2023 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm) | 978-1-6654-5556-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/SmartGridComm57358.2023.10333926

Network-Constrained Reinforcement Learning for Optimal EV Charging Control

Tong Wu, *Member, IEEE*, Anna Scaglione, *Fellow, IEEE*, Adrian Petru Surani, *Student Member, IEEE*, Daniel Arnold, *Member, IEEE*, Sean Peisert, *Senior Member, IEEE*

***Abstract*—This paper introduces a comprehensive control model that integrates aggregate electric vehicle (EV) charging demand with power grid systems operations, capitalizing on the flexible nature of EV charging. This innovative approach allows us to model and manage electrical loads in a scalable manner. The main contribution is the study of a constrained reinforcement learning (CRL) method for the predictive control of optimal power flow, paired with EV charging control. The CRL-based control method operates with the understanding that future EV arrivals are uncertain, while ensuring the feasibility of control actions. Our case studies, conducted on IEEE standard systems, highlight the superior performance of our approach that dynamically adapts to the evolving EV environment while consistently upholding safety constraints.**

I. Introduction

A. Background and Motivation

The transition to electric vehicles (EVs) plays a critical role in combating climate change by reducing greenhouse gas emissions that contribute to global warming [1]. As the electricity used to charge EVs is generated from renewable energies such as wind or solar power, they emit zero tailpipe pollutants, thereby significantly reducing carbon emissions [2]. In addition, EV charging offers a significant degree of spatio-temporal flexibility, allowing for adaptable charging locations and times [3]. Furthermore, the batteries within EVs can also be leveraged to offer grid services, enhancing power grid stability and security [4]. This potential is unlocked through the deployment of Vehicle-to-Grid (V2G) technology, which enables bidirectional power transfer, thereby transforming electric vehicles into dynamic energy resources that can contribute to a more resilient and reliable power grid [5].

B. Related Works

a) Control of EVs in Power Grids: In the domain of EVs, research on controlling them by leveraging their flexibility for the electricity market bifurcates into two primary directions. The first stream delves into harnessing the aggregated EVs for participation in demand response (DR) within the wholesale market [4, 6, 7]. Such methodologies illuminate various traits of DR device populations, including aggregate energy demand, flexibility, and ramping capabilities. The second area centers on real-time pricing with demand response, as seen in [8–10]. Here, studies target households with appliances, Plug-in hybrid electric vehicles (PHEVs), and batteries. They highlight a utility-maximization framework for demand response, linking appliance utility to power consumption. By adjusting power consumption, households maximize benefits within consumption limits. Dynamic pricing, with its time-varying nature, aligns individual and societal benefits, guiding demand responses for overall system efficiency.

Harnessing the flexibility of EVs in the power grid on a large scale poses challenges. Traditional optimization methods are not ideal for real-time EV control due to their reliance on future data like EV arrivals [11]. Reinforcement learning offers a solution by enabling agents to make decisions based on environment interactions [12]. With V2G technology and deep reinforcement learning (DRL), it is possible to optimize both EV charging and power generation. This approach aims to reduce fuel costs through optimal power flow methods that account for EV charging, leading to a more efficient and eco-friendly energy system [11].

b) Constrained Reinforcement Learning: DRL techniques, as emphasized by [13, 14], are promising for dynamic optimization. However, they sometimes yield unfeasible results, especially regarding power flow and EV charging limits. Addressing this, Constrained Reinforcement Learning (CRL) is emerging as a pivotal approach to ensure feasibility in power systems decision-making [15, 16]. Techniques like Lagrangian relaxation are especially noteworthy in this context, as mentioned in [15]. Furthermore, [17] explored chance-constrained learning, and [18] proposed an innovative algorithm delineating regret and constraint boundaries. While much CRL research for EV charging, such as [12], focuses on voltage magnitude, it often overlooks vital network constraints. It's essential to understand that power grid safety norms require consistent adherence, given their immediate nature.

C. Contributions and Organization

In this study, we leverage the potential of a network-constrained reinforcement learning framework for optimal control of EV charging and power flow. The primary contributions of this paper can be summarized as follows:

- We capitalize on the inherent flexibility of aggregated EV charging models for real-time control. We treat this as a stochastic control problem, given the uncertainties in future EV arrivals, renewable energy generation, and demand patterns.
- Building upon our previous research [16], we employ constrained reinforcement learning for the predictive control of aggregated EV charging demands. These demands can subsequently be disaggregated into specific control actions through slack-charging models. This approach en-

Tong Wu, Anna Scaglione and Adrian Petru Surani are with the Department of Electrical and Computer Engineering, Cornell Tech, Cornell University, 10044 USA (e-mail: {tw385, as337, as3259}@cornell.edu). Daniel Arnold and Sean Peisert are with Lawrence Berkeley National Laboratory. This research was supported in part by the NSF under Grant NSF ECCS # 2210012. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors of this work.

978-1-6654-5556-5/23/$31.00 ©2023 IEEE

sures scalability for large-scale EV charging applications and adheres to network and EV charging constraints.

The rest of the paper is organized as follows. Section II presents the aggregated EV charging models and Section III introduce the V2G model for EV control. Section IV reviews the constrained reinforcement learning method. We then implement our primal-dual constrained reinforcement learning method for a specific case study in Section V. Finally, Section VI draws some conclusions.

II. Aggregate EV Charging Model

In this section we show how the flexibility of a population of EVs can be expressed through an integer linear model, following the ideas in [4, 6]. In the context of Demand Response (DR), EV charging models demonstrate inherent flexibility with various potential energy consumption profiles. Using these models we describe a novel approach for managing optimally the charging decisions for the population.

Before delving deeply into the slack-charging EV model details, it is useful to present some of the conventions in our notation. We use the italic font for discrete variables, symbolized as $z(t)$ and boldface for vectors or matrices, indicated as $\boldsymbol{z}$. We denote time-dependent finite differences as $\partial z(t) = z(t+1) - z(t)$. The unit step is marked by $u(t)$, and the Kronecker delta function is expressed by $\delta(t)$, which is 1 when the argument is zero and 0 in other cases. Consider $t \in \mathcal{T} = \{0, \cdots, T\}$ as the set of $\mathcal{T}$ equally spaced. Discrete time indices distanced by δT allow us to use the index $t \in \mathbb{Z}$ such that $t = t\delta T$. We refer to the feasible set $\mathcal{L}$, that will be defined in (6) as containing different instances of $\ell_{ev}(t)$ with t as the argument, where each instance is a unique set element.

Each EV has attributes: $(t_p, X_p, E_p, \chi_p, \rho_p)$, where t_p is the charging start time, X_p is the initial energy units, E_p is the maximum charge capacity, χ_p is the deadline for battery p, and ρ_p is the fixed charge rate based on the charging site's voltage level. We standardize the quantization step to $\delta T = 1$. Expressing these five parameters directly requires complex clustering to form the aggregated EV demand models. In the next subsection, we aim to reform this model with fewer clusters while retaining sufficient flexibility to manage EV charging demands.

Before introducing the aggregate EV charging model, we first introduce two elements for EV user: the charging time, which is when the EV car is effectively charged at the rate ρ_p, and the slack time, during which the EV car remains at the charging station without receiving any charge. Let R_p be the total time required to fully charge an EV p when it arrives at the charging station, given by:

$$R_p = (E_p - X_p)/\rho_p. \tag{1}$$

Let S_p denote the slack time or laxity of EV p, defined as

$$R_p + S_p = \chi_p - t_p. \tag{2}$$

where t_p indicates the arrival time of the p-th EV and ρ_p represents the charging rate, which we presume is constant for all EVs within a specific cluster. The feasible set defined by $(t_p, X_p, E_p, \chi_p, \rho_p)$ can be represented using a set of parameters: (R_p, S_p, ρ_p). For any quantized charging time, $[R_p]$ can range from 0 to $N_r - 1$, denoted as $[R_p] \in \{0, \ldots, N_r - 1\}$. Similarly, the quantized slack time $[R_s]$ can range from 0 to $N_s - 1$, expressed as $[R_s] \in \{0, \ldots, N_s - 1\}$.

This common characteristic forms the foundation for creating a unified model. Within this model, the status of a load at any given time, symbolized as $t \in \mathcal{T}$, is denoted by the pair $(r, s) \in \mathcal{U}_{rs} = \{0, \ldots, N_r - 1\} \times \{0, \ldots, N_s - 1\} \subset \mathbb{N}_+^2$. In this context, r indicates the remaining service time needed, while s represents the remaining slack time. We want to make it clear to the reader that in the following discussions, we will be using $\boldsymbol{x} = (r, s) \in \mathcal{U}_{rs}$. $\boldsymbol{x}$ represents a two-dimensional EV state. Our main goal is to regulate the dynamics of $n_{\boldsymbol{x}}(t)$, representing the count of EVs in state $\boldsymbol{x} = (r, s)$ at a specific time t. To achieve this, we consider a step function that details the arrival of EV t at the moment t_p^a:

$$a_p(t) = u(t - t_p^a) \tag{3}$$

As such, we can define an arrival process that increments the count by adding new cars in a particular state $n_{\boldsymbol{x}}^t$:

$$a_{\boldsymbol{x}}(t) = \sum_{p \in \mathcal{P}_{\boldsymbol{x}}} \delta\left([R_p] - r\right) \delta\left([S_p] - s\right) a_p(t), \text{ where } \boldsymbol{x} = (r, s), \tag{4}$$

where the process $a_{\boldsymbol{x}}(t)$ is typically defined as a non-stationary Poisson process, and $\mathcal{P}_{\boldsymbol{x}}$ represents the set of EVs that uniformly share the same state $\boldsymbol{x} = (r, s)$. Given the discretization of time, the number of arrivals within the time span from t to $t+1$ follows a Poisson distribution:

$$\mathbb{P}(a_{\boldsymbol{x}}(t) = n) = \frac{\lambda_{\boldsymbol{x}}(t)}{n!} e^{-\lambda_{\boldsymbol{x}}(t)} \tag{5}$$

where $\lambda_{\boldsymbol{x}}(t)$ is non-uniform, as the average number of cars

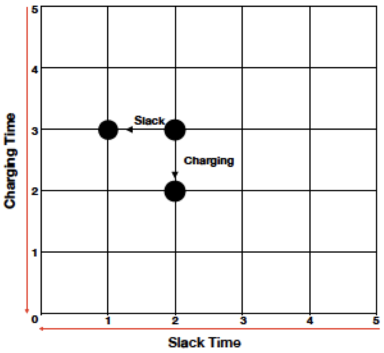

Figure 1. An illustration of the slack-charging EV model.

The symbol $\partial d_{\boldsymbol{x},\boldsymbol{x}'}(t)$ represents the count of EVs transitioning from state $\boldsymbol{x}$ to state $\boldsymbol{x}'$ at time t. Hence, the term $\big((\boldsymbol{x}' - \boldsymbol{x})\, \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t)\big)_r$ signifies the quantity of EVs charging (without slack) from state $\boldsymbol{x}$ to state $\boldsymbol{x}'$ at time t. Moreover, the expression $\rho\big((\boldsymbol{x}' - \boldsymbol{x})\, \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t)\big)_r$ symbolizes the aggregated charging power, which corresponds to the consolidated demands of the electric vehicles. Additionally, the total count of EVs transitioning from state $\boldsymbol{x}$, i.e., $\sum_{\boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}} \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t)$, accounting for both charging and slack choices, is equal to the existing number of EVs at state $\boldsymbol{x}$, i.e., $n_{\boldsymbol{x}}(t)$. This equality is a consequence of the persistent reduction in the total time, indicated by $\chi_i - t_i$. Therefore, the aggregate feasible set for EVs with the same rate ρ:

$$\begin{aligned}\mathcal{L} = \Big\{ & \ell_{ev}(t) \mid \ell_{ev}(t) = \sum_{\forall \boldsymbol{x}} \sum_{\boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}} \rho\big((\boldsymbol{x}' - \boldsymbol{x})\, \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t) \big)_{\boldsymbol{r}}, \\ & 0 \le \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t) \le \overline{d}_{\boldsymbol{x},\boldsymbol{x}'}, \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t) \in \mathbb{Z}^+, \\ & \sum_{\boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}} \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t) = n_{\boldsymbol{x}}(t), \forall t \in \mathcal{T} \Big\}\end{aligned} \tag{6}$$

where $(\cdot)_{\boldsymbol{r}}$ represents the charging-related dimension value, $\overline{d}_{\boldsymbol{x},\boldsymbol{x}'}$ is the maximum capability and the set of $\mathcal{U}_{\boldsymbol{x}}$ is defined as

$$\begin{aligned}\mathcal{U}_{\boldsymbol{x}} = \Big\{ & \boldsymbol{x}' \mid \|\boldsymbol{x}' - \boldsymbol{x}\|_1 = \min\left(\|\boldsymbol{x}\|_1, 1\right), \\ & (\boldsymbol{x} - \boldsymbol{x}')_{\boldsymbol{r}} \ge 0, (\boldsymbol{x} - \boldsymbol{x}')_{\boldsymbol{s}} \ge 0 \Big\}\end{aligned} \tag{7}$$

where the EV charging only has two choices, i.e., charging or slack. As illustrated in Fig. 1, $\mathcal{U}_{\boldsymbol{x}}$ demonstrates that each EV car has only two choices at any given moment: either to charge or to wait (slack). In this scenario, the equality constraint mainly focuses on transitions to states at a distance of 1, unless we are close to the origin with the distance less than one. On the other hand, the inequality constraints ensure that we stay within the upper right quadrant of the state space. These constraints also stipulate that any movement can only be directed towards one of the two axes.

The state population tensor $n_{\boldsymbol{x}}(t)$ and $d_{\boldsymbol{x},\boldsymbol{x}'}(t)$ are:

$$\begin{aligned}& n_{\boldsymbol{x}}(t) = a_{\boldsymbol{x}}(t) + \sum_{\boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}} \left[d_{\boldsymbol{x}',\boldsymbol{x}}(t) - d_{\boldsymbol{x},\boldsymbol{x}'}(t) \right], \forall \boldsymbol{x} \\ & \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t) = d_{\boldsymbol{x},\boldsymbol{x}'}(t) - d_{\boldsymbol{x},\boldsymbol{x}'}(t-1), \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}.\end{aligned} \tag{8}$$

where $n_{\boldsymbol{x}}(t)$ also indicates the control capacity to monitor the number of EVs that exist in the system at time t with state $\boldsymbol{x}$. Let $\mathcal{N}$ denote the set of all buses with cardinality N, $\mathcal{G}$ represent the set of all buses with generators installed (having G generators), $\mathcal{G}_s$ and $\mathcal{G}_n$ stand for the sets of slack and non-slack buses with generation, respectively, and $\mathcal{B}$ be the set of all EV stations with a cardinality of B. In the notation that follows, we define the vector $\boldsymbol{z}$ such that $\boldsymbol{z} \triangleq (z_i)_{\forall i}$, where z_i represents each individual element. In this context, we employ the notation $\boldsymbol{d}^i_{\boldsymbol{x},\boldsymbol{x}'}(t)$ to signify the vector representing the cumulative state of EV charging for every pair of vectors $\boldsymbol{x}, \boldsymbol{x}'$ belonging to the set $\mathcal{U}_{\boldsymbol{x}}$, located on bus $i \in \mathcal{N}$ at time t. Similarly, $\boldsymbol{n}^i_{\boldsymbol{x}}(t)$ is used to denote the vector of all available $\boldsymbol{x}$, aggregated EVs.

III. Problem Formulation

We consider a stochastic optimal control problem:

$$\min_{\pi(\boldsymbol{s}(t-1))} \mathbb{E}_{\boldsymbol{\ell}_d(t)} \left[\varsigma(\boldsymbol{s}(t), \boldsymbol{A}(t), \boldsymbol{\ell}_d(t)) \right] \tag{9a}$$

$$\boldsymbol{s}(t) = f_t(\boldsymbol{s}(t-1), \boldsymbol{A}(t-1), \boldsymbol{\ell}_d(t)), \tag{9b}$$

$$\boldsymbol{A}(t) = \pi(\boldsymbol{s}(t-1)), \quad (\boldsymbol{s}(t), \boldsymbol{A}(t)) \in \chi(t), \tag{9c}$$

where $\boldsymbol{s}(t)$ represent the state vector at a given time t, and $\boldsymbol{A}(t)$ symbolizes the control vector at that same time, encapsulating all controllable components within power networks. The dynamic function is represented by $f(\cdot)$, while $\boldsymbol{\ell}_d(t)$ signifies the combined demands and sustainable energy sources. The expression $\varsigma(\boldsymbol{s}(t), \boldsymbol{A}(t), \boldsymbol{\ell}_d(t))$ captures the associated cost function, and $\chi(t)$ encapsulates the limits for the network and its devices. The term $\pi(\boldsymbol{x}(t-1))$ corresponds to a stochastic policy. Our study's primary objective is to optimize the distribution of power within network boundaries by effectively managing power generation, EVs, and energy storage devices.

To further elaborate, $\boldsymbol{s}(t) = [\boldsymbol{\vartheta}(t), \boldsymbol{soc}(t), (\boldsymbol{n}^i_{\boldsymbol{x}}(t))_{\forall i})_{\forall i}]^\top$ comprises voltage angles, denoted by $\boldsymbol{\vartheta}(t)$, and the collective state of charge (SOC) vector $\boldsymbol{soc}(t)$ pertaining to every battery in the infrastructure. The control vector $\boldsymbol{A}(t) = [\boldsymbol{g}(t); \boldsymbol{\ell}_{ev}(t), \boldsymbol{p}_{ch}(t); \boldsymbol{p}_{dis}(t)]^\top$ encompasses power outputs represented by $\boldsymbol{g}(t) = (g_i(t))_{\forall i \in \mathcal{G}}$, consolidated EV energy $\boldsymbol{\ell}_{ev}(t) = (\ell^i_{ev}(t))_{\forall i \in \mathcal{B}}$, and the battery charge and discharge velocities given by $\boldsymbol{p}_{dis}(t) = (p^i_{dis}(t))_{\forall i \in \mathcal{B}}$ and $\boldsymbol{p}_{ch}(t) = (p^i_{ch}(t))_{\forall i \in \mathcal{B}}$.

We consider the coupling of EVs with the power grids, taking into account both power flow constraints and slack-charging EV models. The objectives of OPF with aggregated EVs include the fuel costs as follows:

$$\varsigma_f(t) = \sum_{i \in \mathcal{G}} (\alpha_i g_i^2(t) + \beta_i g_i(t) + \gamma_i), \tag{10}$$

where α_i, β_i and γ_i are positive. We aim to leverage demand response to facilitate a smoother power demand curve and reduce peak loads. To accomplish this objective, we introduce two additional loss function:

$$\varsigma_s(t) = \sum_{i \in \mathcal{B}} \left\| \ell^i_{ev}(t) - \ell^i_{ev}(t-1) \right\|_2^2, \tag{11}$$

where $[x]_+ = \max(0, x)$. The reward for action is the complement of the objectives in (11):

$$r(t) = -\varsigma_f(t) - \varsigma_s(t). \tag{12}$$

In this study, we employ the DC power flow approximation to uphold the power-flow constraints. The OPF issue, when considering aggregated EVs, is defined as follows:

$$\max_{\boldsymbol{A}(t)} \mathbb{E}_{\boldsymbol{\ell}_d(t)} \left[r(t) \right] \tag{13a}$$

$$\mathbf{M}_b(\boldsymbol{\ell}_{ev}(t) + \boldsymbol{p}_{dis}(t) - \boldsymbol{p}_{ch}(t)) + \mathbf{M}_g \boldsymbol{g}(t) - \boldsymbol{\ell}_d(t) = \mathbf{B}\boldsymbol{\vartheta}(t), \tag{13b}$$

$$\underline{\boldsymbol{g}} \le \boldsymbol{g}(t) \le \overline{\boldsymbol{g}}, \quad |\mathbf{K}\boldsymbol{\vartheta}(t)| \le \boldsymbol{s}_{\max} \tag{13c}$$

$$\ell^i_{ev}(t) = \sum_{\forall \boldsymbol{x}} \sum_{\boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}} \left(\rho\left(\boldsymbol{x}' - \boldsymbol{x}\right) \partial d^i_{\boldsymbol{x},\boldsymbol{x}'}(t) \right)_{\boldsymbol{r}}, \tag{13d}$$

$$\boldsymbol{soc}_{min} \le \boldsymbol{soc}(t) \le \boldsymbol{soc}_{max}, \; 0 \le \boldsymbol{p}_{ch}(t), \boldsymbol{p}_{dis}(t) \le \overline{\boldsymbol{p}}_b \tag{13e}$$

$$\boldsymbol{soc}(t) = \boldsymbol{soc}(t-1) + \frac{\Delta t}{E_{cap}} \Big(\eta_{ch} \boldsymbol{p}_{ch}(t) - \frac{\boldsymbol{p}_{dis}(t)}{\eta_{dis}} \Big) \tag{13f}$$

$$\ell^i_{ev}(t) \in (6), \forall i, \forall t \tag{13g}$$

Here, we consider the demand vector, inclusive of renewable energy, as $\boldsymbol{d}(t) = [d_1(t), \cdots, d_N(t)]^\top$. The susceptance matrix is represented by $\mathbf{B}$. We define the matrix $\mathbf{K}$ as $\mathbf{K} \triangleq \mathbf{BI}$, with $\mathbf{I} \in \mathbb{R}^{m \times n}$ being the directed graph incidence matrix for the network, where m denotes the number of network lines. The grid state within the DC power flow approximation is represented by $\boldsymbol{\vartheta}(t) = [\vartheta_1(t), \cdots, \vartheta_N(t)]^\top$.

Lastly, $\mathbf{M}_g$, a $\{0,1\}^{N\times G}$ matrix, associates the generation vector $\boldsymbol{g}(t)\in\mathbb{R}^{|\mathcal{G}|}$ to $\mathbb{R}^N$, and is described as:

$$\begin{aligned}&[\mathbf{M}_g\boldsymbol{g}]_i = 0, \quad \forall i\in\mathcal{N}\setminus\mathcal{G}\\&[\mathbf{M}_g\boldsymbol{g}]_i = g_j, \quad \forall i\in\mathcal{G}, \quad \forall j\in[1,\cdots,G]\end{aligned} \tag{14}$$

and similarly $\mathbf{M}_b$ the matrix that maps the vectors $\boldsymbol{\ell}_{ev}(t)$ (and $\boldsymbol{p}_{ch}(t)$ and $\boldsymbol{p}_{dis}(t)$) onto the entire network, adding zero in the buses that do not have EV charging stations. Specifically, Eq. (13b) describes power flow constraints that enable the integration of power grids, aggregate EV demands, and battery storage, with each EV demand assumed to have battery storage. Concurrently, Eq. (13c) encapsulates power flow limit constraints.

IV. Constrained Reinforcement Learning

The task of directly resolving the optimization problem in Eq. (13) is highly complex. This complexity arises due to the engagement of multiple aggregate EV demands across different buses, with each aggregate EV demand incorporating integer variables. To address the above challenge, in this section, we apply our CRL methodology in [16] for real-time predictive control of the OPF problem, incorporating aggregate EVs as described in equation (13).

A. Constrained Twin Delayed Deep Deterministic Policy Gradient (TD3)

Within the actor-critic structure, the TD3 approach modifies policy function parameters, guided by an approximate value or critic function [19]. The actor, symbolized as π_θ, determines actions, and the critic, represented by Q_ξ, assesses them. Temporal difference learning in Q-learning [20] extracts the value function based on the Bellman equation [21].

1) Critic Design: In the framework of deep reinforcement learning, the critic section is fundamental. It comprises the target network, which has been meticulously designed in pursuit of two primary objectives: ensuring stability and significantly reducing the error associated with function approximation. Alongside the target network, there are Critic Networks. These Critic Networks actively engage in the regular updating of parameters pertinent to their individual networks, as highlighted in the study by [16].

a) Target and Critic Networks: We integrate two target networks, specifically $Q_{\xi_1'}$ and $Q_{\xi_2'}$, determining y by taking the minimum between the two value estimates:

$$y = r + \gamma \min_{i=1,2} Q_{\xi_i'}(s,\boldsymbol{A}), \tag{15}$$

where actions are chosen from a target actor network π_ϕ, and r is as defined in (12). Using this y, the critic networks update their parameters:

$$\xi_i \leftarrow \arg\min_{\xi_i}\frac{1}{N}\sum(y-Q_{\xi_1}(s,\boldsymbol{A}))^2, \forall i=1,2, \tag{16}$$

where N denotes the batch size. The weights of these target networks are then adjusted:

$$\xi_i' \leftarrow \tau\xi_i + (1-\tau)\xi_i', \forall i=1,2, \tag{17}$$

with ξ_1 and ξ_2 indicating critic network parameters and ξ_1' and ξ_2' signifying target network parameters. The networks reciprocally update one another.

2) Constrained Actor Design: After appropriately configuring the critic, the actor network can be set up, incorporating the constrained action space. Traditionally, the aim is to train the action network to optimize the output from the critic network.

$$\phi \leftarrow \arg\max_{\phi} Q_{\xi_1}(s(t-1),\pi_\phi(s(t-1))). \tag{18}$$

where ϕ denotes the parameters of the action network. Either Q_{ξ_1} or Q_{ξ_2} can be utilized to guide the updates in ϕ via $\pi_\phi(\cdot)$. An action, represented as $\boldsymbol{A}(t)=\pi_\phi(s(t-1))$, is considered valid when it aligns with its constraints, denoted by $\chi(t)$. Thus, the policy π_ϕ is framed by optimizing the critic network and concurrently respecting $\chi(t)$.

$$\max_{\phi} Q_{\xi_1}(s(t-1),\pi_\phi(s(t-1))) \quad s.t.\ \boldsymbol{A}(t)\in\chi(t). \tag{19}$$

where $\boldsymbol{A}(t)$ is taken by policy $\boldsymbol{A}(t)=\pi_\phi(s(t-1))$.

B. Primal-Dual OPF Formulation

In this subsection, we aim to develop the constrained policy function, denoted as $\pi_\phi(\cdot)$, for the OPF issue. We detail the normalized power generations as $\boldsymbol{g}(t)$, the normalized powers for both discharging and charging as $\boldsymbol{p}_{dis}(t)$ and $\boldsymbol{p}_{dis}(t)$, and the standardized aggregated EV demand represented by $\boldsymbol{\ell}_{ev}(t)$. These are delineated using the actions $\boldsymbol{a}(t)$ in the subsequent manner:

$$\begin{aligned}&\boldsymbol{A}(t)\triangleq\pi_\phi(s(t-1)), \quad \boldsymbol{A}(t)\triangleq[\boldsymbol{g}(t),\boldsymbol{\ell}_{ev}(t),\boldsymbol{p}_{ch}(t),\boldsymbol{p}_{dis}(t)]^\top,\\&\hat{\boldsymbol{\ell}}_{ev}(t)\triangleq\boldsymbol{\ell}_{ev}(t)L_{\max}, \quad \hat{\boldsymbol{g}}(t)\triangleq(1-\boldsymbol{g}(t))\underline{\boldsymbol{g}}+\boldsymbol{g}(t)\overline{\boldsymbol{g}},\\&\hat{\boldsymbol{p}}_{ch}(t)=\boldsymbol{p}_{ch}(t)\overline{\boldsymbol{p}}_b, \quad \hat{\boldsymbol{p}}_{dis}(t)=\boldsymbol{p}_{dis}(t)\overline{\boldsymbol{p}}_b\end{aligned} \tag{20}$$

where the original power generation, in addition to the raw discharging and charging powers and the initial aggregated EV demand, influence the environment.

We present dual variables, denoted as $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, associated with Eq. (13), and include the enhanced penalty parameters:

$$\boldsymbol{\lambda}=\begin{bmatrix}\boldsymbol{\lambda}_1\\\boldsymbol{\lambda}_2\\\boldsymbol{\lambda}_3\end{bmatrix}, \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_1\\\boldsymbol{\mu}_2\\\boldsymbol{\mu}_3\end{bmatrix}, \boldsymbol{\alpha}_\lambda=\begin{bmatrix}\boldsymbol{\alpha}_1\\\boldsymbol{\alpha}_2\\\boldsymbol{\alpha}_3\end{bmatrix}, \boldsymbol{\alpha}_\mu=\begin{bmatrix}\boldsymbol{\alpha}_5\\\boldsymbol{\alpha}_6\\\boldsymbol{\alpha}_7\end{bmatrix}, \tag{21}$$

Additionally, we condense the equality and inequality constraints from Eq. (13) for brevity:

$$\boldsymbol{\omega}_\lambda(t)=\begin{bmatrix}\mathbf{M}_b(\boldsymbol{\ell}_{ev}(t)+\boldsymbol{p}_{dis}(t)-\boldsymbol{p}_{ch}(t))+\mathbf{M}_g\boldsymbol{g}(t)-\boldsymbol{\ell}_d(t)-\mathbf{B}\boldsymbol{\vartheta}(t)\\\left(\ell_{ev}^i(t)-\sum_{\forall\boldsymbol{x}}\sum_{\boldsymbol{x}'\in\mathcal{U}_{\boldsymbol{x}}}\left(\rho\left(\boldsymbol{x}'-\boldsymbol{x}\right)\partial d_{\boldsymbol{x},\boldsymbol{x}'}^i(t)\right)_r\right)_{\forall i}\\\boldsymbol{soc}(t)-\boldsymbol{soc}(t-1)+\frac{\Delta t}{E_{cap}}\left(\eta_{ch}\boldsymbol{p}_{ch}(t)-\frac{\boldsymbol{p}_{dis}(t)}{\eta_{dis}}\right)\end{bmatrix} \tag{22}$$

$$\boldsymbol{\omega}_\mu(t)=\begin{bmatrix}|\mathbf{K}\boldsymbol{\vartheta}(t)|-s_{\max}\\\boldsymbol{soc}(t)-\boldsymbol{soc}_{max}\\\boldsymbol{soc}_{min}-\boldsymbol{soc}(t)\end{bmatrix}_+ \tag{23}$$

With the definition of (20), the augmented Lagrangian is:

$$\min_{\phi}\mathcal{L}_\phi=-Q_{\xi_1'}(s(t-1),\pi_\phi(s(t-1)))+\Bigg(\boldsymbol{\lambda}^\top\boldsymbol{\omega}_\lambda(t)+\boldsymbol{\mu}^\top\boldsymbol{\omega}_\mu(t)+\left\|\begin{bmatrix}\mathrm{diag}(\boldsymbol{\alpha}_\lambda)&\mathbf{0}\\\mathbf{0}&\mathrm{diag}(\boldsymbol{\alpha}_\mu)\end{bmatrix}\begin{bmatrix}\boldsymbol{\omega}_\lambda(t)\\\boldsymbol{\omega}_\mu(t)\end{bmatrix}\right\|_2^2\Bigg) \tag{24}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ represent the vectors of dual variables, while $\boldsymbol{\alpha}_\lambda$ and $\boldsymbol{\alpha}_\mu$ are positive scalars penalizing the augmented terms.

To perform the primal-dual update, one must first minimize the Lagrangian function followed by maximizing the dual function. Based on the definition provided in (20), the update for the dual variables' gradient is as follows:

$$\begin{aligned} \boldsymbol{\lambda}^{k+1} &\leftarrow \boldsymbol{\lambda}^{k} + \mathrm{diag}(\boldsymbol{\alpha}_\lambda)\boldsymbol{\omega}_\lambda(t), \\ \boldsymbol{\mu}^{k+1} &\leftarrow \boldsymbol{\mu}^{k} + \mathrm{diag}(\boldsymbol{\alpha}_\mu)\boldsymbol{\omega}_\mu(t), \end{aligned} \tag{25}$$

where$\boldsymbol{\lambda}^{k+1}$ and $\boldsymbol{\mu}^{k+1}$ are refined using batch samples. We alternate between primal and dual updates to hone the ϕ of $\pi_\phi(\cdot)$, ensuring adherence to both equality and inequality constraints. A convergence proof is available in our earlier research [16], which also examines the Lagrangian function's attributes and local optimality. The time and memory requirements for the suggested method are elaborated upon in the same study [16].

C. Projection of Aggregated EV Demands

The CRL predicts control actions for EVs, represented as $\ell_{ev}(t+1)$, given the state $s(t)$. To achieve the disaggregation of the EV charging control, denoted as $\partial d^i_{\boldsymbol{x},\boldsymbol{x}'}(t), \forall \boldsymbol{x}, \forall \boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}$, it is feasible to address a small-scale dis-aggregation problem individually for each i. More specifically, $\forall i, t$, the corresponding dis-aggregation problem is independently resolved.

$$\min \left\| \ell^i_{ev}(t) - \sum_{\forall \boldsymbol{x}} \sum_{\boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}} \left(\rho\left(\boldsymbol{x}' - \boldsymbol{x}\right) \partial d^i_{\boldsymbol{x},\boldsymbol{x}'}(t)\right)_r \right\|_2^2 \tag{26}$$
$$s.t. \quad \partial d^i_{\boldsymbol{x},\boldsymbol{x}'}(t) \in (6)$$

Simultaneously, during batch training, we ensure that the forecasted $\ell_{ev}(t+1)$ closely approximates the value of $\sum_{\boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}} \left(\rho\left(\boldsymbol{x}' - \boldsymbol{x}\right) \partial d^i_{\boldsymbol{x},\boldsymbol{x}'}(t)\right)_r)_{\forall i}$ through the dual process as defined by equation (22). Here, $\partial d^i_{\boldsymbol{x},\boldsymbol{x}'}(t)$ represents the dis-aggregated feasible solutions provided by equation (26).

D. Constrained Reinforcement Learning Algorithm

The above-described constrained reinforcement learning algorithm is encapsulated in Algorithm 1. Initially, in Steps 1-3, we set the parameters for both the double critic (Q_{ξ_1}, Q_{ξ_2}) and target networks ($Q_{\xi_1'}$, $Q_{\xi_2'}$), as well as the actor network (π_ϕ). The procedure from Steps 5-8 pertains to data collection and transition tuple storage, capturing ($\boldsymbol{s}(t-1)$, $\boldsymbol{A}(t)$, $r(t)$, $\boldsymbol{s}(t)$). The critic networks are revised in Steps 9-11, the actor network in Steps 12-13, and the dual variables in Steps 15-16. Steps 17-24 focus on refining $\partial d_{\boldsymbol{x},\boldsymbol{x}'}$, constraining it according to (6).

V. Case Studies

Experiments use the IEEE 14-bus system with an EV station on Bus 8. The Poisson parameter, $\lambda(t)$, is set as $\lambda(t) = 12 * (1 + 0.5 * \sin(2\pi * t/24))$. Charging and slack times are limited to 6 hours. Time is considered in hourly increments, and constrained DRL training uses PyTorch. As shown in Figs. 2 and 3, demand data comes from the Texas power grid, with three scaled wind sources integrated into the 14-bus system.

Algorithm 1: Constrained Reinforcement Learning for OPF with aggregate EVs

```
1  Begin by setting the critic network $Q_{\xi_1}$, $Q_{\xi_2}$, and actor
     network $\pi_\phi$ with random parameters $\xi_1$, $\xi_2$, and $\phi$;
2  Set the target networks: $\xi_1'$ to $\xi_1$, $\xi_2'$ to $\xi_2$, and $\phi'$ to $\phi$;
3  Initiate the replay buffer as $\mathcal{B}$ and determine the primal and
     dual update frequencies as $pu$ and $du$;
4  for $t = 1 : T$ do
5      Select action based on policy: $\boldsymbol{A}(t) \sim \pi_\phi(\boldsymbol{s}(t-1))$;
6      Compute the reward using Eq. (12);
7      etermine the new state by taking the action:
         $\boldsymbol{s}(t) = \mathbf{env}(\boldsymbol{A}(t))$ by taking action $\boldsymbol{A}(t)$;
8      Store the transition tuple: $(\boldsymbol{s}(t-1), \boldsymbol{A}(t), r(t), \boldsymbol{s}(t))$ in
         $\mathcal{B}$;
9      Sample a mini-batch of $N$ transitions from $\mathcal{B}$:;
10     $y \leftarrow r(t) + \gamma \min_{i=1,2} Q_{\xi_i'}(\boldsymbol{s}^n, \pi_\phi(\boldsymbol{s}^n))$;
11     Update critics: $\xi_{i=1,2} \leftarrow$
         $\arg\min_{\xi_{i=1,2}} \frac{1}{N} \sum (y - Q_{\xi_{i=1,2}}(\boldsymbol{s}^{n-1}, \boldsymbol{A}^n))^2$;
12     if t mod pu then
13         Update $\phi$ using the deterministic policy gradient:
             $\phi \leftarrow \phi - \eta \nabla \mathcal{L}_\phi(\boldsymbol{s}^{n-1}, \boldsymbol{s}^n)$, where $\eta$ denotes the
             learning rate and $\mathcal{L}_\phi$ is elaborated in Eq. (24);
14         Update target networks as per Eq. (17);
15     if t mod du then
16         Update the dual variables by (25).
17 Function env ($\boldsymbol{A}(t)$)
18     Choose the action, i.e.,
         $\boldsymbol{A}(t) = [\boldsymbol{g}(t), \boldsymbol{\ell}_{ev}(t), \boldsymbol{p}_{ch}(t), \boldsymbol{p}_{dis}(t)]^\top$;
19     for $i \in \mathcal{B}$ do
20         $\hat{\ell}^i_{ev}(t)$ is mapped to the set (6) by Eq. (26);
21         Compute the aggregated EV demands
             $\ell^i_{ev}(t) = \sum_{\boldsymbol{x}} \sum_{\boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}} \rho\big((\boldsymbol{x}' - \boldsymbol{x})\, \partial d_{\boldsymbol{x},\boldsymbol{x}'}(t)\big)_r$;
22         Update $\boldsymbol{n}^i_{\boldsymbol{x}}(t)$ and $\boldsymbol{d}^i_{\boldsymbol{x},\boldsymbol{x}'}(t)$ by using the projected
             $\partial \boldsymbol{d}^i_{\boldsymbol{x},\boldsymbol{x}'}(t)\ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{U}_{\boldsymbol{x}}$ ;
23     Obtain $\boldsymbol{\vartheta}(t)$ by solving the DC power flow equation
         with $\boldsymbol{g}(t)$ and $\boldsymbol{\ell}_{ev}(t)$ fixed;
24     Return $\boldsymbol{s}(t) = [\boldsymbol{\vartheta}(t), \boldsymbol{soc}(t), (\boldsymbol{n}^i_{\boldsymbol{x}}(t))_{\forall i}]^\top$
```

A tri-layer neural network with 512 neurons/layer forms our critic, target, and actor networks. Parameters are updated using Adam optimization with a learning rate of 10^{-3}. After each step, networks train on a mini-batch of 128 transitions from the agent's entire history, shown as tuples ($\boldsymbol{s}(t-1)$, $\boldsymbol{A}(t)$, $r(t)$, $\boldsymbol{s}(t)$).

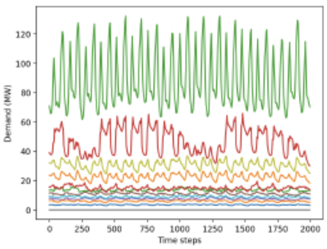

Figure 2. The power demands over 14 buses.

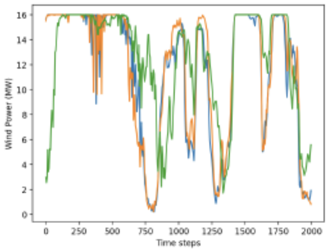

Figure 3. The wind power generations over 3 buses.

In Fig. 4, we display the learning curves derived from our experiments. The vertical axis represents rewards, as described in Eq. (12), observed at each point in time, while the horizontal axis corresponds to the time step. As evident in Fig. 4, the learning policy adheres closely to the optimal curve, thereby

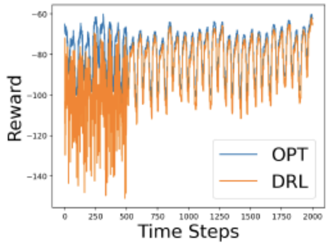

Figure 4. The training curves of the Constrained DRL and its corresponding optimal rewards by the optimization method.

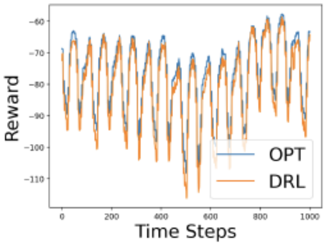

Figure 5. The testing curves of the Constrained DRL and its corresponding optimal rewards by the optimization method.

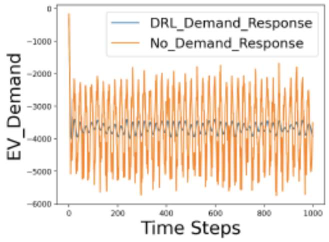

Figure 6. A comparison of the aggregate EV demand prediction curves generated by the CRL method with demand responses, juxtaposed without demand responses.

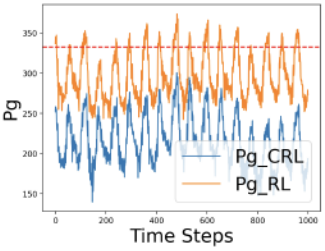

Figure 7. Power generations of slack buses by the proposed constrained DRL and DRL without constraints. The red line represents the upper and lower bounds.

suggesting that our method is adept at forecasting optimal actions even in the absence of future information. In fact, the average Normalized MAE (NMAE) between the DRL solution and the oracle optimization method is a mere $1.8779e-02$. Using the policy trained by the constrained DRL, we test it against future demands for 1000 samples. Comparing with baseline optimal results and globally optimal results with future demand knowledge, we observe in Fig. 5 that our policy closely approximates optimal actions, even without future information regarding the demands and EV arrivals.

As illustrated in Fig. 6, we conduct a comparative evaluation between the forecasted aggregate EV demand with demand response strategies and the same without these strategies. The results demonstrate that the CRL method can effectively regulate the demands, resulting in smoother power usage and significantly reduced peak loads. We conducted tests on the average fuel cost, given by $\frac{\sum_{t=1}^{1000}\sum_{i\in\mathcal{G}}(\alpha_i g_i^2(t)+\beta_i g_i(t)+\gamma_i)}{1000}$. By introducing the DRL with control EV demand response, we observed a reduction of 31.35% in the fuel cost when compared to scenarios without DRL demand response. The demand response of EVs significantly reduces the average fuel cost, as the control policy leverages renewable energy to meet EV consumption, thereby minimizing fuel expenses. We evaluate the feasibility of our policy, emphasizing the risk of infeasibility without dual updates. For instance, extremely small $g_i(t), i \in \mathcal{G}_n$ predictions by RL can lead to upper bound violations for $g_i(t), i \in \mathcal{G}_s$. Fig. 7 showcases a comparison between the violation rates of CRL and RL. Our constrained DRL consistently achieves 100% feasibility, surpassing the traditional DRL's rate of 83.87%. Additionally, within Fig. 7, the wind powers are fully utilized, making $g_i(t)$ appear conservative, which aids in minimizing fuel generations.

VI. Conclusion

Our study proposed an advanced control model that adeptly integrates collective EV charging demands into power grid systems. By leveraging the flexibility of EV charging and slack times, our model efficiently manages electrical loads in a scalable fashion. Our use of CRL techniques demonstrates their effectiveness in handling unpredictable future EV arrivals while ensuring control actions' feasibility. Our numerical studies on IEEE standard systems underscore our approach's outstanding performance as it dynamically adapts to the changing EV environment while always maintaining safety constraints.

References

[1] M. Tran, D. Banister, J. D. Bishop, and M. D. McCulloch, "Realizing the electric-vehicle revolution," *Nature climate change*, vol. 2, no. 5, pp. 328–333, 2012.

[2] H. Nazaripouya, B. Wang, and D. Black, "Electric vehicles and climate change: Additional contribution and improved economic justification," *IEEE Electrification Magazine*, vol. 7, no. 2, pp. 33–39, 2019.

[3] M. Alizadeh, A. Scaglione, J. Davies, and K. S. Kurani, "A scalable stochastic model for the electricity demand of electric and plug-in hybrid vehicles," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 848–860, 2013.

[4] K. Hreinsson, A. Scaglione, and V. Vittal, "Aggregate load models for demand response: Exploring flexibility," in *2016 GlobalSIP*. IEEE, 2016, pp. 926–930.

[5] B. K. Sovacool, J. Kester, L. Noel, and G. Z. de Rubens, "Actors, business models, and innovation activity systems for vehicle-to-grid (v2g) technology: A comprehensive review," *Renewable and Sustainable Energy Reviews*, vol. 131, p. 109963, 2020.

[6] M. Alizadeh, A. Scaglione, A. Applebaum, G. Kesidis, and K. Levitt, "Reduced-order load models for large populations of flexible appliances," *IEEE Trans. Power Syst.*, vol. 30, no. 4, pp. 1758–1774, 2014.

[7] K. Hreinsson, A. Scaglione, M. Alizadeh, and Y. Chen, "New insights from the shapley-folkman lemma on dispatchable demand in energy markets," *IEEE Transactions on Power Systems*, vol. 36, no. 5, pp. 4028–4041, 2021.

[8] N. Li, L. Chen, and S. H. Low, "Optimal demand response based on utility maximization in power networks," in *2011 IEEE power and energy society general meeting*. IEEE, 2011, pp. 1–8.

[9] W. Shi, N. Li, X. Xie, C.-C. Chu, and R. Gadh, "Optimal residential demand response in distribution networks," *IEEE journal on selected areas in communications*, vol. 32, no. 7, pp. 1441–1450, 2014.

[10] M. H. Christensen, C. Ernewein, and P. Pinson, "Demand response through price-setting multi-agent reinforcement learning," in *Proceedings of the 1st International Workshop on Reinforcement Learning for Energy Management in Buildings & Cities*, 2020, pp. 1–5.

[11] O. Sundstrom and C. Binding, "Flexible charging optimization for electric vehicles considering distribution grid constraints," *IEEE Transactions on Smart grid*, vol. 3, no. 1, pp. 26–37, 2011.

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

[13] T. Wu, I. L. Carreño, A. Scaglione, and D. Arnold, "Spatio-temporal graph convolutional neural networks for physics-aware grid learning algorithms," *IEEE Trans. Smart Grid*, 2023.

[14] T. Wu, A. Scaglione, and D. Arnold, "Reinforcement learning using physics inspired graph convolutional neural networks," in *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2022, pp. 1–8.

[15] Y. Liu, A. Halev, and X. Liu, "Policy learning with constraints in model-free reinforcement learning: A survey." in *IJCAI*, 2021, pp. 4508–4515.

[16] T. Wu, A. Scaglione, and D. Arnold, "Constrained reinforcement learning for stochastic dynamic optimal power flow control," *arXiv preprint arXiv:2302.10382*, 2023.

[17] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, "Learning safe policies via primal-dual methods," in *2019 IEEE 58th CDC*. IEEE, 2019, pp. 6491–6497.

[18] S. Qiu, X. Wei, Z. Yang, J. Ye, and Z. Wang, "Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss," *NeurIPS*, vol. 33, pp. 15 277–15 287, 2020.

[19] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *ICML*, 2018, pp. 1587–1596.

[20] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[21] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.