Multi-Armed Bandits Learning for Optimal Decentralized Control of Electric Vehicle Charging

Sharyal Zafar, Raphaël Féraud, Anne Blavette, Guy Camilleri, H. Ben Ahmed

Multi-Armed Bandits Learning for Optimal Decentralized Control of Electric Vehicle Charging

Sharyal Zafar
SATIE Laboratory
École Normale Supérieure de Rennes
Bruz, France
sharyal.zafar@ens-rennes.fr

Raphaël Féraud
Orange Labs
Orange
Lannion, France
raphael.feraud@orange.com

Anne Blavette
CNRS, SATIE Laboratory
École Normale Supérieure de Rennes
Bruz, France
anne.blavette@ens-rennes.fr

Guy Camilleri
IRIT Laboratory
Paul Sabatier University
Toulouse, France
guy.camilleri@irit.fr

Hamid Ben Ahmed
SATIE Laboratory
École Normale Supérieure de Rennes
Bruz, France
benahmed@ens-rennes.fr

***Abstract*—Optimal control of new grid elements, such as electric vehicles, can ensure an efficient, and stable operation of distribution networks. Decentralization can result in scalability, higher reliability, and privacy (which may not be present in centralized or hierarchical control solutions). A decentralized multi-agent multi-armed combinatorial bandits system using Thompson Sampling is presented for smart charging of electric vehicles. The proposed system utilizes the concepts of bandits reinforcement learning to manage the uncertainties in the choice of other players' actions, and in the intermittent photovoltaic energy production. This proposed solution is fully decentralized, real-time, scalable, model-free, and fair. Its performance is evaluated through comparison with other charging strategies i.e., basic charging, and centralized optimization.**

***Index Terms*—Electric vehicles, Active distribution networks, Smart charging, Multi-agent reinforcement learning, Combinatorial multi-armed bandits**

I. Introduction

The increasing number of electric vehicles (EVs) and photovoltaics (PVs) can introduce new challenges, such as congestion, and peak load demand, in the distribution networks. Optimal control of these new grid elements, while satisfying the local constraints of each EV prosumer, and the global constraints of the distribution system operator (DSO), can help in mitigating the mentioned challenges. In [1], [2], the real-time charging of EVs is controlled to assist the distribution network. However, such systems may suffer from a lack of scalability, a single point of failure, and data privacy concerns, as they are centralized. Multi-agent systems for smart grid applications are proposed in [3], [4] to tackle the challenges of centralized systems. However, these systems are hierarchical, and may still suffer from the drawbacks of centralization. Furthermore, an accurate distribution network model is assumed to be known in the majority of the mentioned systems, which is not always the case in real life. Thus, a decentralized, and model-free control algorithm can be useful for efficient smart grid operations.

Reinforcement learning (RL) has also found its application in active distribution networks in recent years, as it can help in designing a decentralized multi-agent system. In [5], [6], multi-agent systems combined with RL have been proposed to control the functioning of a variety of elements in distribution networks. In standard RL, an agent learns the optimal policy through interactions with the environment. Each action of the agent changes the state of the environment, and the agent receives a reward from the environment. The agent's goal is to maximize the running sum of these observed rewards [7]. These RL algorithms can be useful to minimize the objective function cost, but the inclusion of multi-level distribution network constraints (local constraints of the EV prosumers, and global constraints of the DSO) can still be a challenge [8]. More importantly, no known Oracle can evaluate each action's performance, thus the cost of such a system would be directly linked to the learning time of the agent in real-life. Stability problems, and limited theoretical convergence results are also some of the challenges of commonly used RL algorithms with function approximations (e.g., DQN learning).

Multi-armed bandits (MAB) is a simpler subclass of reinforcement learning [11]. Being a simpler subset of Markov Decision Processes enables MAB algorithms to converge relatively faster (compared to DQN or Q-learning). This is a significant advantage for smart grid applications, as the learning is purely online, and no perfect Oracle to evaluate the performance of each action is available. The simpler nature of MAB algorithms also results in well-defined theoretical guarantees. Bandit algorithms have been used for smart charging [9], [10]. However, the mentioned systems are not decentralized. Selfish bandits have also been utilized to optimize modern communication networks in a decentralized manner with excellent results [11]–[13].

This selfish MAB approach is extended here for the smart charging of EVs. The goal of each EV agent is to minimize its daily charging cost, in the presence of dynamic electricity

pricing, and uncertain PV energy production, while satisfying a set of constraints. This optimization problem is modeled as a combinatorial multi-armed bandits (CMAB) problem [14]. The main contributions of this paper include a fully decentralized smart charging MAS using CMAB, the selfish heuristic for handling multi-player bandits, and Bayesian estimation of PV energy production. We call our proposed system decentralized because each network entity that encounters an issue (e.g., a node with under-voltage or transformer with congestion) deals with this issue by sending messages to the flexible entities (EVs here), and each EV optimizes its charging strategy. It is in contrast to centralized control by a DSO identifying itself where the grid issues are and how to react. The presented system is also scalable, and each agent can find the estimated best arm to play in $O(m)$. The proposed system is also real-time, keeps fairness among all agents in check, and is generic (adaptable to other smart grid applications). The paper is organized as follows: optimization formulation is described in section II, and section III will present the CMAB formulation of the smart charging problem. A detailed case study with results will be presented in section IV, and finally, the conclusion will be made in section V.

II. Optimization Formulation

The objective of the smart charging problem is to minimize the total charging cost of the EVs in the presence of daily dynamic electricity pricing, and uncertain PV energy production. It is defined as:

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

where J is the total number of EVs, m is the total number of decision instants, $C_j(d)$ is the total charging cost of the j-th EV on day d, $c(i)$ is the normalized electricity price at i-th instant, $P_j(i)$ is the charging power of the EV agent j at i-th instant of the day, and Δi is the duration of each charging instant. The decision variable of the problem is $P_j(i) \in [0, P_{max}]$. The first term in (1) minimizes the sum of daily charging costs of all EVs, while the second term ensures fairness by minimizing the differences among per-unit charging costs (cost per energy unit) of all EVs. The constraints of the optimization problem are described below.

Network's Physical Constraints: These constraints satisfy the network's physical constraints and result in correct power flow results. The active power of bus a at instant i is:

$$P_a(i) = P_{a,gen}(i) - P_{a,dem}(i) \tag{2}$$

where $P_{a,gen}(i)$ is the generated power at bus a at instant i. This term can include electrical generators (in the case of grid bus), and PV energy generation as well. Term $P_{a,dem}(i)$ is the power demand at bus a at instant i. This includes both household loads as well as EVs demand. The reactive power equation can be written in a similar way. The power flow is linked to the instantaneous bus voltages as:

$$P_{ab}(i) + jQ_{ab}(i) = V_a(i)(V_a^*(i) - V_b^*(i))Y_{ab}^* \tag{3}$$

where $P_{ab}(i)$, and $Q_{ab}(i)$ are the active and reactive powers respectively, flowing from bus a to b at instant i. Term $V_a(i)$ is the instantaneous voltage at bus a, and Y_{ab}^* is the admittance of electrical line connecting bus a and bus b.

DSO's Constraints: These constraints include no electrical current congestion, and no voltage limit violation in the system. These are defined as:

$$I_{ab}(i) < I_{ab,max} \tag{4}$$

$$V_{a,min} < V_a(i) < V_{a,max} \tag{5}$$

where $I_{ab}(i)$ is the instantaneous current flowing from bus a to bus b, $I_{ab,max}$ is the rated current value, $V_{a,min}$ is the allowed minimum instantaneous voltage at bus a, and $V_{a,max}$ is the allowed maximum voltage at bus a.

Electric Vehicle's Constraints: These are the local constraints of each EV prosumer, which include that the state of charge (SoC) of the EV should always remain within the specified range, and it should be greater than the specified value at EV's departure time. Furthermore, the state of health (SoH) of all EVs should be positive. These constraints are defined as:

$$SoC_{a,min} < SoC_a(i) < SoC_{a,max} \tag{6}$$

$$SoC_a(i_{depart}) \geq SoC_{a,depart} \tag{7}$$

$$SoH_a(i) > 0 \tag{8}$$

where $SoC_{a,min}(i)$ is the instantaneous SoC of EV connected to bus a, $SoC_{a,min}$ is the minimum allowed SoC, $SoC_{a,max}$ is the maximum allowed SoC, $SoC_a(i_{depart})$ is the SoC at the departure time of EV a, $SoC_{a,depart}$ is the desired SoC at the departure time, and $SoH_a(i)$ is the instantaneous SoH of EV. Both SoC, and SoH are defined in [15].

A. Linearization

The above-mentioned formulation is non-linear due to the product of voltages in (3). To solve this formulation as a mixed integer linear programming (MILP) optimization problem, linearization of (3) is performed. Small angle approximation (i.e., $sin(\vartheta_a - \vartheta_b) \approx (\vartheta_a - \vartheta_b)$) is assumed here. Furthermore, it is assumed that the magnitude of per-unit voltages is sufficiently close to 1. After applying these approximations, (3) is linearized and the following two equations are obtained for active and reactive powers respectively:

$$P_{ab}(i) = G_{ab}(V_a(i) - V_b(i)) + B_{ab}(\vartheta_a(i) - \vartheta_b(i)) \tag{9}$$

$$Q_{ab}(i) = B_{ab}(V_a(i) - V_b(i)) + G_{ab}(\vartheta_b(i) - \vartheta_a(i)) \tag{10}$$

where G_{ab}, and B_{ab} are the conductance, and the susceptance of the electrical line between bus a and bus b respectively. Term $\vartheta_a(i)$ is the instantaneous phase angle at bus a. A lower bound can be obtained using the centralized MILP

formulation when PV energy production is assumed to be accurately known for the day. This lower bound is used to evaluate the performance of the proposed decentralized CMAB system (which considers the uncertainty in daily PV production). Also, this MILP formulation belongs to the NP (non-deterministic polynomial time) complexity class, and hence may not be scalable [16].

III. Combinatorial Multi-Armed Bandits Formulation

In CMAB, a combination of base arms (defined as the *super arm*) with unknown distributions is selected. Based on this selection, a reward is observed. The estimated return of each base arm is updated based on the observed reward. Each agent tries to find the best super arm, i.e. the combination that minimizes the cost of the agent, while satisfying the constraints [14]. For smart charging, each day d is divided into $m \in [m]$ equally spaced instants. Each instant $i \in [m]$ acts as a base arm in this CMAB formulation, and is linked to the instantaneous electricity cost $c(i)$.

Congestion (Reward) Model: The transformer agent is modeled to handle global electrical current congestions, while each bus agent is responsible for managing local voltage congestions (voltage limit violations) in the distribution network. As the cause(s) of both types of congestions can be coupled, a collaborative framework among the mentioned agents is proposed, shown in Fig. 1, to tackle all congestions.

First, the transformer agent calculates the instantaneous reward $Rew_l(i)$ for the set of EVs $[E]$, present in the network at instant i. If there is no current congestion in the system (instantaneous current $I(i)$ through the transformer is lower than the rated current I_{rated}), each charging EV gets a positive reward based on the instantaneous electricity price (i.e., $1-c(i)$). In case of congestion ($Cong(i) = 1$), $[X]$ (a set of EV agents uniformly sampled from $[E]$ to avoid congestion) will receive a positive reward, and the remaining elements in $[E]$ will obtain a negative reward. The transformer agent forwards this information to each local bus agent. This current congestion management model is described in Algorithm 1.

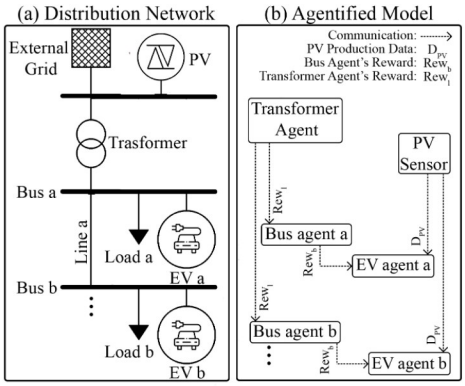

Fig. 1. (a) Sub-section of a distribution network (b) "Agentified model" of the distribution network in (a).

Algorithm 1 Current Congestion Model (Transformer Agent)

Require: I_{rated} := rated line current
Observe electrical current $I(i)$ from the sensor
Observe $[E]$
$i := i$−th instant of the day
if $I(i) < I_{rated}$ **then**
    $Rew_l(i) := (1 - c(i)) \ \forall \ e \in [E]$
    $Cong(i) := 0$
else
    $[X] \sim U(0, E)$
    $Rew_l(i) := (1 - c(i)) \ \forall \ [X] \cap [E]$
    $Rew_l(i) := (-1) \ \forall \ [E] - [X]$
    $Cong(i) := 1$
end if
Forward ($Rew_l(i)$, $Cong(i)$) to each bus agent

The bus reward signal depends on the type of local voltage congestion (over-voltage or under-voltage). This bus reward, $Rew_b(i)$ is calculated by the bus agent. Priority is given to the global electrical current congestion by transferring the reward generated by the current congestion management algorithm directly to the connected EV in case of current congestion in the system. Otherwise, if there is only local voltage congestion, then the generated bus reward is forwarded to the connected EV agent. The functioning of this voltage congestion model for each bus agent is presented in Algorithm 2.

Action Space: The action space of each EV agent consists of two actions $\mathcal{A} = \{0, 1\}$ for each base arm, i.e., an EV agent makes a binary decision of picking (or not picking) each base arm in $[m]$, which corresponds to the EV agent deciding at what instants it will charge from the grid (at rated power) and at what instants it will not charge. It is assumed that super arms follow a linear structure. Then, the expected reward for super arm S can be written as:

$$\mathbb{E}[r(S)] = S^{*\top}\theta \tag{11}$$

where $S^* = \arg\max S^T\theta$, and $\theta \in R^m$ is an unknown parameter. The optimal super arm S^* is obtained when θ is completely known. This unknown vector is learned by each agent using linear Thompson Sampling [17]. It is a Bayesian learning approach, in which the estimation of each element in θ is updated based on the observed reward.

Algorithm 2 Voltage Congestion Model (Bus Aget)

i := i-th instant of the day
$Rew_l(i)$:= Bus agent's i-th instant reward
Observe ($Rew_l(i)$, $Cong(i)$) from Algorithm 1
$Rew_b(i)$:= 1 **if** over-voltage
$Rew_b(i)$:= -1 **if** under-voltage
if ($Rew_b(i)$ = 0 or $Cong(i)$ = 1) **then**
    Forward $Rew_l(i)$ to the connected EV agent
else
    Forward $Rew_b(i)$ to the connected EV agent
end if

The estimated best d-th day super arm S_d (consisting of best instants to charge from the grid), based on the d-th day estimation of the unknown vector $\hat{\theta}_d$ is defined as:

$$S_d = \arg \max_{S \in \{0,1\}^m} S^\top \hat{\theta}_d \tag{12}$$

The pseudo-regret after D days of learning is defined as:

$$\mathbb{E}[R(D)] = \sum_{d=1}^{D} {S^*}^\top \theta - \sum_{d=1}^{D} S_d^\top \hat{\theta}_d \tag{13}$$

Uncertainties: In the tackled problem, there are two sources of uncertainties. First, in the choice of super arms of other agents, which is addressed using selfish Thompson Sampling [12]. Second, in the free energy available through PV production plants connected to the grid side [18]. Here, it is assumed that this PV energy can be utilized by the EVs in the network without any cost. Thus, the EVs should be motivated to learn about the uncertainties in this PV generation (to utilize this free energy while also avoiding congestion in the system). Let $\hat{\phi}_d \in \mathbb{R}^m$ denote the instantaneous estimation of freely available PV energy at day d. This unknown vector is also learned by each agent through Thompson Sampling. Each EV updates its estimation of daily PV production based on the production data obtained from the PV sensors. At each instant i, the EV agent updates the required number of charging instants K_f to achieve the desired SoC SoC_f as follows:

$$k_f = \left\lceil \frac{60 E_{bat}(SoC_f - SoC_s)}{\Delta i P_{max} \eta_{chrg}} - \frac{\sum_{i=t_{start}}^{t_{depart}} \hat{\phi}_i}{P_{max} \eta_{chrg}} \right\rceil - k_p \tag{14}$$

where $\lceil . \rceil$ is the ceiling function. Term SoC_s is EV's SoC at the time of its connection, and K_p is the number of instants EV has already charged from the grid during the day. Here, at each instant i, the EV agent is subtracting the number of already charged instants from the total required grid charging instants (which is the subtraction of the total charging instants required and the estimated free PV charging instants). This proposed linear combinatorial multi-armed bandits with Thompson Sampling (D-LC2AB-TS) algorithm is presented in Algorithm 3, where $\mathbb{1}\{.\}$ is the indicator function, and $||.||_1$ gives the total number of selected base arms.

Remark 1. *The proposed D-LC2AB-TS algorithm (Algorithm 3) is fair and computationally scalable.*

As uniform sampling is done in case of congestion, fairness among all EV agents is maintained. Also, the linear structure of the super arm allows evaluation of the best super arm in $O(m)$, which makes the algorithm computationally scalable.

Algorithm 3 D-LC2AB-TS (EV Agent)

Require: $\alpha \in \mathbb{R}_+, \beta \in \mathbb{R}_+$
$A, Y := I_{m,m}$, $\hat{\theta}, \hat{\phi} := 0_m$, $b, z := 0_m$
for $d = 1, 2, 3, ...$ **do**
  $R := 0_m$, $P := 0_m$, $M := 0_m$, $k_p := 0$
  $\tilde{\theta} \sim \mathcal{N}(\hat{\theta}, \alpha^2 A^{-1})$, $\tilde{\phi} \sim \mathcal{N}(\hat{\phi}, \beta^2 Y^{-1})$
  for $i = 1, 2, 3, ..., m \ \forall \ t_{start} \leq i \leq t_{depart}$ **do**
    Calculate k_f using (14)
    Play S_d using (12) s.t. $\sum_{l>i,d} S_{l,d} = ||S_d||_1 = k_f$
    R_i is received from Algorithm 2
    P_i is the PV sensor data of the i-th instant
    $M_i := \mathbb{1}\{i \in S_d\}$, $k_p := ||M_i||_1$
  end for
  $A := A + MM^\top$; $b := b + R$; $\hat{\theta} := A^{-1} b$
  $Y := Y + I_k I_k^\top$; $z := z + P$; $\hat{\phi} := Y^{-1} z$
end for

IV. Evaluation

A. Simulation Settings

To demonstrate the scalability of the proposed system, two case studies are presented, i.e., a small-scale case study and a large-scale case study. The topologies of both networks are shown in Fig. 2. Each sub-district (SD), in the studied distribution networks, is modeled as the IEEE low voltage test feeder (LVTF) [19]. The small-scale network consists of 55 EVs, whereas there are 10,175 EVs in the large-scale network. The arrival and departure times of the EVs are set based on a real-life dataset [20]. Terms P_{max}, η_{chrg}, $\eta_{dischrg}$, E_{bat}, and SoC_f are set to 7 kW, 0.95, 0.96, 52 kWh, and 0.8. The irradiance data from the national renewable energy laboratory database are used to calculate the instantaneous PV production $P_{PV}(i)$ as [21]:

$$P_{PV}(i) = A \eta_{PV} Irr(i) \tag{15}$$

where $Irr(i)$ is the instantaneous irradiance value, A is the area of the PV panels, and η_{PV} is the efficiency of the PV panels. In the presented studies, the duration of each instant is 1 minute i.e., $m = 1$ in the proposed CMAB formulation. A higher value of m would decrease the system's optimality. In our studied problem, $m < 5$ is generally a good selection, as the optimality gap starts saturating after that. For comparison, the following charging strategies are studied:

- **Basic Charging Strategy:** In this non-optimal charging strategy, the EV starts charging at its rated power as soon as it is plugged-in for charging.
- **Centralized Charging Strategy:** This strategy is presented in section II. In the presented case studies, a perfect daily PV production profile is assumed to be

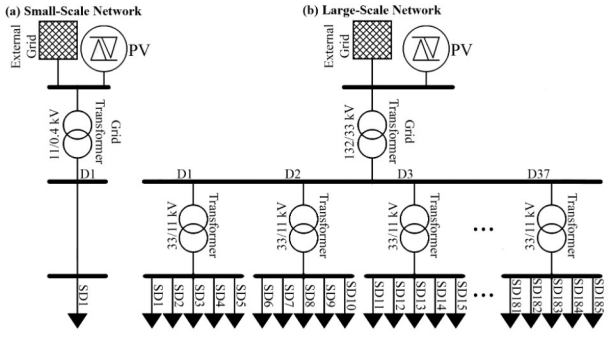

Fig. 2. Topology of the studied (a) Small-scale distribution network (b) Large-scale distribution network.

known, and hence the obtained solution can be considered as the lower bound while evaluating the performance of the proposed decentralized algorithm. It should be noted that due to the uncertainties and variability in the daily PV irradiance, building a perfect forecaster is an extremely challenging task [18]. Hence, the centralized optimal solution would not be realistic for real-life scenarios or for methodologies that consider PV uncertainties (such as the proposed D-LC2AB-TS algorithm).

- **CMAB (no PV estimation) Charging Strategy:** It is a variation of the proposed D-LC2AB-TS algorithm with no PV production estimation i.e., $\phi = 0$. This would highlight the improvement in the performance of the proposed D-LC2AB-TS algorithm, in which the PV production is also learned.
- **D-LC2AB-TS Charging Strategy:** This is the proposed CMAB algorithm, presented in section III.

To compare the fairness among all participating EVs, a set consisting of per-unit charging costs of each EV $[D]$, is calculated as: $\frac{\sum_{i=1}^{m} c(i)P_j(i)\Delta i}{\sum_{i=1}^{m} P_j(i)\Delta i}$. The fairness index value of this set is calculated using the following formula: $\mathcal{F}(D) = \frac{1}{1+(\frac{\sigma_D}{\overline{D}})^2}$. Here, the standard deviation of the set D is denoted by σ_D, and $\overline{D}$ represents the mean value of the set D. This fairness index ranges from 0 (completely unfair i.e., $\sigma_D = \infty$) to 1 (completely fair i.e., $\sigma_D = 0$).

B. Results

1) Small-Scale Study: The average learning rewards (mean of the average rewards of all EV agents in the network) for the D-LC2AB-TS, and the CMAB (no PV estimation) strategies are shown in Fig. 3.

Convergence can be observed within 30 simulation days. The next 30 days are considered as the evaluation period. The daily transformer current and the voltage on the last bus of the network, on the last day of the training, are shown in Fig. 4. In the case of basic charging, both voltage and current constraints are violated due to the peak load demand in the evening. In the case of CMAB (no PV learning), no constraints are violated but the EVs are charging during the early hours of the day (low price instants) and not benefiting from the freely available PV production during the day. The profiles of both the D-LC2AB-TS and the optimal centralized optimization are very similar,

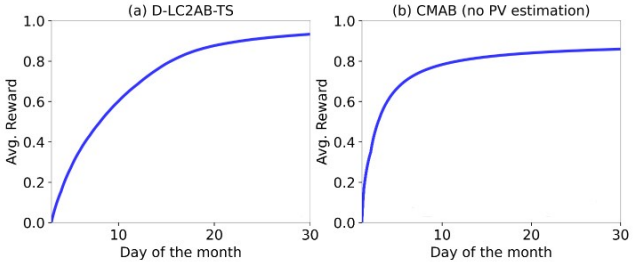

Fig. 3. Average learning reward of the total network for (a) D-LC2AB-TS strategy (b) CMAB (no PV estimation) strategy.

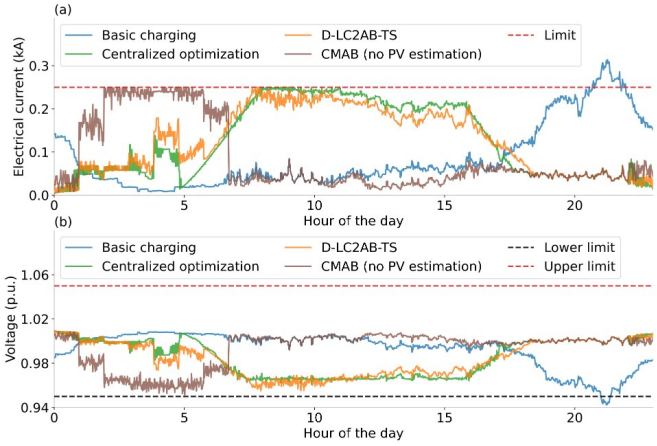

Fig. 4. Profiles for one day (a) Grid transformer current (b) Last bus' voltage.

as shown in Fig. 4. This confirms that the proposed algorithm efficiently utilizes the freely available PV energy production.

The performance of each strategy during the evaluation period is summarized in Table I. Constraints violations are observed in the basic charging strategy, and not in the optimization strategies. All EVs attain the desired final SoC in the studied charging strategies. The optimality gap (% increment compared to the optimal centralized optimization lower bound) is highest for the basic charging strategy. The CMAB (no PV) strategy also remains far from optimal, as it does not utilize the freely available PV production. The proposed D-C2AB-TS strategy learns the available PV production trend and significantly reduces the optimality gap. The fairness index is close to 1 for all the studied optimization strategies.

2) Large-Scale Study: Centralized optimization cannot be performed for the large-scale network due to a large number of agents in the system. However, the D-LC2AB-TS and the CMAB (no PV estimation) strategies work. The average learning rewards of both strategies are shown in Fig. 5.

Similar to the small-scale study, 30 simulation days after the training phase are used for evaluation. Voltage and transformer electrical current violations can be observed in the basic charging strategy in Fig. 6. Whereas, these violations are not present in the proposed D-LC2AB-TS and the CMAB (no PV estimation) strategies.

The D-LC2AB-TS outperforms other strategies in terms

TABLE I
SMALL-SCALE CASE STUDY PERFORMANCE EVALUATION

Charging Strategy	Optimality Gap (%)	Current Constraint Violation (%)	Voltage Constraint Violation (%)	Fairness Index
Basic	187.035	5.417	1.528	-
Centralized	0	0	0	0.999
CMAB (no PV)	133.887	0	0	0.992
D-LC2AB-TS	12.218	0	0	0.994

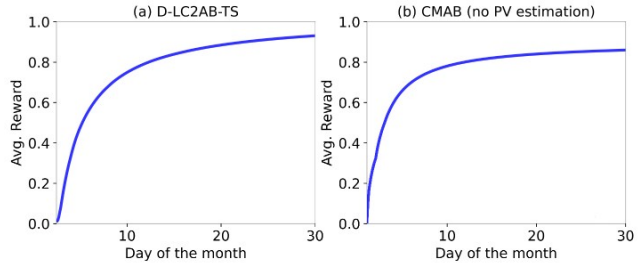

Fig. 5. Average learning reward of the total network for (a) D-LC2AB-TS strategy (b) CMAB (no PV estimation) strategy.

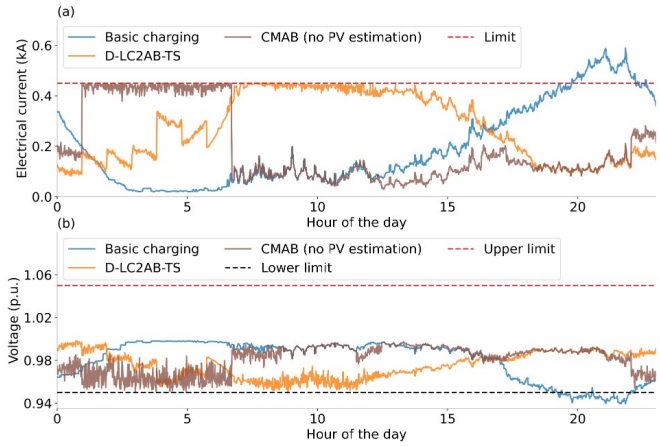

Fig. 6. Profiles for one day (a) Grid transformer current (b) Last bus' voltage.

of cost reduction (reduction compared to the basic charging strategy), as presented in Table II. Fairness index values are also close to 1 for both decentralized optimization strategies. It should be noted that the proposed system is adaptable to the topology of the distribution network. Also, the proposed methodology can be used to manage the current congestion at multiple locations by installing sensors at critical locations.

V. Conclusion

A multi-agent multi-armed bandits system is proposed in this work for smart charging. The mathematical formulation, and the proposed combinatorial multi-armed bandits framework for smart charging are discussed. The experimental evaluations show that the presented D-LC2AB-TS algorithm efficiently manages the uncertainties in the studied problem and performs significantly better compared to the basic charging strategy. The proposed system satisfies all the required constraints while also managing to obtain near-optimal solutions (obtained through centralized optimization). In future works, optimality gap reduction by including PV forecasts as contextual data in the bandits formulation, and comparison with other decentralized control algorithms can be studied.

TABLE II
Large-Scale Case Study Performance Evaluation

Charging Strategy	**Cost Reduction (%)[a]**	**Current Constraint Violation (%)**	**Voltage Constraint Violation (%)**	**Fairness Index**
Basic	0	6.458	11.944	-
CMAB (no PV)	20.711	0	0	0.992
D-LC2AB-TS	77.901	0	0	0.993

[a]Considering basic charging strategy as the reference.

References

[1] L. Bitencourt, B. Dias, T. Abud, B. Borba, M. Fortes, and R. S. Maciel, "Electric Vehicles Charging Optimization Considering EVs and Load Uncertainties," 2019 IEEE Milan PowerTech, 2019, pp. 1-6.

[2] F. Carere et al., "Electric Vehicle Charging Rescheduling to Mitigate Local Congestions in the Distribution System," 2021 IEEE Madrid PowerTech, 2021, pp. 1-6.

[3] M. Habibidoost, and S. M. Taghi Bathaee, "A self-supporting approach to EV agent participation in smart grid," 2018 International Journal of Electrical Power & Energy Systems, 2018, pp. 394-403.

[4] J. Hu, H. Morais, M. Lind, and H. W. Bindner, "Multi-agent based modeling for electric vehicle integration in a distribution network operation," Electric Power Systems Research, 2016, pp. 341-351.

[5] S. Aladdin, S. El-Tantawy, M. M. Fouda, and A. S. Tag Eldien, "MARLA-SG: Multi-Agent Reinforcement Learning Algorithm for Efficient Demand Response in Smart Grid," in IEEE Access, vol. 8, pp. 210626-210639, 2020.

[6] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management," in IEEE Transactions on Smart Grid, vol. 11, no. 4, pp. 3201-3211, July 2020.

[7] R. S. Sutton, and A. G. Barto, "Reinforcement Learning: An Introduction (second ed.)," The MIT Press, 2018.

[8] Y. Liu, A. Halev, and X. Liu, "Policy Learning with Constraints in Model-free Reinforcement Learning: A Survey," in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 4508–4515, Survey Track.

[9] Y. Liu et al., "Privacy-Preserving Context-Based Electric Vehicle Dispatching for Energy Scheduling in Microgrids: An Online Learning Approach," in IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, vol. 6, no. 3, pp. 462-478.

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

[11] L. Besson, and E. Kaufmann, "Multi-Player Bandits Revisited," in Proceedings of Algorithmic Learning Theory (Proceedings of Machine Learning Research, 218 vol. 83, pp. 56–92.

[12] R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot, "Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings,", CoRR, 2018.

[13] H. Dakdouk, E. Tarazona, R. Alami, R. Féraud, G. Z. Papadopoulos, and P. Maillé, "Reinforcement Learning Techniques for Optimized Channel Hopping in IEEE 802.15.4-TSCH Networks", in MSWIM, 2018, pp. 99–107.

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

[15] S. Zafar, V. Maurya, A. Blavette, G. Camilleri, H. B. Ahmed, and M. Gleizes, "Adaptive Multi-Agent System and Mixed Integer Linear Programming Optimization Comparison for Grid Stability and Commitment Mismatch in Smart Grids," 2021 IEEE PES Innovative Smart Grid Technologies Europe, 2021, pp. 01-05.

[16] R. Combes, M. Lelarg, A. Proutière, and M. S. Talebi, "Stochastic and Adversarial Combinatorial Bandits," in CoRR abs/1502.03475, 2015.

[17] S. Agrawal, and N. Goyal, "Thompson Sampling for Contextual Bandits with Linear Payoffs," in CoRR abs/1209.3352, 2012.

[18] H. Ye, B. Yang, Y. Han, and N. Chen, "State-Of-The-Art Solar Energy Forecasting Approaches: Critical Potentials and Challenges," in Frontiers in Energy Research, vol. 10, 2022.

[19] K. P. Schneider et al., "Analytic Considerations and Design Basis for the IEEE Distribution Test Feeders," in IEEE Transactions on Power Systems, vol. 33, no. 3, pp. 3181-3188, May 2018.

[20] Test an EV project, http://smarthg.di.uniroma1.it/Test-an-EV

[21] National Solar Radiation Database, https://nsrdb.nrel.gov/data-viewer