Letter

Intelligent Electric Vehicle Charging Scheduling in Transportation-Energy Nexus With Distributional Reinforcement Learning

Tao Chen and Ciwei Gao

Dear Editor,

This letter is concerned with electric vehicle (EV) charging scheduling problem in transportation-energy nexus using an intelligent decision-making strategy with probabilistic self-adaptability features. In order to accommodate the coupling effects of stochastic EV driving behavior on transport network and distribution network, a risk-captured distributional reinforcement learning solution is presented by using explicit probabilistic information for action and reward function in Markov decision process (MDP) model, where the Bellman equation is extended to a more generalized version. Scheduling EV charging in a transportation-energy nexus, according to both transport and distribution network conditions, is an important topic in recent studies to improve the driving and charging energy efficiency, especially considering the high penetration rate of EV nowadays and even more extremely higher one in the future [1]. In order to accommodate the coupling effects of stochastic EV driving behavior and battery state-of-charge (SoC) on transport and distribution network, various methods have been developed for designing the smart charging scheduling strategy with consideration of electricity price, renewable energy adoption, road conditions and many others.

However, it can be pointed out that most of existing works are dependent heavily on the optimization-based solutions with assumption of convex characteristics and various pre-defined forecasting information in a deterministic manner. In practices, the transportation-energy nexus is close to a complex system without holding such good model characteristics and well-structured given input parameters for highly stochastic driving and charging behaviors. Thus, it is desirable to address EV charging scheduling problem in transportation-energy nexus environment using an distributional reinforcement learning-based strategy with probabilistic and self-adaptability features. Many EV and ordinary vehicle navigation and routing applications using deep reinforcement learning (DRL) framework are briefed and summarized in [2]. Less works study the joint transport routing and energy charging problems due to the resultant complex coupled constraints of congestion management, traffic flow overlap, energy allocation and many other issues that are not incurred in separated system [3]. A few works tried DRL framework to solve EV charging and navigation problems at the same time in the coordinated smart grid and intelligent transportation system [4], [5]. However, most of these works just exploit conventional DRL algorithms (e.g., soft actor-critic (SAC), deep deterministic policy gradient (DDPG), deep Q-network (DQN)) that are heavily dependent on deterministic reward value feedback and hard to capture the joint uncertain and distributional probability information, especially causal risks, in the coupled transport and distribution network system model.

Motivated by the above observation, this letter aims to develop a risk-captured distributional reinforcement learning solution for a joint routing and charging problem in the transportation-energy nexus. The main contributions of this letter can be summarized as: 1) An intelligent decision-making strategy with probabilistic self-adaptability features is designed to capture the system dynamics of coordinated transportation and distribution network. 2) Some key characteristics of C51 algorithm are analyzed and derived to ensure the good enough performance for the joint EV routing and charging problem in the uncertain MDP environment.

Problem formulation: An expanded transportation network presented in [6] is used to model the EV driving and routing behavior, which could be further coordinated with the distribution network with consideration of promoting renewable energy as charging power source. Taking as example a simple network $G(\mathcal{V},\mathcal{E})$ in Fig. 1.

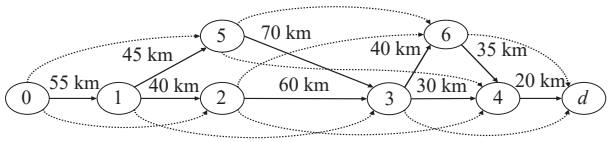

Fig. 1. The original (solid line) and expanded (dashed line) network.

It only has a single origin-destination (OD) pair, $g=(o,d)$. The edges in edge set $\mathcal{E}$ are unidirectional arcs with distance marked besides. The vertices in vertex set $\mathcal{V}$ are indexed by numbers in the circles. They denote transportation links (roads) and transportation nodes with charging station available, respectively. The energy consumption in the selected travelling path g is denoted as λ_g in the O-D travelling tuple (o,d,λ_g^{od}). Further more, this network can be expanded to a new network $G(\mathcal{V},\hat{\mathcal{E}})$ by connecting any two nodes by a pseudo edge if they are neighbored to the same node. The transportation network constraints are written as follows:

$$\sum_{j|(i,j)\in\hat{\mathcal{E}}}\lambda_{g,ij}-\sum_{j|(j,i)\in\hat{\mathcal{E}}}\lambda_{g,ji}=\begin{cases}\lambda_g^{od}, & i=o_g\\ -\lambda_g^{od}, & i=d_g\\ 0, & i\neq d_g\end{cases}\quad\text{AND}\quad i\neq o_g$$
$$\forall g\in\mathcal{G},\quad \forall i\in\mathcal{I} \tag{1}$$

$$\lambda_g^{ij}\geq 0,\quad \forall g\in\mathcal{G},\quad \forall(i,j)\in\hat{\mathcal{E}} \tag{2}$$

where λ_g^{ij} is the portion of traffic flow driving on edge (i,j) with the associated energy demand λ_g^{od}. It is assumed some transport nodes with available charging stations are supplied by distributed renewable energy resources, which would prioritize such route passing green energy nodes and charging options using clean energy. This assumption enables the modification of [6]. By mapping the decision variables to action status in MDP model, we can let action $\mathbf{a}=[\mathbf{p}_c,\mathbf{p}_r,\lambda_g]$, where $\mathbf{p}_c$ stands for the vector of power output from conventional energy resources, $\mathbf{p}_r$ for vector of adopted power from renewable energy resources, and λ_g for vector of nodal traffic flow. By using similar symbols in [6], we let state $\mathbf{s}=[\mathbf{v}_{ev},\mathbf{P},\mathbf{Q},\mathbf{V},\hat{\mathbf{p}}_r]$, where $\mathbf{v}_{ev}$ stands for the vector of variable available driving speed of EVs depending on road conditions, $\mathbf{P}$ and $\mathbf{Q}$ stand for the vector of real and reactive power injection respectively in the distribution network, $\mathbf{V}$ for the vector of node voltage, and $\hat{\mathbf{p}}_r$ for the vector of predicted renewable energy power output at each possible node. The reward function in MDP is designed similar to the objective function in [6] except for the add-on term indicating the credit token $\mathbf{c}_r$ to award and prioritize the renewable energy adoption

Corresponding authors: Tao Chen and Ciwei Gao.

Citation: T. Chen and C. W. Gao, "Intelligent electric vehicle charging scheduling in transportation-energy nexus with distributional reinforcement learning," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 11, pp. 2171–2173, Nov. 2023.

The authors are with the School of Electrical Engineering and Jiangsu Provincial Key Laboratory of Smart Grid Technology and Equipment, Southeast University, Nanjing 210096, China (e-mail: taoc@seu.edu.cn; ciwei.gao@seu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.123285

$$\mathbf{r} = c^t\Big(\frac{1}{v} + \frac{\xi}{\eta p_{rc}}\Big)\sum_{g\in\mathcal{G}} \mathbf{L}^T\lambda_g + \mathbf{p}_c^T\mathbf{Q}\mathbf{p}_c + (\mathbf{c}+\mathbf{c}_{CO2})^T\mathbf{p}_c - \mathbf{c}_r^T\mathbf{p}_r \quad (3)$$

where c^t stands for the per-unit time cost, ξ for the energy consumption of each route distance, η and p_{rc} for charging efficiency and rated charging power, $\mathbf{Q}$ and $\mathbf{c}$ for the quadratic and linear cost coefficients of power supply, $\mathbf{c}_{CO2}$ for the carbon tax, and $\mathbf{c}_r$ for the renewable promotion credit (e.g., certified emission reduction).

To quantify the implicit uncertainty characteristics of reward function $\mathbf{r}$ caused by the stochastic power output $\mathbf{p}_r$ of distributed renewable energy resources and the driving behavior of routine choice for O-D travelling tuple (o,d,λ_g^{od}), the reward function $\mathbf{r}$ in conventional Markov decision process (MDP) model should be replaced by a random return function Z.

Definition 1: Z is denoted as the random return or reward value, whose expectation is the normal Q value function, in the modified MDP model within distributional reinforcement learning framework.

By using such representations, the random return function Z instead of deterministic straightforward reward value will also be linked to the volatile transition probabilities on top of stochastic definitions of action space and state status space. Following the recursive equation to describe Q value function, the distribution of random variable Z is characterized by the interaction of three other random variable R, the next state-action pair (S',A') and its random return $Z(S',A')$. This quantity is called value distribution as the following equation:

$$Z(s,a) := R(s,a) + \gamma Z(S',A'). \quad (4)$$

Main results: In this section, some sufficient conditions are derived to ensure the applicability of distributional reinforcement learning framework for the well-defined reward maximization problem, where the characteristics of distributional Bellman operator are described.

Next, we state the following main result. In normal MDP model, the conventional Bellman operator $\mathcal{T}^\pi$ and optimality operator $\mathcal{T}^*$ are dependent on the expectation calculation and usually defined as the following:

$$\mathcal{T}^\pi Q(x,a) = \mathbb{E}R(x,a) + \gamma\mathbb{E}_{P,\pi}Q(x',a') \quad (5)$$

$$\mathcal{T}^* Q(x,a) = \mathbb{E}R(x,a) + \gamma\mathbb{E}_P \max_{a'\in\mathcal{A}} Q(x',a'). \quad (6)$$

However, in the distribution reinforcement learning framework, the expectations inside Bellman's equations could be taken away with consideration of the full distribution of any random variable appeared in MDP [7]. Here, we can define a random variable Z^π similar to reward function $R(s,a)$ to indicate a mmapping from state-action pairs (s,a) to distributions over returns or value distribution. To help analyze the theoretical behavior of the distributional analogues of the Bellman operators, the probability space $(\Omega,\mathcal{F},Pr)$ is used to help derivation. We will write $\|\mathbf{u}\|_p$ to denote the L_p norm of a vector $\mathbf{u}\in\mathbb{R}^{\mathcal{X}}$ for $1\le q\le\infty$. The L_p norm of a random vector $\mathcal{U}:\omega\to\mathbb{R}^{\mathcal{X}}$ is then $\|\mathcal{U}\|_p = [\mathbb{E}[\|\mathcal{U}(\omega)\|_p^p]]^{1/p}$, and for $p=\infty$ we have $\|\mathcal{U}\|_\infty = \sup\|\mathcal{U}(\omega)\|_\infty$. We will denote the cumulative distribution function (c.d.f.) of a random variable U by $F_U(y) = Pr\{U\le y\}$, and its inverse c.d.f. by $F_U^{-1}(q) = \inf\{y : F_U(y)\ge q\}$.

With these definitions, we can write the policy evaluation in MDP following such distributional settings of value distribution Z^π similar to instantiated *reward* function.

Theorem 1: Convergence of value distribution. Let $Z_k = \mathbb{T}Z_{k-1}$ with $Z_0\in\mathcal{Z}$, $\mathcal{X}$ be measurable and suppose that $\mathcal{A}$ is finite. Then,

$$\lim_{k\to\infty}\inf_{Z^*\in\mathcal{Z}^*} d_p(Z_k(x,a),Z^*(x,a)) = 0 \quad \forall x,a. \quad (7)$$

It could be claimed that Z_k converges to Z^* uniformly when $\mathcal{X}$ is finite. This claim also applies for any $Z^*\in\mathcal{Z}^*$ if there is a total ordering $<$ on Π^*.

$$\mathcal{T}Z^* = \mathcal{T}^\pi Z^* \text{ with } \quad \pi\in\mathcal{G}_{Z^*}, \quad \pi < \pi' \forall \pi' \in = \mathcal{G}_{Z^*}\setminus\{\pi\} \quad (8)$$

then $\mathcal{T}$ has a unique fixed point $Z^*\in\mathcal{Z}^*$.

Proof: For every state x, there is a time k after which the greedy policy w.r.t, Q_k is mostly optimal; a unique and therefore deterministic optimal policy π' is assumed. For rotational convenience, it can be written as $Q_k := \mathbb{E}Z_k$ and $\mathcal{G}_k := \mathcal{G}_{Z_k}$. Let $B := 2\sup_{Z\in\mathcal{Z}}\|Z\|_\infty < \infty$ and let $\epsilon_k := \gamma^k B$. We can define the set of states $\mathcal{X}_k\subseteq\mathcal{X}$ whose values must be sufficiently close to Q^* at time k

$$\mathcal{X}_k := \Big\{x : Q^*(x,\pi^*(x)) - \max_{a\ne\pi^*(x)} Q^*(x,a) > 2\epsilon_k\Big\}. \quad (9)$$

By the characteristics of linearity of expectation, we can write $\mathcal{T}_D$ for the distributional operator and $\mathcal{T}_E$ for the usual operator as following:

$$\begin{aligned}\|\mathbb{E}\mathcal{T}_D Z_1 - \mathbb{E}\mathcal{T}_D Z_2\|_\infty &= \|\mathcal{T}_E\mathbb{E}Z_1 - \mathcal{T}_E\mathbb{E}Z_2\|_\infty \\ &\le \gamma\|Z_1 - Z_2\|_\infty. \end{aligned} \quad (10)$$

Thus, after k iterations

$$|Q_k(x,a) - Q^*(x,a)| \le \gamma^k|Q_0(x,a) - Q^*(x,a)| \le \epsilon_k. \quad (11)$$

For $x\in\mathcal{X}$, let $a^* := \pi^*(x)$. The following can be deduced for any $a\in\mathcal{A}$:

$$Q_k(x,a^*) - Q_k(x,a) \ge Q^*(x,a^*) - Q^*(x,a) - 2\epsilon_k. \quad (12)$$

If $x\in\mathcal{X}_k$, it can be written as that $Q_k(x,a^*) > Q_k(x,a')$ for all $a'\ne\pi^*(x)$: thus, the greedy policy $\pi_k(x) := \text{argmax}_a Q_k(x,a)$ corresponds to the optimal policy π^* for these states. ■

By using the distributional version of Bellman operator, we can model the value distribution using a discrete distribution parameterized by $N\in\mathbb{N}$ and $V_{\min},V_{\max}\in\mathbb{R}$, and whose support is the set of atoms $\{z_i = V_{\min} + i\Delta z : 0\le i < N\}$, $\Delta z = (V_{\max}-V_{\min})/(N-1)$. The atom probabilities are given by a parrametric model $\theta : \mathcal{X}\times\mathcal{A}\to\mathbb{R}_N$ as follows:

$$Z_\theta(x,a) = z_i \qquad \text{w.p.} \qquad p_i(x,a) := \frac{e_{\theta_i(x,a)}}{\sum_j e_{\theta_j(x,a)}}. \quad (13)$$

However, using such a discrete distribution will pose a problem that the Bellman update $\mathcal{T}Z_\theta$ and the parametrization Z_θ always have disjoint supports. Thus, the sample Bellman $\hat{\mathcal{T}}Z_\theta$ needs to update onto the support of original Z_θ as illustrated in Fig. 2.

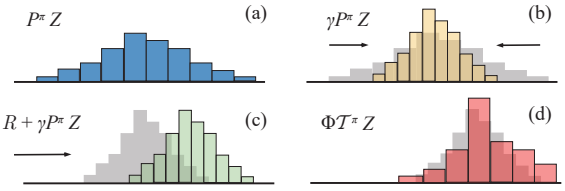

Fig. 2. Illustration of distributional Bellman operator for a deterministic reward function: (a) Next state distribution update policy π; (b) Discounting shrinks the value distribution; (c) Reward shifts the value distribution; (d) Projection to the pre-defined support.

In the reward shrinking and shifting process the Bellman update will be reduced to multi-class classification. Let π be the greedy policy w.r.t. $\mathbb{E}Z_\theta$. Given (x,a,r,x') as a transition sample, the Bellman update $\hat{\mathcal{T}}z_j = r + \gamma z_j$ can be computed for each atom z_j. The probability $p_j(x',\pi(x'))$ should also be distributed to the immediate neighbours of $\hat{\mathcal{T}}z_j$. In the projected update $\Phi\hat{\mathcal{T}}Z_\theta(x,a)$, the ith component is computed as follows:

$$(\Phi\hat{\mathcal{T}}Z_\theta(x,a))_i = \sum_{j=0}^{N-1}\left[1 - \frac{|[\hat{\mathcal{T}}z_j]_{V_{\min}}^{V_{\max}} - z_i|}{\Delta Z}\right]_0^1 p_j(x',\pi(x')). \quad (14)$$

where $[\cdot]_a^b$ indicates the bounds within $[a,b]$. The next state distribution can be parametrized using $\tilde{\theta}$. The cross-entropy term of the Kullback-Leibler (KL) divergence is used to define $\mathcal{L}_{x,a}(\theta)$ as sample loss

$$D_{KL}(\Phi\hat{\mathcal{T}}Z_{\tilde{\theta}}(x,a)\|Z_\theta(x,a)). \quad (15)$$

Then, the sample loss can be readily minimized using gradient descent. The solution algorithm following such choice of loss and distribution is called categorical algorithm or *C51* when $N = 51$ chosen for the number of support atoms. The particular C51 algorithm

for the coordinated EV routing and charging problem is presented in Algorithm 1 based on the standard distributional DQN algorithm [7].

Algorithm 1 C51 Algorithm for Coordinated EV Routing and Charging

Input A transition $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0,1]$
  $Q(x_{t+1}, a) = \sum_i z_i p_i(x_{t+1}, a)$ for random EV SoC
  $a^* \leftarrow \text{argmax}_a Q(x_{t+1}, a)$
  $m_i = 0, \quad i = 0, \ldots, N-1$
  for $j = 0, \ldots, N-1$ **do**
    $\hat{\mathcal{T}} z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\min}}^{V_{\max}}$ with physical constraints
    $b_j \leftarrow (\hat{\mathcal{T}} z_j - V_{\min})/\Delta z$
    $l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$
    $m_l \leftarrow m_l + p_j(x_{t+1}, a^*)(u - b_j)$
    $m_u \leftarrow m_u + p_j(x_{t+1}, a^*)(b_j - l)$
  for
Output $-\sum_i m_i \log p_i(x_t, a_t)$

Numerical example: In the numerical results, a 22-node highway transport network with 6-node of available on-site renewable energy resources is considered in couple with a 14-node 110 kV high voltage distribution network. The transport network is modified from the original 25-node version with detailed information in [8] and similar coupling relationship. The transport network the particular system step-up and model parameters in [6] are used for the simulation with emphasis on the performance of learning-based methods. Some key parameters for C51 algorithm are provided as follows: Set discounting rate $\gamma = 0.99$, learning rate $\alpha = 0.001$, number of atoms $N_{\text{atoms}} = 51$, $V_{\max}/V_{\min} = \pm 20$ and three-layer fully connected neural networks. The simulation results out of multiple runs are summarized in Table 1 with learning performance using C51 algorithm presented in Fig. 3.

Table 1. The Cost Comparison of Different Solution Methods

		Optimization	DQN algorithm	C51 algorithm
Cost	Avg	4840.4	5255.5	4962.3
(CNY)	Max	–	5640.3	5133.2
	Min	–	4988.4	4843.6

We can easily observe that similar to most reinforcement learning-based methods, the distributional categorical method also needs training steps to gradually improve its performance with incremental average return values by sampling the distributional information. As shown in Table 1, although the C51 algorithm mostly outperforms conventional DQN algorithm, it hardly exceeds the upper bound limits calculated from the well-defined optimization method. It can be explained by the facts that in the simulations, we feed the learning algorithm much less input information (e.g., deterministic per-unit time cost) as a prior or assume no ideal prediction (accuracy less than 90%) for the future state estimation (e.g., accurate renewable power output forecasting and guaranteed shortest path). Compared with the results reported in [6] and [8] using similar system setup, the proposed method has slightly higher cost ($\leq 0.5\%$) but with an ultimate gradually improved economic performance in the long-term operation and much less computational cost (≤ 60 s) if using pre-trained reinforcement learning (RL) agent model for online operation directly. Additionally, the C51 algorithm only has an insignificant increase in computational time cost compared with DQN algorithm, costing roughly 30 000 s for 14 000 steps with about 12% more computational load.

In Fig. 4, it is shown that most EVs actually indeed give priority to the transportation nodes with renewable energy source powered charging options (e.g., node 5, 9, 14). By tuning the value of green credit token, the weighting of appropriate environmental friendly charging options can overcome the possible tension caused by the increased per-unit time cost due to traffic congestion.

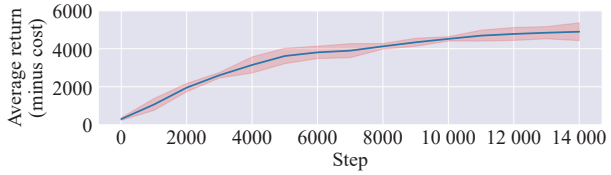

Fig. 3. Learning curve for the coordinated EV routing and charging benefit.

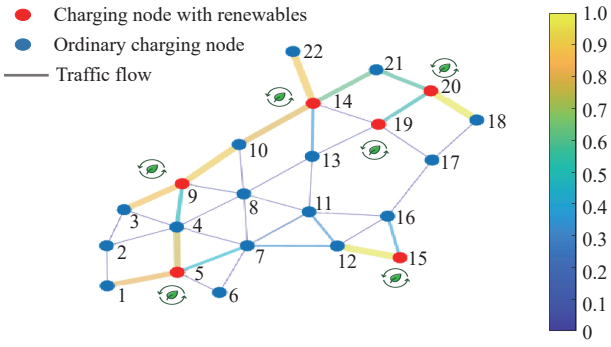

Fig. 4. EV traffic flows in the coupled transport and distribution network.

Conclusion: In this letter, the coordinated EV routing and charging scheduling problem in transportation-energy nexus is investigated, particularly using an intelligent decision-making strategy with probabilistic self-adaptability features. In order to accommodate the effect of stochastic EV driving and charging behavior on transport network and distribution network, a risk-captured distributional reinforcement learning solution is presented by using explicit probabilistic information for action and reward function in MDP model.

Acknowledgments: This work was supported by National Natural Science Foundation of China (52107079), Natural Science Foundation of Jiangsu Province (BK20210243), and the Open Research Project Program of the State Key Laboratory of Internet of Things for Smart City (University of Macau) (SKL-IoTSC(UM)-2021-2023/ORPF/A14/2022).

References

[1] T. Chen, B. Zhang, H. Pourbabak, A. Kavousi-Fard, and W. Su, "Optimal routing and charging of an electric vehicle fleet for high-efficiency dynamic transit systems," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3563–3572, 2016.

[2] A. Haydari and Y. Yilmaz, "Deep reinforcement learning for intelligent transportation systems: A survey," *IEEE Trans. Intelligent Transportation Systems*, vol. 23, no. 1, pp. 11–32, Jan. 2022.

[3] W. Wei, L. Wu, J. Wang, and S. Mei, "Network equilibrium of coupled transportation and power distribution systems," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6764–6779, 2017.

[4] Y. Liang, Z. Ding, T. Ding, and W.-J. Lee, "Mobility-aware charging scheduling for shared on-demand electric vehicle fleet using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1380–1393, 2020.

[5] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, 2019.

[6] H. Zhang, Z. Hu, and Y. Song, "Power and transport nexus: Routing electric vehicles to promote renewable power integration," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3291–3301, 2020.

[7] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. Int. Conf. Machine Learning*, 2017, pp. 449–458.

[8] H. Zhang, S. J. Moura, Z. Hu, W. Qi, and Y. Song, "A second-order cone programming model for planning PEV fast-charging stations," *IEEE Trans. power Systems*, vol. 33, no. 3, pp. 2763–2777, 2017.