Heendeniya and Nespoli *Energy Informatics* 2022, **5**(Suppl 1):28
https://doi.org/10.1186/s42162-022-00197-5

Energy Informatics

RESEARCH Open Access

A stochastic deep reinforcement learning agent for grid-friendly electric vehicle charging management

Charitha Buddhika Heendeniya* and Lorenzo Nespoli

From The 11th DACH+ Conference on Energy Informatics 2022
Freiburg, Germany. 15-16 September 2022

*Correspondence:
charitha.heendeniya@supsi.ch

Scuola Universitaria Professionale della Svizzera Italiana (SUPSI), Istituto sostenibilità applicata all'ambiente costruito, Via Flora Ruchat Roncati 15, 6850 Mendrisio, Switzerland

Abstract

Electrification of the transportation sector provides several advantages in favor of climate protection and a shared economy. At the same time, the rapid growth of electric vehicles also demands innovative solutions to mitigate risks to the low-voltage network due to unpredictable charging patterns of electric vehicles. This article conceptualizes a stochastic reinforcement learning agent that learns the optimal policy for regulating the charging power. The optimization objective intends to reduce charging time, thus charging faster while minimizing the expected voltage violations in the distribution network. The problem is formulated as a two-stage optimization routine where the stochastic policy gradient agent predicts the boundary condition of the inner non-linear optimization problem. The results confirm the performance of the proposed architecture to control the charging power as intended. The article also provides extensive theoretical background and directions for future research in this discipline.

Keywords: Electric-mobility, Charging control, Voltage management, Optimization, Reinforcement learning, Smart-grid

Introduction

Electrification of the mobility sector is at the top of the decarbonization agenda for many countries. Several countries have already taken policy steps to either heavily restrict or ban internal combustion vehicles within the next decade (Cellina et al. 2021). It also enables further innovations in the transportation sector, such as one-way electric car sharing that further acts in favor of reducing emissions and air pollution (Mounce and Nelson 2019). However, the wide-spread adoption of Electric vehicle (EV)s and Autonomous Electric vehicle (AEV)s and their simultaneous charging may result in increased peak loads, voltage limit violations, sustained under-voltage conditions, and supply imbalances (Dubey and Santoso 2015).

Algorithms such as rule-based approaches (e.g., Rauf and Salam (2018)), heuristics (e.g., Alonso et al. 2014), and central optimization methods (e.g., Richardson et al. 2011; Sun et al. 2018) have been tested to achieve the goal of effective EV charging management. However, in the presence of high stochasticity and the absence of perfect foresightedness, the methods mentioned above cannot converge to the optimal charging behavior (Abdullah et al. 2021). As a result, there has been an increasing interest towards more flexible data-driven approaches to model and manage the EV charging process. Data-driven approaches can be used without assumptions regarding the underlying model. They are also capable of representing the inherent stochasticities in the environment and consequently suggest probabilistic strategies that perform better than deterministic strategies over long time horizons even in adversarial settings (Wang et al. 2016).

This article presents a method and a case study that demonstrate the application of Deep reinforcement learning (DRL) to control the charging power at an AEV charging node. We demonstrate the capability of DRL to learn the optimal charging policy in a highly stochastic environment with multiple charging objectives. Moreover, we derive the state vectors based on the observations that are readily available through standard metering infrastructure in a Low-voltage (LV) network and perform online learning via policy gradient update. The main contributions of the article are as follows.

- We present a DRL solution based on the actor-critic architecture to regulate charging power at an AEV charging node considering both minimizing charging time and voltage limit violations.
- The proposed solution makes use of voltage magnitude measurements from standard metering infrastructure and learns a stochastic policy that is optimal in the limit as $time \rightarrow \infty$.
- To improve the scalability of the method to much larger use-cases, we impose partial observability in the form of a local actor with a global critic.
- We present a case study and based on our results, discuss the broader implications of AEV charging and the potential for future research work.

State of the art

Application of Reinforcement learning (RL) in the electro-mobility domain has attracted a lot of interest recently, leading to several published use-cases such as charging load forecasting (e.g., Zhang et al. 2021; Zhu et al. 2019), fleet assignment (e.g., Shi et al. 2020), charging station recommendation (e.g., Blum et al. 2021), and charging management (e.g., Chang et al. 2019; Wan et al. 2019; Ding et al. 2020; Dorokhova et al. 2021).

Table 1 shows the summary of some exemplary studies using RL for AEV charging management. We see that the temporal resolution of the previous RL studies related to electro-mobility in Table 1 is in the hourly range. Indeed, the choice of temporal resolution depends mainly on the modeling objective. However, (Bucher et al. 2013) studied the effect of temporal-averaging in the context of LV power systems and recommends one-minute resolution for studies that make use of steady-state voltages and power flows.

Table 1 Summary of some exemplary studies using RL for EV charging management

References	Temporal resolution	Objective	Stochastic policy	Voltage violation	Method
Chang et al. (2019)	30 mins	Cost, expected SOC at the end	No	No	Q-learning
Wan et al. (2019)	1 h	Cost, incl. battery degradation	No	No	DQN
Ding et al. (2020)	1 h	DSO profits	Yes	Yes	DDPG
Dorokhova et al. (2021)	1 h	PV self consumption	No	Yes	DDQN, DDPG, PDQN

The authors in the past have applied both the value-based [e.g., Q-learning, Deep Q-network (DQN), Deep double Q-network (DDQN)] and policy-based [e.g., Deep deterministic policy-gradient (DDPG)] RL methods to solve for an optimal charging strategy. (Abdullah et al. 2021) presents a review of RL-based charging management strategies available in the published literature.

In a nutshell, in value-based methods, the agent learns an approximate Q-value function through continuous interaction with the environment. Often the Q-value function is represented as a kernel function or a parameterized function approximator like a neural network. As such, the learning process converges when we iteratively update the parameters to minimize the error between the predicted and target Q-values. Policy-based methods (synonymously referred to as policy-gradient methods), by contrast, directly learn the optimal policy (denoted by $\pi^{\star}$). Policy gradient methods have shown better convergence properties compared to value-based methods (Sutton et al. 1996); they are often capable of handling imperfect state information and able to learn stochastic policies (Peters and Bagnell 2016; Sutton et al. 1996), which are more robust than deterministic policies. One key drawback of policy-gradient methods is their sample complexity (Peters and Bagnell 2016). However, (Wang et al. 2016) shows that experience replay, first introduced during the early stages of RL, can significantly improve the sample efficiency of policy-gradient problems as well.

Actor-critic is a family of policy-gradient algorithms where two function approximators (the critic and the actor) are used simultaneously to learn the value function and optimal policy. The actor-network is a parameterized representation of the agent's current policy π. At each iteration, the agent takes an action based on the state of the environment, and its current policy, i.e., $a^t = \pi(s^t)$. The critic evaluates the value of the action at the given state and updates the value function's parameters using a temporal difference update. Finally, the actor updates the policy in the policy-gradient direction, calculated using the critic's value estimate.

There are a variety of policy gradient algorithms published in the literature. The algorithm used in our case study is called Proximal policy optimization (PPO), which was first published in 2017 (Schulman et al. 2017). The main advantages of PPO are its simplicity and general applicability. Moreover, PPO is an off-policy learning algorithm and it is sample efficient. The original implementation of the PPO algorithm demonstrated superior performance in solving tasks with high-dimensional continuous action spaces such as half-cheetah and running humanoid robot (Schulman et al.

2017). We provide a brief mathematical introduction of the PPO algorithm for the benefit of the reader in the paragraph below.

Proximal policy optimization According to the policy gradient theorem, the gradient of a stochastic policy objective J with respect to the policy parameter θ is given by $\nabla J(\theta) = \mathbb{E}\,\nabla_\theta \log \pi_\theta(s \mid a) Q_\pi(s \mid a)$ (Sutton and Barto 2018). A state-of-the-art way to reduce the variance of the policy gradient is to subtract a baseline function that does not depend on the action to not introduce bias. A common baseline function is the value function and then we can rewrite the gradient as $\nabla J(\theta) = \mathbb{E}\,\nabla_\theta \log \pi_\theta(s \mid a) A(s \mid a)$ where $A(s \mid a) = Q_\pi(s \mid a) - V(s)$.

The convergence stability of policy gradient algorithms depends on the iterative gradients updates on the policy parameters. PPO is a trust-region method that uses a clipped surrogate objective that penalizes excessively large policy parameter updates (Schulman et al. 2017).

$$r^t(\theta) = \frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)} \tag{1}$$

$$J^{\text{CLIP}}(\theta) = \mathbb{E}\left[\min(r^t(\theta) A^t_{\theta_{\text{old}}}(s,a), \text{clip}(r^t(\theta), 1-\epsilon, 1+\epsilon) A^t_{\theta_{\text{old}}}(s,a))\right] \tag{2}$$

$r^t(\theta)$ in Eq. 1 is the probability ratio between the new policy and the old policy. The clipped surrogate objective (in Equation 2) clips the probability ratio outside the interval $[1-\epsilon, 1+\epsilon]$ where ϵ is a hyper-parameter (Schulman et al. 2017).

Problem formulation

Our objective is to regulate the AEV charging power to minimize the charging time and voltage limit violations at the charging node. The power flow equations describe the relationship between power and voltage in an electrical distribution network. For simplicity, we do not consider reactive power control in our use case. However, it is important to note that the European LV grid benchmark has R/X ratios of 0.7–11.0 (Ayaz et al. 2018), which are relatively high, and at high R/X ratios, active power has the most significant influence on voltage (Blažič and Papič 2008).

The mathematical form of the objective function is given by Eq. 3. In Eq. 3, P_{max} is the maximum charging load (maximum charging power of a charging point times the number of charging points at the node), $\alpha^t = P^t_c / P_{max}$ is the ratio between the charging load at time t and P_{max}, $\mathcal{N}$ is the set of nodes in the LV grid, V_m is the voltage magnitude at the charging node, and V_{lb} is the statutory voltage limit. G_{ij} and B_{ij} are the real and imaginary parts of the bus admittance matrix corresponding to the $(i,j)^{th}$ element. δ_{ij} is the voltage angle difference between the i^{th} and j^{th} buses. P_i and Q_i are the real and reactive power injections at node i.

$$\begin{aligned}
\max_{\alpha^t} \quad & \mathbb{E}_{t\in\mathcal{T}}\left(\mathbb{1}^{|V_m^t - V_{lb}|\leq\zeta} + \mathbb{1}^{V_m^t > V_{lb}+\zeta}\alpha^t\right) \\
\text{s.t.} \quad & P_i^t = \left|V_i^t\right| \sum_{j\in\mathcal{N}} \left|V_j^t\right| \left(G_{ij}\cos\delta_{ij}^t + B_{ij}\sin\delta_{ij}^t\right) \\
& Q_i^t = \left|V_i^t\right| \sum_{j\in\mathcal{N}} \left|V_j^t\right| \left(G_{ij}\sin\delta_{ij}^t - B_{ij}\cos\delta_{ij}^t\right) \\
& P_c^t = \alpha^t P_{max}
\end{aligned} \tag{3}$$

Equation 3 is a concise way to combine both the charging power and voltage objectives. The statutory voltage limit is imposed as a soft constraint with a small allowable margin of error of ζ. In the case study, we set V_{lb} and ζ to 0.95 and 0.01 respectively.

To solve the optimization problem with DRL, we need to define the states, actions, and reward function of the RL agent. Moreover, we employ the stochastic policy gradient approach that enables us to create an RL agent that learns the optimal stochastic policy directly from observations.

States We define two state vectors, one for the critic and another for the actor. The state vector of the actor is a local subset of the state vector available to the critic, which imposes the partial observability condition.

Critic's state, denoted by S_c, is a discrete-transformed vector of voltage magnitudes (in p.u.) at each load and generator-connected bus. Given the number of load-connected buses is m, n is the number of generator-connected buses, and l is the number of bins, the critic state at time t is a vector of the shape $(1, m+n, l)$. We impose partial observability by limiting the actor's state to the b nearest load or generator buses from the charging node, including the charging node itself. Therefore, the actor state is a vector of the shape $(1, b, l)$[1].

Transformation from continuous to a finite discrete domain is a simple but powerful state abstraction that reduces the size of the state space, improves convergence, and improves generalization properties of the model to unseen data. We recommend (Kirk et al. 2021) for more information related to the generalization of DRL models.

Actions The agent's policy yields an action at each time-step t that regulates the charging power. Therefore, we define the action of the stochastic charging agent as $\alpha^t = \pi(s^t)$. Clearly, α^t is a real value in the range [0, 1] that can be represented as a random realization of a beta policy, i.e., $\alpha^t \sim Beta(a, b)$. In other words, we can write the optimal stochastic policy $\pi^\star = Beta(a^\star, b^\star)$ where $a^\star, b^\star$ are the optimal parameter values of the beta policy.

Reward function Reward functions require careful engineering. Efficient reward functions help guide the RL agent find the optimal policy by avoiding local optimal and improving the convergence speed (Dorokhova et al. 2021). Our problem has multiple objectives that should simultaneously minimize charging time and expected voltage violations. Therefore, following Eq. 3, we define the reward function as;

[1] In our case study, the bins are equally spaced in the interval [0.9, 1.1). An additional bin accounts for any voltage magnitude less than 0.9, i.e., l = 42. Moreover, we set b = 3, ensuring observations from both an upstream and a downstream node, whenever possible. Values of m and n are introduced with the case study.

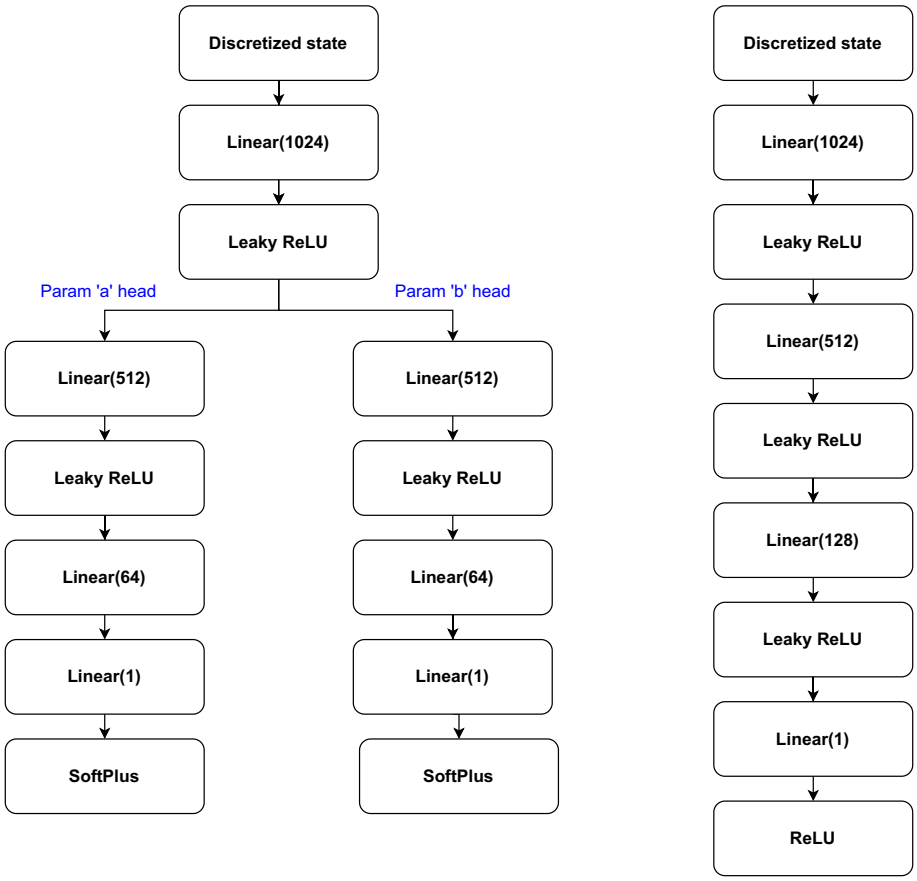

Fig. 1 The deep neural network architecture of the actor (left) and critic (right). Parameter heads *a* and *b* refers to the two branches of the actor network that returns the beta policy parameters *a* and *b*

$$R(s^t, a^t) = \mathbb{1}^{|V_m - V_{lb}| \leq \zeta} + \alpha^t \left(\mathbb{1}^{V_m > V_{lb} + \zeta} \right) \tag{4}$$

Model architecture The policy network parameterizes the stochastic charging policy π_θ that returns policy parameters a and b of a beta distribution. Beta distribution is a bounded distribution between 0 and 1; therefore, it is well-suited for representing the stochastic charging action $\alpha^t = \pi_\theta(s^t)$ of the agent. We encourage the reader to refer to the motivating examples (Chou et al. 2017; Petrazzini and Antonelo 2022) that describe the use of beta policy for solving policy gradient problems with bounded action spaces.

Value-network (the critic) is updated based on the mean-squared error (MSE) of the critic prediction and the immediate true reward. In other words, the agent's interactions with the environment at each time-step is an episode consisting of only one step. Moreover, we implement a replay-buffer to improve the sample efficiency of the training process.

The architectures of the deep neural network that implement the actors and the centralized critic are depicted in Figure 1. The actor-network has two heads corresponding to the two parameters of the beta distribution of the stochastic policy that we need to estimate. The number of layers, layer dimensions, and layer activation functions are design choices based on hyper-parameter tuning.

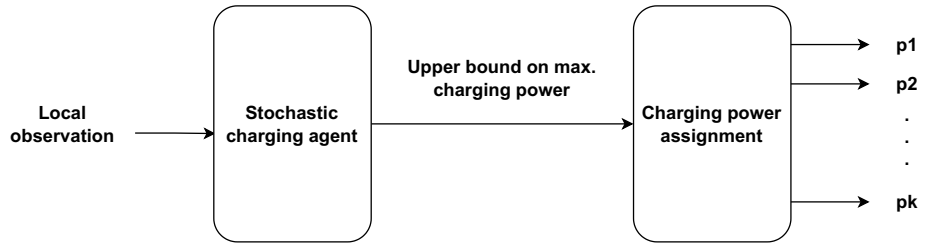

Fig. 2 Flow diagram that shows the interconnection of outer and inner optimization problems at a given time step. We run this process in iteration for each time step of the simulation. *p*1... *pk* in the figure are the charging power at each charging point for a given time step, which can be also written as $p^k = \alpha^{k,t=t_s} P^k_{max}$

Charging power assignment So far, we have designed a mathematical formulation that enables us to optimally control the total charging power at a node minimizing expected voltage violations and charging time. The assignment problem that we discuss now answers the question of the equitable allocation of the total charging power between the multiple vehicles that require charging simultaneously. We define equity as minimizing the sum of instantaneously evaluated charging times for all vehicles. This definition allows us to prioritize more depleted AEVs and charge them faster. Consequently, we expect more AEVs to be available for users, leading to better mobility services. The nonlinear optimal power assignment problem can be written as in Eq. 5, where K' is the set of active charging points. Furthermore, $\alpha^{k',t}$ is the charge rate of the charging point k' at time t, and it is a real value in the range $[\epsilon, 1]$. The lower-bound ϵ is a very small real value introduced for numerical stability.

$$\begin{aligned} \min_{\alpha^{k',t}} \; & \frac{1 - SOC^{k',t}}{\alpha^{k',t} + \epsilon} \\ \text{s.t. } & 0 \leq \alpha^t P_{max} - \sum_{k' \in K'} \alpha^{k',t} P^{k'}_{max} \\ & \alpha^{k',t} \leq \epsilon \quad if SOC^{k',t} = 1 \\ & \epsilon \leq \alpha^{k',t} \leq 1 \end{aligned} \tag{5}$$

Figure 2 shows the combined optimization problem that we solve in iteration for each time step of the simulation.

Case study

To demonstrate the concept and methodology described earlier in the context of a shared taxi fleet, we set up a synthetic example using both real and synthetic data.

The case study consists of 216 trips within the Swiss municipality Lugano within a day. The travel data is synthetically generated using MATsim (http://www.matsim.org), an agent-based micro-simulation framework for mobility systems simulations (Horni et al. 2016). The road network extracted from OpenStreetMaps as a graph contains all roads and links in Lugano with the importance level either residential or higher. The metadata includes distance and maximum travel speed for each edge of the graph. The resulting network has 1122 nodes and 3602 edges.

To simulate the power system impacts, we use a modified CIGRE LV benchmark grid (Fig. 3) with representative residential load profiles. The environment consists of

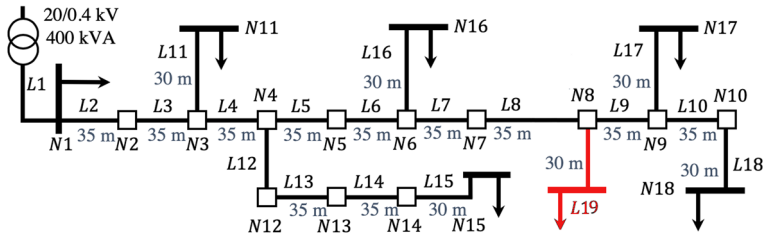

Fig. 3 Modified CIGRE LV grid used in the case study. The charging stations are connected at the bus R19. Figure adapted from (CIGRE'Task Force C6.04.02)

one charging station with 11 charging points connected to the charging node (L19 in the CIGRE benchmark grid). Each charging point has a maximum charging power of 11 kW. The aggregate residential load profiles are obtained by simulating typical household appliances and devices (heat pumps and boilers, rooftop Photovoltaic (PV) generation, and non-dispatchable demand). The medium-voltage side of the transformer is connected to a constant slack bus. Given that we want to observe the effect of the charging controller in isolation, we deactivate the transformer tap changer in our simulations.

The simulated appliances and the corresponding modeling methods are as follows:

- Heat-pump and boilers: To obtain a representative dataset for Switzerland, we used the STASCH6 standard (Afjei et al. 2002) and its variants as a reference for the heating system and the control logic. The STASCH6 standard comprehends three main components: a heat-pump, a water tank used as an energy buffer, and a heating element delivering heat to the building. The heat-pump control logic is based on two temperature sensors placed at different heights of the water tank, while the circulation pump connecting the tank with the building's heating element is controlled by an hysteresis on the temperature measure by a sensor placed inside the house. More details on the hydronic system modeling can be fund in (Nespoli 2019). The models' parameters, as households equivalent thermal resistance and capacitance, were tuned using data from a local pilot project, the Lugaggia Innovation Community (LIC)[2].
- Rooftop-mounted PV power plants: These were modeled using the Sandia National Laboratories' PV Collaborative Toolbox (Stein 2012), using typical inverter data. Data for the type of panels, inclinations and nominal power were taken from LIC.
- Non-dispatchable demand: Non-dispatchable demand was modeled using the Load Profile Generator tool[3], which uses a full behavioural modeling approach to generate residential load profiles. As an input of the tool we have used the same typical meteorological year used to generate the PV power plant profiles and as an input to the households' thermal models.

Note that the input to the simulation model are aggregate profiles. Consequently, the power flow model of the LV grid consists of only load (PQ) buses and we set the

[2] https://lic.energy/

[3] https://loadprofilegenerator.de/

Table 2 Hyper-parameters of the PPO model

Hyper-parameter	Value
Layers and layer dims.	Figure 1
Activation functions	Figure 1
Learning rate	Actor: 1×10^{-6}
	Critic: 1×10^{-5}
Loss function	Actor: Eq. 2
	Critic: MSE
Optimizer	Adam
ϵ	0.2
Batch size	64
Soft update rate	0.001

parameters m and n introduced in section Problem formulation to seven and zero, respectively.

A discrete-time simulation environment with one minute time resolution based on SimPy (Matloff 2008) is developed to simulate the fleet of shared AEV servicing the travel requests. The fleet consists of 11 AEVs, and they are randomly located at the start of the simulation. A python generator pops a travel request when the environment time reaches the start time of a trip. A free AEV can accept that request and initiate a series of processes to service the request by (1) routing to the pickup location, (2) picking up the customer, and (3) routing to the destination. En route, an AEV can decide to charge the batteries if it senses a chance of battery depletion. Similarly, an AEV can leave the charging station during the charging process when it senses sufficient State of charge (SOC) to serve an incoming travel request. The routing is based on the shortest path algorithm, weighted by the travel time. "Go to charge" is a binomial decision based on the current SOC.

The training dataset consists of 20 days of residential load profiles covering all four seasons of the year. We add a small Gaussian noise to each residential load profile during model training to assist the stochastic charging agent to learn from similar but not identical observations at each iteration. The customer travel demand profile is identical in each day. The validation dataset consists of 10 days of residential load profiles (without added Gaussian noise) and the customer demand profile identical to the one in the training data. The model training is performed in batches of 64 randomly sampled observations from the replay buffer.

The Table 2 describes the set of hyper-parameters used in the PPO model.

Results

In this section, we present the results of the simulations we carried out and compare the performance of the stochastic RL charge controller with a simple benchmark controller. The benchmark controller is one that regulates charging power based on a droop strategy given by the function below. ζ is set to 0.01 in our case study.

Fig. 4 Performance of the contextual stochastic charging agent in terms of the expected reward function. The variance bounds are set at $\pm\sigma$ distance from the mean reward

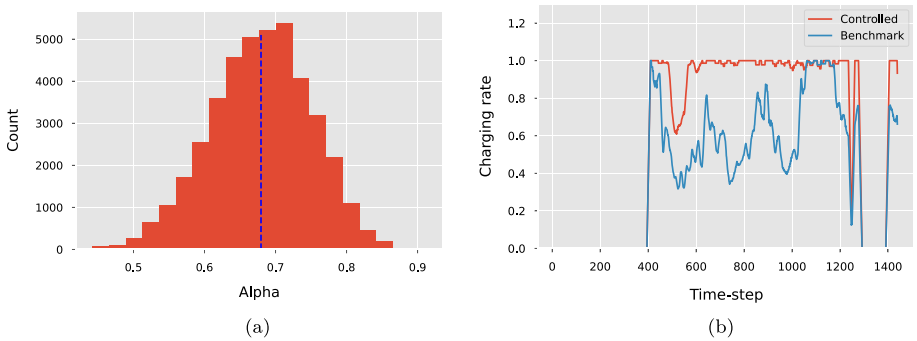

Fig. 5 **a** The distribution of the stochastic charging agent's predictions over the validation period, **b** The charging power upper-bound forecasted by the stochastic agent for one validation day

$$\alpha^t = \begin{cases} \frac{V_m - V_{lb} - \zeta}{V_{max} - V_{lb} - \zeta} & V_{lb} - \zeta \leq V_m \\ 0.5 & otherwise \end{cases}$$

After running the training loop over 15 epochs, we observe a relatively smooth convergence of the stochastic charging agent as shown in Fig. 4. The variance bounds indicate variability of the expected reward that is high at the start of the training and then stabilizes at roughly 0.1 after 15 epochs. Note that we stopped agent training after 15 epochs, although even longer training time could have resulted in tighter variance bounds.

The stochastic charging agent predicts a charging power upper bound with a mean of approximately 68% of the maximum charging power of the station (Fig. 5a).

The peak shaving effect takes place only at specific times of the day when the charging power demand exceeds the upper bound forecast of the stochastic charging agent, as shown in Fig. 6a. We also observe, in comparison to the benchmark strategy, that the stochastic charging agent enforces higher charging rates when possible (Fig. 5b). The voltage impact of peak-shaving is depicted in Fig. 6b. Over the 10-day validation period, the stochastic control strategy results in 17 instances of voltage dead-band violations (0.1% of the total observed time steps), whereas the benchmark strategy

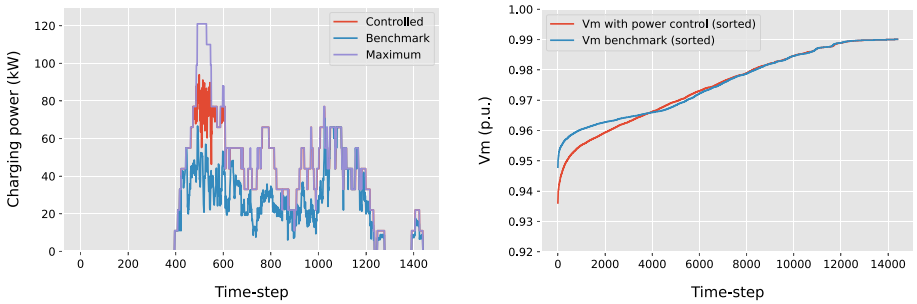

Fig. 6 **a** The peak shaving effect of the stochastic charging agent, **b** The voltage magnitudes at the charging node with and without charging control over the 10 day validation period, sorted in the ascending order

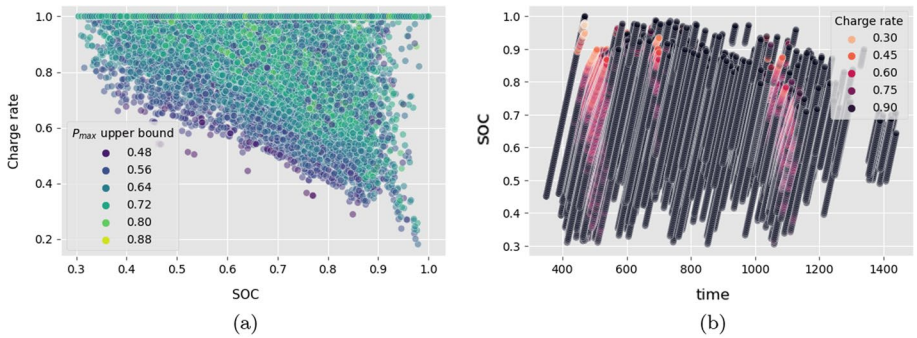

Fig. 7 **a** Relationship between the SOC and charging rates, **b** Charging trajectories of the EVs. Observe that there is a reduction of the charging rate as the SOC of the vehicle increases, particularly when the charging demand is high (e.g., between time steps 400 and 600)

results in zero violations. However, the proposed strategy provides a 7.4% extra charging rate during the same period, on average. Furthermore, between the peak charging times (time steps 400–1000 of each day), the proposed strategy provides an additional 39.07% average charging rate compared to the benchmark strategy.

Figure 7a is a graphical depiction of how the SOC, charging rates, and α^t are related to each other. Firstly, we observe that the charging rates increase when the SOC are lower, which is the expected behavior of the inner optimization. However, the sensitivity of this relationship is governed by α^t. If the constraint is strict (low α^t), the charging rate becomes more sensitive to the changes in SOC. Conversely, if the charging power constraint is lenient, the sensitivity of the charging rate to SOC gets lower.

The charging trajectories (profiles) describe the change of SOC of a vehicle over time (Fig. 7b). Due to the negative dependency of the charging rates on SOC, the charging profiles of the AEVs are, by default, non-linear. Charging trajectories can progress linearly only when the total charging power requirement is less than the constraint set by the stochastic charging agent and as the SOC increases towards 100%, the charging rate slows down. The non-linearity of the charging profiles exacerbates when the charging power constraint is more stringent, for example, between time steps 400–600. Figures 6a, 7b jointly enable us to visualize that when charging power demand is higher, there is more non-linearity in charging trajectories.

As a result, as the SOC of a vehicle increases beyond a certain threshold, it may become unproductive for an AEV to remain connected to the charging point, given the diminishing charging rates. As a result, this behavior provides an additional degree of freedom for intelligent decision-making and optimization. For example, we can argue that in a sharing economy, it is much better to have two vehicles at 70% SOC levels than to have one vehicle fully charged and the other one at, say, 40%. The additional degree of freedom encourages faster turnover of vehicles and can improve the use of limited charging resources. While we do not address this question in the current article, we would like to present it to the research community as a promising area to investigate.

Conclusion

This article presents a policy gradient RL based strategy to solve the optimal electric-vehicle charging problem considering both charging rates and voltage violations. We formulate the problem as an optimization problem with two levels. To solve the outer-level optimization problem, we train a stochastic agent using PPO. The inner-level is a non-linear optimization problem, subject to the boundary condition evaluated by the PPO agent. The case study presented in the article serves as a proof of concept for the applicability of stochastic RL controllers for AEV charging management in a smart-grid.

Comparison against the benchmark controller with a droop strategy illustrates that both control schemes can shave the peak demand and manage statutory voltage limit violations. In addition, the stochastic RL controller also optimizes the charging rate, reducing the total charging time. However, we observe some instances (0.1% of the entire time duration) when the statutory voltage limit gets violated under the stochastic RL control scheme. This observation highlights the critical detail that due to the probabilistic nature of decision making, there is a non-zero chance for a stochastic RL agent to make a decision that leads to an undesirable state. Since our case study is not safety-critical, we can allow a small number of instances when the voltage constraint is violated. But, it is an essential consideration for integrating stochastic RL controllers in weaker grids, which require further investigation.

There is a multitude of open research avenues extending from our work. One apparent future step is to investigate the impacts of stochastic charging control under different circumstances, such as fast charging and more complex grid topologies. Moreover, estimating the benefits to the upstream network, especially under different formulations of the control objective, is also a promising avenue for future research. Such problems are challenging for the learning process of the PPO agent, which may call for better feature extraction and state-space representations.

From an algorithmic and architectural viewpoint, understanding the benefits and drawbacks of different RL model architectures in high-resolution and partially observable environments has many practical advantages. Most current work focuses on prediction problems at low temporal resolutions. However, applying RL for real-time control problems in the smart-grid domain requires robust models that handle highly stochastic time series data.

Optimal control of AEV charging has broader consequences. If appropriately designed optimal charge controllers can be used to improve energy security, quality of

mobility services, economic efficiency, and social equity, as pointed out in the case study. However, as of now, the energy, social, and economic nexus of AEV management and control is a largely untouched topic.

We believe that such research directions have tremendous value because while the smart-grid future is at our doorstep, we often need to build solutions with technical, economic, and social relevance based on partial data.

Acknowledgements
The authors wish to acknowledge the contribution of Matteo Salani of Dalle Molle Institute for Artificial Intelligence (IDSIA) at USI and SUPSI and Clarissa Livingston of Institute for Transport Planning and Systems, ETH Zürich for providing MATsim generated transportation data. We also acknowledge valuable comments from Fabrizio Sossan of MINES ParisTech—PSL University.

About this supplement
This article has been published as part of Energy Informatics Volume 5 Supplement 1, 2022: Proceedings of the 11th DACH+ Conference on Energy Informatics. The full contents of the supplement are available online at https://energyinformatics.springeropen.com/articles/supplements/volume-5-supplement-1.

Author contributions
Conceptualization: CBH; Methodology: CBH, LN; Modeling, testing, and validation: CBH; Writing: CBH, LN; Review and Editing: CBH, LN. All authors read and approved the final manuscript.

Funding
This project has received funding in the framework of the joint programming initiative ERA-Net Smart Energy systems' focus initiative Integrated Regional Energy Systems, with support from the European Union's Horizon 2020 research and innovation program under grant agreement No. 775970, in the context of the EVA project.

Availability of data and materials
The data sources used for the production of this article are described in the section Case study. The models developed within the context of this article and EVA project are not available via open-source channels at this time. Following software libraries and packages are used for modeling and simulation work: PyTorch (Reinforcement learning) (Paszke et al. 2019), SimPy (Discrete-event simulation) (Matloff 2008), Pandapower (Steady-state power flow simulation) (Thurner et al. 2018), and GEKKO (optimization) (Beal et al. 2018).

Declarations

Competing interests
The authors declare that they have no competing interests.

Published: 7 September 2022

References

Abdullah HM, Gastli A, Ben-Brahim L (2021) Reinforcement learning based EV charging management systems—a review. IEEE Access 9:41506–41531

Afjei T, Schonhardt U, Wemhöner C, Erb M, Gabathuler HR, Mayer H, Zweifel G, Achermann M, von Euw R, Stöckli U (2002) Standardschaltungen für Kleinwärmepumpenanlagen Teil 2: Grundlagen und Computersimulationen. Schlussbericht, Technical report

Alonso M, Amaris H, Germain JG, Galan JM (2014) Optimal charging scheduling of electric vehicles in smart grids by heuristic algorithms. Energies 7(4):2449–2475

Ayaz MS, Azizipanah-Abarghooee R, Terzija V (2018) European LV microgrid benchmark network: Development and frequency response analysis. 2018 IEEE International Energy Conference, ENERGYCON 2018, 1–6

Beal L, Hill D, Martin R, Hedengren J (2018) Gekko optimization suite. Processes 6(8):106

Blažič B, Papič I (2008) Voltage profile support in distribution networks - Influence of the network R/X ratio. 2008 13th International Power Electronics and Motion Control Conference, EPE-PEMC 2008, 2510–2515

Blum C, Liu H, Xiong H (2021) CoordiQ: Coordinated Q-learning for Electric Vehicle Charging Recommendation

Bucher C, Betcke J, Andersson G (2013) Effects of variation of temporal resolution on domestic power and solar irradiance measurements. 2013 IEEE Grenoble Conference PowerTech, POWERTECH 2013 (June 2011)

Cellina F, Bettini A, Eva D, Rudel R (2021) Literature review regarding future mobility scenarios. Technical report, SUPSI, Mendrisio. https://evaproject.eu/wp-content/uploads/2021/04/EVA_D31.pdf

Chang F, Chen T, Su W, Alsafasfeh Q (2019) Charging Control of an Electric Vehicle Battery Based on Reinforcement Learning. 10th International Renewable Energy Congress, IREC 2019 (March) (2019)

Chou PW, Maturana D, Scherer S (2017) Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. 34th International Conference on Machine Learning, ICML 2017 2, 1386–1396

CIGRE' Task Force C6.04.02: Benchmark systems for network integration of renewable and distributed energy resources. Technical report, CIGRE International Council on large electric systems (July 2009)
Ding T, Zeng Z, Bai J, Qin B, Qin B, Yang Y, Shahidehpour M, Shahidehpour M (2020) Optimal electric vehicle charging strategy with markov decision process and reinforcement learning technique. IEEE Transactions on Industry Applications
Dorokhova M, Martinson Y, Ballif C, Wyrsch N (2021) Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation. Appl Energy 301(August):117504
Dubey A, Santoso S (2015) Electric vehicle charging on residential distribution systems: impacts and mitigations. IEEE Access
Horni A, Nagel K, Axhausen KW (2016) Introducing MATsim. Ubiquity Press, London
Kirk R, Zhang A, Grefenstette E, Rocktäschel T (2021) A survey of generalisation in deep reinforcement learning, 1–43. 2111.09794
Matloff N (2008) Introduction to discrete-event simulation and the simpy language. Davis, CA. Dept of Computer Science. University of California at Davis. Retrieved on August 2(2009), 1–33
Mounce R, Nelson JD (2019) On the potential for one-way electric vehicle car-sharing in future mobility systems. Transp Res Part A Policy Pract 120:17–30
Nespoli L (2019) Model based forecasting for demand response strategies. PhD thesis
Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, KilleenT, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A,Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S,Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperativestyle, high-performance deep learning library. In: Wallach, H.,Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E.,Garnett, R. (eds.) Advances in Neural Information Processing Systems32, pp. 8024–8035. Curran Associates, Inc., http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
Peters J, Bagnell JA (2016) Systems, I.A.: Policy Gradient Methods, 1–4
Petrazzini IGB, Antonelo EA (2022) Proximal policy optimization with continuous bounded action space via the beta distribution, 1–8
Rauf A, Salam Z (2018) A rule-based energy management scheme for uninterrupted electric vehicles charging at constant price using photovoltaic-grid system. Renewable Energy 125:384–400
Richardson P, Flynn D, Keane A (2011) Optimal charging of electric vehicles in low voltage distribution systems. IEEE Trans Power Syst 27(1):268–279
Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms, 1–12
Shi J, Gao Y, Wang W, Yu N, Ioannou PA (2020) Operating electric vehicle fleet for ride-hailing services with reinforcement learning. IEEE Trans Intell Transp Syst 21(11):4822–4834
Stein JS (2012) The photovoltaic Performance Modeling Collaborative (PVPMC). Conference Record of the IEEE Photovoltaic Specialists Conference, 3048–3052
Sun B, Huang Z, Tan X, Tsang DHK (2018) Optimal scheduling for electric vehicle charging with discrete charging levels in distribution grid. IEEE Trans Smart Grid 9(2):624–634
Sutton RS, Barto AG (2018) Reinforcement learning, Second Edition: an introduction, 2nd edn Adaptive Computation and Machine Learning series, 2nd edn. MIT Press, Massachusetts
Sutton RS, Mcallester D, Singh S, Mansour Y, Avenue P, Park F (1996) Policy gradient methods for reinforcement learning with function approximation. Adv Neural Inf Proc Syst 12
Thurner L, Scheidler A, Schäfer F, Menke J-H, Dollichon J, Meier F, Meinecke S, Braun M (2018) Pandapower-an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. IEEE Trans Power Syst 33(6):6510–6521
Wan Z, Li H, He H, Prokhorov DV (2019) Model-free real-time EV charging scheduling based on deep reinforcement learning. IEEE Transactions on Smart Grid
Wang Z, Bapst V, Heess N, Mnih V, Munos R, Kavukcuoglu K, de Freitas N (2016) Sample efficient actor-critic with experience replay
Zhang X, Chan KW, Li H, Wang H, Qiu J, Wang G (2021) Deep-learning-based probabilistic forecasting of electric vehicle charging load with a novel queuing model. IEEE Trans Cybern 51(6):3157–3170
Zhu J, Yang Z, Mourshed M, Guo Y, Zhou Y, Chang Y, Wei Y, Feng S (2019) Electric vehicle charging load forecasting: a comparative study of deep learning approaches. Energies 12(14):1–19

Publisher's Note