition favors the UCB and the EXP3 learning strategies, thus achieving the same average reward value. The relatively small difference in terms of asymptotic value between the three algorithms (Thompson sampling, UCB and EXP3) when PV estimation is performed, seem to show that this specific problem may be relatively insensitive to the type of bandit algorithm used, in terms of asymptotic value. This should be confirmed with additional studies.

5.2.2. Large-Scale Study (10,175 Agents):

Analysis of Thompson sampling-based approaches. The centralized strategy becomes intractable due to the large number of agents in the studied system, and fails to provide computation results. Unlike the centralized approach, the proposed D-MAB2AB-TS algorithm is able to provide results. The mean reward of the electrical network (mean of all EV's average rewards) for the proposed strategy and its variant without PV estimation strategy is shown in Figures 12 and 13.

The proposed algorithm converges in a similar amount of time as the studied smaller-scale electrical network (30 days).

In Figure 14, frequency distributions of the electrical current are shown. For the proposed D-MAB2AB-TS and its variant without PV estimation, no congestion is observed. However, congestion is observed for approximately 2% of the total evaluation period if the basic charging strategy is followed.

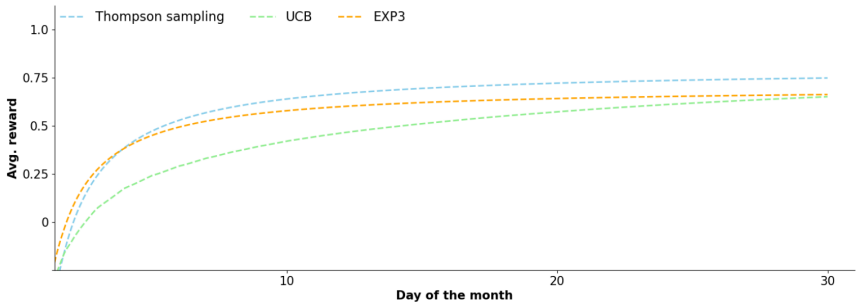

Figure 12: Average reward comparison between Thompson sampling, EXP3 and UCB (without PV prevision)

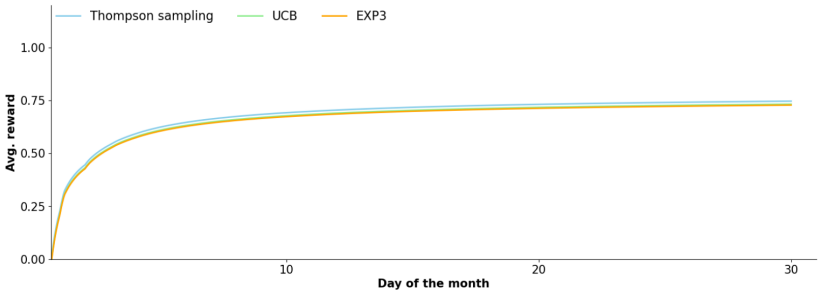

Figure 13: Average reward comparison between Thompson sampling, EXP3 and UCB (with PV prevision)

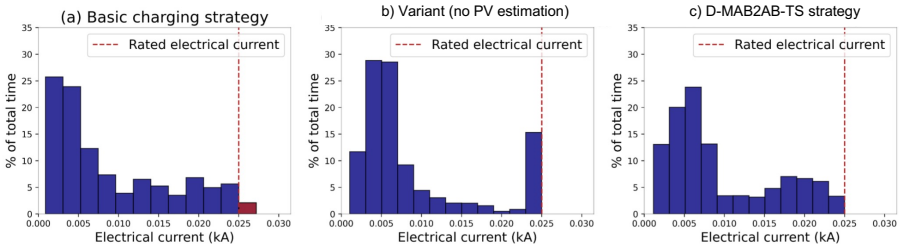

Figure 14: Frequency distribution of the electrical current at the network transformer for (a) Basic Charging (b) variant (no PV estimation) (c) D-MAB2AB-TS.

The average total costs and fairness comparisons are presented in Figure 15. The proposed algorithm has the lowest average total cost while basic charging has the highest. The local state-of-charge constraints are satisfied in all strategies. The value of the fairness index $\mathcal{F}(D)$ for basic charging is not calculated

as it is not an optimization strategy, while for the other two shown strategies, this value is 99%.

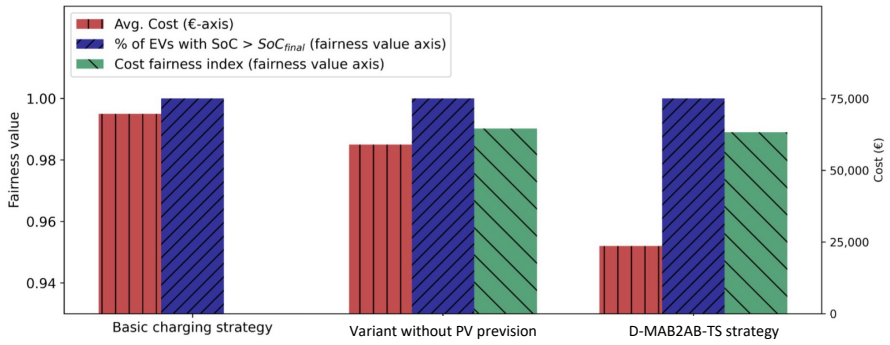

Figure 15: Average total cost and fairness comparison based on index $\mathcal{F}(D)$ for the basic charging strategy, the variant and the D-MAB2AB-TS strategy.

Comparative analysis of Thompson sampling-based approaches with EXP3 and UCB. Figures 12 and 13 show that similar performance trends can be observed here as well compared to the small-scale studies, i.e., Thompson sampling outperforms UCB and EXP3 when no PV estimation is done, and all learning strategies are converging to approximately the same average reward value when PV estimation is performed. The behaviour is thus similar for the small-scale and the large-scale study, which shows that the performance of the proposed algorithm is independent from the number of EV agents.

6. Discussion

Future work will address some additional aspects, some of which may represent limitations to our study. These aspects are discussed in this section.

6.1. Conditions diversity

Our proposed method is evaluated against days that present a significant level of similarity. Although we have shown that performance was independent on the number of EV agents through a comparison between the small-scale and the large-scale study, additional sensitivity analyses could be performed. In particular, different solar irradiance conditions could be investigated, as well as different network characteristics (topology, etc.). Also, no major EV energy demand change is modeled as all EVs are present for each day. However, in real-life, variations in the number of EVs may be observed, such as between working weeks and holidays. Considering these aspects is the object of current on-going work where the type of day (i.e. holidays, sunny day, etc.) is included as a context in contextual bandits. In addition, a sensitivity analysis on the number of arms (time granularity) could be of interest.

6.2. Communications overhead and latency

Regarding the communication overhead, a congested agent (e.g. agent corresponding to the highest voltage transformer) is linked to EV agents located downstream only (and not to all EV agents in the system), as only those located downstream may have a beneficial action to solve the congestion. The congestion signal consists in a small message, containing potentially only a binary number, which could be sent only in case of a congestion. Also, it is important to remind that there is no inter-EV communication, as each EV acts as an independent learner. Hence, the daily communication burden may be relatively light, even for large-scale EV fleets (tens of thousands of agents and more).

Regarding latency, it is assumed that, if the system operates with wired telecommunications (e.g. Ethernet protocol), the impact of latency on the algorithm performance should negligible. Agents make their decisions every minute. Hence, at the beginning of every minute, they receive the information directly from the congested agents (in case a congestion occurs), assumingly within hundreds of milliseconds latency maximum. In this context, the potential congestion information would arrive far before the minute is over. Therefore, the performance of our system should not suffer from latency under these conditions. However, should the system be based on different technologies (power line carrier (PLC) or wireless (e.g. WiFi)), latency effects may have to be considered.

6.3. Synchronization issues

Regarding synchronization issues, several aspects must be discussed.

- Synchronization in terms of agents choosing the same arms (and so, instants): Thompson sampling is a stochastic algorithm, so the arms selected at each round are random variables. In this sense, agents cannot be synchronized, contrary to the UCB algorithm which is a deterministic and for which we show that performance may be less efficient than Thompson sampling and EXP3.

- Synchronization in terms of agents making their decisions simultaneously: EV agents do not communicate with each other, therefore not requiring synchrony. In addition, the time step was selected as sufficiently small (1 minute) compared to the allowed time for congestion (few minutes for an overhead line to tens of minutes for cables or transformers presenting a significant thermal inertia). This may leave the opportunity for the agents to solve congestion issues quickly on this timescale, whether or not their decision making is synchronized or may experience small asynchrony (e.g. in the order of a minute). In the case of a larger asynchrony, the reliability of the method remains to be investigated.

6.4. Fairness

The uniform sampling approach used for ensuring fairness in the congestion management mechanism is relatively simple and potentially suboptimal

compared to more sophisticated fairness-preserving algorithms [58, 59, 60, 61]. Although the simple mechanism led to excellent results in the considered case studies, future work should consider integrating also such sophisticated algorithms for comparison.

6.5. Resilience

The study has covered decision making in uncertain environments where the perturbations are limited to usual ones (e.g. uncertainty in PV production). Larger perturbations in the form of HILP event (high impact, low probability) were out of the scope of this paper. However, it would be interesting to complete the study in future work by considering the performance of the considered algorithms under HILP conditions. It is possible that EXP3, designed for adversarial settings, may perform better. The choice of a given algorithm should therefore be based on an arbitrage between its performance under normal and HILP conditions.

6.6. Charging power modulation

In the current work, the charging power is binary (i.e. charging at full power or not charging). This is representative of current practices where power modulation is currently rarely used. However, future work will consider finer power modulation with a choice between several thresholds of charging power. The extension of the proposed methodology to a non-binary charging power (i.e. including several levels of non-null charging power) is straightforward, as it would only require to add more base-arms (e.g. a set of base-arms {0;1;2} per bandit.

6.7. Mechanisms for voltage and reactive power regulation, and topology changes

The base electrical grid model considered in this work is the IEEE European low voltage test feeder. In this network, only the low voltage part is modeled. No voltage and reactive power regulation mechanisms, nor topology change mechanisms are included. The large-scale study is based on a larger network model where these mechanisms may be present although they were not considered in this work, as a first stage approximation. However, this does not change the proposed methodology, as the congestion signal sent by a congested element may be sent before or after the other regulation mechanisms are activated, depending on the preferences of the distribution system operator. In other words, it will be the responsibility of the distribution system operator to determine the coordination between different mechanisms (including flexibility from EVs), but this does not influence our proposed methodology.

7. Acknowledgments

This work has been carried out as part of the "Opt-Real" project funded by the Brittany regional council and the Ecole Normale Supérieure de Rennes (ENS Rennes) which are gratefully acknowledged.

8. Conclusion

To tackle the challenges of existing smart charging methodologies, we propose a decentralized smart charging multi-agent system using 2-armed bandits with Thompson sampling. The proposed algorithm manages the uncertainties in PV electricity production and in the selected action(s) of other agents in a decentralized system. The scalability of the proposed D-MAB2AB-TS algorithm as demonstrated experimentally. The performance comparison of the proposed system against a basic charging strategy and a strategy without PV estimation highlights a significant reduction in the total cost. Several directions for future work were discussed to extend this preliminary study.

References

[1] M. D. Tabone, D. S. Callaway, Modeling variability and uncertainty of photovoltaic generation: A hidden state spatial statistical approach, IEEE Transactions on Power Systems 30 (6) (2015) 2965–2973.

[2] R. Fachrizal, M. Shepero, D. van der Meer, J. Munkhammar, J. Widén, Smart charging of electric vehicles considering photovoltaic power production and electricity consumption: A review, eTransportation 4 (2020) 100056.

[3] D. van der Meer, G. R. C. Mouli, G. M-E Mouli, L. R. Elizondo, P. Bauer, Energy management system with pv power forecast to optimally charge evs at the workplace, IEEE Transactions on Industrial Informatics 14 (1) (2018) 311–320.

[4] J. F. Franco, M. J. Rider, R. Romero, A mixed-integer linear programming model for the electric vehicle charging coordination problem in unbalanced electrical distribution systems, IEEE Transactions on Smart Grid 6 (5) (2015) 2200–2210.

[5] B. Amirhosseini, S. M. H. Hosseini, Scheduling charging of hybrid-electric vehicles according to supply and demand based on particle swarm optimization, imperialist competitive and teaching-learning algorithms, Sustainable Cities and Society 43 (2018) 339–349.

[6] N. Mehboob, C. Cañizares, C. Rosenberg, Day-ahead dispatch of pev loads in a residential distribution system, in: 2014 IEEE PES General Meeting | Conference & Exposition, 2014, pp. 1–5.

[7] E. L. Karfopoulos, N. D. Hatziargyriou, A multi-agent system for controlled charging of a large population of electric vehicles, IEEE Transactions on Power Systems 28 (2) (2013) 1196–1204.

[8] J. Hu, H. Morais, M. Lind, H. W. Bindner, Multi-agent based modeling for electric vehicle integration in a distribution network operation, Electric Power Systems Research 136 (2016) 341–351.

[9] M. Habibidoost, S. M. T. Bathaee, A self-supporting approach to ev agent participation in smart grid, International Journal of Electrical Power & Energy Systems 99 (2018) 394–403.

[10] P. Papadopoulos, N. Jenkins, L. M. Cipcigan, I. Grau, E. Zabala, Coordination of the charging of electric vehicles using a multi-agent system, IEEE Transactions on Smart Grid 4 (4) (2013) 1802–1809.

[11] S. Mocci, N. Natale, F. Pilo, S. Ruggeri, Multi-agent control system to coordinate optimal electric vehicles charging and demand response actions in active distribution networks, in: 3rd Renewable Power Generation Conference (RPG 2014), 2014, pp. 1–6.

[12] P. Goli, W. Shireen, Pv powered smart charging station for phevs, Renewable Energy 66 (2014) 280–287.

[13] J. García-Villalobos, I. Zamora, J. San Martín, F. Asensio, V. Aperribay, Plug-in electric vehicles in electric distribution networks: A review of smart charging approaches, Renewable and Sustainable Energy Reviews 38 (2014) 717–731.

[14] M. Conoscenti, A. Vetrò, J. C. De Martin, Peer to peer for privacy and decentralization in the internet of things, in: 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C), 2017, pp. 288–290.

[15] M. Conoscenti, A. Vetrò, J. C. De Martin, Peer to peer for privacy and decentralization in the internet of things, in: 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C), 2017, pp. 288–290.

[16] Does reinforcement learning outperform deep learning and traditional portfolio optimization models in frontier and developed financial markets?, Research in International Business and Finance 65 (2023) 101936.

[17] M. Moosmann, M. Kaiser, J. Rosport, F. Spenrath, W. Kraus, R. Bormann, M. F. Huber, Performance comparison of supervised and reinforcement learning approaches for separating entanglements in a bin-picking application, in: N. Kiefl, F. Wulle, C. Ackermann, D. Holder (Eds.), Advances in Automotive Production Technology – Towards Software-Defined Manufacturing and Resilient Supply Chains, Springer International Publishing, Cham, 2023, pp. 158–167.

[18] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, C. S. Lai, A multi-agent reinforcement learning-based data-driven method for home energy management, IEEE Transactions on Smart Grid 11 (4) (2020) 3201–3211.

[19] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, X. Guan, Multi-agent deep reinforcement learning for hvac control in commercial buildings, IEEE Transactions on Smart Grid 12 (1) (2021) 407–419.

[20] Y. Yang, J. Hao, Z. Wang, M. Sun, G. Strbac, Recurrent deep multiagent q-learning for autonomous agents in future smart grid, AAMAS '18, International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[21] M. Tan, J. Zhao, X. Liu, Y. Su, L. Wang, R. Wang, Z. Dai, Federated reinforcement learning for smart and privacy-preserving energy management of residential microgrids clusters, Engineering Applications of Artificial Intelligence 139 (B) (2025).

[22] D. Choudhary, R. N. Mahanty, N. Kumar, Demand management of plug-in electric vehicle charging station considering bidirectional power flow using deep reinforcement learning, Engineering Applications of Artificial Intelligence 139 (A) (2025).

[23] V. Bui, S. Mohammadi, S. Das, A. Hussain, G. Vieira Hollweg, W. Su, A critical review of safe reinforcement learning strategies in power and energy systems, Engineering Applications of Artificial Intelligence 143 (2025).

[24] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, 2nd Edition, The MIT Press, 2018.

[25] N. Ebell, M. Gütlein, M. Pruckner, Sharing of energy among cooperative households using distributed multi-agent reinforcement learning, in: 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), 2019, pp. 1–5.

[26] H. M. Abdullah, A. Gastli, L. Ben-Brahim, Reinforcement learning based ev charging management systems–a review, IEEE Access 9 (2021).

[27] X. Chen, G. Qu, Y. Tang, S. Low, N. Li, Reinforcement learning for selective key applications in power systems: Recent advances and future challenges, IEEE Transactions on Smart Grid 13 (4) (2022).

[28] Y. Liu, A. Halev, X. Liu, Policy learning with constraints in model-free reinforcement learning: A survey, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, survey Track.

[29] H. V. Hasselt, Y. Doron, F. Strub, M. Hessel, N. Sonnerat, J. Modayil, Deep reinforcement learning and the deadly triad, CoRR (2018).

[30] J. Achiam, E. Knight, P. Abbeel, Towards characterizing divergence in deep q-learning, CoRR (2019).

[31] T. Lu, D. Schuurmans, C. Boutilier, Non-delusional q-learning and value-iteration, in: Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018.

[32] D. Huh, P. Mohapatra, Multi-agent reinforcement learning: A comprehensive survey, CoRR (2023).

[33] A. Slivkins, Introduction to multi-armed bandits, Foundations and Trends in Machine Learning 12 (1-2) (2019).

[34] H. Dakdouk, R. Féraud, P. Varsier, N.and Maillé, R. Laroche, Massive multi-player multi-armed bandits for iot networks: An application on lora networks, Ad Hoc Networks 151 (2023).

[35] R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, J. Palicot, Multi-armed bandit learning in iot networks: Learning helps even in non-stationary settings, CoRR (2018).

[36] L. Besson, E. Kaufmann, Multi-Player Bandits Revisited, in: Proceedings of Algorithmic Learning Theory, Vol. 83 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 56–92.

[37] H. Dakdouk, E. Tarazona, R. Alami, R. Féraud, G. Z. Papadopoulos, P. Maillé, Reinforcement learning techniques for optimized channel hopping in ieee 802.15.4-tsch networks, MSWIM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 99–107.

[38] R. Kerkouche, R. Alami, R. Féraud, N. Varsier, P. Maillé, Node-based optimization of lora transmissions with multi-armed bandit algorithms, 2018 25th International Conference on Telecommunications (ICT) (2018) 521–526.

[39] H. Ou, C. Siebenbrunner, J. A. Killian, M. B. Brooks, D. Kempe, Y. Vorobeychik, M. Tambe, Networked restless multi-armed bandits for mobile interventions, AAMAS '22, International Foundation for Autonomous Agents and Multiagent Systems, 2022.

[40] Z. Hu, Z. Wang, Z. Li, S. Hu, S. Ruan, J. Zhang, Fraud regulating policy for e-commerce via constrained contextual bandits, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[41] Y. Liu, P. Zhou, L. Yang, Y. Wu, Z. Xu, K. Liu, X. Wang, Privacy-preserving context-based electric vehicle dispatching for energy scheduling in microgrids: An online learning approach, IEEE Transactions on Emerging Topics in Computational Intelligence 6 (3) (2022).

[42] Z. Yu, Y. Xu, L. Tong, Large scale charging of electric vehicles: A multi-armed bandit approach, in: 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2015, pp. 389–395.

[43] O. Chapelle, L. Li, An empirical evaluation of thompson sampling, in: Advances in Neural Information Processing Systems, Vol. 24, Curran Associates, Inc., 2011.

[44] E. Kaufmann, N. Korda, R. Munos, Thompson sampling: An asymptotically optimal finite-time analysis, in: ALT 2012 - International Conference on Algorithmic Learning Theory, Vol. 7568, 2012, pp. 199–213.

[45] R. Combes, M. Lelarg, A. Proutière, M. S. Talebi, Stochastic and adversarial combinatorial bandits, CoRR (2015).

[46] S. Agrawal, N. Goyal, Analysis of thompson sampling for the multi-armed bandit problem, in: Proceedings of the 25th Annual Conference on Learning Theory, Vol. 23 of Proceedings of Machine Learning Research, PMLR, Edinburgh, Scotland, 2012, pp. 39.1–39.26.

[47] M. Amjad, A. Ahmad, M. H. Rehmani, T. Umer, A review of evs charging: From the perspective of energy optimization, optimization approaches, and charging techniques, Transportation Research Part D: Transport and Environment 62 (2018) 386–417.

[48] S. Agrawal, N. Goyal, Thompson sampling for contextual bandits with linear payoffs, CoRR (2012).

[49] H. Ye, B. Yang, Y. Han, N. Chen, State-of-the-art solar energy forecasting approaches: Critical potentials and challenges, Frontiers in Energy Research 10 (2022).

[50] C. H. Papadimitriou, K. Steiglitz, Combinatorial Optimization : Algorithms and Complexity, Dover Publications, 1998.

[51] S. Zafar, V. Maurya, A. Blavette, G. Camilleri, H. B. Ahmed, M. Gleizes, Adaptive multi-agent system and mixed integer linear programming optimization comparison for grid stability and commitment mismatch in smart grids, in: 2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), 2021, pp. 01–05.

[52] K. P. Schneider, B. A. Mather, B. C. Pal, C.-W. Ten, G. J. Shirek, H. Zhu, J. C. Fuller, J. L. R. Pereira, L. F. Ochoa, L. R. de Araujo, R. C. Dugan, S. Matthias, S. Paudyal, T. E. McDermott, W. Kersting, Analytic considerations and design basis for the ieee distribution test feeders, IEEE Transactions on Power Systems 33 (3) (2018).

[53] Test-an-EV project, Electrical vehicle (ev) data, Tech. rep., SEAS-NVE (216).
URL `http://smarthg.di.uniroma1.it/Test-an-EV`

[54] B. Kroposki, D. Mooney, T. Markel, B. Lundstrom, Energy systems integration facilities at the national renewable energy laboratory, in: 2012 IEEE Energytech, 2012, pp. 1–4.

[55] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The nonstochastic multiarmed bandit problem, SIAM Journal on Computing (2002).

[56] P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multiarmed bandit problem, Machine Learning (2002).

[57] R. Jain, D. Chiu, W. Hawe, A quantitative measure of fairness and discrimination for resource allocation in shared computer systems, CoRR (1998).

[58] V. Patil, G. Ghalme, V. Nair, Y. Narahari, Achieving fairness in the stochastic multi-armed bandit problem, Journal of Machine Learning Research (2021).

[59] I. Bistritz, T. Z. Baharav, A. Leshem, N. Bambos, My fair bandit: distributed learning of max-min fairness with multi-player bandits, in: Proceedings of the 37th International Conference on Machine Learning, ICML'20, JMLR.org, 2020.

[60] S. Hossain, E. Micha, N. Shah, Fair algorithms for multi-agent multi-armed bandits, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Inc., 2021, pp. 24005–24017.

[61] C. Herlihy, A. Prins, A. Srinivasan, J. P. Dickerson, Planning to fairly allocate: Probabilistic fairness in the restless bandit setting, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, ACM, 2023, p. 732–740.