The Dissertation Committee for Kyung-bin Kwon
certifies that this is the approved version of the following dissertation:

Reinforcement Learning for Enhancing the Stability and Management of Power Systems with New Resources

Committee:

Hao Zhu, Supervisor

Surya Santoso

Alex Q. Huang

Sandeep Chinchali

Grani A. Hanasusanto

Reinforcement Learning for Enhancing the Stability and Management of Power Systems with New Resources

by

Kyung-bin Kwon

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2023

To my family.

Acknowledgments

First and foremost, I wish to express deep gratitude to my advisor, Dr. Hao Zhu. Working with Dr. Zhu over the past several years has been a remarkable privilege. Her profound impact on my development as an independent researcher is immeasurable, and I extend heartfelt thanks for her unwavering guidance and support.

Appreciation also goes to my committee members, who generously aided me throughout this journey. I would like to thank Dr. Surya Santoso for his expertise in power quality. I am thankful to Dr. Grani A. Hanasusanto for his exceptional instruction on optimization under uncertainty. Special appreciation is due to Dr. Alex Q. Huang and Dr. Sandeep Chinchali for their constructive advices on my dissertation.

I express my gratitude to Dr. Vijay Gupta and Dr. Lintao Ye for their collaboration on risk-aware LQR works. The financial support from NSF grants 1802319 and 1952193 is acknowledged with deep gratitude for making this research possible.

Heartfelt thanks to all the researchers at Pacific Northwest National Laboratory (PNNL), including Dr. Thanh Long Vu, Dr. Sayak Mukherjee, Dr. Veronica Adetola, Dr. Soumya Kundu, Dr. Kaustav Chatterjee, and Dr. Sameer Nekkalapu, with whom I collaborated during my internship. The experience was invaluable,

leading me to choose PNNL as my next step after graduation.

I am indebted to my fellow students and friends at the University of Texas at Austin. Special thanks to Yuqi Zhou, Shaohui Liu, Shanny Lin, Jeehyun Park, Pablo Paz Salazar, Fabricio Espinoza, Young-ho Cho, Ayat Albuali, Mohamad Fares El Hajj Chehade, Wei-Chun Chang, Nora Agah, Taehyung Kim, Woosung Kim and Dr. Hyunkoo Kang for the cherished memories we have shared.

Gratitude is extended to my family, whose love and support made this dissertation possible. Their encouragement, patience, and belief in my abilities have been indispensable.

Lastly, I want to express heartfelt appreciation to Dr. Minkyung Sung, my life partner and lifelong mentor, affectionately referred to as Ella. From the beginning of my Ph.D. journey, her unwavering support, extending beyond the academic realm to every facet of life, has been a source of immeasurable strength and inspiration. Her belief in my abilities, even during the most challenging moments, kept me moving forward. Beyond her intellectual contributions, her love and understanding created a nurturing environment that allowed me to focus on my research with a clear mind. I am profoundly grateful to Ella for her continuous presence as both my life partner and mentor.

Reinforcement Learning for Enhancing the Stability and Management of Power Systems with New Resources

Publication No. ____________________

Kyung-bin Kwon, Ph.D.
The University of Texas at Austin, 2023

Supervisor: Hao Zhu

Modern power systems face numerous challenges due to uncertainties arising from factors such as renewable energy source intermittency, stochastic load demand, and evolving grid dynamics. These uncertainties can lead to imbalances in power supply and demand, resulting in frequency and voltage deviations and, in extreme cases, blackouts. To address these challenges, advanced control and optimization techniques, particularly reinforcement learning (RL), have gained significant interest in ensuring efficient and reliable power system operations. RL offers a promising approach for decision-making under uncertainty, enabling agents to learn optimal policies without explicit uncertainty modeling. This thesis explores the application of RL to two classes of operational problems within power systems.

The first class focuses on power system resource management, including optimal battery control (OBC) and electric vehicle charging station (EVCS) opera-

tion. Challenges arise when formulating these problems as Markov Decision Process (MDP) to adopt RL. For example, incorporating cycle-based degradation costs into the MDP for OBC is not straightforward due to its dependence on past state of charge (SoC) trajectories. Similarly, the state and action spaces in EVCS problem scale with the number of EVs, leading to high-dimensional MDP formulations. This thesis proposes RL-based solutions for these resource management problems, while addressing the challenges by incorporating precise battery degradation model and efficient aggregation schemes to MDP.

The second class of problems deals with wide-area dynamics control for power system stability enhancement. Here, it is crucial for RL approaches to account for risk measures in offline-trained RL policies, considering uncertainties and perturbations in practice. The thesis focuses on load frequency control (LFC), which is vulnerable to variability due to high load perturbations, especially in small-scale systems like networked microgrids. Additionally, wide-area damping control (WADC) relies on communication networks, and communication delays can negatively impact its performance, given its fast time-scale. Moreover, the increasing integration of grid-forming inverters (GFMs) poses challenges in accurately modeling the overall system dynamics, which results in high variability in the system. To address these uncertainties and perturbations, this thesis integrates a mean-variance risk constraint into classic linear quadratic regulator (LQR) problems with linearized dynamics, limiting deviations of state costs from their expected values and reducing system variability in worst-case scenarios. In addition, structured feedback controllers need to be considered to match specific information-exchange

graphs, which complicates the geometry of feasible region.

To design risk-aware controllers for constrained LQR problems, a stochastic gradient-descent with max-oracle (SGDmax) algorithm is developed. This algorithm ensures convergence to a stationary point with a high probability, making it computationally efficient as it solves the inner loop problem of a dual problem easily and utilizes zero-order policy gradients (ZOPG) to estimate unbiased gradients, eliminating the need to compute first-order values. The policy gradient nature of SGDmax also allows the incorporation of structure by considering only non-zero entries in the ZOPG.

In summary, this thesis presents RL applications for effectively managing emerging energy resources and enhancing the stability of interconnected power systems. The analytical and numerical results offer efficient and reliable solutions to address uncertainty, supporting the transition towards a sustainable and resilient electricity infrastructure.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modern power systems face numerous uncertainties, posing significant challenges in ensuring a reliable and cost-effective power supply. The inherent variability and intermittency of renewable energy sources (RES), coupled with the stochastic nature of load demand, can lead to imbalances in the supply-demand equation, resulting in frequency and voltage deviations, and, in extreme cases, even blackouts. To tackle these challenges, there is a growing interest in applying advanced control and optimization techniques, particularly reinforcement learning (RL), to enable efficient and reliable operations of power systems [1].

RL presents as a promising approach for decision-making under uncertainty, allowing an agent to learn the optimal policies without the need for explicit uncertainty modeling [2]. Specifically, RL can be applied to address two classes of operational problems within power systems. The first one pertains to power system resource management, encompassing the optimal battery control (OBC) [3, 4] and the scheduling of electric vehicle charging station (EVCS) [5, 6]. The second one involves wide-area dynamics control for stability enhancement, by utilizing emerging grid-connected resources like voltage source converters (VSC) [7, 8] or grid-forming inverters (GFM) [9].

Nevertheless, adopting RL to address these emerging problems with the integration of new energy resources presents several challenges. First, formulating the problems into the form of Markov Decision Process (MDP) is a necessary step, but the MDP representations for problems with complicated and large number of resources can be complex and high-dimensional. Taking the example of the OBC problem, incorporating cycle-based degradation cost into the MDP is not straightforward, as this cost relies on the past state of charge (SoC) trajectory, not just the instantaneous SoC [10]. Additionally, for the EVCS scheduling problem, the state and action spaces scale up with the number of electric vehicles (EVs), resulting in a large and time-varying dimensionality issue in its MDP formulation [11, 12].

For the second class of grid dynamics control problem, it is of high importance for RL approaches to account for the risk measures of the offline trained RL policies, due to potential uncertainty and large perturbation factors in practice like communication delays and modeling mismatches. For example, the multi-area load frequency control (LFC) problem is vulnerable to variability caused by high load perturbations, particularly for small-scale systems like networked microgrids (MGs) [13]. As for the wide-area damping control (WADC) problem, it is known to rely on dedicated communication networks and thus the communication delays therein could negatively affect the WADC performance with the latter's very fast time-scale [14]. Last but not least, considering the increasing integration of GFMs for grid dynamics control, the lack of accurate modeling information poses as a significant issue for perfectly representing the overall system dynamics [15, 16]. All of these aforementioned uncertainty/perturbation factors can adversely affect the worst-case performance of grid dynamics control in terms of increasing the oscillation level, thus significantly reducing the stability margin of next-generation power systems.

Our proposed RL approaches for power system resource management and dynamics control problems will address these domain-specific challenges faced by a generic RL framework. First, this thesis proposes RL-based approaches for operating emerging energy resources, including batteries and EVCS, which can provide valuable flexibility to grid operations. Specifically, we consider a utility-scale battery that participates in both the real-time electricity market and frequency regulation typed ancillary services. To incorporate precise battery degradation modeling, we develop a new representation of cycle-based degradation cost based on the rainflow algorithm, that can easily deal with the past SoC trajectory issue. Furthermore, we solve the EVCS scheduling problem, by minimizing the total electricity cost while meeting the EV charging demands. Here, we propose state and action aggregation scheme based on a least-laxity first (LLF) rule and come up with an efficient and equivalent MDP representation with fixed and low problem dimensions.

For the second class of dynamics control problems, we develop a risk-aware RL framework to systematically address the various uncertainty factors arising in practical implementations such as high load perturbations in LFC, communication delays in WADC, and modeling errors in GFM problem, as previously discussed. To mitigate the increased system variability resulting from these uncertainty factors, we integrate a mean-variance risk constraint after formulating the problem as classic linear quadratic regulator (LQR) with linearized dynamics. Bounding the mean-variance risk limits the deviations of state cost from its expected value, thus mitigating the high system variability particularly in worst-case scenarios. Furthermore, we need to consider a structured feedback controller that follows the specific information-exchange graph, which is practically important due to limited communication links. This structured constraint leads to a complicated geometry of the feasible region, which makes the analysis much more difficult than the full feedback case. In order to design a risk-aware controller by solving the constrained LQR problems, we develop a stochastic gradient-descent with max-oracle (SGDmax) algorithm which can guarantee convergence to a stationary point with a high probability. The algorithm is computationally efficient as it easily solve the inner loop problem of a dual problem and utilizes zero-order policy gradient (ZOPG) to estimate unbiased gradients without the need to compute first-order values. Additionally, the policy gradient nature of SGDmax makes it easy to incorporate structure by only considering the non-zero entries in the structure in the ZOPG.

The contributions of this thesis are three-fold: First, we integrate RL into energy resource control, with the goal of developing efficient and equivalent representations that can enable effective RL training. Specifically, we introduce an approach to compute the cycle-based degradation cost as instantaneous rewards in the OBC problem, and propose an equivalent state and action aggregation to achieve a time-invariant state/action formulation in the EVCS problem. Numerical tests validate that the proposed methods lead to significant reduction of the testing costs, attributed to the benefits of our proposed equivalent modeling. Second, we design risk-aware RL strategies to address the increasingly variability in power system dynamics control problems. We focus on LFC, WADC, and GFM control problems, accounting for uncertainty factors such as high load variability, communication delays, and modeling errors. To enhance worst-case performance, we incorporate a mean-variance risk constraint which effectively reduces state deviations. Numerical results demonstrate that the worst-case performance is significantly improved by mitigating the system variability. Third, the RL approaches proposed for both energy resource control and power system stability control are versatile and can be applied to various problems within the power system domain. Especially, we develop the SGDmax algorithm, which can effectively solve the risk-constrained LQR problem with a high probability of convergence and computational efficiency. This algorithm can be adopted to various power system dynamics problems that can be formulated as LQR problems, thereby providing risk-aware control policies through the integration of RL.

Overall, this thesis presents useful RL methods for managing emerging energy resources and for enhancing the stability of interconnected power systems. The analytical and numerical results in this thesis provide efficient and reliable solutions to address uncertainty, heterogeneity, and complexity factors arising from the transition to an sustainable and resilient electricity infrastructure.

The dissertation is organized as follows: Chapter 2 develops an RL-based battery control strategy that considers the cycle-based battery degradation cost. This chapter begins with an introduction to the optimal battery control (OBC) problem and outlines the motivation behind this research. Key variables necessary for modeling the battery control problem as a Markov Decision Process (MDP) are defined.

Additionally, the cycle-based degradation cost is modeled using the rainflow algorithm, and a novel approach for representing it as an instantaneous cost through state augmentation is introduced. The OBC problem is formalized, and we present the RL solution technique, known as the deep Q-network (DQN) method. Chapter 3 focuses on the development of a control policy for EVCS using an efficient MDP representation. It starts by introducing the EVCS problem and the motivation behind this work in the introductory section. Following that, the EVCS operations problem is formulated as an MDP. The chapter then continues by developing the least-laxity first (LLF)-based action reduction and introducing a novel equivalent state aggregation approach to address concerns related to dimensionality. Building upon these developments, the RL approach is presented, which utilizes policy gradient and linear Gaussian policy parameterization.

From Chapter 4 on, the class of grid dynamics control problems are considered. Chapter 4 designs a risk-aware LFC controller that takes into consideration a mean-variance risk constraint and structured feedback. It begins with an introduction to the LFC problem and outlines the motivation behind this research. The LFC problem is then formulated based on a radially-connected networked microgrid (MG) system. In a subsequent section, a general infinite-horizon risk-constrained linear quadratic regulation (LQR) problem is formulated, incorporating structured feedback control. The chapter also introduces a dual-related minimax reformulation and analyzes the convergence of the Gradient Descent with max-oracle (GDmax) algorithm. Furthermore, the chapter extends this framework to model-free learning by introducing the Stochastic (S)GDmax algorithm through zero-order policy gradient. Chapter 5 develops a risk-aware wide-area damping controller using RL, to tackle communication delays within the information-exchange network. It initiates with an introduction to the WADC problem and outlines the underlying motivation for this research. A linearized system model is then formulated by combining the dynamics of both synchronous generators and voltage source converters (VSCs). Additionally, the chapter addresses the modeling of communication networks, the analysis of delay impacts, and the formulation of a risk-constrained LQR problem. Finally, in Chapter 6, we explore the development of a risk-aware GFM controller while considering model parameter mismatch of synchronous generators (SGs) and GFMs. It commences with an introduction to the GFM problem and outlines the motivation behind this research. Subsequently, the chapter formulates a system model that considers the dynamics of both synchronous generators and GFMs, with their interactions accounted for through network coupling. Furthermore, we design a risk-constrained GFM problem, incorporating a mean-variance risk constraint to mitigate frequency oscillations resulting from parametric mismatch in the system model.

Chapter 2

RL-based Optimal Battery Control

This chapter develops an RL-based battery control considering a cycle-based battery degradation cost. Section 2.1 introduces an optimal battery control (OBC) problem and the motivation of this work. Section 2.2 formulates the key variables for modeling the battery control problem into the Markov Decision Process (MDP) form. In Section 2.3, we model the cycle-based degradation cost using the rainflow algorithm, and develop a new approach to represent it as instantaneous cost through state augmentation. Section 2.4 formalizes the OBC problem and presents the deep Q-network (DQN) method as the RL solution technique. Numerical results using real-world data are presented in Section 2.5 to validate the performance improvement of the proposed degradation model, as compared to earlier approach using linearized approximation.

This chapter is based on the following publications:

K. Kwon and H. Zhu, "Reinforcement Learning-Based Optimal Battery Control Under Cycle-Based Degradation Cost," *IEEE Transactions on Smart Grid,* vol. 13, no. 6, pp. 4909-4917, 2022, doi: 10.1109/TSG.2022.3180674.

2.1 Optimal Battery Control Problem

Battery energy storage systems as flexible resources are a key technology to enable the decarbonization of electricity infrastructure in future [17, 18]. Particularly, utility-level battery systems can be used to increase the payoff from electricity market via energy arbitrage [19], while contributing to the grid's power balance through participating in ancillary services [20]. It is crucial to develop effective strategies for real-time battery operations in order to utilize its flexibility potentials to mitigate the increasing uncertainty introduced by renewable or non-controllable loads.

The optimal battery control (OBC) problem for determining the (dis)charging policies has been popularly considered to reduce a combination of battery operational costs. It aims to reduce the net cost for electricity usage and frequency regulation (FR) penalty, as well as possible violations of network constraints; see e.g., [3, 4, 21, 22]. In addition, battery's cycle life as characterized by the degradation cost is especially needed when participating in FR or other fast services [3,21]. Unlike other costs that mostly depend on the instantaneous battery status, the modeling of battery degradation is *cycle-based* according to the full trajectory of battery's state of charge (SoC). It requires the identification of all charging/discharging cycles using the so-termed *rainflow* algorithm [10]. Thus, degradation-aware OBC problem results in increased complexity as shown by [3, 21].

Due to the fast dynamics in prices or load demands, the OBC solution can be greatly affected by the uncertainty of future information. To address this issue, a model-predictive control (MPC) framework has been widely used by optimizing the current action according to the predicted input values for a fixed time window; see e.g., [23–25]. Nonetheless, the FR signal exhibits very minimal temporal correlation [26], leading to significant difficulty in predicting it and thus applying MPC for reducing FR penalty. Furthermore, even though battery health has been considered in MPC-based OBC work [27, 28], the cycle-based degradation model is largely missing.

The goal of our work is to develop a modeling approach to precisely represent the battery degradation cost and use it for the design of RL-based OBC algorithm. The overall objective includes the net electricity cost, FR penalty, and cycle-based degradation cost. The main modeling challenge lies in the latter as it is determined by the battery's full SoC trajectory. Based on the rainflow algorithm, the complex process of material fatigue is associated with the stress level of each individual charging or discharging cycle [29]. Thus, the degradation cost is an exponentially increasing function of cycle depth [10], and the latter strongly depends on the past trajectory of battery status. This leads to a pronounced mismatch with the Markov Decision Process (MDP) form used by RL algorithms, as the latter would represent the problem objective as functions of instantaneous states and actions only. The aforementioned approach of linearizing the degradation cost as in [30, 31] fails to recognize this exponential relation with the cycle depth, and unfortunately can lead to deep (dis)charging cycles that may not be overall profitable.

To this end, we have analytically shown that it is possible to keep track of the battery cycles by augmenting the state with the more recent switching points (SPs) along the SoC trajectory. These critical transition points between charging and discharging sessions are extremely useful for identifying the correct cycle depth according to the rainflow condition. In addition, they allow for decomposing the degradation cost of a full (dis)charging cycle into incremental differences between consecutive time instances in the form of instantaneous cost. This proposed representation of battery degradation cost helps to deploy state-of-the-art RL algorithms to learn the OBC policy. We have used the DQN technique to search for the parameters of the action-value function, or Q-function, associated with the resultant MDP form.

2.2 System Modeling

This work considers the optimal battery control (OBC) for maximizing the economic pay-off while accounting for the battery degradation. The pay-off is from energy market participation and also the provision of FR service, as discussed later. One notable feature of the present work is the consideration of battery degradation cost, which can greatly increase the life-cycle under any general pay-off model [3].

To determine the battery's effective (dis)charging power $b_t \in [\underline{b},\ \bar{b}]$ at each discrete-time instance $t = 0, 1, \ldots$, we introduce a list of state variables based on battery status or external inputs.

- $c_t \in [\underline{c},\ \bar{c}]$: normalized state of charge (SoC) of the battery;
- p_t: electricity market price;
- f_t: frequency regulation (FR) signal.

Note that the SoC is normalized by the maximum capacity; i.e., $c_t \in [0, 1]$. It is also an internal battery state affected by the past actions $\{b_\tau\}$, whereas the other states are received from grid operators and thus are not directly action-dependent.

To leverage reinforcement learning algorithms for this problem, we consider it as a Markov Decision Process (MDP) [2, Ch. 3] denoted by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, as detailed here.

State space $\mathcal{S}$ contains the set of feasible values for the system state s_t, including both the SoC c_t, and the other inputs p_t and f_t which affect the economic benefits. Additional state variables will be specified in Section 2.4 for representing cycle-based degradation cost. State dynamics need to follow the Markov property as discussed soon.

Action space $\mathcal{A}$ includes the set of decisions that battery can take. We consider a discrete multi-level set with a total of $|\mathcal{A}|$ actions, as

$$a_t \in \mathcal{A} = \{a^{(1)}, a^{(2)}, \cdots, a^{(|\mathcal{A}|)}\} \tag{2.1}$$

with normalized actions $a^{(n)} \in [-1,\ 1]$. Accordingly, the normalized (dis)charging power $b_t \in [\underline{b},\ \overline{b}]$ is set to be

$$b_t = \begin{cases} \min\{\overline{c} - c_t, \overline{b}a_t\} & \text{if } a_t \geq 0, \\ \max\{\underline{c} - c_t, \underline{b}a_t\} & \text{if } a_t < 0. \end{cases} \tag{2.2}$$

Continuous action space that directly determines $b_t = a_t$ is also possible. While this work focuses on a discrete $\mathcal{A}$, the RL algorithm can be generalized to continuous a_t as well.

Transition kernel $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ captures the system dynamics under the Markov property [2, Ch. 3]. For the input states such as price p_t, we assume $\Pr(p_{t+1}|\{p_\tau\}_{\tau=1}^{t}) = \Pr(p_{t+1}|p_t)$; and similarly for f_t. This is reasonable as the market price has very short-term memory [32], while FR signal f_t can be modeled as a white noise sequence of no memory [33]. A longer memory is possible too; such as the prices that follow $\Pr(p_{t+1}|\{p_\tau\}_{\tau=1}^{t}) = \Pr(p_{t+1}|p_t,\ p_{t-1})$. In this case, both p_t and p_{t-1} are included as the part of the state per time t to satisfy the Markov transition property.

Using Eq. (2.2), the SoC state c_t transitions as

$$c_{t+1} = c_t + b_t, \text{ with } b_t \text{ given in (2.2).} \tag{2.3}$$

For general action space with any $b_t \in [\underline{b}, \bar{b}]$, c_t is updated by

$$c_{t+1} = \begin{cases} \overline{c} & \text{if } \overline{c} - c_t \le b_t, \\ \underline{c} & \text{if } \underline{c} - c_t \ge b_t, \\ c_t + b_t & \text{otherwise.} \end{cases} \tag{2.4}$$

Reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ captures the learning objective. Notably, it is always the accumulated reward consisting of *instantaneous* terms, where per time t the latter only depends on the current state and action as

$$r_t = r_t(s_t, a_t) \tag{2.5}$$

In the following we will minimize the objective cost function h_t as negative reward, where its instantaneous property will be ensured after introducing additional state variables as detailed in Section 2.4.

Discount factor $\gamma \in (0, 1]$ is a constant to accumulate the total reward along the time horizon. Smaller γ values imply that future rewards are less important than current ones at a discounted rate [2, Ch. 3]. As we adopt a finite exploration time-horizon $\mathcal{T} = [1, \ldots, T]$ for the OBC problem, for simplicity $\gamma = 1$ will be used.

2.3 Modeling of Battery Degradation Cost

We consider three types of operational cost related to battery management. The energy cost relates to the electricity price according to (dis)charging, while the FR cost is based on its fast-varying flexibility. Under a contract of providing FR service, the battery would follow the f_t signal sent by the market operator as much as possible [3]. These two costs can be simply obtained by the state variables discussed so far. First, the net cost for electricity usage under (dis)charging power b_t can be represented as

$$h_t^e(p_t, b_t) = p_t b_t, \quad \forall\, t \in \mathcal{T}. \tag{2.6}$$

Second, using a penalty coefficient δ for deviation from FR signal f_t, one can form

$$h_t^f(f_t, b_t) = \delta |f_t - b_t|, \quad \forall t \in \mathcal{T}. \tag{2.7}$$

The energy cost in (2.6) is typically negative due to the energy arbitrage capability, while the FR penalty in (2.7) is always positive. This is because the additional economic benefit by participating in the FR contract is not included here. Overall, a battery should receive positive pay-off from these two tasks.

Remark 1. (Frequency regulation signal) *In practice, the FR signal is much faster than other system dynamics. For example, the real-time price is typically updated* *every 5 minutes, while the FR signal may be at 2-second rate [34]. To reduce the complexity of the training computation later on, we will down-sample the FR signal to attain $\{f_t\}$ at a slower rate for searching the policy. In testing and implementing the resultant policy, the original fast FR signal will be instead used to realistically evaluate the performance of the RL approach.*

As for the battery degradation cost, there are several stress factors affecting the battery lifetime such as temperature, high C-rates, average SoC, and Depth of Discharge (DoD) [10,35]. During daily battery operations, the DoD stress model is considered the most relevant while other factors may be minimally affected. This will be shown numerically in Section 2.5. According to the DoD stress model, the aging of battery cells mainly depends on material fatigue as a result of (dis)charging cycles of the SoC trajectory, especially due to following the FR signal [26]. Since this cycle-based degradation constitutes as a key battery lifetime consideration [36], the proposed OBC formulation to reduce it can greatly increase the battery's lifetime revenue.

Fig. 2.1 illustrates an example of battery SoC trajectory which consists of several charging and discharging cycles. The switching points (SPs), labeled by $A - E$, correspond to the transitions between charging and discharging and will be used for identifying the cycles by rainflow algorithm. For example, the trajectory $A - B - C - D$ consists of a long charging cycle with a small discharging part from $B - C$. The respective depths of these two cycles, defined as the absolute SoC differences between the start and end SPs, are d_0 and d_1. As d_1 is smaller than the difference between $A - B$ and that between $C - D$, this trajectory is thus

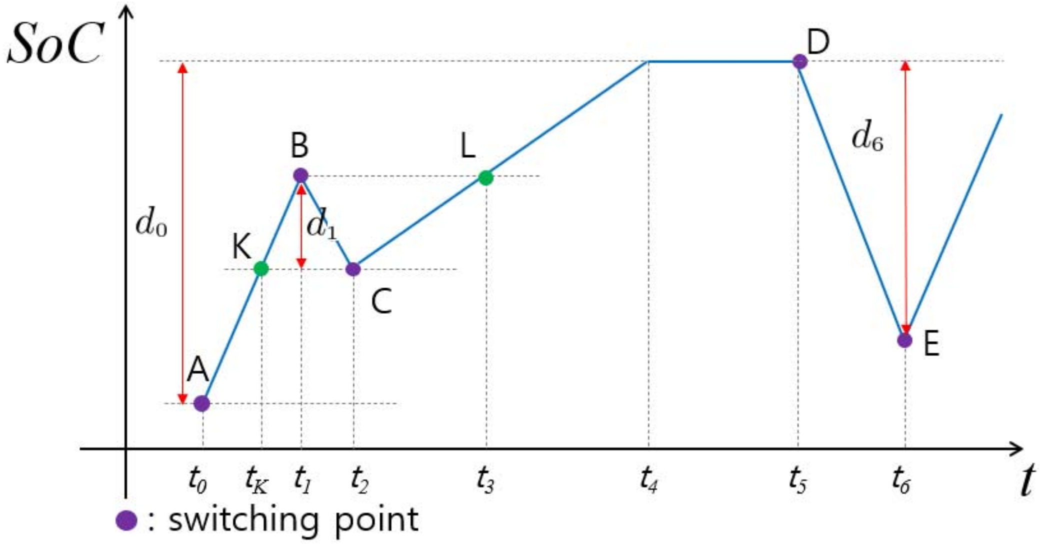


Figure 2.1: An example of battery SoC trajectory used for modeling the battery degradation cost based on the rainflow algorithm.

divided into the *full cycle* from $K - B - C$ of depth d_1 and the other *half cycle* from $A - K(C) - D$ of depth d_0. This is the so-called rainflow condition as stated in Lemma 1; see e.g., [3].

Lemma 1. *The SoC values of the last three SPs by time t are sufficient for evaluating the rainflow condition and determining the depth of (dis)charging cycles.*

Based on the cycle depth $d > 0$, the associated degradation cost is given by

$$\Phi(d) = \alpha_d e^{\beta d} \tag{2.8}$$

with positive constant coefficients α_d and β based on battery types [10, 31]. Recalling the normalized SoC $c_t \in [0, 1]$, we have the cycle depth $d \in [0, 1]$ as well. Note that for any pair in $\mathcal{D} := \{(d_1, d_2) : \ d_1, \ d_2 \geq 0, \ d_1 + d_2 \leq 1\}$, we can show that $e^{(d_1+d_2)} \leq e^{d_1} + e^{d_2}$. This is because the maximum value of the function $g(d_1, d_2) := e^{(d_1+d_2)} - e^{d_1} - e^{d_2}$ for the simplex $\mathcal{D}$ equals to $(e - 2e^{0.5}) < 0$, which is attained at $(d_1, d_2) = (0.5, 0.5)$. Thus, to reduce the degradation cost a single (dis)charging cycle that is longer and deeper is typically preferred, as opposed to the combination of multiple shorter cycles. This intuitive rule for cycle-based degradation model will be demonstrated later on in numerical tests. Unfortunately, this cycle-based degradation cost depends on the past SoC trajectory, and unfortunately, it does not follow the accumulated form of instantaneous terms as in Eq. (2.5).

Linearized degradation model has been developed in [30] to compute the averaged degradation coefficient from past SoC trajectory. Specifically, a degradation coefficient α_d is first determined using a given SoC trajectory over $\mathcal{T}$ as

$$a_d = \frac{\sum_{i=0}^{\bar{N}} \Phi(\bar{d}_i)}{\sum_{t=0}^{T} |\bar{b}_t|} \tag{2.9}$$

by averaging the total degradation costs of the $(\bar{N} + 1)$ cycles over the accumulative absolute charging power throughout the sample trajectory. This way, the instantaneous degradation cost for any new SoC trajectory is approximated by

$$h_t^d(b_t) \approxeq -a_d |b_t|. \tag{2.10}$$

This linearized degradation cost model can be easily computed once a_d is known. However, this approximation inexplicitly assumes that the new trajectory should be very similar to the given sample trajectory for computing a_d. To implement the RL algorithm later on, the coefficient a_d will be updated using the most recent trajectory during the sampling process. Nonetheless, as an approximation it does not represent

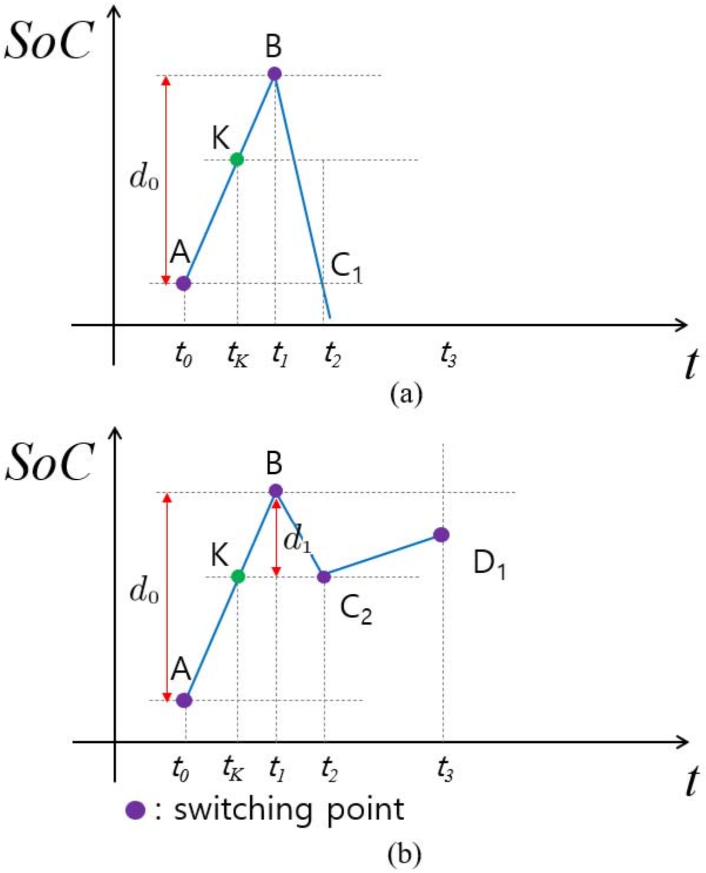


Figure 2.2: Two cases of rainflow condition not satisfied: (a) case NR_a and (b) case NR_b.

the actual cycle-based degradation cost and thus limits the RL algorithm's search for the best SoC trajectory.

Cycle-based degradation model will be pursued instead to address the approximation issue by augmenting the state s_t with the last three SPs before time t. As stated in Lemma 1, they are sufficient information for checking the rainflow condition. The state s_t now includes three additional variables, $c_t^{(0)}$, $c_t^{(1)}$, and $c_t^{(2)}$, as the SoC from the oldest SP to the latest one. Note that they may overlap if there are less than three SPs before time t. For example, at point K in Fig. 2.1, these three SP states all equal to the SoC of point A; and similarly for point C, we have $c_t^{(1)} = c_t^{(2)}$ equal to the SoC of B. The latest SP's SoC $c_t^{(2)}$ can be used to identify if the current instance t is a new SP, using the rule

$$b_t(c_t - c_t^{(2)}) < 0. \tag{2.11}$$

If Eq. (2.11) holds, we have a new SP and will update $\{c_t^{(i)}\}$ based on whether the rainflow condition is satisfied.

To update $\{c_t^{(i)}\}$, Fig. 2.2 illustrates two cases where the rainflow condition is not satisfied. Fig. 2.2(a) shows the SoC of the point C_1 is not within the range between A and B, while Fig. 2.2(b) indicates the SoC of the new SP D is within the range between C_2 and B. These cases are denoted by cases NR_a and NR_b, respectively. In either case, the oldest SP A will be removed while the remaining SPs will be used to update $\{c_{t+1}^{(i)}\}$, as listed in Table 2.1. In addition, case RA denotes the scenario of rainflow condition being satisfied, such as the point D in Fig. 2.1. This is because at SP D: i) the third SP C lies between the first two SPs A and B; and ii) the current SoC at SP D exceeds the range between the latest two SPs B and C. Hence, the trajectory $A - B - C - D$ is divided in to the long half-cycle $A - K - C - L$, and another full-cycle $K - B - C$ of depth d_1. After the RA case is satisfied, the two SPs B and C will be removed from the record. The SoC state updates for all three cases are summarized in Table 2.1.

Interestingly, the state transitions in Table 2.1 also allow for decomposing

Table 2.1: State transitions at a new SP identified at time t

Next state	NR_a	NR_b	RA
$c_{t+1}^{(0)}$	$c_t^{(1)}$	$c_t^{(1)}$	$c_t^{(0)}$
$c_{t+1}^{(1)}$	$c_t^{(1)}$	$c_t^{(2)}$	$c_t^{(0)}$
$c_{t+1}^{(2)}$	$c_t^{(1)}$	c_t	$c_t^{(0)}$

the cycle-based degradation cost into instantaneous difference term for each instance t. As the cycle depth changes according to b_t only, the degradation cost in Eq. (2.8) can be modeled by accumulating the following incremental term per time t:

$$h_t^d(b_t, c_t, c_t^{(2)}) = \alpha_d e^{\beta|c_t+b_t-c_t^{(2)}|} - \alpha_d e^{\beta|c_t-c_t^{(2)}|}. \tag{2.12}$$

Basically, the instantaneous degradation model in Eq. (2.12) accounts for difference of $\Phi(\cdot)$ due to the change of cycle-depth, which can be computed based on the latest SP state $c_t^{(2)}$. Therefore, summing up all instantaneous terms in Eq. (2.12) yields the total degradation cost, as formally stated in **Proposition 1** with the proof provided in the Appendix.

Proposition 1. *Under Lemma 1, the summation of the instantaneous terms in Eq.* (2.12) *throughout the time-horizon* $\mathcal{T} = [1, \cdots, T]$ *is exactly equivalent to the total cycle-based degradation cost along* $\mathcal{T}$.

2.4 Optimal Battery Control Algorithm

Thanks to our proposed model of instantaneous degradation cost, we can define the MDP form for the OBC problem.

First, each state is given by

$$s_t = [p_t, f_t, c_t, c_t^{(0)}, c_t^{(1)}, c_t^{(2)}], \quad \forall t \in \mathcal{T} \tag{2.13}$$

which is used to determine the action a_t based on the policy of interest. The transition kernel $\mathcal{P}$ now includes the updates in Table 2.1, while the instantaneous reward is given by

$$r_t(s_t, a_t) = -h_t^e - h_t^f - h_t^d, \quad \forall t \in \mathcal{T}. \tag{2.14}$$

The battery control problem now becomes to determine the best policy π for forming the action as $a_t \sim \pi(s_t)$ with s_t given in Eq. (2.13). To simplify the policy search, we are particularly interested in the set of parameterized policies given by $\pi_\theta(\cdot) = \pi(\cdot; \theta)$, with parameter θ optimized through

$$\max_\theta \ \mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{T} \gamma^t r_t(s_t, \ a_t) \right]. \tag{2.15}$$

To solve Eq. (2.15), we can adopt certain RL algorithms to search for the optimal parameter θ; see e.g., [1]. We use the deep Q-networks (DQNs) [2, Ch. 9] here as a popular RL approach based on nonlinear neural network modeling. Accordingly, the parameter θ represents the DQN weights to be learned, and the DQN is used to obtain the so-termed Q-network that models the MDP's *action-value, or Q-function*, namely the expected total future reward under a given pair of state and action:

$$Q(s_t, a_t) := \mathbb{E}_{\pi_\theta} \left[\sum_{\tau=t}^{T} \gamma^{(\tau - t)} r_\tau(s_\tau, a_\tau) \middle| s_t, a_t \right]. \tag{2.16}$$

For the optimal Q-function, the Bellman optimality condition [1] states that:

$$Q^*(s_t, a_t) = r_t(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}} \left[\max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}) \middle| s_t, a_t \right]. \tag{2.17}$$

To find the optimal Q^*, we parameterize the action-value Q-function using θ as the NN weights, as denoted by $Q(s_t, a_t; \theta)$. The Bellman optimality in Eq. (2.17) can be used to develop iterative *gradient descent* updates to obtain the best θ. At each update, the Q-network on the right-hand side of Eq. (2.17) is kept constant as the *target network*, whereas the other one is varied to minimize the difference between both sides. Letting θ' denote the latest NN weights, we design the loss function for DQN training as the expected squared difference:

$$\mathcal{L}(\theta) = \mathbb{E}_{\{s_t, a_t, s_{t+1}\}} \Big[\big(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta') - Q(s_t, a_t; \theta) \big)^2 \Big]. \tag{2.18}$$

To minimize $\mathcal{L}(\theta)$, one can need to compute its gradient over the parameter θ given by

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta) = & \mathbb{E}_{\{s_t, a_t, s_{t+1}\}} \Big[-2 \Big(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta') \\ & - Q(s_t, a_t; \theta) \Big) \nabla_\theta Q(s_t, a_t; \theta) \Big]. \end{aligned} \tag{2.19}$$

Each gradient-based update relies on the estimate from sampling the trajectory such that the expectation in Eq. (2.19) is replaced by the sample average. To this end, the action a_t is sampled for given state s_t based on θ' as $a_t^* = \arg\max_{a_t} Q(s_t, a_t; \theta')$, $\forall t$. To ensure adequate exploration of the state space, the ϵ-greedy method [2, Ch. 2] can be used to randomize the action by selecting a_t^* with probability $(1-\epsilon)$ at every time. The value of ϵ would decrease as the DQN updates continue, typically at an exponential decreasing rate $\kappa \in (0, 1)$. This method can improve the exploration process at the beginning phase while eventually picking the optimal actions to attain convergence.

To improve the efficiency and stability of DQN implementation, we introduce two additional techniques. First, we implement the *experience replay* method [37] to efficiently use the past samples by storing all the past samples in the memory $\mathfrak{D} := \{(s_t, a_t, s_{t+1}, r_t)\}$ along the trajectory. When computing the loss function Eq. (2.18), a subset of samples denoted by mini-batch $\mathfrak{J}$ is randomly picked from $\mathfrak{D}$ and used as the samples for gradient estimation. This method can improve the training efficiency by selectively reusing past samples. In addition, we advocate the *fixed target network* approach [38] by keeping the target network parameter fixed for several updates. To this end, let θ^- denote the target network parameter, which is only updated once every N_o iterations. This technique could mitigate any potential instability issue by changing the DQN target weights less frequently. By adopting *experience replay* method and *fixed target network* approaches, we obtain the estimates of loss function and its gradient as

$$\begin{aligned}\nabla_\theta \hat{\mathcal{L}}(\theta) = (1/|\mathfrak{J}|) \sum_{t \in \mathfrak{J}} \Big[-2\Big(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-) \\ - Q(s_t, a_t; \theta) \Big) \nabla_\theta Q(s_t, a_t; \theta) \Big]. \qquad (2.20)\end{aligned}$$

The detailed algorithmic steps for DQN-based OBC algorithm are tabulated in Algorithm 1. As mentioned earlier, the state variables p_t and f_t are not action dependent. Thus, their transitions are obtained from the profiles given as the algorithm input, such as real data provided by the market operators. For the convergence of DQN algorithms, the total number of episodes N is typically chosen to be large enough in practice. For each episode n, there are total T samples from $t = 1$ to $t = T$. Note that Algorithm 1 can be used to search for the best policy under the linearized degradation cost as well, by using this simpler degradation cost model in Eq. (2.10). The ensuing section will compare these two degradation models numerically.

2.5 Numerical Tests

We have compared the proposed RL-based battery control algorithm under cycle-based degradation cost with the linearized approximation one [30]. Actual data of electricity market prices and FR signals have been used, respectively from the ERCOT's market data depository [39] and PJM's ancillary service datasets [40]. Each time instance corresponds to a 5-minute interval. The FR signal is normalized to indicate either maximum charging or discharging for the battery. As mentioned in Remark 1, the fast FR signal at 2-second rates is averaged over a 10-second interval for the training phase, while the original data rate is maintained for the testing phase. We have used a 200kWh-capacity battery with (dis)charging rate of 120kW and minimum SoC of 20kWh, which takes 90 minutes to fully (dis)charge. The multiple discrete action space is adopted with overall 11 actions, as $\mathcal{A} = \{-1, -0.8, \cdots, 0.8, 1\}$. The parameters associated with battery degradation are set to $\alpha_d = 4.5 \times 10^{-3}$ and $\beta = 1.3$, as used in [31].

The DQN **Algorithm 1** has been implemented in Python with the popular NN toolboxes Tensorflow and Keras [41]. Table 2.2 lists the parameter settings

Algorithm 1: DQN-based Optimal Battery Control

```
Hyperparameters: discount factor $\gamma = 1$, learning rate $\eta > 0$,
  $\epsilon$-greedy coefficient $\kappa \in (0, 1)$, mini-batch size $|\mathfrak{J}|$, target network
  update interval $N_o$, and maximum number of episodes $N$.
Input: training profiles of prices and FR signals with the exploration
  time horizon $T$.
Initialize: the $\epsilon$-greedy probability $\epsilon \in (0, 1)$, replay memory $\mathfrak{D} = \emptyset$,
  initial action-value function $Q(s, a; \theta')$ with a random $\theta'$ and the target
  network parameter $\theta^{-} = \theta'$ at episode $n = 0$.
while $n \leq N$ do
    for t=1, $\cdots$ ,T do
        Select a random action $a_t$ with probability $\epsilon$; otherwise, use the
          action $a_t^{*} = \arg\max_{a_t} Q(s_t, a_t; \theta')$.
        Implement the action $a_t$ to obtain the ensuing state $s_{t+1}$ based
          on the transitions of both Eq. (2.4) and Table 2.1, and by using
          the input profiles of $p_{t+}$ and $f_{t+1}$.
        Compute the instantaneous reward $r_t$ in Eq. (2.14).
        Store the tuple $(s_t, a_t, s_{t+1}, r_t)$ in $\mathfrak{D}$.
        Select a random mini-batch $\mathfrak{J}$ with size $|\mathfrak{J}|$ from $\mathfrak{D}$.
        Compute the gradient estimate using Eq. (2.20).
        Update the parameter $\theta' \leftarrow \theta' - \eta \nabla \hat{\mathfrak{L}}(\theta')$.
        if $t/N_o$ is an integer then
            Update the target network parameter $\theta^{-} \leftarrow \theta'$.
        end
        Update $\epsilon = \kappa\epsilon$.
    end
    Update the episode number $n \leftarrow n + 1$.
end
```

for the DQN training, which uses 7 daily profiles of $\{p_t, f_t\}$. There are a total of $T = 8,640$ time instances for each exploration episode. Upon the convergence of Q-network, it is used for determining the optimal (dis)charging actions for each 2-second interval of 60 days of testing data, while each testing trajectory having

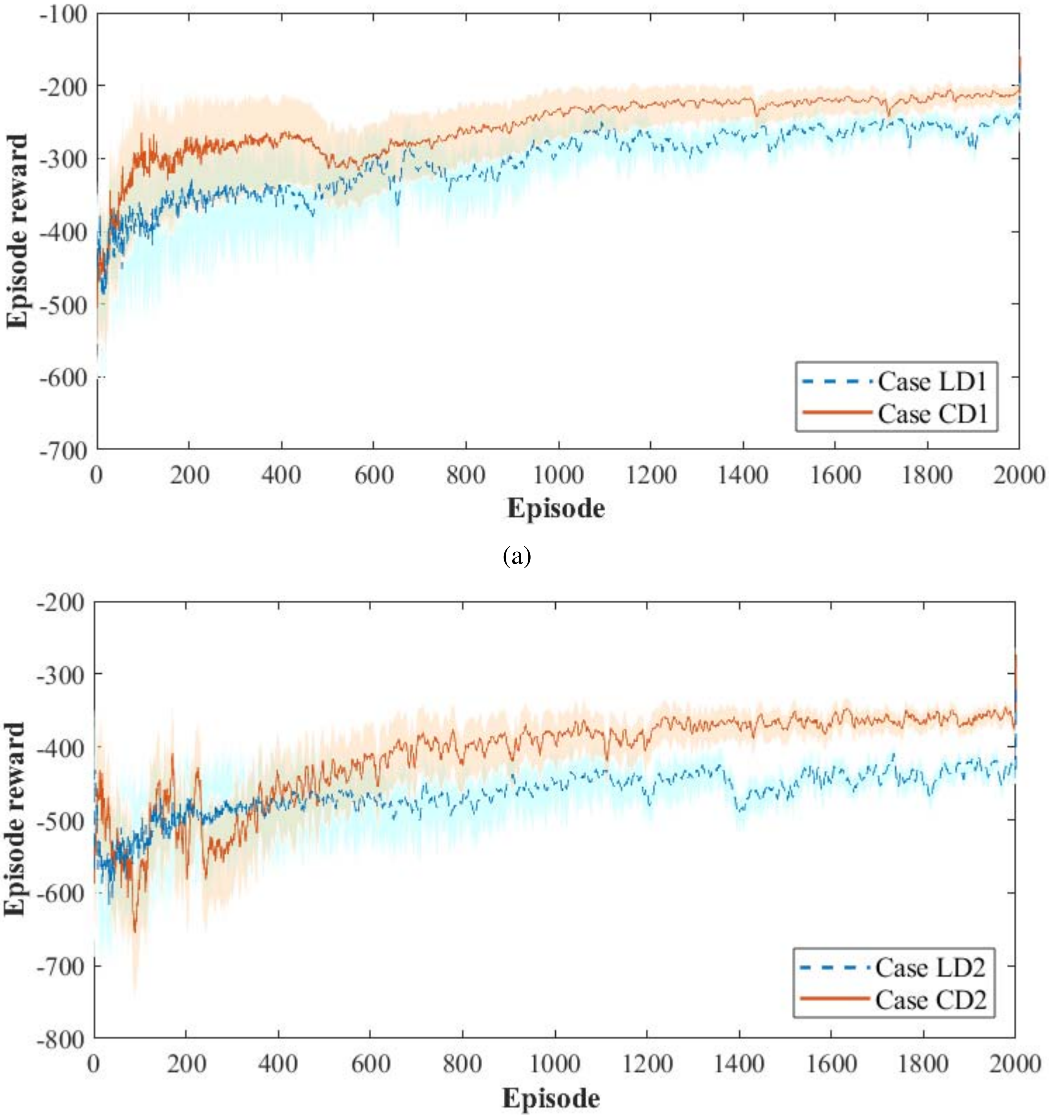
Table 2.2: Parameter settings for DQN training

Parameter	**Value**
Number of hidden layers	2
Number of nodes	[128, 32]
Activation function	ReLU
Learning rate	0.001
Optimizer	Adam
Epsilon (ϵ)	0.001
Batch size (J)	256
Maximum number of episodes (N)	2000
Number of daily profiles	7

43,200 time instances.

Training Comparisons. To compare the proposed cycle-based degradation cost with the linear one (denoted by CD and LD, respectively) in terms of battery control performance, We have considered two levels of degradation coefficients, at α_d and $2\alpha_d$, respectively. Fig. 2.3 illustrates the training comparisons of the actual episode rewards (all based on cycle-based degradation) for all the three cases. Clearly, all the DQN iterations are convergent as the total reward trajectories tend to be non-decreasing till reaching the highest values. Moreover, while the CD cases using our proposed *instantaneous cycle-based degradation* modeling outperform the LD counterparts, especially at larger degradation cost. This comparison validates the advantages of the proposed degradation model in terms of accurately representing the battery cost and thus leading to effective control policies.

Testing Comparisons. We have further compared the testing performance of the learned Q-networks using both CD and LD based models. Fig. 2.4 plots



(a)

(b)

Figure 2.3: Comparisons of the total reward trajectory between (a) cases LD1 and CD1 (α_d) and (b) case LD2 and CD2 ($2\alpha_d$)

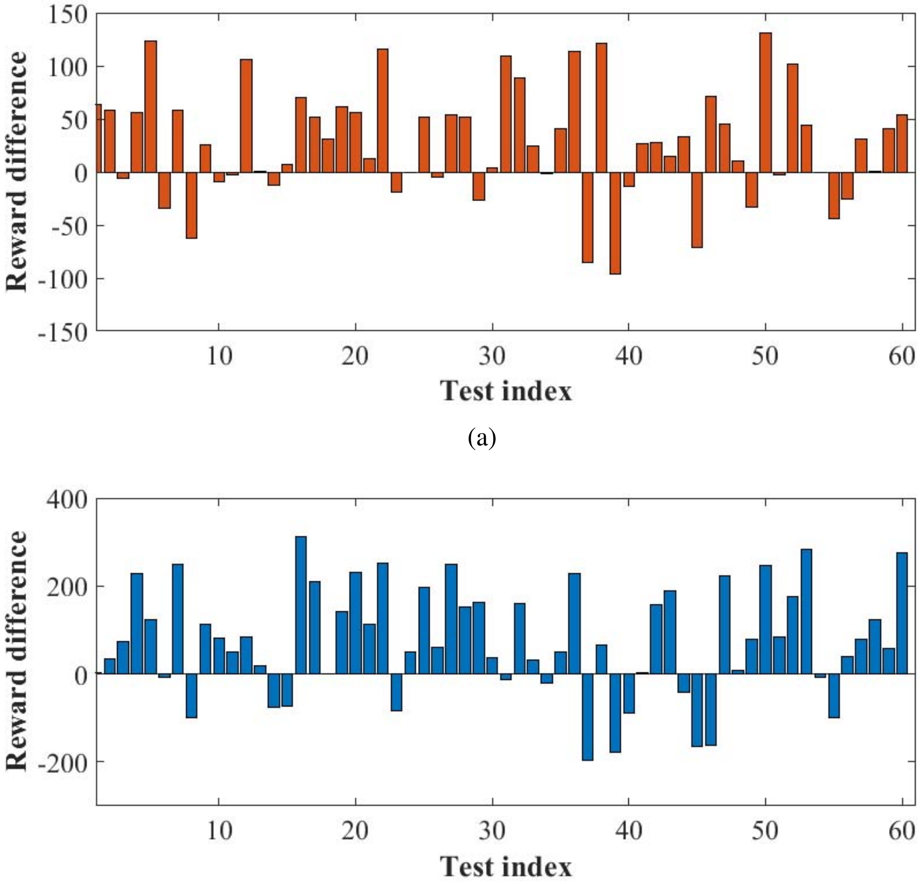
the total reward differences between the CD and LD solutions (positive differences indicating higher reward for CD) for each test trajectory under the two levels of

Table 2.3: Reward differences between CD and LD

Cases (A-B)	**Mean**	**Max**	**Min**	**Mean (A > B)**	**Mean (A < B)**
CD1-LD1	84.45	133.09	-97.04	111.38	-72.12
CD2-LD2	176.51	315.15	-198.85	172.79	-134.19

degradation coefficients. The proposed CD based control leads to higher total reward for at least 73.33% or 81.67% of test scenarios, respectively for the two α_d levels. This result confirms the earlier observations in training phase that proposed solution is more attractive for larger degradation cost. Table 2.3 indicates the total mean, maximum, minimum values and average values of the cases when CD has better reward than LD, and vice versa. As shown in the table, the overall mean value increases as the degradation coefficient increases and the battery degradation cost affects more in total cost accordingly. In addition, the maximum, minimum and mean values show that even though there are some cases that LD show better performance than CD, the number of these cases is very small. Similar comparison is also observed in differences of battery degradation performance only (again, positive differences indicating lower degradation cost for CD), as illustrated by Fig. 2.5. Clearly, the proposed CD solutions overwhelmingly improve the battery degradation performance and accordingly the total reward, as compared to the existing LD-based approximation. In addition, Fig. 2.4 and Fig. 2.5 share very similar pattern, which implies that the increase in total reward is mostly caused by the decrease in the battery degradation cost and has least impacts on the decrease in the rewards regarding the net cost for electricity usage or frequency regulation penalty.

Fig. 2.6 plots the selected testing SoC trajectory along with the electricity

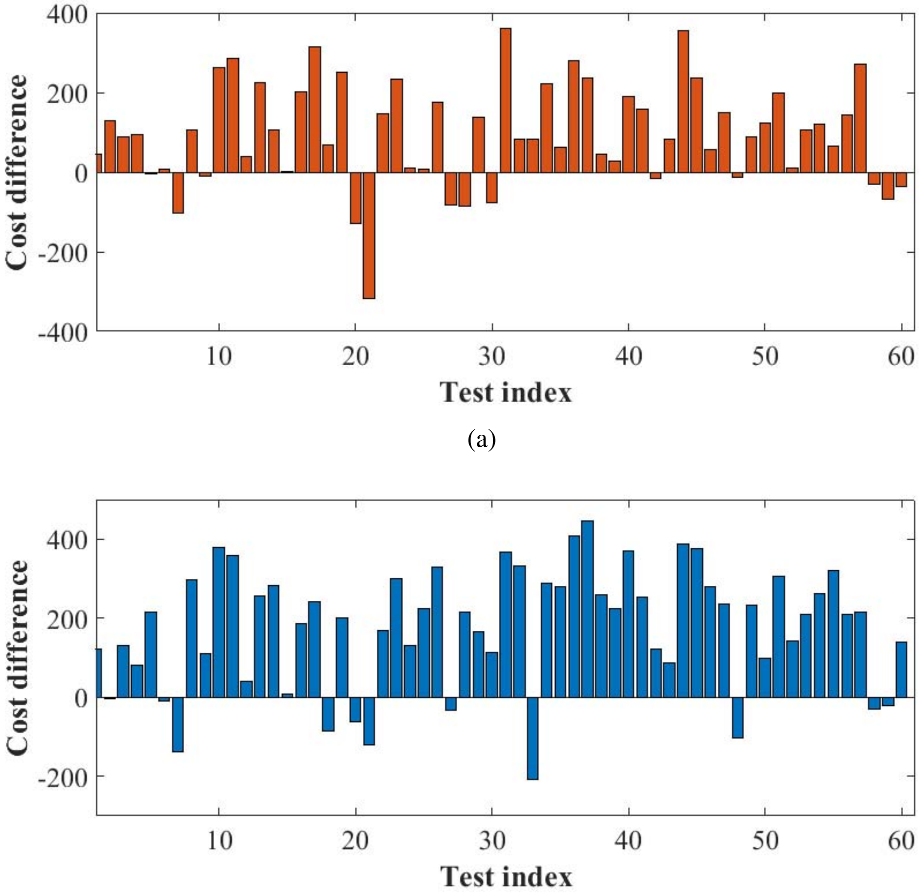


(a)

(b)

Figure 2.4: The total reward differences (positive difference indicating higher reward for CD) between (a) cases LD1 and CD1 (α_d) as well as (b) case LD2 and CD2 ($2\alpha_d$)

price to better illustrate the improvement of the proposed CD-based policy, corresponding to the two choices of degradation parameter (α_d and $2\alpha_d$). Clearly, both trajectories show that the CD-based policy leads to less number of cycles with long depth as compared to the LD one, which reduces the overall degradation cost espe-

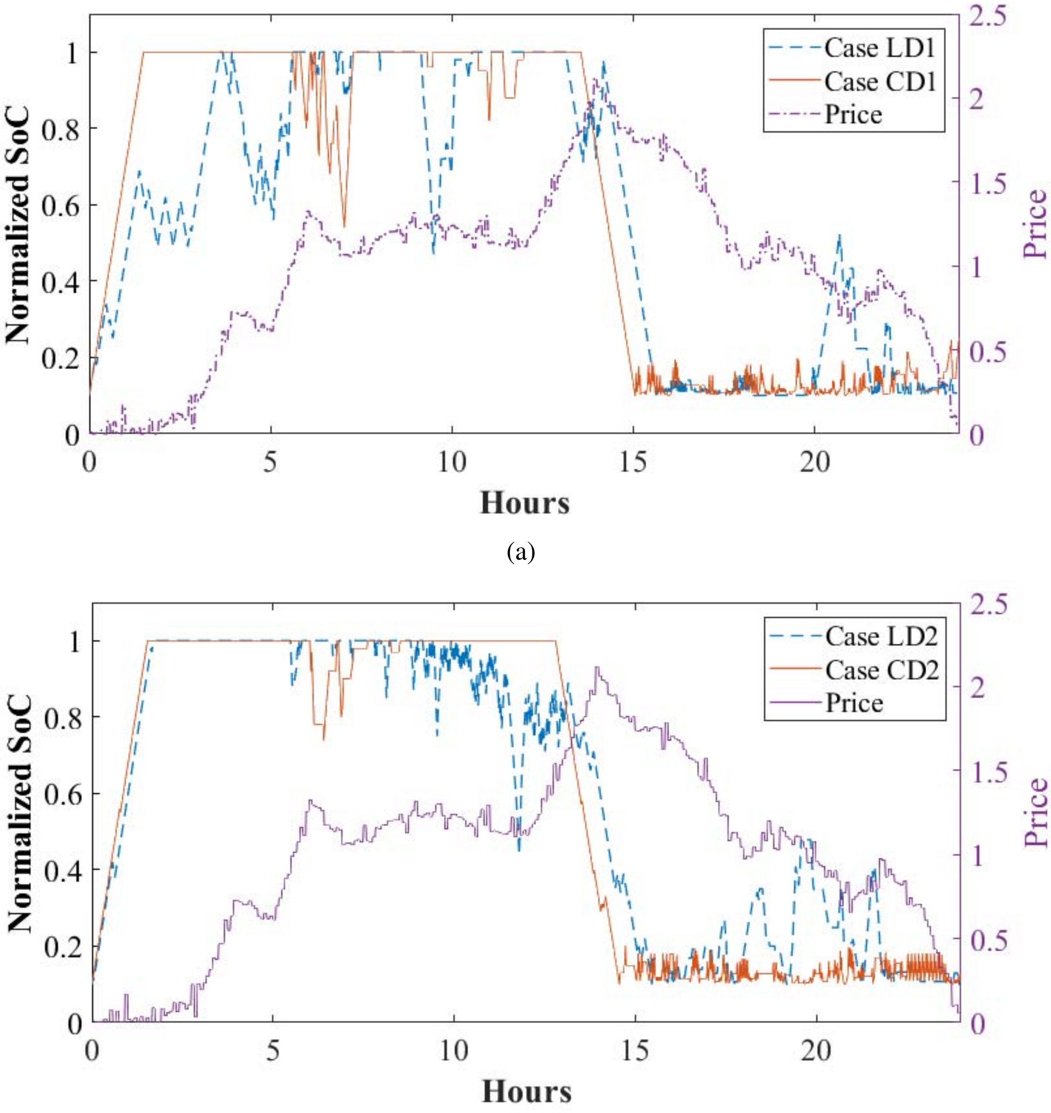


(a)

(b)

Figure 2.5: The battery degradation cost differences (positive differences indicating lower cost for CD) between (a) cases LD1 and CD1 (α_d) as well as (b) case LD2 and CD2 ($2\alpha_d$)

cially for hours between $[3,\ 13]$. In addition, during high-price hours in $[12,\ 15]$, the CD trajectory has one smooth and long discharging cycle and this pattern is amendable to mitigating battery degradation. In contrast, the LD one has frequent, noticeable fluctuations during this period. Because of the linearized approximation,



(a)

(b)

Figure 2.6: Comparison of selected SoC trajectories in testing between (a) cases LD1 and CD1 (α_d) and (b) cases LD2 and CD2 ($2\alpha_d$)

the LD-based policy leads to eight more noticeable (dis)charging cycles of considerate depth than the CD one. This speaks for the capability of the proposed CD

Table 2.4: Degradation comparisons between CD and LD

Degradation factor	**LD1**	**CD1**	**LD2**	**CD2**
High C-rates	0.0563	0.0342	0.0463	0.0369
SoC stress	0.0141	0.0108	0.0182	0.0107

model in effectively removing some unnecessary cycles of moderate depth, thanks to the accurate representation of rainflow-based degradation. In addition, in the post-peak hours $[15,\ 20]$, the LD based policy produces a couple of cycles of moderate depth which are not very profitable. The proposed CD based policy is able to successfully remove these nonprofitable cycles and does not lead to any considerate cycles.

Interestingly, by mitigating cycle-based degradation the proposed CD approach can potentially contribute to the improvement of other degradation factors too. Table 2.4 compares the proposed CD with the LD approach on the degradation related to high C-rates [35] and SoC stress [10], both of which have been numerically improved by the CD-based policies. The high C-rate based degradation depends on the total DoD summed over all cycles of the trajectory. Intuitively, a concise list of smooth and long (dis)charging cycles attained by CD-based policy can reduce both the number of cycles and their DoD, thus beneficial for the high C-rate metric. Similarly, as CD-based policy has also been observed to remove unnecessary cycles in the post-peak hours, the average SoC level decreases which relieves the SoC stress. These intuitions corroborate the claim in Section 2.3 that the cycle-based DoD stress model is most relevant for the fast battery control problem.

To sum up, the numerical results have validated the performance improvement attained by the proposed *instantaneous* cycle-based degradation model, by exactly representing the rainflow conditions. The proposed approach effectively leads to battery control trajectories that reduce unnecessary fluctuations or improve the overall economical profits.

This work proposes an accurate model of cycle-based degradation cost in order to allow for efficient battery control designs using reinforcement learning (RL). In order to model the degradation which depends on the full cycle, we introduce additional state variables to judiciously keep track of important switching points of SoC trajectory for effectively identifying (dis)charging cycles. This way, the actual degradation cost is separated into instantaneous terms along with other operation costs such as the net cost for electricity usage and FR penalty, such that powerful DQN based RL algorithms are readily applicable. Numerical tests confirm the effectiveness of proposed cycle-based degradation model and demonstrate the performance improvements in effectively mitigating battery degradation over existing linearized approximation approach.

Chapter 3

RL-based Electric Vehicle Charging Station Operation

This chapter focuses on developing electric vehicle charging station operation policy with the efficient Markov Decision Process (MDP) representation. Section 3.1 introduces an electric vehicle charging station (EVCS) problem and the motivation of this work. Section 3.2 formulates the EVCS operations problem as a MDP. Section 3.3 develops the least-laxity first (LLF)-based action reduction and our proposed equivalent state aggregation to deal with dimensionality issues. Based on this, Section 3.4 presents the reinforcement learning approach using policy gradient and linear Gaussian policy parameterization. Numerical tests using real-world data are studied in Section 3.5 to demonstrate the performance improvement of the proposed algorithm.

This chapter is based on the following publication:
K. Kwon and H. Zhu, "Efficient Representation for Electric Vehicle Charging Station Operations using Reinforcement Learning." *Proceedings of the Hawaii International Conference on System Sciences,* 2022.

3.1 Electric Vehicle Charging Station Operation Problem

Electrified transportation is drastically reshaping worldwide urban mobility as a key technology to enable a future low-carbon energy society. The number of electric vehicles (EVs) continues to grow rapidly [42], thanks to their high efficiency [43] and low pollution emissions [44]. This has propelled the popularity of EV charging stations (EVCS) in metropolitan areas, as supported by significant investment in urban electricity infrastructure.

Solving the problem of optimal operational strategies is crucial for maximizing the economic profit of EVCS owners while ensuring the quality-of-service for EV charging. In general, this problem aims to find the optimal policy for determining EV charging schedules to reduce the total electricity cost by utilizing the flexibility of EV charging needs [5, 6, 45, 46]. In addition, several papers have accounted for co-located renewable generation or energy storage [12, 47, 48] or the coupling between EV traffic and electricity flow [49–51]. Nonetheless, one key challenge in formulating the EVCS problem lies in the randomness and uncertainty of EV arrivals and other inputs such as electricity market prices. It is possible to develop probabilistic models from actual data, such as the Gaussian distribution model of EV parking time and require demand in [11], or the representation of the charging demand as a mixed Gaussian model to be estimated in [12]. Although these models have led to efficient stochastic programming approaches for the EVCS problem, they could be prone to potential modeling mismatches or fail to capture the problem dynamics therein.

To tackle this challenge, this work aims to develop a data-driven framework to solve the EVCS operation problem by leveraging reinforcement learning (RL) techniques [2]. Using actual data samples, RL has shown some success in solving this problem with no need for stochastic modeling [52–54].

Nonetheless, most existing approaches use the original problem representation of individual EVs' status and charging action. This leads to very *high* and *time-varying* dimensionality for both the state and action spaces, significantly affecting the efficiency and convergence of policy search by generic RL algorithms. By transforming the EV status to the so-termed *laxity* that measures the emergency level of its charging need, the work in [54] has proposed to consider the total charging power across the EVCS as the action instead. Furthermore, a least-laxity first (LLF) rule has been advocated to recover individual EVs' actions from the aggregated one, which can maintain the feasibility of the charging solutions. The dimensionality issue of state space is solved by approximating the action-value function, or Q-function, which lacks approximation guarantees.

To this end, our work has proposed a new state representation by aggregating the individual EV status into the number of EVs in each laxity group. We have analytically shown that this aggregation scheme is equivalent to the original one and thus can lead to the same optimal policy by an RL algorithm. The main contribution of the present work is two-fold. First, we have developed a comprehensive representation for both the state and action spaces of the EVCS operations problem, with guaranteed equivalence to the original model. Second, the proposed representation enjoys fixed and low problem dimensions, developing an efficient algorithm to search for the optimal policy. Our numerical results have validated the performance improvement of the proposed state representation compared to the existing approach of Q-function approximation and suggested additional state aggregation by further grouping the higher-laxity EVs with minimal performance degradation.

3.2 System Modeling

Consider the operations of an EV charging station (EVCS) as depicted in Fig. 3.1 over the time period $\mathcal{T} = [0, \ldots, T]$. For each time $t \in \mathcal{T}$, let $\mathcal{I}_t$ denote the set of parked EVs, with $\mathcal{J}_t$ and $\mathcal{L}_t$ denoting the sets of arriving and departing EVs, respectively. Hence, the set of EVs is updated by $\mathcal{I}_{t+1} = (\mathcal{I}_t \cup \mathcal{J}_{t+1}) \backslash \mathcal{L}_{t+1}$, thus time-varying. Upon the arrival of EV $i \in \mathcal{J}_t$, its remaining demand $d_{i,t}$ and parking time $p_{i,t}$ are determined by the owner. The goal of EVCS operations is to determine the charging action $a_{i,t}$ for every parked EV $i \in \mathcal{I}_t$, based on the real-time electricity prices $\{\rho_t\}$ received from the market operator. Each EV's status is updated according to the $\{a_{i,t}\}$ sequence, until its departure time $\tau \in \mathcal{T}$ such that either $d_{i,\tau} = 0$ or $p_{i,\tau} = 0$. For simplicity, all EVs are assumed to have the same charging power, with the possibility of extension to different charging rates as analyzed in [55]. In addition, this work assumes the charging actions will ensure each EV to be fully charged before departure; i.e., the departure time τ is the first slot with $d_{i,\tau} = 0$. This assumption is reasonable because the EVCS can always increase the total charging budget to meet all EV demands. In future, we will extend it to the general case of non-fully charged EVs by introducing a penalty cost.

This work aims to develop efficient reinforcement learning (RL) algorithm for the EVCS operation problem. To this end, we model it as a Markov Decision

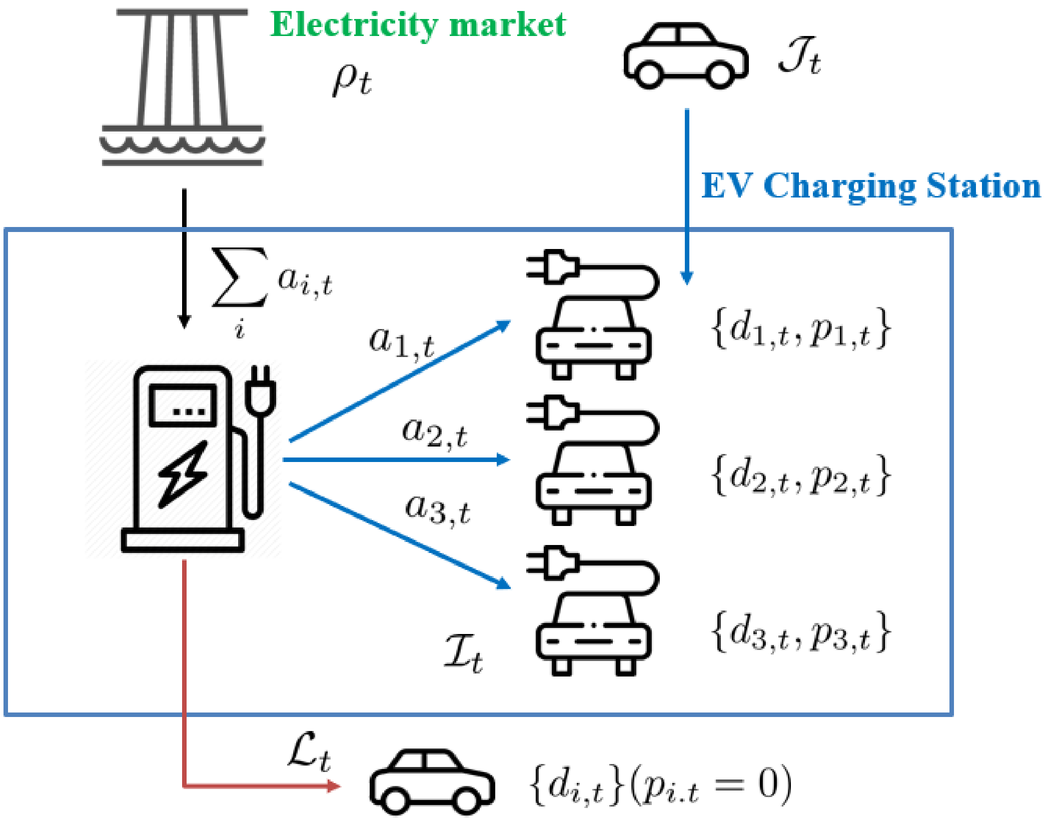


Figure 3.1: System Model of EV Charging Station

Process (MDP) [2, Ch. 3] denoted as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, as detailed here.

State space $\mathcal{S}$ contains the set of feasible values for both the EV-internal and external status variables. This includes the remaining demand and parking time for each EV, as well as the electricity market price ρ_t. Hence, the state per time t is given by $s_t = [\rho_t, \{d_{i,t}, p_{i,t}\}_{i\in\mathcal{I}_t}]$.

Action space $\mathcal{A}$ includes the set of decisions that the active EVs can take. Without loss of generality (Wlog), consider a simple binary decision rule for each

EV as given by $a_{i,t} \in \{0 \text{ (do nothing)}, 1 \text{ (charge)}\}$. It can be extended to a multi-level charging rate with $|\mathcal{A}| > 2$ or a continuous charging action. For simplicity, this work focuses on the case of binary action.

Transition kernel $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ captures the system dynamics under the Markov property [2, Ch. 3]. In the case of stochastic electricity market prices, we assume $\Pr(\rho_{t+1}|\{\rho_\tau\}_{\tau=1}^t) = \Pr(\rho_{t+1}|\rho_t)$. This is reasonable since the market price has short-term memory [32]. A longer memory is possible too; such as the prices that follow $\Pr(\rho_{t+1}|\{\rho_\tau\}_{\tau=1}^t) = \Pr(\rho_{t+1}|\rho_t,\ \rho_{t-1})$. In this case, both ρ_t and ρ_{t-1} are included as the part of the state per time t to satisfy the Markov transition property.

In addition, the EV status is updated according to the charging action in a deterministic fashion. For simplicity, let $d_{i,t}$ and $p_{i,t}$ denote the number of time slots for EV i to attain full charging and stay parked at time t, respectively. This way, their transitions are given by

$$d_{i,t+1} = d_{i,t} - a_{i,t}, \text{ and } p_{i,t+1} = p_{i,t} - 1. \tag{3.1}$$

This update rule also holds for general action spaces if $a_{i,t}$ is not binary.

Reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ indicates the instantaneous reward used for defining the optimal actions. Wlog, assume all EVs have the same charging rate and thus the reward related to the total charging cost in time t is given by $r_t(s_t,\ a_t) = -\rho_t(\sum_{i \in \mathcal{I}_t} a_{i,t})$. The reward objective can also consider other economic factors such as peak demand reduction and load shaping benefits.

Discount factor $\gamma \in (0, 1]$ is a constant to accumulate the total reward along the time horizon. Smaller γ values imply that future rewards are less important than current ones at a discounted rate [2, Ch. 3]. For this finite time-horizon problem, $\gamma = 1$ will be used for simplicity.

For the MDP-based model, we can formulate the EVCS operation problem. The goal is to find the optimal policy π for mapping $a_t \sim \pi(s_t)$ with s_t. To simplify the policy search, we are particularly interested in the set of parameterized policies given by $\pi_\mu(\cdot) = \pi(\cdot; \mu)$, which optimizes over parameter μ as given by

$$\max_{\mu} J(\mu) = V^{\pi}(s_0) := \mathbb{E}_{a_t \sim \pi_\mu(s_t), \mathcal{P}} \left[\sum_{t=0}^{T} \gamma^t r_t(s_t,\ a_t) \Big| s_0 \right] \tag{3.2}$$

where $V^{\pi}(s_0)$ denotes the value function for given initial state s_0. The formulation (3.2) allows for adopting popular RL algorithms. The parameterized model and problem set-up will be discussed with more details in Section 3.4 along with the policy gradient (PG) solution method [56]. Notably, the dimensions of state and action in (3.2) can be very high and are time-varying, making it challenging to search for an effective policy using RL. The following section will develop efficient state/action representation for the EVCS problem.

3.3 Efficient MDP Representation

Solving the MDP problem is challenged by the state/action representations of high dimension and time-varying. As the policy maps from state to action, the number of parameters in μ would grow with both state/action dimensions. This increasing rate would significantly slow down the search for an effective policy by

Time period		$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$
a_t		2	1	0	2	0
EV_1	$d_{1,t}$	3	2	1	1	0
	$p_{1,t}$	4	3	2	1	0
	$\ell_{1,t}$	1	1	1	0	0
	$a_{1,t}$	1	1	0	1	0
EV_2	$d_{2,t}$	2	1	1	1	0
	$p_{2,t}$	4	3	2	1	0
	$\ell_{2,t}$	2	2	1	0	0
	$a_{2,t}$	1	0	0	1	0

Table 3.1: Two-EV example by following LLF rule.

generic RL algorithms. To tackle these issues, we propose considering the action reduction using the least-laxity first (LLF) rule and proposing an equivalent state aggregation through laxity-based grouping.

We can reduce the action space to $\mathcal{A}'$ that only consists of the total charging action $a_t = \sum_{i \in \mathcal{I}_t} a_{i,t}$. This way, the instantaneous reward becomes $r_t = -\rho_t \cdot a_t$. To recover each $a_{i,t}$ from a_t, we adopt the LLF rule proposed in [54] to rank the priority of EVs according to the laxity, as defined by $\ell_{i,t} := p_{i,t} - d_{i,t}$. The smaller $\ell_{i,t}$ is, the fewer flexible slots EV i can use to skip charging before departure, and thus the more emergent it is at time t compared to other EVs. If $\ell_{i,t} = 0$, or $p_{i,t} = d_{i,t}$, then EV i needs to be charged throughout its remaining parking time to be fully charged before departure. The LLF rule aims to increase the flexibility of EV charging by serving the least flexible ones first.

To demonstrate the advantage of LLF-based action recovery, we use a simple example of only two EVs in the charging station as indexed by EV_1 and EV_2, respectively. A total horizon of $T = 4$ is considered, and a possible initial state is

Time period		$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$
a_t		2	1	0	2	0
EV_1	$d_{1,t}$	3	2	2	2	1
	$p_{1,t}$	4	3	2	1	0
	$\ell_{1,t}$	1	1	0	-1	-1
	$a_{1,t}$	1	0	0	1	0
EV_2	$d_{2,t}$	2	1	0	0	0
	$p_{2,t}$	4	3	2	1	0
	$\ell_{2,t}$	2	2	0	0	0
	$a_{2,t}$	1	1	0	0	0

Table 3.2: Two-EV example not following the LLF rule.

given in Table 3.1. Under a given sequence of total charging actions a_t, Table 3.1 lists the individual charging actions following the LLF rule, while Table 3.2 shows one case of not following it. In Table 3.2, EV_2 is charged at $t = 1$ instead of EV_1 even though $\ell_{2,1} > \ell_{1,1}$. As a result, EV_1 is not fully charged at the end, while the total charging sequence $\{a_t\}$ has led to both EVs being fully charged in Table 3.1. This comparison points out the importance of having the LLF rule in disaggregating the total a_t. With the given total charging budget a_t, **Algorithm 2** demonstrates a procedure for selecting EVs to charge at time t according to the LLF rule.

The LLF based action reduction allows to recover feasible individual EV schedules, as shown in [54] and restated here for completeness.

Proposition 2. *If the EVCS total charging schedule $\{a_t\}_{t\in\mathcal{T}}$ is feasible, i.e., there exist corresponding feasible charging schedules for individual EVs that ensure each EV to be fully charged before departure, then the LLF procedure in **Algorithm 2** can produce such a feasible charging schedule for all the EVs.*

Algorithm 2: Least-laxity first (LLF) rule

```
**Inputs:** Total charging power $a_t$, the set of EVs in $\mathcal{I}_t$ along with their remaining demand $d_{i,t}$ and parking time $p_{i,t}$.
**Initialize:** the allocated charging budget $a = 0$.
Compute the laxity for each EV $i \in \mathcal{I}_t$ as $\ell_{i,t} := p_{i,t} - d_{i,t}$ and set $a_{i,t} = 0$ to indicate that it is not yet selected for charging.
**while** $a \leq a_t$ **do**
    Search for the least-laxity EV $k = \arg\min_{i:a_{i,t}=0} \ell_{i,t}$ from the remaining unchosen EVs by arbitrarily breaking the tie if there is any.
    Set $a_{k,t} = 1$.
    $a \leftarrow a + 1$
**end**
```

Instead of formally showing Proposition 2, we provide some intuition behind it. For given $\{a_t\}_{t\in\mathcal{T}}$, if there exist corresponding feasible individual EV schedules that do not follow the LLF rule, then we can transform the latter to feasible individual schedules that follow the LLF rule. Consider an arbitrary feasible EV schedule $\{a_{i,t}\}_i$ for each EV i that corresponds to the given $\{a_t\}_{t\in\mathcal{T}}$, i.e., it holds that $\sum_i a_{i,t} = a_t$ at every $t \in \mathcal{T}$. If the former does not follow the LLF rule, then there exist two EVs, say j and k, that violate the LLF rule at certain time t'. Specifically, we have $a_{j,t'} = 1$, and $a_{k,t'} = 0$ with the laxity $\ell_{j,t'} > \ell_{k,t'}$. The feasibility implies that $\ell_{j,t} \geq 0$ and $\ell_{k,t} \geq 0$, $\forall t \in \mathcal{T}$. Hence, let us switch the charging for those two EVs at time t', i.e., instead we pick EV k to charge by setting $a_{j,t'} = 0$, and $a_{k,t'} = 1$. First, this switch does not change the total charging action. Second, as $\ell_{j,t'} > \ell_{k,t'}$ at time t', this change still ensures feasibility or that the laxity values are always non-negative throughout the horizon $\mathcal{T}$. Hence, this example shows that by following the LLF rule, one can always recover the feasible individual EV schedules. Detailed proof for this result can be found in [54].

In addition to action reduction, we also develop a state aggregation scheme to address the variable and high dimensionality issues of $\mathcal{S}$. We pursue the ideal *equivalent state aggregation* [57] such that the new state space $\mathcal{S}'$ can maintain the necessary information in $\mathcal{S}$. The aggregation needs to ensure that both $\mathcal{S}$ and $\mathcal{S}'$ attain the same value functions $V^{\pi}(\cdot)$ and thus the same optimal policies for any given action in $\mathcal{A}'$. Two conditions need to hold [57], as defined here.

Definition 1. *A state aggregation scheme $\mathcal{S} \to \mathcal{S}'$ satisfies* ***reward homogeneity*** *if for any pair of original states $\{s_t^{(i)}, s_t^{(j)}\}$ that will be aggregated into the same new state in $\mathcal{S}'$, it holds that*

$$r_t(s_t^{(i)}, a_t) = r_t(s_t^{(j)}, a_t),\ \forall a_t \in \mathcal{A}' \tag{3.3}$$

Definition 2. *A state aggregation scheme $\mathcal{S} \to \mathcal{S}'$ satisfies* ***dynamic homogeneity*** *if for any pair of original states $\{s_t^{(i)}, s_t^{(j)}\}$ that will be aggregated into the same new state in $\mathcal{S}'$, it holds that*

$$\Pr(s_{t+1}|s_t = s^{(i)}, a_t) = \Pr(s_{t+1}|s_t = s^{(j)}, a_t),\ \forall s_{t+1} \in \mathcal{S},\ a_t \in \mathcal{A}' \tag{3.4}$$

To achieve these homogeneity conditions, we propose to aggregate parked EVs at time t into the number of EVs for every integer-valued laxity level in $[0,\ L]$, where $L := \max_{i,t} \ell_{i,t}$ denotes the maximally possible laxity level at the EVCS. Note that as all EVs are assumed to be fully charged before departure, the laxity is always non-negative with the minimum equal to zero. Upon determining each EV's laxity as in Section 3.3, we define the aggregated state

$$s'_t = [\rho_t, n_t^{(0)}, n_t^{(1)}, \cdots, n_t^{(L)}] \in \mathcal{S}' \tag{3.5}$$

with $n_t^{(\ell)}$ denoting the number of EVs with laxity equal to ℓ. In order to show the new MDP is equivalent to the original one, let us consider the two homogeneity conditions. First, the reward homogeneity is easily satisfied as $r_t = -\rho_t a_t$ is not affected by the aggregation. Second, dynamic homogeneity also holds due to the LLF rule for action reduction. Upon recovering the individual EV actions $\{a_{i,t}\}_i$ from a_t, the original MDP transition in (3.1) states that $(d_{i,t+1}, p_{i,t+1}) = (d_{i,t} - a_{i,t}, p_{i,t} - 1)$ for each $i \in \mathcal{I}_t$. For the new MDP through aggregation, the state transition instead depends on the allocation of a_t to each subset of EVs of the same laxity. Specifically, if $a_{i,t} = 1$ or EV i is charged at time t, its laxity stays unchanged as $\ell_{i,t+1} = \ell_{i,t}$. Otherwise, its laxity is reduced by one as $\ell_{i,t+1} = \ell_{i,t} - 1$. We can update the subset of EVs with laxity ℓ for time $t+1$ based on those of laxity ℓ at time t that are charged, those of laxity $(\ell + 1)$ that are not charged, along with the new arrival or departure at time $(t+1)$, as given by

$$n_{t+1}^{(\ell)} = a_t^{(\ell)} + [n_t^{(\ell+1)} - a_t^{(\ell+1)}] + x_{t+1}^{(\ell)} - y_{t+1}^{(\ell)},\ \forall l \tag{3.6}$$

where $a_t^{(\ell)}$ denotes the number of EVs of laxity ℓ that are charged in time t, while $x_{t+1}^{(\ell)}$ and $y_{t+1}^{(\ell)}$ representing the number of EVs of laxity ℓ that arrive/depart at time $(t+1)$, respectively. Similar to the LLF-based action recovery in Section 3.3, we allocate the total charging budget a_t into each $a_t^{(\ell)}$ in an ordered fashion, as given by

$$a_t^{(\ell)} = \min\left\{ n_t^{(\ell)}, \min\left\{ a_t - \sum_{\ell=0}^{\ell-1} a_t^{(\ell)}, 0 \right\} \right\},\ \forall \ell. \tag{3.7}$$

Basically, starting from the smallest laxity level $\ell = 0$, we set $a_t^{(\ell)} = n_t^{(\ell)}$ until the total charging budget is met. Based on the two homogeneity conditions, we can formally establish the following proposition using the result from [57].

Proposition 3. *Consider the original MDP* $(\mathcal{S}, \mathcal{A}', \mathcal{P}, \mathcal{R}, \gamma)$ *and the new MDP* $(\mathcal{S}', \mathcal{A}', \mathcal{P}, \mathcal{R}, \gamma)$*. If* $\mathcal{S}'$ *is aggregated through* $s'_t = [\rho_t, n_t^{(0)}, n_t^{(1)}, \cdots, n_t^{(L)}]$ *with the transition following* (3.6) *and* (3.7)*, then it satisfies both reward homogeneity and dynamic homogeneity and thus the two MDPs are equivalent. As a result, the new MDP through aggregation can be used to obtain the optimal policies (determine the optimal actions) that are equivalent to the original ones.*

By guaranteeing the equivalence of the two MDPs, the aggregation maintains the same value function for any initial state as mentioned earlier. Hence, the optimal policy obtained by an RL algorithm for the new MDP would be the same as the original one. This state aggregation scheme can efficiently search for the best $\pi(\cdot)$, at no sacrifice of optimality.

Note that the state aggregation can be further simplified in practice by merging the higher-laxity groups. If the maximum laxity L is very large, the equivalent aggregation can still be of quite large dimension. Our numerical experiences suggest that the groups of higher laxity values play similar role in determining the optimal action, as the LLF rule implies that the recovered action (or the transition) would mostly depend on the groups of smaller laxity values. Hence, we can cap the number of laxity groups at a value $L_{\max} < L$ such that $n^{(L_{\max})} = \sum_{\ell \geq L_{\max}} n^{(\ell)}$. Although this further simplification may not be equivalent, it can be effective in addressing the immense value of laxity in practice.

3.4 Learning the Optimal Policy

The proposed efficient MDP representation has successfully handled the dimensionality issue for state/action, and will be leveraged to efficiently solve for the optimal policy π in (3.2) using general RL algorithms. Recall that the unknown policy $\pi(\cdot)$ is assumed to follow certain parameterized model, and thus the problem is to find the optimal parameter μ for the mapping $a \sim \pi_\mu(s')$. The choice of parameterized model can affect the performance of RL algorithms. Without loss of generalizability, we consider a simple model of π_μ and adopt the policy gradient (PG) method [56] to search for the best μ. We use the linear Gaussian policy [58], which is popular for continuous spaces, as defined by the conditional distribution

$$a \sim \pi_\mu(s') = \pi_\mu(a|s') = \mathcal{N}(\mu_s^\top s' + \overline{\mu}, \sigma^2) \tag{3.8}$$

with parameter $\mu = [\mu_s;\ \overline{\mu}]$ relating s' to the mean for the Gaussian distributed action a. The variance σ^2 can be either part of the parameter or pre-determined as exploration noise. Equivalently, the random action in (3.8) can be simply generated by the following *linear policy*

$$a = \mu_s^\top s' + \overline{\mu} + e \tag{3.9}$$

where the additive noise $e \sim \mathcal{N}(0, \sigma^2)$. Using (3.8), the total reward function in (3.2) now becomes

$$J(\mu) = \int_{a \in \mathcal{A}'} \pi_\mu(a|s') Q_\mu(s', a) \mathrm{d}a, \tag{3.10}$$

with the Q-function, or action-value function, given by

$$Q^\pi(s', a) := \mathbb{E}_{a_t \sim \pi_\mu(s_t), \mathcal{P}} \left(\sum_{t=0}^{T} \gamma^t r_t | s_0 = s', a_0 = a \right). \tag{3.11}$$

Before discussing the PG method, it is worth mentioning that other choices of π_μ can be readily applied as well. For example, one can use a nonlinear neural network to parameterize the Q-function, known as the Deep Q-Network (DQN) approach [2, Ch. 20]. The proposed state/action aggregation would be powerful for accelerating these nonlinear policy based RL methods too, which can be greatly affected by the dimensionality issue.

To maximize $J(\mu)$, we are interested to find its gradient over μ following from the *log-derivative trick*, as

$$\nabla_\mu J(\mu) = \mathbb{E}_{a \sim \pi_\mu(s)} \left[Q^\pi(s', a) \nabla_\mu \ln \pi_\mu(a|s') \right]. \tag{3.12}$$

Interestingly, this gradient computation boils down to that of the logarithmic term only, which can be easily obtained for Gaussian distribution as

$$\nabla_{\mu_s} \ln \pi_\mu(a|s') = \frac{a' - (\mu_s^\top s' + \overline{\mu})}{\sigma^2} s', \tag{3.13}$$

$$\nabla_{\overline{\mu}} \ln \pi_\mu(a|s') = \frac{a' - (\mu_s^\top s' + \overline{\mu})}{\sigma^2}. \tag{3.14}$$

To estimate this gradient, one can replace the expectation in (3.12) by the sample mean obtained from the trajectory $\{s'_0, a_0, s'_1, a_1, \cdots, s'_T, a_T\}$:

$$\hat{\nabla}_\mu J(\mu) \propto \sum_{t=1}^{T} \hat{Q}_\mu(s'_t, a_t) \nabla_\mu \ln \pi_\mu(a_t|s_t). \tag{3.15}$$

with the samples $\hat{Q}_\mu(s'_t, a_t) = \sum_{\tau=t}^{T} \gamma^{\tau-t} r_\tau(s'_\tau, a_\tau)$ estimated from the trajectory. Note that the time window for approximating $\hat{Q}_\mu(s'_t, a_t)$ decreases as t increases under the finite time-horizon setting of $\mathcal{T}$. For larger t values, fewer samples are used and the scale of Q-value is expected to decrease. To cope with this issue, one can normalize the approximated Q-function by subtracting the mean and dividing it with the standard deviation of all episode rewards [59]. This can generally improve the training stability under the high variance of the policy gradient estimator.

With a given learning rate (step-size) α, the policy gradient method uses the estimated gradient in (3.15) and implements the iterative gradient ascent updates of μ. Per iteration n, the update becomes

$$\mu^{n+1} = \mu^{n} + \alpha \hat{\nabla}_{\mu} J(\mu^{n}), \tag{3.16}$$

until the parameters converge. To improve the gradient update, we can incorporate multiple training samples, each of which will produce a gradient estimate. Accordingly, the sum (or average) of the gradients estimated from each training sample will be used for the update in (3.16).

Algorithm 3 has detailed steps for solving the proposed MDP representation under LLF-based action reduction and the equivalent state aggregation.

3.5 Numerical Tests

We have tested the proposed **Algorithm 3** to demonstrate the effectiveness of our new MDP representation. To set up the EVCS operation problem, we have used the hourly data of electricity market prices from the ERCOT market portal [39] and the vehicle arrival data collected at the Richards Ave Station near downtown Davis, CA [60]. Three categories of EVs are considered: emergent, normal and residential uses, each having different initial demand and parking time distribution.

Algorithm 3: Optimal EVCS policy search

```
Hyperparameters: discount factor $\gamma$, step-size $\alpha$, and exploration time
  period $T$.
Inputs: the price sequence $\{\rho_t\}_{t=0}^{T}$, and the EV arrivals in $\{\mathcal{J}_t\}_{t=0}^{T}$
  along with the initial states of EVs
Initialize: $\mu^0$ at iteration $n = 0$.
while $\mu^n$ not converged do
    Initialize $t = 0$ with the original state $s_0$.
    for $t = 0, \cdots, T-1$ do
        Find the aggregated state $s'_t$ using (3.5);
        Sample $a_t \sim \pi_{\mu_n}(s'_t)$ using (3.8);
        Use the LLF rule in Algorithm 2 to recover the individual EV
          charging actions $\{a_{i,t}\}_{i \in \mathcal{J}_t}$;
        Compute the instantaneous reward $r_t$;
        Update the new state $s_{t+1}$ using (3.1).
    end
    Use the sample trajectory to estimate gradient $\hat{\nabla}_\mu J(\mu^n)$ and
      perform the update in (3.16);
    Update iteration $n \leftarrow n + 1$.
end
```

Fig. 3.2 shows an example of the number of EVs in each category for a typical workday. Accordingly, the RL exploration time is the full-day period at 15-minute intervals, leading to a total horizon of $T = 96$. The EV data show the maximum laxity $L = 12$, and thus there are a total of 14 variables in s'.

We have compared **Algorithm 3** to the existing approach by estimating an approximate Q-function in [54], denoted by **Algorithm *QE***. In [54], the same LLF-based action reduction is used while four binary feature functions approximate the Q-function to deal with the state dimensionality issue. These feature functions correspond to the charging cost or constraints on EV charging for the EVCS problem,

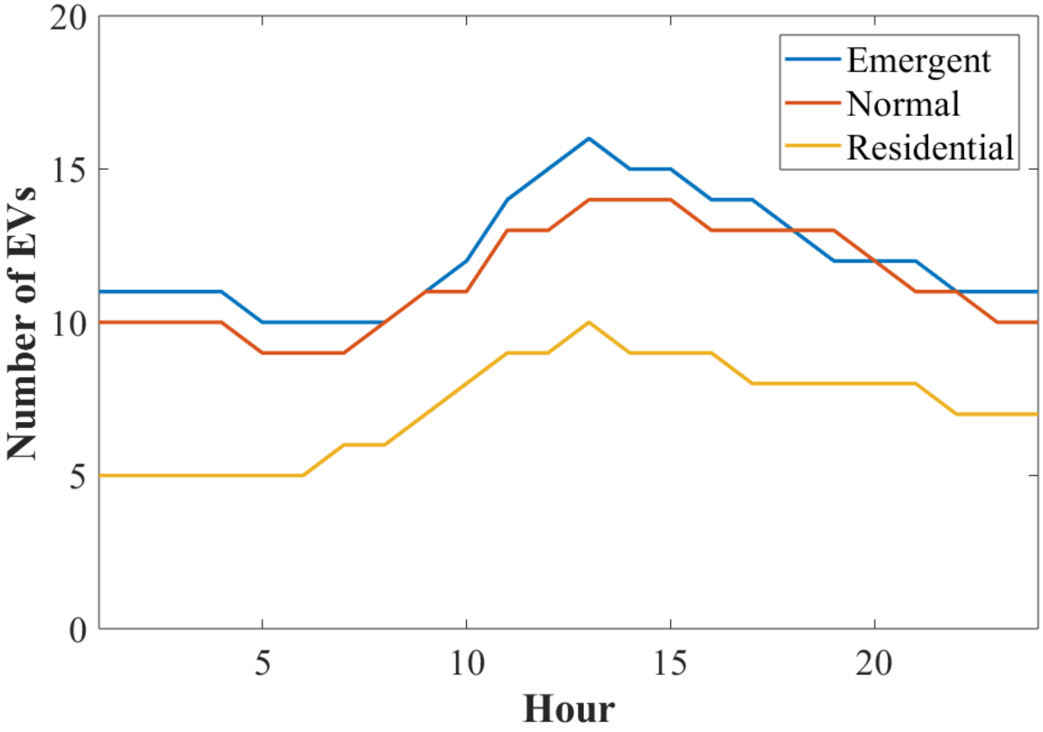
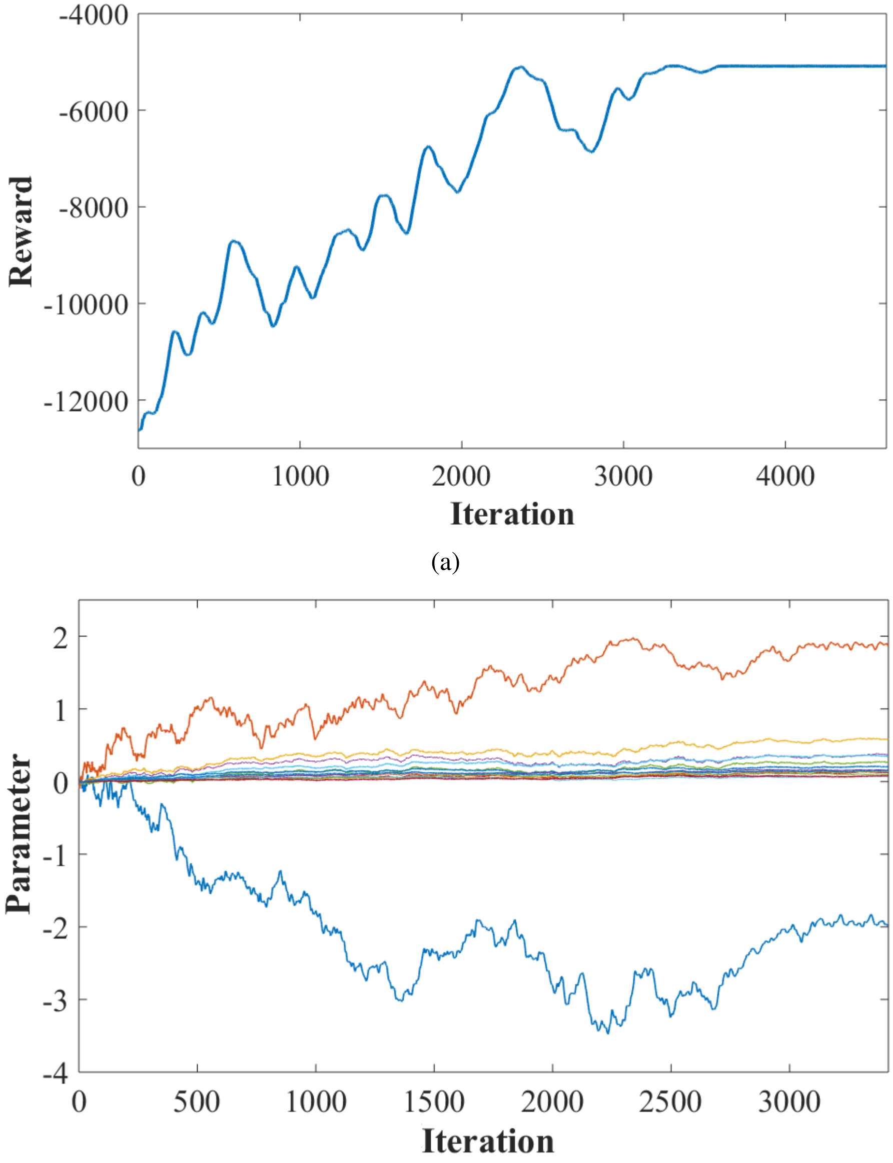


Figure 3.2: Hourly arrivals for the three categories of EVs during one day.

while the total Q-function is assumed to be a linear combination of them. Hence, the RL problem becomes to estimate the best linear coefficients as the parameter based on the Bellman optimality condition for Q-function. Although this approach can deal with time-varying states, the approximation therein is heuristic and could be inaccurate.

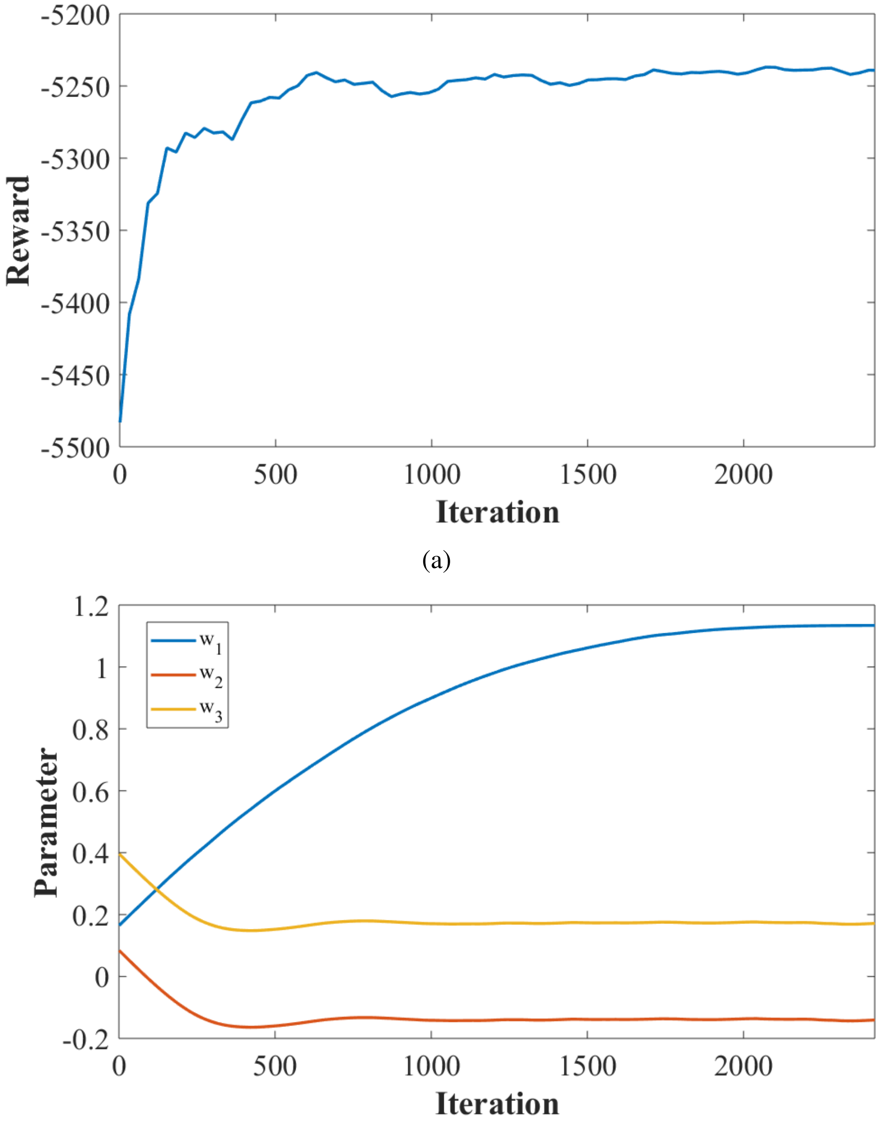
We have used 20 different daily profiles to train the RL algorithms. Fig. 3.3 and Fig. 3.4 plot the episode rewards and parameter values for the proposed **Algorithm 3** and **Algorithm *QE***, respectively. Clearly, both RL algorithms are shown to converge as rewards gradually increasing and parameter values stabilizing.



(a)

(b)

Figure 3.3: (a) The episode reward and (b) episode parameter values for **Algorithm 3**.



(a)

(b)

Figure 3.4: (a) The episode reward and (b) episode parameter values for Algorithm QE.

One important observation from the episode parameter values in Fig. 3.3 is that they are almost zero for most states, except for state ρ_t and $n_t^{(0)}$. Specifically,

State	Parameter	**State**	Parameter
ρ_t	-1.9735	$n_t^{(6)}$	0.2021
$n_t^{(0)}$	1.8628	$n_t^{(7)}$	0.1404
$n_t^{(1)}$	0.5772	$n_t^{(8)}$	0.1386
$n_t^{(2)}$	0.3674	$n_t^{(9)}$	0.1592
$n_t^{(3)}$	0.2651	$n_t^{(10)}$	0.0975
$n_t^{(4)}$	0.3485	$n_t^{(11)}$	0.0693
$n_t^{(5)}$	0.1191	$n_t^{(12)}$	0.0797

Table 3.3: Parameter values obtained by **Algorithm 3**.

the negative most parameter is for ρ_t as the total charging budget a_t should decrease when the price is high. In addition, the positive most parameter is for $n_t^{(0)}$ as a_t should increase when there are many EVs with emergent charging needs. Compared to these two parameters, the states for other laxity groups have minimal parameter values, with the parameter value decreasing at larger laxity ℓ, as listed in Table 3.3. This learning result is very reasonable as this problem depends mainly on the EVs approaching their department deadlines. As mentioned in Section 3.3, it is possible to further reduce the number of states by merging the high-laxity EVs (larger than a threshold $L_{\max}$) into one single group. This simplification may violate the dynamic homogeneity condition, but it may not affect much the optimality of the resultant RL solution for practical systems based on this observation on minimal parameter values for high-laxity group states.

Using the two policies obtained by the RL training, we have compared their testing performances using five additional daily profiles. Table 3.4 lists the total reward values attained by each of the two policies for each test trajectory. Clearly, the solution by **Algorithm 3** achieves higher total reward values, increasing those

	Test 1	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Average**
Alg. 2	-5016.2	-5022.6	-5009.5	-5012.8	-5007.8	-5013.8
Alg. QE	-5240.1	-5240.3	-5234.2	-5239.3	-5230.6	-5236.9
Increase (%)	4.27	4.15	4.29	4.32	4.26	4.26

Table 3.4: Testing reward values and percentage reward increases of the solution obtained by **Algorithm 3**, as compared to **Algorithm *QE***.

acquired by **Algorithm *QE*** by around 4.15% to 4.32%. Thanks to the equivalent state aggregation, **Algorithm 3** can effectively reduce the total charging cost for the EVCS. It enjoys high modeling accuracy as compared to the Q-function approximation in [54].

To better illustrate the improvement of **Algorithm 3**, Fig. 3.5 plots the daily total charging action comparisons along with the electricity market price. Interestingly, **Algorithm 3** is very sensitive to the price peaks and has chosen to dramatically reduce a_t. Meanwhile, **Algorithm QE** fails to reduce the charging needs over the peak-price period, as highlighted by the shaded area. This example further verifies that our proposed EVCS operation can improve the cost performance while enjoying efficient RL solution time by considering the equivalent MDP problem.

This work has developed a practical modeling approach for the optimal EV charging station operation problem, allowing for efficient solutions using reinforcement learning (RL). To deal with the high and variable dimensions of states/actions, we propose to design efficient aggregation schemes by utilizing the EV's laxity that

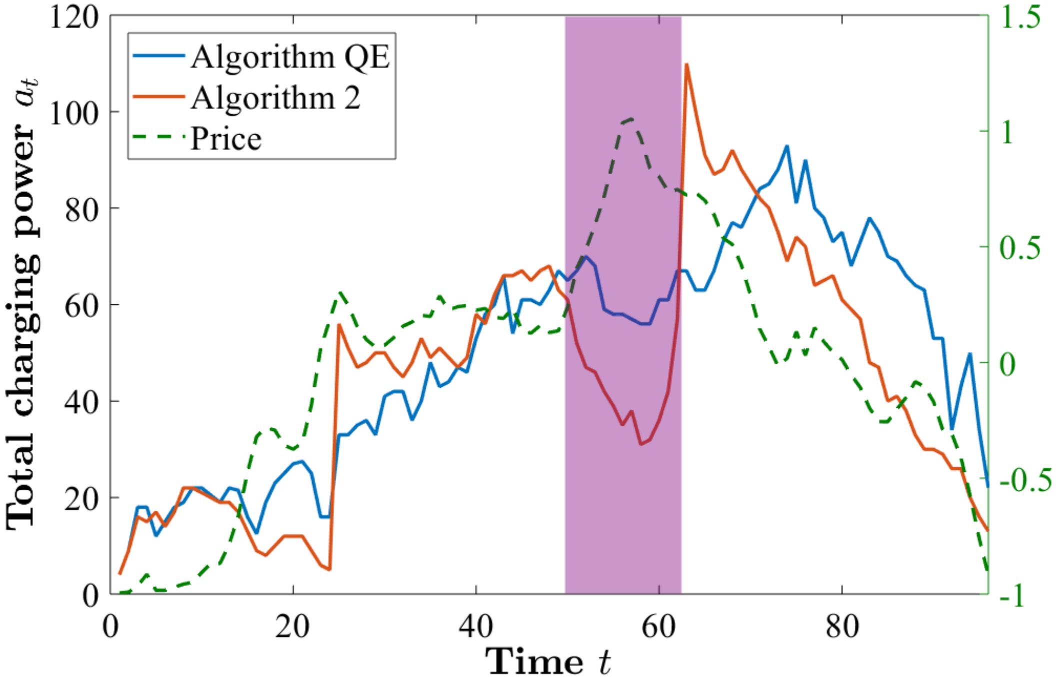


Figure 3.5: The daily profiles of total charging power respectively produced by Algorithms 3 and QE for one testing day as compared to the electricity market prices.

measures the emergency level of its charging need. First, the least-laxity first (LLF) rule has made it possible to consider only the total charging action across the EVCS, which is shown to recover feasible individual EV charging schedules if existing. Second, we propose aggregating the state into the number of EVs in each laxity group, which satisfies reward and dynamic homogeneities and thus leads to equivalent policy search. We have developed the policy gradient method based on the proposed MDP representation to find the optimal parameters for the linear Gaussian policy. Case studies based on real-world data have demonstrated the performance improvement of the proposed MDP representation over the earlier approximation-based approach for the EVCS problem. The RL parameter results imply that further state aggregation can deal with many laxity levels in practical systems at a minimal loss of optimality.

Chapter 4

RL-based Load Frequency Control

This chapter focuses on the a risk-aware load frequency control considering a risk-constraint and structured feedback. Section 4.1 introduces a load frequency control (LFC) problem and the motivation of this work. Section 4.2 formulates the LFC problem based on a radially-connected networked microgrid (MG) system. In Section 4.3, we formulate a LFC problem as a general infinite-horizon risk-constrained LQR problem with structured feedback control. Section 4.4 introduces the dual-related minimax reformulation and analyzes the convergence of the Gradient Descent with max-oracle (GDmax) algorithm. Section 4.5 extends it to model-free learning by developing the Stochastic (S)GDmax via zero-order policy gradient. Section 4.6 presents the numerical results in a networked LFC problem.

This chapter is based on the following publications:

K. Kwon, L. Ye, V. Gupta, and H. Zhu, "Model-free Learning for Risk-constrained Linear Quadratic Regulator with Structured Feedback in Networked Systems," *2022 IEEE 61st Conference on Decision and Control (CDC)*, Cancun, Mexico, 2022, pp. 7260-7265, doi: 10.1109/CDC51059.2022.9993178,

K. Kwon, S. Mukherjee, H. Zhu, and T. L. Vu, "Reinforcement Learning-based Output Structured Feedback for Distributed Multi-Area Power System Frequency Control," *2023 American Control Conference (ACC)*, San Diego, CA, USA, 2023, pp. 4483-4488, doi: 10.23919/ACC55779.2023.10156618.

4.1 Load Frequency Control Problem

Load frequency control (LFC) is one of the most important control problems in power system operations. The objective of LFC is to maintain the frequency of each area in an interconnected power system by adjusting the output of generators with automatic generation control (AGC) regulator or excitation controller [61]. The LFC has been studied in various research to cope with conventional generators [62], distributed energy resources [63], and electric vehicles (EVs) [64]. However, most research considers a centralized framework, implying that a centralized dispatch center controls all generators [65]. As a modern power system relies more on distributed generation (DG) and requires resilience to cyberattacks, the existing centralized control paradigm with one-point-failure faces many challenges. Distributed or decentralized frameworks can be considerable frameworks to improve the stability and security of the power system. Consequently, there is a lot of research that considers distributed or decentralized LFC problems, assuming the limited information exchanges among the interconnected areas [35, 66–68]. Especially in the case of peer-to-peer (P2P) based LFC, generation control in one area is determined by only the information from the areas connected in the information-exchange graph [66].

The general LFC problem can be represented as a linear quadratic regulator (LQR) problem, which minimizes the frequency deviations and other factors such as power outputs, power inflow between interconnected areas, and control efforts such as AGC control signal [69, Ch. 2]. When the model is known, this problem can be easily solved by adopting the Algebraic Riccati equation (ARE) [70] or by applying gradient-based methods [71, 72]. However, finding the feedback gain will become complicated as we consider the model uncertainties, the constraints on the optimization problem, or the structure of the feedback gain. The recent works solves

This work aims to solve the LFC problem, while dealing with these three challenges at once by considering uncertainty from the environment, the risk constraints, and structured output feedback. In particular, reinforcement learning (RL) will be applied to solve this problem. There are two advantages that RL has [2]. First, RL is model-free learning, i.e., we do not need to consider the model parameters. Instead, we will generate the trajectory, observe the reward and update the controller toward increasing the total reward. Second, RL is data-driven learning, i.e., instead of generating a probability distribution for the uncertainty and performing the Monte-Carlo method, we can directly use the data gathered from the grid and use it to train the controller. Recent research papers such as [73–75] have investigated distributed implementation of RL. While RL has also been adopted in [76] and [77] to solve the LFC problem in a model-free setting, the statistical risk of the resultant controllers therein has not been considered.

To solve this problem, we develop a general model-free learning algorithms for risk-constrained LQR problem under sparse feedback structure that arises in networked systems. The structured feedback is incorporated by considering the sparse non-zero entries only, and thus the gradient computation and updates can be performed without accounting for such structured constraint. Nonetheless, it leads to convergence to only a stationary point. As for the constraint function, it is similar to the LQR cost with the mean-variance risk as a special case as shown by [78, 79]. To deal with this constraint, we consider the dual problem which shares the stationary point (SP) with the minimax problem for the Lagrangian function. The resultant nonconvex-concave minimax reformulation motivates us to adopt Gradient-Descent max-oracle (GDmax) and the stochastic (S)GDmax algorithms in [80] to solve the outer minimization problem via GD updates. More specifically, the SGDmax relies on the zero-order policy gradient (ZOPG) [81] which has bounded noise variance.

Nonetheless, the key challenge in establishing the convergence results lies in the LQR cost function, which is shown to exhibit local-only Lipschitz and smoothness properties with location-dependent constants [71, 82]. To tackle this, we can introduce a compact sublevel set within which the upper bounds of Lipschitz and smoothness constants hold everywhere. Such analysis enables us to carefully design the stepsize and related parameters to establish the convergence to SP, while the convergence of SGDmax in a model-free setting can be attained with a high probability. Numerical results have validated the convergence of our algorithms and demonstrated the impact of having risk constraint and structured feedback in learning LQR policy. The SGDmax algorithm have attained satisfactory optimality gap compared to the classical LQR control, especially for the full feedback case.

Notations: Let $\|\cdot\|$ denotes the L_2-norm, $\nabla_{\mathcal{K}}\mathcal{L}$ the gradient of $\mathcal{L}$ that admits the structure defined in $\mathcal{K}$, $\{X^j\}$ a sequence of $\{X^0, X^1, \ldots\}$, $\mathcal{P}_{\mathcal{Y}}(\cdot)$ the projection onto the set $\mathcal{Y}$, and the operator $\otimes$ the Kronecker product of matrices. Last, $\mathbb{E}(\cdot)$ denotes the expectation while $\mathbb{P}(\cdot)$ the probability of an event.

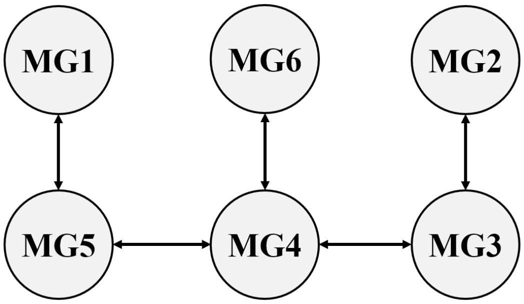


Figure 4.1: A radially connected networked microgrid system.

4.2 Problem Formulation

We consider the load frequency control (LFC) problem in a low-inertia networked microgrid (MG) system with a risk constraint on the frequency states. Fig. 4.1 depicts a radially connected system with $N = 6$ MGs, while Table 4.1 lists the model information which follows from [13]. Consider the communication graph to be the same as the MG network show in Fig. 4.1. Thus, each MG a can only exchange information with their neighboring MGs that are physically connected by tie-lines, and the structured feedback $\mathcal{K}$ is specified accordingly.

Each MG a is assumed to follow linearized power-frequency dynamics including turbine swing and primary control based on the automatic generation control (AGC) signal. Thus, the following symbols all correspond to the deviation from steady-state values as denoted by Δ, with the parameters listed in Table 4.1. First, the primary frequency control in each MG a is proportional to frequency deviation as $\Delta P_{f,a} = -(1/R_a)\Delta f_a$ based on the given droop R_a. Second, the secondary AGC signal $\Delta P_{C,a}$ constitutes as the control action u_t in (4.5) to be designed. The two controls jointly determine the power output of MG a as denoted by $\Delta P_{G,a}$.

Last, Δf_a is also affected by the unknown load demand deviation $\Delta P_{L,a}$ and the total power inflow $\Delta P_{tie,a}$, in addition to $\Delta P_{G,a}$. Note that $\Delta P_{tie,a}$ is the total tie-line power inflow from all neighboring MGs due to their frequency differences, as

$$\Delta P_{tie,a} = \int \sum_{a \leftrightarrow b} K_{tie,a}(\Delta f_a - \Delta f_b)dt, \tag{4.1}$$

where $a \leftrightarrow b$ indicates two MGs are connected to each other. In addition to the MG dynamics, the Area Control Error (ACE) defined as $z_a := \beta_a \Delta f_a + \Delta P_{tie,a}$ is also a state variable as an integral control input with the bias factor $\beta_a = D_a + 1/R_a$ [83].

Hence, MG a has the state vector $x_a = [\Delta f_a, \Delta P_{G,a}, \Delta P_{tie,a}, \int z_a]^\top$ and the control action $u_a = \Delta P_{C,a}$, with load disturbance $w_a = \Delta P_{L,a}$. Assuming all MGs having the same parameter values, we can drop the parameter index a and represent the aggregated network dynamics by:

$$\dot{x} = (I_N \otimes A_1 + L \otimes A_2)x + (I_N \otimes B_u)u + (I_N \otimes B_w)\tilde{w} \tag{4.2}$$

with each variable collecting all MGs' respective state, action, and disturbance. In addition, the system matrices are given by

$$A_1 = \begin{bmatrix} -\frac{1}{T_p} & \frac{K_p}{T_p} & -\frac{K_p}{T_p} & 0 \\ -\frac{K_t}{RT_t} & -\frac{1}{T_t} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \beta & 0 & 1 & 0 \end{bmatrix},$$
$$A_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ K_{tie} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, B_u = \begin{bmatrix} 0 \\ \frac{K_t}{T_t} \\ 0 \\ 0 \end{bmatrix}, B_w = \begin{bmatrix} -\frac{K_p}{T_p} \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{4.3}$$

For the aggregated dynamics, the LQR objective cost is specified by

$$Q = I_{N_L} \otimes Q_a, \text{ and } R = I_{N_L} \otimes R_a \tag{4.4}$$

where the matrices Q_a and R_a are same for every MG a and aim to penalize the deviation of both state and action from steady-state values. As discussed in Section 4.3, we further consider a risk constraint $R_c(\cdot)$ in (4.8) for reducing the mean-variance risk in order to improve frequency regulation.

4.3 General LQR Formulation with Structured Feedback

As seen in Section 4.1, LFC problem can be represented as an infinite-horizon LQR problem for a linear time-invariant system. In this section, we formulate a general LQR formulation with structured feedback. First, the dynamics of the linear time-invariant system can be represented as below;

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t = 0, 1, \ldots \tag{4.5}$$

with the state $x_t \in \mathbb{R}^n$, action $u_t \in \mathbb{R}^m$, and random noise $w_t \in \mathbb{R}^n$ that is uncorrelated across time. In addition, the model parameters $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ can be unknown. The constrained LQR problem with structured feedback aims to find an optimal linear feedback gain $K \in \mathbb{R}^{m\times n}$ for the control policy $u_t = -Kx_t$ to:

$$\begin{aligned} \min_{K\in\mathcal{K}} \; R_0(K) &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}[x_t^\top Q x_t + u_t^\top R u_t] \qquad (4.6)\\ \text{s.t. } R_i(K) &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}[x_t^\top Q_i x_t + u_t^\top R_i u_t] \le c_i, \forall i \end{aligned}$$

where matrices $\{Q, R\}$ and $\{Q_i, R_i\}_{i\in\mathcal{I}}$ are all positive (semi-)definite, with $\mathcal{I}$ representing the set of the constraints. The feasible set $\mathcal{K}$ enforces a structured policy, as

$$\mathcal{K} = \{K : K_{a,b} = 0 \text{ if and only if } (a, b) \notin \mathcal{E})\} \tag{4.7}$$

Here, the structure pattern $\mathcal{E}$ is specified by the edges of a given communication or information-exchange graph. Hence, the action for agent a, denoted as $u_{a,t}$, is determined as $u_{a,t} = -K_a x_{a,t}$, where K_a is a row vector with only non-zero elements in a-th row of K and $x_{a,t}$ is a sub-vector of x_t according to $\mathcal{E}$. An example of the communication graph is illustrated in Fig. 4.1. The structured $\mathcal{K}$ is motivated by a multi-agent setting for networked control, where individual agents can access partial feedback only depending on communication links. Notably, this structured constraint will lead to a complicated geometry of the feasible region [71,84]. While the structured $\mathcal{K}$ makes the analysis more difficult than the full feedback case, it does not increase the complexity of computing the gradient as denoted by $\nabla_{\mathcal{K}}$ later on. This is because one can represent the cost as a function of only non-zero entries in K which can eliminate this structured constraint [71]. Accordingly, the $\nabla_{\mathcal{K}}$ operation needs no projection onto $\mathcal{K}$, and can be thought of as the gradient for an unstructured K. Therefore, gradient-based methods are ideal for learning a structured policy.

As for the quadratic constraint in (4.6), one can consider the mean-variance risk as a special instance, represented by

$$R_c(K) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} \left(x_t^\top Q x_t - \mathbb{E}[x_t^\top Q x_t | h_t] \right)^2 \leq \delta$$

with the system trajectory $h_t := \{x_0, u_0, \ldots, x_{t-1}, u_{t-1}\}$ and a risk tolerance δ. This risk measure limits the deviation from the expected cost given the past trajectory, and thus can mitigate extreme scenarios due to the uncertainty in the noisy dynamics. Interestingly, under a finite fourth-order moment of noise w_t, [78, 79] has developed a tractable reformulation $R_c(K)$, as

$$R_c(K) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} \left(4x_t^\top QWQx_t + 4x_t^\top QM_3 \right) \leq \bar{\delta} \tag{4.8}$$

with $\bar{\delta} = \delta - m_4 + 4\text{tr}\{(WQ)^2\}$ and the (weighted) noise statistics given as

$$\bar{w} = \mathbb{E}[w_t], \tag{4.9}$$

$$W = \mathbb{E}[(w_t - \bar{w})(w_t - \bar{w})^\top], \tag{4.10}$$

$$M_3 = \mathbb{E}[(w_t - \bar{w})(w_t - \bar{w})^\top Q(w_t - \bar{w})], \tag{4.11}$$

$$m_4 = \mathbb{E}[(w_t - \bar{w})^\top Q(w_t - \bar{w}) - \text{tr}(WQ)]^2. \tag{4.12}$$

With known noise statistics, this risk constraint shares the quadratic form in (4.6) with an additional linear term, which does not affect our proposed gradient-based learning. The ensuing section first develops the deterministic algorithm for problem (4.6), which can provide insights on the model-free extension later on.

4.4 A Primal Gradient Descent (GD) Approach

To deal with constraints in (4.6), consider its Lagrangian function by introducing the multiplier vector $\lambda = \{\lambda_i \geq 0\}$, as

$$\begin{aligned} \mathcal{L}(K, \lambda) &= R_0(K) + \textstyle\sum_{i \in \mathcal{I}} \lambda_i [R_i(K) - c_i] \\ &= \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \sum_{t=0}^{T-1} [x_t^\top Q_\lambda x_t + u_t^\top R_\lambda u_t] - c_\lambda \end{aligned} \tag{4.13}$$

where we define $Q_\lambda := Q + \sum_{i \in \mathcal{I}} \lambda_i Q_i$, and likewise for R_λ and c_λ. Clearly, $\mathcal{L}(K, \lambda)$ shares the same structure as an unconstrained LQR cost which is suitable for first-order algorithms. For simplicity, consider that the problem (4.6) is feasible and thus λ is finite [85, Sec. 5.2]. We consider the bounded set $\mathcal{Y} := [0,\ \Lambda]^{|\mathcal{I}|}$ for λ with a large enough $\Lambda \in \mathbb{R}$, which can be set based on a feasible K_0. Using the dual function $\mathcal{D}(\lambda) := \min_{K\in\mathcal{K}} \mathcal{L}(K,\lambda)$, the dual problem becomes

$$\max_{\lambda\in\mathcal{Y}} \mathcal{D}(\lambda) = \max_{\lambda\in\mathcal{Y}} \min_{K\in\mathcal{K}} \mathcal{L}(K,\lambda). \tag{4.14}$$

As $\mathcal{L}(K,\lambda)$ is related to LQR cost, the inner minimization problem is not convex. Recent works [71, 72, 82] have extensively analyzed the LQR cost which can be used to establish the local Lipschitz and smoothness properties of $\mathcal{L}(K,\lambda)$. Specifically, it is possible to find related constants that hold within a subset $\mathcal{G}^0 \subset \mathcal{K}$. This compact sublevel set will be defined later on, but is first introduced here for bounding the constants as stated below.

Lemma 2 (Lipschitz and smoothness)**.** *For any λ and $K \in \mathcal{G}^0$, the function $\mathcal{L}(K,\lambda)$ is locally L_0-Lipschitz within a radius ψ_K; i.e., for $\forall K' \in \mathcal{G}^0$ such that $\|K-K'\| \leq \psi_K$, we have $\|\mathcal{L}(K,\lambda) - \mathcal{L}(K',\lambda)\| \leq L_0\|K-K'\|$. In addition, it is also locally ℓ_0-smooth within a radius β_K, such that for $\forall K' \in \mathcal{G}^0$ that satisfies $\|K-K'\| \leq \beta_K$, we have $\|\nabla\mathcal{L}_{\mathcal{K}}(K,\lambda) - \nabla\mathcal{L}_{\mathcal{K}}(K',\lambda)\| \leq \ell_0\|K-K'\|$.*

Strictly speaking, the recent LQR analysis [72, 82] asserts that Lipschitz and smoothness are only local properties, and thus the corresponding constants L_K and ℓ_K depend on K. Nonetheless, using a compact set $\mathcal{G}^0$, we can obtain the bounds that can hold for any $K \in \mathcal{G}^0$, as given by

$$L_0 := \sup_{K\in\mathcal{G}^0} L_K, \text{ and } \ell_0 := \sup_{K\in\mathcal{G}^0} \ell_K. \tag{4.15}$$

We can also determine a general neighborhood radius as

$$\rho_0 := \inf_{K \in \mathcal{G}^0} \min\{\beta_K, \psi_K\} \tag{4.16}$$

that holds for any $K \in \mathcal{G}^0$ as well.

Interestingly, the KKT conditions for problem (4.14) is related to the stationary point (SP) of a reformulated minimax problem. Recent results have shown that nonconvex-concave minimax problems can be solved using the so-termed Gradient Descent with max-oracle (GDmax) algorithm [86]. To this end, consider the problem

$$\min_{K \in \mathcal{K}} \Phi(K) \;\text{ where }\; \Phi(K) := \max_{\lambda \in \mathcal{Y}} \mathcal{L}(K, \lambda), \tag{4.17}$$

which is essentially the minimax counterpart of problem (4.14). As the Lagrangian function is linear in λ, it is possible to directly find the best λ in (4.17). Specifically, its i-th element, namely λ_i, depends on the feasibility of constraint i under given K; i.e., λ_i equals to 0 if constraint i is satisfied and Λ otherwise. Unfortunately, the function $\Phi(K)$ is not differentiable everywhere. To tackle this issue, we consider its *Moreau envelope* $\Phi_\mu(\cdot)$ for a given $\mu > 0$, defined as

$$\Phi_\mu(K) := \min_{K' \in \mathcal{K}} \Phi(K') + \frac{1}{2\mu}\|K' - K\|^2, \;\; \forall K \in \mathcal{K}. \tag{4.18}$$

It can be used for defining the SP of the non-differentiable $\Phi(K)$, following from [80, Lemma 3.6].

Lemma 3. *As $\mathcal{L}(K, \lambda)$ is concave in λ and $\mathcal{Y}$ is convex and bounded, Lemma 2 asserts that $\Phi(K)$ is ℓ_0-weakly convex and L_0-Lipschitz within the compact set*

$\mathcal{G}^0$. *Accordingly, its Moreau envelope* $\Phi_{\mu_0}(K)$ *is convex by setting* $\mu_0 := 1/(2\ell_0)$. *Hence, the* ϵ*-SP of* $\Phi(K)$*, namely* K_ϵ*, satisfies* $\|\nabla\Phi_{\mu_0}(K_\epsilon)\| \leq \epsilon$.

The properties of $\Phi(K)$ in Lemma 3 follow from its relation to $\mathcal{L}(K,\lambda)$, as detailed in [80]. Even though it is non-differentiable, one can define the SP here based on $\Phi_{\mu_0}(K)$ which will be used for the convergence analysis of GD updates later on. Notably, the ϵ-SP of $\Phi(K)$ is equivalently related to the stationarity conditions for $\mathcal{L}(K,\lambda)$. According to [80, Prop. 4.12], one can utilize K_ϵ from Lemma 3 to generate the following pair $(\tilde{K}_\epsilon, \tilde{\lambda}_\epsilon)$ by performing an additional $O(\epsilon^{-2})$ number of gradient updates:

$$\begin{aligned}
&\|\nabla_{\mathcal{K}}\mathcal{L}(\tilde{K}_\epsilon, \tilde{\lambda}_\epsilon)\| \leq \epsilon,\\
&\left\|\mathbb{P}_{\mathcal{Y}}\left(\tilde{\lambda}_\epsilon + (1/\ell_0)\nabla_\lambda\mathcal{L}(\tilde{K}_\epsilon, \tilde{\lambda}_\epsilon)\right) - \tilde{\lambda}_\epsilon\right\| \leq \epsilon/\ell_0,
\end{aligned}$$

where $\mathbb{P}_{\mathcal{Y}}$ stands for the projection onto $\mathcal{Y}$. Clearly, when $\epsilon \to 0$ this represents the Lagrangian optimality conditions for problem (4.14), and thus the pair $(\tilde{K}_\epsilon, \tilde{\lambda}_\epsilon)$ can be viewed as the ϵ-SP for $\mathcal{L}(K,\lambda)$.

We can solve (4.17) using iterative GD updates, as tabulated in Algorithm 4. With an initial K^0, we need to find the subgradient of $\Phi(K^j)$ at every iteration j. Interestingly, this is equivalent to the gradient of $\mathcal{L}$ over K^j [80]; i.e., $\partial\Phi(K^j) = \nabla_{\mathcal{K}}\mathcal{L}(K^j, \lambda^j)$ with λ^j being the optimal multiplier for the given K^j. Hence, the Lagrangian $\mathcal{L}$ will be used to perform the GD updates for $\Phi(K)$ minimization. The convergence of Algorithm 4 can be established below, with the detailed proof in Appendix A.

Algorithm 4: Gradient Descent with max-oracle (GDmax)

```
Inputs: A feasible policy $K^0$, upper bound $\Lambda$ for $\lambda$, threshold $\epsilon$, and
  the initial iteration index $j = 0$.
Determine $L_0, \ell_0$, and $\rho_0$ using the set $\mathcal{G}^0$ and compute the stepsize as
  in (4.19).
while $\|\nabla_{\mathcal{K}}\mathcal{L}(K^j, \lambda^j)\| > \epsilon$ do
    Obtain $\lambda^j \leftarrow \arg\max_{\lambda\in\mathcal{Y}} \mathcal{L}(K^j, \lambda)$
    Update $K^{j+1} \leftarrow K^j - \eta\nabla_{\mathcal{K}}\mathcal{L}(K^j, \lambda^j)$;
    Set $j \leftarrow j + 1$.
end
Return: the final iterate $K^j$.
```

Theorem 1. *With an initial $K^0 \in \mathcal{K}$ and by setting stepsize*

$$\eta \leq \min\left\{\frac{\epsilon^2}{4\ell_0 L_0^2}, \rho_0\right\}, \tag{4.19}$$

Algorithm 4 is guaranteed to converge to K_ϵ for $\Phi(K)$, which can be used to obtain an ϵ-SP for the dual problem (4.14)*. The number of iterations required for attaining K_ϵ is $O(\ell_0 L_2^2 \Phi_{\mu_0}(K^0)/\epsilon^4)$.*

As discussed in Appendix A, we can bound the iterative changes in $\Phi_{\mu_0}(K^j)$, which ensures that the sequence $\{\Phi_{\mu_0}(K^j)\}$ is non-increasing. Thus, if we define the sublevel set to be

$$\mathcal{G}^0 := \{K \in \mathcal{K} | \Phi_{\mu_0}(K) \leq \Phi_{\mu_0}(K^0)\}, \tag{4.20}$$

then the iterates $\{K^j\}$ are guaranteed to be within $\mathcal{G}^0$. This is exactly how one can bound the constants L_0 and ℓ_0 as given by (4.15). Of course, the choice of μ_0 in the sublevel set $\mathcal{G}_0$ depends on ℓ_0, which may not be known before $\mathcal{G}^0$ is constructed. This issue is discussed in the following remark.

Remark 2 (Sublevel set)**.** *With initial K^0 given, the set $\mathcal{G}^0$ is defined with the value μ_0, which depends on the upper bound of ℓ_K within $\mathcal{G}^0$ as shown in* (4.15)*. This dependence can be addressed by determining the value of μ_0 in an adaptive fashion. Starting with a rough estimate of ℓ_0 and μ_0, one can first construct a $\mathcal{G}^0$ and compare the resultant bound with the original estimate on ℓ_0. If the latter is larger, then $\mathcal{G}^0$ works well. Otherwise, one can gradually increase the ℓ_0 estimate to achieve that condition. Our experimental experience suggests some conservative choice of stepsize can ensure the convergence in practice.*

4.5 Stochastic GD for Model-free Learning

To account for unknown system dynamics, we extend the GDmax approach to a model-free setting. The iterative gradient will be obtained via the zero-order optimization [81]. Unfortunately, this stochastic gradient update can complicate the convergence analysis as detailed later, mainly due to the aforementioned issue on local properties of LQR cost.

Zero-order policy gradient (ZOPG) has been popularly developed in recent years for model-free gradient-based learning. It provides an unbiased gradient estimate in an efficient manner. For the function $\Phi(K)$, ZOPG aims to evaluate the function value at any K under a structured, random perturbation from the set $\mathcal{S}_{\mathcal{K}} = \{U \in \mathcal{K} : \|U\| = 1\}$, as detailed in Algorithm 5. Note that the structure of perturbation U is the same to that of K with non-zero entries randomly sampled from e.g., the uniform distribution, followed by a normalization step to ensure unity norm. Given a smoothing radius $r > 0$, the ZOPG is estimated using the resultant

```
Algorithm 5: Zero-Order Policy Gradient (ZOPG)
Inputs: smoothing radius $r$, the policy $K$ and its perturbation $U \in \mathcal{S}_{\mathcal{K}}$, both of $n_{\mathcal{K}}$ non-zeros.
Obtain $\lambda' \leftarrow \arg\max_{\lambda \in \mathcal{Y}} \mathcal{L}(K + rU, \lambda)$;
Estimate the gradient $\hat{\nabla}_{\mathcal{K}}\mathcal{L}(K; U) = \frac{n_{\mathcal{K}}}{r}\mathcal{L}(K + rU, \lambda')U$.
Return: $\hat{\nabla}_{\mathcal{K}}\mathcal{L}(K; U)$.
```

$\Phi(K + rU)$ from this perturbation by finding the corresponding optimal λ in (4.17). We denote $n_{\mathcal{K}}$ as the total number of nonzero entries in $\mathcal{K}$, which is used to scale the gradient estimate. Since the estimated $\hat{\nabla}_{\mathcal{K}}\mathcal{L}$ follows from matrix U, it maintains the same sparse structure given by $\mathcal{K}$.

The stochastic ZOPG will make it more difficult to maintain the iterative updates to stay within a sublevel set, and likewise for bounding Lipschitz and smoothness constants. Fortunately, [82] has developed an approach to attain this condition with a high probability. Specifically, one can set up a ten-fold sublevel set, given by

$$\mathcal{G}^1 := \{K \in \mathcal{K} | \Phi_{\mu_0}(K) \leq 10\ \Phi_{\mu_0}(K^0)\}. \tag{4.21}$$

Using $\mathcal{G}^1$, one can determine L_0, ℓ_0, and ρ_0 over the set $\mathcal{G}^1$ similar to (4.15)-(4.16), and they will be used for the convergence analysis. Note that the choice of μ_0 in $\mathcal{G}^1$ depends on the ℓ_0 value, which can be addressed as discussed in Remark 2.

Algorithm 6 tabulates the ZOPG-based model-free learning approach for solving (4.14), termed as stochastic gradient descent with max-oracle (SGDmax) [86]. Its convergence guarantee can be established with the detailed proof in Appendix B.

```
Algorithm 6: Stochastic Gradient Descent with max-oracle (SGDmax)
Inputs: A feasible policy $K^0$, upper bound $\Lambda$ for $\lambda$, threshold $\epsilon$, and number of ZOPG samples $M$.
Determine $L_0, \ell_0$, and $\rho_0$ with the set $\mathcal{G}^1$ and compute $r$, $\eta$, and $J$ as in (4.22);
for $j = 0, 1, \ldots, J-1$ do
    for $s = 1, \ldots, M$ do
        Sample the random $U_s \in \mathcal{S}_{\mathcal{K}}$;
        Use Algorithm 5 to return $\hat{\nabla}\mathcal{L}_{\mathcal{K}}(K^j; U_s)$.
    end
    Update $K^{j+1} \leftarrow K^j - \eta \left(\frac{1}{M}\sum_{s=1}^{M} \hat{\nabla}\mathcal{L}(K^j; U_s)\right)$.
end
Return: the final iterate $K^J$.
```

Theorem 2. *With an initial $K^0 \in \mathcal{K}$ and a given $\epsilon > 0$, we can set the parameters as*

$$r \leq \min\left\{\rho_0, \frac{L_0\sqrt{M}}{\ell_0}\right\},\ \eta \leq \frac{\epsilon^2}{\alpha\ell_0(L_0^2 + \ell_0^2 r^2/M)}, \text{and} \quad J = \frac{2\sqrt{10\alpha}\Phi_{\mu_0}(K^0)}{\eta\epsilon^2} \tag{4.22}$$

with L_0, ℓ_0 and ρ_0 being specified using $\mathcal{G}^1$, and a large constant α. This way, Algorithm 6 converges to the ϵ-SP K_ϵ with probability of at least $(0.9 - \frac{4}{\alpha} - \frac{4}{\sqrt{10\alpha}})$.

Last, the proposed algorithms can be easily extended to the case of full feedback K, with computational advantages over existing solutions as discussed below.

Remark 3 (Full feedback K)**.** *For the full feedback case, we can directly implement the proposed Algorithms 4-6 by dropping the structured set $\mathcal{K}$. This setting has*

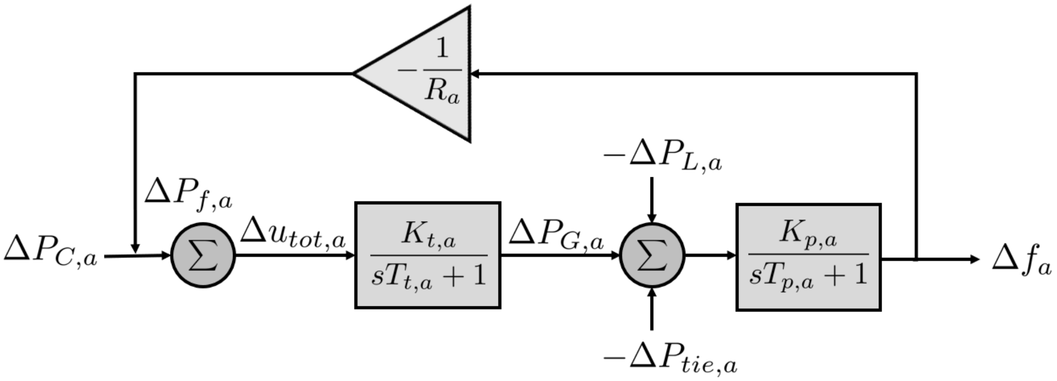


Figure 4.2: Block representation of load frequency control in i-th microgrid

been considered in [67] by using a dual-ascent based double-loop scheme where the inner-loop minimizes K till convergence for any fixed λ. In contrast, our proposed algorithms eliminate this inner-loop, which is more computationally efficient. Investigating the global convergence property of our proposed SGDmax algorithm for the full feedback case constitutes as an interesting future direction.

4.6 Numerical Tests

To demonstrate the effectiveness of the proposed model-free learning approach, we consider a LFC problem based on the system as discussed in Fig. 4.1. The block representation of LFC in i-th MG and the corresponding parameters are listed in Table 4.1, respectively. As we already discussed in Section 4.2, we use the dynamics represented in (4.2) with each matrix given by (4.3). To design a structured feedback controller K, we construct a risk-constrained optimization problem in (4.6), with the positive (semi-)definite matrices Q and R are given by (4.4).

We consider the following three cases to demonstrate the impact of struc-

Table 4.1: List of parameter and their values

Parameter	Symbol	Value	Units
Damping Factor	D	16.66	MW/Hz
Speed Droop	R	1.2×10^{-3}	Hz/MW
Turbine Static Gain	K_t	1	MW/MW
Turbine Time Constant	T_t	0.3	s
Area Static Gain	K_p	0.06	Hz/MW
Area Time Constant	T_p	24	s
Tie-line Coefficient	K_{tie}	1090	MW/Hz

tured K along with the risk constraint:

- Case 1): Structured K with risk constraint
- Case 2): Full K with risk constraint
- Case 3): Full K without risk constraint

For cases 1 and 2, we implemented Algorithm 6 using SGDmax while a simple ZOPG-based algorithm [82] was used for case 3. For all algorithms, we picked a small stepsize of $\eta = 10^{-4}$ with a smoothing radius $r = 1$ and $M = 100$ samples for ZOPG. All three cases have shown to converge to a steady-state with sufficient updates, as shown by Fig. 4.3. In particular, the LQR cost attained by case 2 is slightly over that by case 3, suggesting a global convergence result for SGDmax in full feedback case as discussed in Remark 3. Case 1 demonstrates the highest steady-state LQR cost out of the three, as it has the most restrictive conditions. However, the minimum LQR cost by case 1 is still pretty close to that by case 3, implying some good optimality gap. Notably, case 1 has shown some large

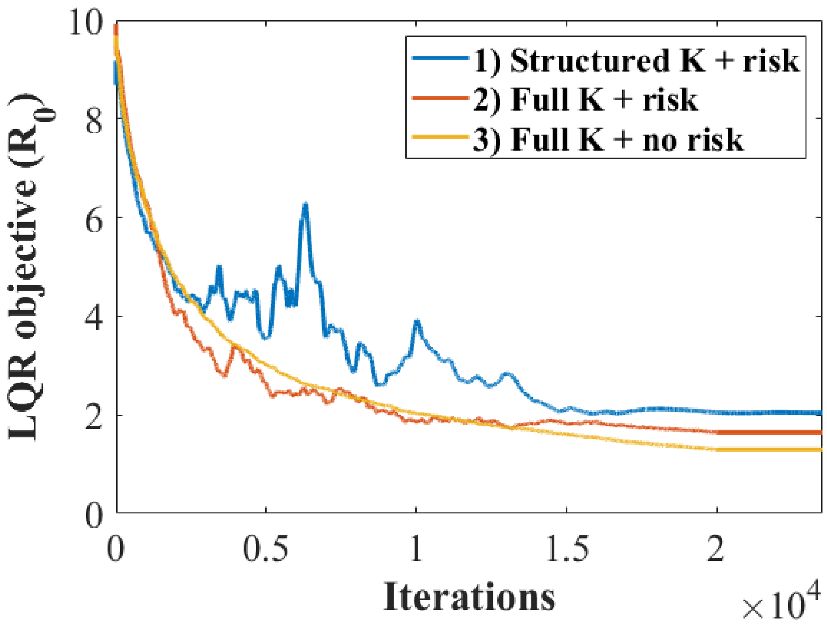


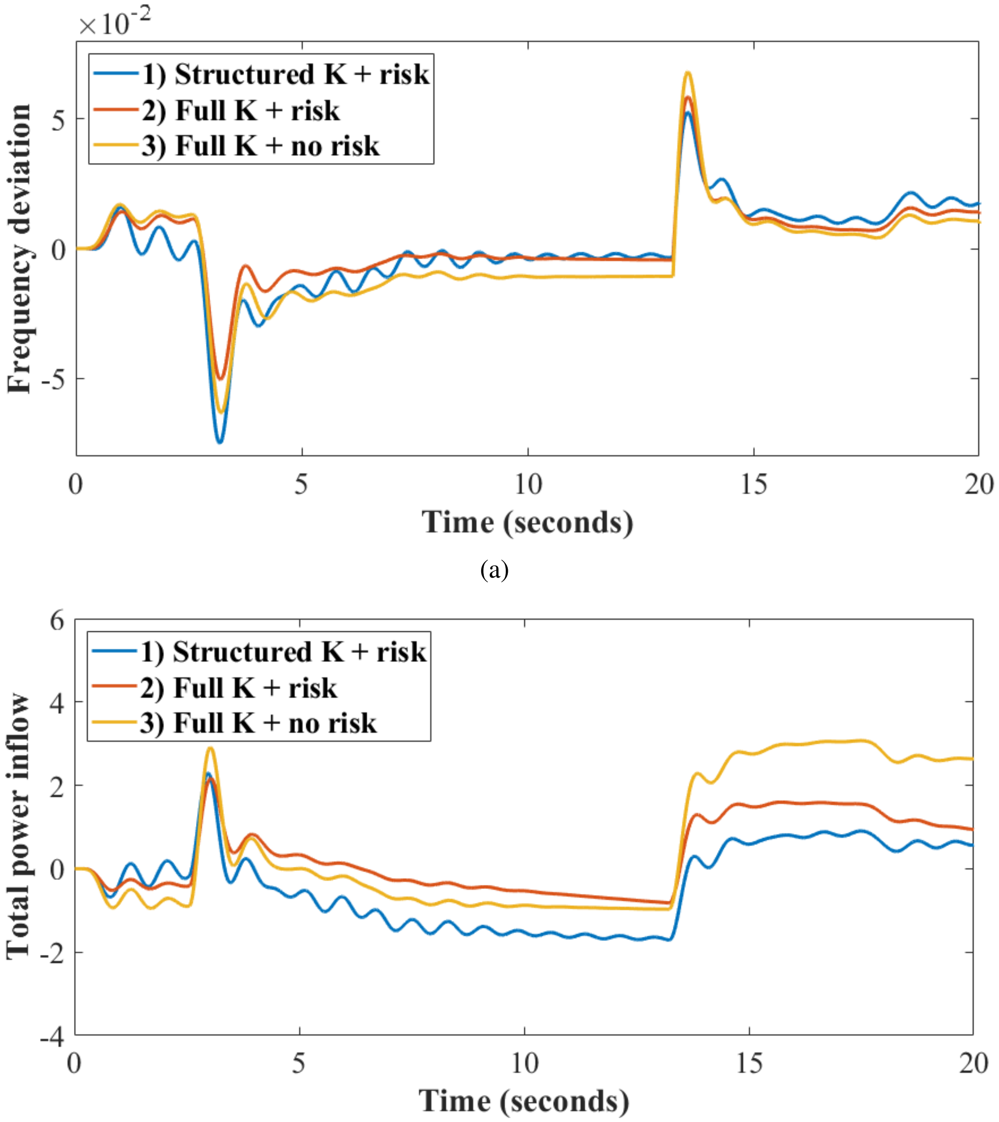
Figure 4.3: Comparison of LQR objective trajectories for the three cases.

fluctuations along the learning process, indicating a complicated geometry that the problem may have.

We also test the converged policy by each case by generating a scenario that all six MGs have some random load changes in a 20-second window. Each area experiences a step load change at a random time. Fig. 4.4 compares the frequency deviation and the total power inflow for MG 2. Clearly, Fig. 4.4(a) demonstrates that the risk constraint can effectively reduce the frequency deviation, as case 2 has the smallest deviation among all three. With the risk constraint, case 1 tends to exhibit great frequency performance as well, but also shows some small oscillations possibly due to the structured feedback policy. This observation points out that limited information exchange can potentially affect the control performance. Similar patterns have been observed in Fig. 4.4(b). While case 1 can maintain the tie-line inflow at the same level as case 2, it still has more noticeable oscillations. As the power inflow is proportional to frequency difference, reducing the risk of frequency deviation can enhance the performance in maintaining the level of power inflow.

To sum up, our numerical tests have validated the convergence performance of the proposed SGDmax based policy gradient method for risk-constrained LQR problem with structured policy. The effectiveness of risk constraint in mitigating large state deviation have been verified, while the sparse structure of K has shown to save communication overhead at the cost of transient oscillations.

This work has developed a practical modeling approach for the optimal EV charging station operation problem, allowing for efficient solutions using reinforcement learning (RL). To deal with the high and variable dimensions of states/actions, we propose to design efficient aggregation schemes by utilizing the EV's laxity that measures the emergency level of its charging need. First, the least-laxity first (LLF) rule has made it possible to consider only the total charging action across the EVCS, which is shown to recover feasible individual EV charging schedules if existing. Second, we propose aggregating the state into the number of EVs in each laxity group, which satisfies reward and dynamic homogeneities and thus leads to equivalent policy search. We have developed the policy gradient method based on the proposed MDP representation to find the optimal parameters for the linear Gaussian policy. Case studies based on real-world data have demonstrated the performance improvement of the proposed MDP representation over the earlier approximation-based approach for the EVCS problem. The RL parameter results imply that further



(a)

(b)

Figure 4.4: Comparison of the (a) frequency deviation and (b) total power inflow at MG 2 for the three cases.

state aggregation can deal with many laxity levels in practical systems at a minimal loss of optimality.

Chapter 5

RL-based Wide-area Damping Control

This chapter focuses developing risk-aware wide-area damping controller using reinforcement learning considering communication delays in the information-exchange network. Section 5.1 introduces an wide-area damping control (WADC) problem and the motivation of this work. Section 5.2 formulates the linearized system model that includes voltage source converters (VSCs). In Section 5.3, we model the communication networks and analyze the impacts of delays, as well as formulate the risk-constrained linear quadratic regulator (LQR) problem. In Section 5.4, numerical tests on the IEEE 68-bus system will be used to demonstrate the performance improvements of the proposed design.

5.1 Wide-area damping control problem

Wide-area damping control (WADC) can greatly enhance power system stability and mitigate inter-area oscillations which are a root cause of large-scale black-

This chapter is based on the following submission:
K. Kwon, L. Ye, V. Gupta, and H. Zhu. "Risk-Constrained Reinforcement Learning for Wide-area Damping Control under Communication Delays." submitted to *IEEE Transactions on Power Systems*, 2023.

outs [87]. Weak tie lines, as well as increasing penetration of low-carbon energy resources, have led to growing concerns over wide-area stability. Despite recent developments in WADC with wide-area measurement system (WAMS), there still exist significant challenges in its implementations, including integrating renewable energy sources (RESs) [88–91], addressing cyberattacks [67, 92–94], and improving the robustness against WAMS' communication delays [95–98].

The optimal WADC problem can be formulated using the linear quadratic regulator (LQR) objective, which minimizes the deviation of state variables such as frequency and angle in addition to the control effort [99]. The underlying communication network of WAMS allows the controller to map from the state data provided by phasor measurement units (PMUs) to the actuation inputs at individual generators. As an important practical aspect, communication links in the WAMS are limited in numbers, making it important to consider either a *structured feedback* framework [100] or a control aggregation scheme per the network connectivity [101]. We consider the former in this work and require the designed controller to follow a prescribed structure in terms of the connectivity between the actuation inputs and the PMU data. More importantly, communication links introduce delays, arising from the fast control timescales and sampling rates. Each link can experience different delays, leading to asynchronous inputs to the distributed actuators [14]. Recent works to address this issue are mainly limited to developing LQR solutions with some heuristic approaches such as hybrid particle swarm optimization [95], robust H_∞ control [97], and adaptive parameter tuning [98]. While the LQR objective is useful for addressing large oscillations due to random faults, it faces a significant gap in systematically mitigating the impacts of communication delays in terms of higher system variability and thus worst-case oscillation levels.

Another practical interest of WADC is to include voltage source converters (VSCs) that can provide additional damping support in the presence of renewable uncertainty [7, 8]. While some approaches [102, 103] have considered to use the flexible power outputs from VSCs, they typically focus on simple synchronous generator (SG) models without the exciter components which are crucial for the WADC dynamics.

Moreover, recent trend in designing WADC is to utilize data-driven controls. In particular, measurement-based approaches have been advocated to first construct the underlying dynamic system models from input-output data; see e.g., [101, 102, 104]. In addition, reinforcement learning (RL) techniques under either model-based or model-free settings have been widely adopted, including Q-learning [105], deep neural networks [106], and actor-critic methods [107]. While model-free RL does not rely on the system knowledge, it often requires a large amount of data samples and exhibits slow learning rates for large-scale problems [108, 109]. Instead, the model-based RL uses either a known or estimated system model from measurement-based approaches, to generate off-line data using simulations to allow for fast policy search and better safety during online control. This trade-off motivates us to develop model-based RL for the WADC problem as fast computation and online safety are very important therein.

This work designs a risk-aware RL-based approach to solve the WADC problem, by explicitly addressing communication delays in WAMS. We linearize the system dynamics with both VSCs and fourth-order SG models around the operating point, as a linearized small-signal regime is sufficient for the WADC design. This way, we can formulate the LQR-based optimal control problem with structured feedback per the communication network's connectivity. Our analysis suggests that larger communication delays introduce higher perturbations to the system model, and inspires us to put forth the so-termed mean-variance risk constraint that can bound the large variability of state cost as a result of delays. To solve this constrained problem, we reformulate it to the dual maximin problem and develop a stochastic gradient-descent with max-oracle (SGDmax) to approach its stationary point. This model-based RL method will use the zero-order policy gradient (ZOPG) to simplify the gradient estimation with guaranteed convergence at a high probability. To sum up, the main contributions of our work are two-fold. First, we have developed a new system model that integrates VSCs into the fourth-order SG dynamics that are needed for the WADC problem. Second, and more importantly, we have proposed to use the mean-variance risk measure to systematically address the impacts of non-negligible delays in terms of increased worst-case damping level. Our risk-constrained WADC design can verifiably enhance the performance under large delays, thereby greatly improving the overall stability of power systems.

5.2 System Modeling

We consider a power system partitioned into N_a areas as shown in Fig. 5.1, which consists of N_g synchronous generators (SGs), as indexed by the set $\mathcal{G} = \{1, 2, \ldots, N_g\}$. We assume that all SGs are equipped with phasor measurements

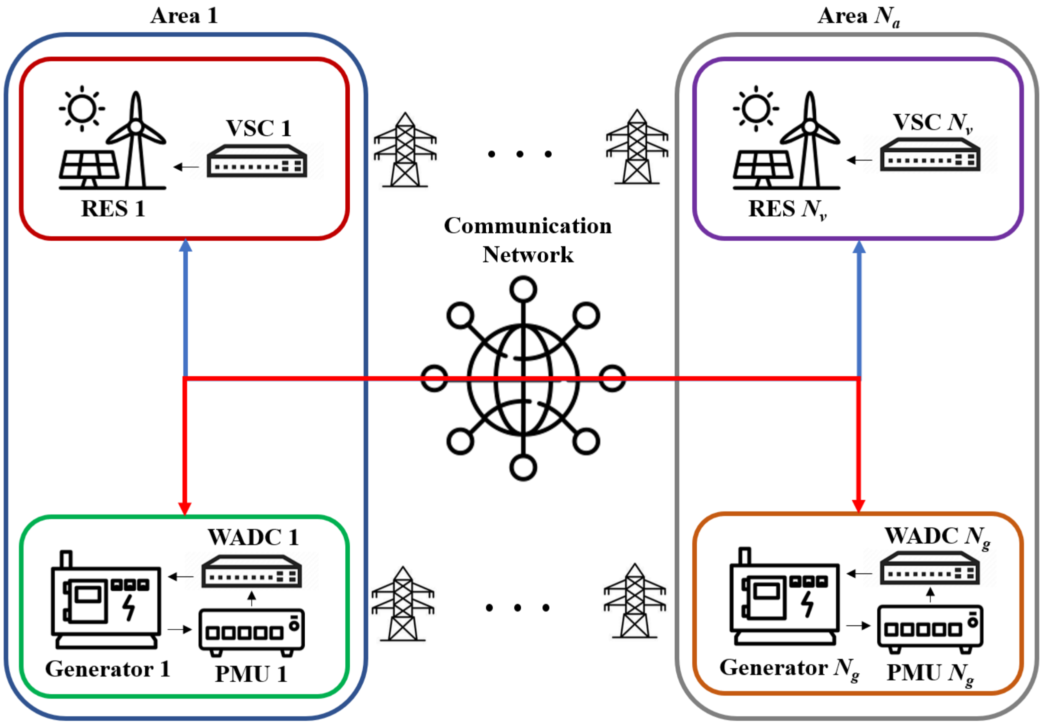


Figure 5.1: System depiction with synchronous generators (SGs) and voltage source converters (VSCs) participating in WADC.

units (PMUs) and wide-area damping controllers (WADC). Only a subset of SGs participating in WADC is also possible, by eliminating certain SGs from the feedback design. In addition, the state of SG i can be measured by its local PMU, denoted by $\mathbf{x}_i = [\delta_i, \omega_i, E_i, E_i^{fd}]^\intercal$. The electro-mechanical (EM) states δ_i and ω_i denote the deviation of internal rotor angle and speed from the operating point, respectively; while the non-EM states, E_i and E_i^{fd}, indicate the generator internal voltage and excitation voltage, respectively [101].

The fourth-order dynamics for SG $i \in \mathcal{G}$ is represented by:

$$\dot{\delta}_i = \omega_i, \tag{5.1a}$$

$$\dot{\omega}_i = \frac{1}{2H_i}\left[P_i^m - P_i^e - D_i\omega_i\right], \tag{5.1b}$$

$$\dot{E}_i = \frac{1}{T_i^{d\prime}}\left[-\frac{x_i^d}{x_i^{d\prime}}E_i + (x_i^d - x_i^{d\prime})I_i^d + E_i^{fd}\right], \tag{5.1c}$$

$$\dot{E}_i^{fd} = \frac{1}{T_i^a}[-E_i^{fd} - K_i^a(E_i - x_i^{d\prime}I_i^d - \bar{V}_i - \Delta\bar{V}_i)]. \tag{5.1d}$$

The mechanical power input P_i^m and reference voltage $\bar{V}_i$ are considered fixed as they are determined by slower operations than WADC. In addition, the electric power output P_i^e and d-axis current I_i^d are algebraic variables depending on the full network nonlinear power flow, which will be linearized later on. Notably, the damping control signal $\Delta\bar{V}_i$ in (5.1d) can quickly adjust the excitation voltage E_i^{fd} and affect other SG states, in order to improve the damping performance. The other terms like H_i and D_i are constant parameters given by generator specifications. Along with SGs, the system also has N_v voltage source converters (VSCs) indexed by the set $\mathcal{V} = \{N_g + 1, N_g + 2, \ldots, N_g + N_v\}$ that can participate in WADC. The VSCs can be modeled as power sinks and sources that can quickly provide supplementary active and reactive power adjustments to their steady-state references [110]. For each VSC $j \in \mathcal{V}$, we control its power injection adjustments ΔP_j^v and ΔQ_j^v.

To formulate the overall network dynamics, we need to consider the power flow coupling between the SG internal nodes and VSC buses. This allows us to express P_i^e and I_i^d in (5.1) as functions of the full system states and control inputs.

By applying Kron reduction [111] and eliminating all other buses, we consider the power flow equations between all SG interval voltages $\{E_i, \delta_i\}_{i\in\mathcal{G}}$ and all

VSC terminal bus voltages $\{V_j, \theta_j\}_{j \in \mathcal{V}}$.

For each SG $i \in \mathcal{G}$, the active and reactive power outputs are given by

$$\begin{aligned} P_i^e &= \textstyle\sum_{\ell=1}^{N_g} E_i E_\ell (G_{i\ell} \cos(\delta_i - \delta_\ell) + B_{i\ell} \sin(\delta_i - \delta_\ell)) \\ &\quad + \textstyle\sum_{\ell=N_g+1}^{N_g+N_v} E_i V_\ell (G_{i\ell} \cos(\delta_i - \theta_\ell) + B_{i\ell} \sin(\delta_i - \theta_\ell)), \\ Q_i^e &= \textstyle\sum_{\ell=1}^{N_g} E_i E_\ell (G_{i\ell} \sin(\delta_i - \delta_\ell) - B_{i\ell} \cos(\delta_i - \delta_\ell)) \\ &\quad + \textstyle\sum_{\ell=N_g+1}^{N_g+N_v} E_i V_\ell (G_{i\ell} \sin(\delta_i - \theta_\ell) - B_{i\ell} \cos(\delta_i - \theta_\ell)). \end{aligned}$$

Similar equations can be written for each VSC node $j \in \mathcal{V}$, namely P_j^v and Q_j^v.

To match with the SG dynamics in (5.1), we should also consider the d-axis current flow $I_i^d = Q_i^e / E_i$ using the reactive power output, namely

$$\begin{aligned} I_i^d &= \textstyle\sum_{\ell=1}^{N_g} E_\ell (G_{i\ell} \sin(\delta_i - \delta_\ell) - B_{i\ell} \cos(\delta_i - \delta_\ell)) \\ &\quad + \textstyle\sum_{\ell=N_g+1}^{N_g+N_v} V_\ell (G_{i\ell} \sin(\delta_i - \theta_\ell) - B_{i\ell} \cos(\delta_i - \theta_\ell)). \end{aligned}$$

To simplify this nonlinear network power flow, these equations will be linearized around the steady-state operating point. A linearized model could well capture the WADC dynamics [102], and we will eventually test the proposed design on the actual nonlinear system model. By using the bold symbols to concatenate all variables into the vector form, we have

$$\begin{bmatrix} \Delta \mathbf{P}^e \\ \Delta \mathbf{I}^d \\ \Delta \mathbf{P}^v \\ \Delta \mathbf{Q}^v \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathbf{P}^e}{\partial \boldsymbol{\delta}} & \frac{\partial \mathbf{P}^e}{\partial \mathbf{E}} & \frac{\partial \mathbf{P}^e}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{P}^e}{\partial \mathbf{V}} \\ \frac{\partial \mathbf{I}^q}{\partial \boldsymbol{\delta}} & \frac{\partial \mathbf{I}^q}{\partial \mathbf{E}} & \frac{\partial \mathbf{I}^q}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{I}^q}{\partial \mathbf{V}} \\ \frac{\partial \mathbf{P}^v}{\partial \boldsymbol{\delta}} & \frac{\partial \mathbf{P}^v}{\partial \mathbf{E}} & \frac{\partial \mathbf{P}^v}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{P}^v}{\partial \mathbf{V}} \\ \frac{\partial \mathbf{Q}^v}{\partial \boldsymbol{\delta}} & \frac{\partial \mathbf{Q}^v}{\partial \mathbf{E}} & \frac{\partial \mathbf{Q}^v}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{Q}^v}{\partial \mathbf{V}} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{\delta} \\ \Delta \mathbf{E} \\ \Delta \boldsymbol{\theta} \\ \Delta \mathbf{V} \end{bmatrix}, \tag{5.2}$$

where the partial derivatives form the Jacobian matrix. This way, we can represent the algebraic variables $\Delta \mathbf{P}^e$ and $\Delta \mathbf{I}^d$ as follows; see the derivations in the Appendix. C.

$$\Delta \mathbf{P}^e = \mathbf{A}_1^P \Delta \boldsymbol{\delta} + \mathbf{A}_2^P \Delta \mathbf{E} + \mathbf{A}_3^P \Delta \mathbf{P}^v + \mathbf{A}_4^P \Delta \mathbf{Q}^v \tag{5.3a}$$

$$\Delta \mathbf{I}^d = \mathbf{A}_1^I \Delta \boldsymbol{\delta} + \mathbf{A}_2^I \Delta \mathbf{E} + \mathbf{A}_3^I \Delta \mathbf{P}^v + \mathbf{A}_4^I \Delta \mathbf{Q}^v \tag{5.3b}$$

The linearized relation in (5.3) will allow to integrate all SGs' dynamics in (5.1) with the VSC power injections. For simplicity, our work does not consider the VSC's reactive power adjustment by fixing $\Delta \mathbf{Q}^v = 0$. This is because the reactive power component has much smaller impact on WADC than the active one, as discussed in [112, 113].

By substituting (5.3) into (5.1), the overall dynamics for $\mathbf{x} := \{\mathbf{x}_i\}_{i \in \mathcal{G}}$ with the full input $\mathbf{u} := [(\Delta \bar{\mathbf{V}})^\intercal \ (\Delta \mathbf{P}^v)^\intercal]^\intercal \in \mathbb{R}^{N_g + N_v}$ can be linearized as:

$$\dot{\mathbf{x}} = \mathbf{A}_c \mathbf{x} + \mathbf{B}_c \mathbf{u} + \boldsymbol{\xi} \tag{5.4}$$

where $\boldsymbol{\xi} \in \mathbb{R}^{4N_g}$ represents the random perturbations to system states from e.g., external disturbances and imperfect modeling, as detailed later on. Note that SG states in $\mathbf{x}$ are actually deviations from the corresponding steady-state values due to the linearization.

Last, we will consider the discrete-time dynamics based on (5.4), given by

$$\mathbf{x}_{t+1} = \mathbf{A} \mathbf{x}_t + \mathbf{B} \mathbf{u}_t + \boldsymbol{\xi}_t, \ \forall t = 0, 1, \ldots \tag{5.5}$$

which has been obtained with a sufficiently small time step.

Remark 4 (WADC dynamics with controllable VSCs)**.** *The key to establishing the overall system model* (5.4) *lies in integrating the VSC power injections with the SG dynamics in* (5.1)*. Different from earlier work [102] using the second-order swing equations for SG dynamics, our work instead employs the fourth-order SG dynamics which are more accurate and better connected to excitation control [114]. Nonetheless, the latter makes it more challenging to integrate the VSCs which can affect both* $\mathbf{P}^e$ *and* $\mathbf{I}^d$ *at the SG terminals. To this end, we have considered the static power flow coupling between the SG internal voltages and the VSC voltages, and used the linearization approach to simplify the coupling. Thus, we have improved the modeling accuracy for representing the VSC-integrated system dynamics and designing the WADC.*

5.3 Risk-constrained WADC Problem

Under the system dynamics in (5.5), the WADC problem is typically cast as a linear quadratic regulator (LQR) design problem to minimize the total cost of state and control input, given by

$$R_0(\mathbf{K}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}[\mathbf{x}_t^\intercal \mathbf{Q}\mathbf{x}_t + \mathbf{u}_t^\intercal \mathbf{R}\mathbf{u}_t] \tag{5.6}$$

where matrices $\{\mathbf{Q}, \mathbf{R}\}$ are given positive (semi-)definite matrices used to weight the corresponding elements. The decision variable here is the feedback gain $\mathbf{K} \in \mathbb{R}^{(N_g+N_v)\times(4N_g)}$ for a linear mapping between $\mathbf{x}_t$ and $\mathbf{u}_t$; i.e., $\mathbf{u}_t = -\mathbf{K}\mathbf{x}_t$.

A structured feedback $\mathbf{K} \in \mathcal{K}$ is very common in WADC due to limited deployment of communication links. Thus, for a given sparse communication graph between any SG $i \in \mathcal{G}$ and control node $\ell \in \mathcal{G} \cup \mathcal{V}$, the feasible set $\mathcal{K}$ becomes

$$\mathcal{K} = \{\mathbf{K} : \mathbf{K}_{\ell,i} = 0 \text{ if and only if } \ell \nleftrightarrow i)\}$$

where $\ell \nleftrightarrow i$ indicates no communication link available from SG i to control node ℓ. Note that $\mathbf{K}_{\ell,i} \in \mathbb{R}^{1\times 4}$ denotes the submatrix of $\mathbf{K}$ mapping from $\mathbf{x}_i$ to $\mathbf{u}_\ell$. While the sparse $\mathbf{K}$ makes it difficult to analyze the feasible region [84], it will not affect the implementation of our proposed gradient-based solutions as detailed later on.

Furthermore, the communication delays through the WAMS are a crucial factor affecting the performance of WADC [14, 96]. Due to the very fast timescale of WADC, typically at 0.01s level, the communication delay effects are more notable than other slower control designs. To model it, we consider that the measured $\mathbf{x}_{i,t}$ at SG i would experience a time-invariant delay h_i when reaching all other control nodes $\ell \neq i$. Per time t, let us denote this delayed state by $\tilde{\mathbf{x}}_{i,t} := \mathbf{x}_{i,t-h_i}$. This way, the local state vector available at each control node $\ell \in \mathcal{G} \cup \mathcal{V}$ becomes $\tilde{\mathbf{x}}_t^{(\ell)} = [\tilde{\mathbf{x}}_{1,t}^\intercal, \ldots, \mathbf{x}_{\ell,t}^\intercal, \ldots, \tilde{\mathbf{x}}_{N_g,t}^\intercal]^\intercal$ which uses all delayed states except for the local SG state if $\ell \in \mathcal{G}$. For simplicity, our model assumes a uniform communication delay for each SG's state. But we can generalize it to the heterogeneous delay setting with different (or even random) delay times for each link and utilize our risk-constrained WADC design to address these more realistic settings.

The communication delays are detrimental to maintaining the WADC performance, as they introduce additional uncertainty and perturbations to the system dynamics in (5.5). Intuitively, with larger delays, the delayed state $\tilde{\mathbf{x}}^{(\ell)}$ received at control node ℓ would incur a higher error difference from the actual state. As a result, the control input $\mathbf{u}_\ell$ formed by $\tilde{\mathbf{x}}^{(\ell)}$ would introduce an increasing perturbation to the system dynamics. Specifically, for the control $\mathbf{u}_{\ell,t}$ per time t, using the delayed state $\tilde{\mathbf{x}}_t^{(\ell)}$ would cause the perturbation as follows:

$$\beta_{\ell,t} = \sum_{i\in\mathcal{G}, i\neq\ell} \mathbf{K}_{\ell,i}(\tilde{\mathbf{x}}_{i,t} - \mathbf{x}_{i,t}). \tag{5.7}$$

In addition, the difference term for each SG's state $(\tilde{\mathbf{x}}_{i,t} - \mathbf{x}_{i,t})$ would accumulate more deviation terms with an increasing delay h_i. Effectively, we need to update the perturbation term in (5.5) to $\boldsymbol{\xi}'_t$ which captures both the original random noise $\boldsymbol{\xi}_t$ and the additional control perturbation caused by communication delays. This system perturbation due to large delays and thus non-negligible $\beta_{\ell,t}$ would increase the level of variability in the system trajectory for (5.5), greatly challenging the LQR-based WADC design which only focuses on the average trajectory performance.

To address this delay-induced perturbation, we propose a risk-constrained LQR formulation for the WADC problem by limiting the mean-variance risk of the state deviation, as

$$\begin{aligned} \min_{\mathbf{K}\in\mathcal{K}} \; R_0(\mathbf{K}) &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}[\mathbf{x}_t^\intercal \mathbf{Q}\mathbf{x}_t + \mathbf{u}_t^\intercal \mathbf{R}\mathbf{u}_t] \\ \text{s.t. } R_c(\mathbf{K}) &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}\left(\mathbf{x}_t^\intercal \mathbf{Q}\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t^\intercal \mathbf{Q}\mathbf{x}_t|\mathcal{F}_t]\right)^2 \leq c, \end{aligned} \tag{5.8}$$

where $\mathcal{F}_t := \{\mathbf{x}_0, \mathbf{u}_0, \ldots, \mathbf{x}_{t-1}, \mathbf{u}_{t-1}\}$ has the system trajectory up to time t, while the scalar c is a risk tolerance parameter. This mean-variance risk constraint aims to reduce the average deviation of the state cost term $(\mathbf{x}_t^\intercal \mathbf{Q}\mathbf{x}_t)$ from its expected value conditioned on the past data, thus mitigating the high system variability due to external perturbations. The benefit of using it for the WADC problem is two-fold, as discussed in the following remark.

Remark 5 (Risk-constrained WADC)**.** *The mean-variance risk constraint can address the increased system perturbation and thus state variability due to the large communication delays in fast WADC problems. Specifically, constraining the* $R_c(\mathbf{K})$ *risk in* (5.8) *effectively limits the variability of state-related cost term from its expected value conditioned on past data, and thus can reduce its worst-case cost. This is very important for the resultant WADC designs to meet the safety operations limits in power system dynamics.*

The risk-constrained optimization problem in (5.8) is solved by utilizing SGDmax with ZOPG in Algorithm 5 and 6, as introduced in Section 4.5.

5.4 Numerical Tests

To demonstrate the effectiveness of our proposed risk-constrained WADC, we have conducted numerical tests on the IEEE 68-bus system [115]. This system is a simplified model of the interconnection between the New York and New England power grids, consisting of five areas with a total of 16 SGs. Each SG is equipped with a WADC controller and a PMU meter. We also add three VSCs at buses 20, 42, and 54, similar to [102]. Based on the area partition in Fig. 5.2, the information is exchanged only between neighboring areas. For example, Area 2 can communicate with all other areas, while Area 1 can only exchange data with Area 2 but cannot

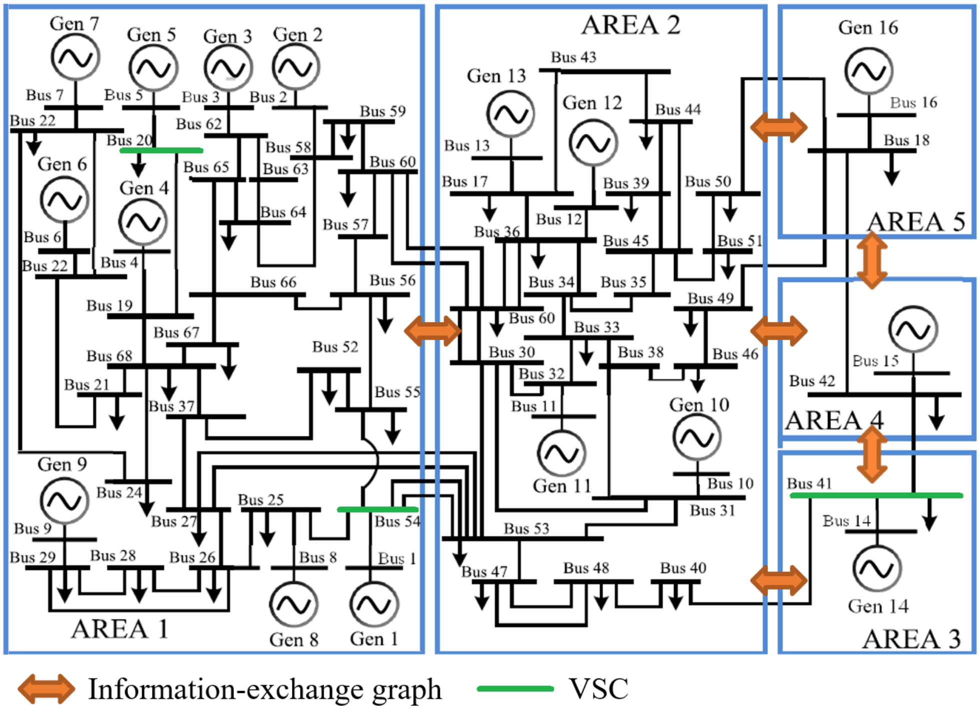


Figure 5.2: IEEE 68-bus system with 16 SGs and 3 VSCs divided into 5 areas.

access to the state measurements in Areas 3, 4, and 5. This remains us to consider the structured feedback following the communication graph.

We have considered four types of WADC designs: SG, SG-Risk, VSC, and VSC-Risk, as listed in Table 5.1. The SG and SG-Risk methods include SGs only for the WADC, while VSC and VSC-Risk have both SGs and VSCs participating in WADC. The risk-constraint has been considered in SG-Risk and VSC-Risk, with the other two developed using the unconstrained LQR cost. For each WADC design, we have used Algorithm 6 with the ZOPG estimated by Algorithm 5 to find the converged feedback gain. We set the parameters as $r = 0.1$, $M = 100$, $\eta = 10^{-4}$

Table 5.1: The four WADC designs considered in numerical tests

	WADC	**Risk constraint**
SG	SGs	Unconstrained
SG-Risk	SGs	Risk-constrained
VSC	SGs, VSCs	Unconstrained
VSC-Risk	SGs, VSCs	Risk-constrained

and risk tolerance $c = 0.5$. Additionally, we set both $\mathbf{Q}$ and $\mathbf{R}$ as identity matrices for the LQR objective in (5.8). The time step for both sensing and control is set to $\Delta t = 0.01$ s. To train the WADC policy for each design, we generate random impulse inputs to each generator at $t = 0$ for each scenario and observe the response for a 20-second time window. The training phase does not consider communication delays which will be investigated during the testing phase.

We first present the training results to verify the convergence of the proposed Algorithm 6. Fig. 5.3 plots the log-scale objective trajectories for the four WADC designs, with the no-WADC cost objective as the baseline. Convergence has been observed for all trajectories, outperforming the baseline. Notably, SG-Risk and VSC-Risk exhibit higher fluctuation than SG and VSC. This is because the risk constraint would complicate the feasible region, and thus the search of optimal λ leads to some oscillations in the trajectory.

The rest of simulation results present the testing performance comparisons for the converged WADC policies obtained by training. In the testing phase, new scenarios with different input disturbances are generated, along with communication delays. As discussed in Section 5.3, we have randomly generated the time-invariant link-specific delays from the uniform distribution with the maximum bound

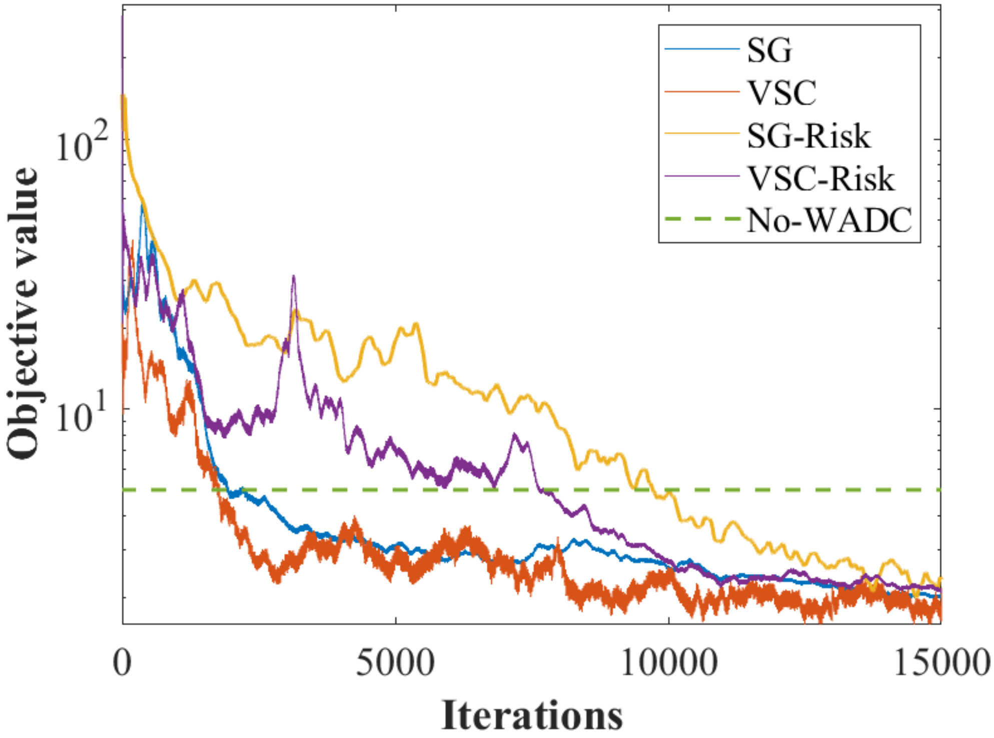


Figure 5.3: Trajectory of the training objective value for different WADC designs.

of 0.02s, 0.06s, and 0.10s, respectively. To illustrate the impact of communication delays, we select one specific scenario for each delay and plot the actual frequency deviation when using the VSC-Risk design, as depicted in Fig. 5.4. Each testing scenario has been selected to have the highest frequency deviation under each delay setting. Clearly, the WADC performance degrades gradually with increasing delays, taking more time to damp the oscillations. However, the damping performance seems to be pretty reasonable even for the highest delay setting, thanks to the risk-constrained design.

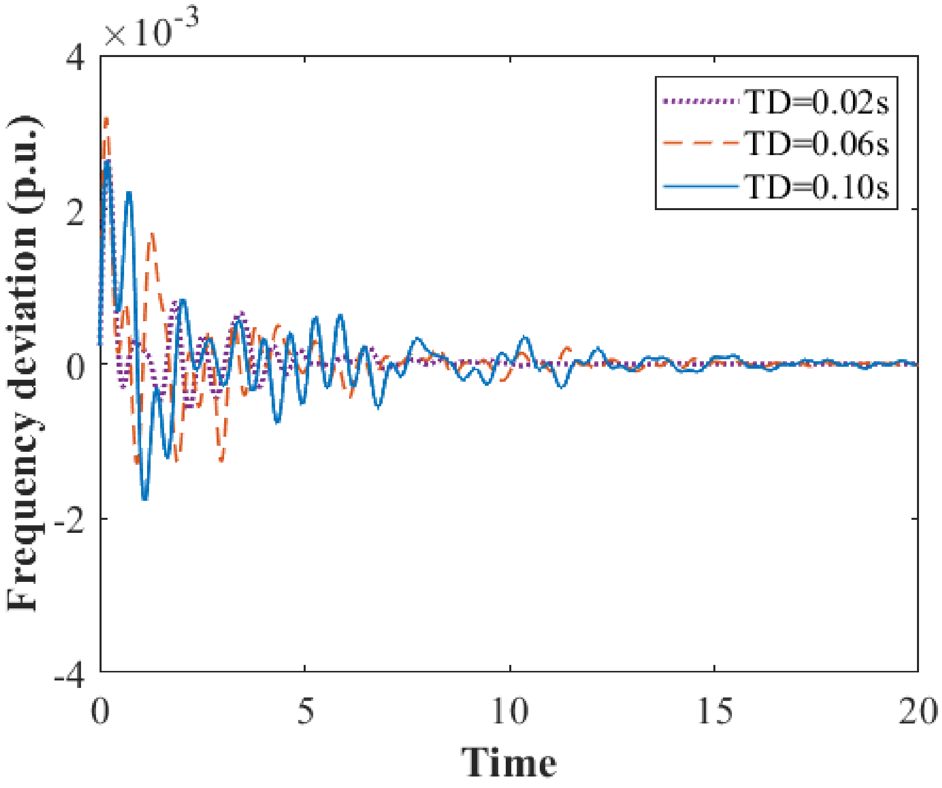


Figure 5.4: Frequency deviation of VSC-Risk for different maximum delays.

To better demonstrate the effectiveness of integrating the risk constraint and VSCs into WADC, we compare the frequency deviation of the scenario with the highest frequency deviation under the setting of a maximum 0.10s delay. First, Fig. 5.5 shows a frequency deviation comparison between VSC and VSC-Risk to demonstrate the effects of risk constraint. It has been observed that VSC-Risk has more damping capability and reaches steady-state faster than VSC. This corroborates the usefulness of risk constraint in mitigating communication delays and maintaining the WADC performance. Second, Fig. 5.6 compares SG-Risk and VSC-Risk to showcase the improvement of using VSCs for WADC. We can observe that the

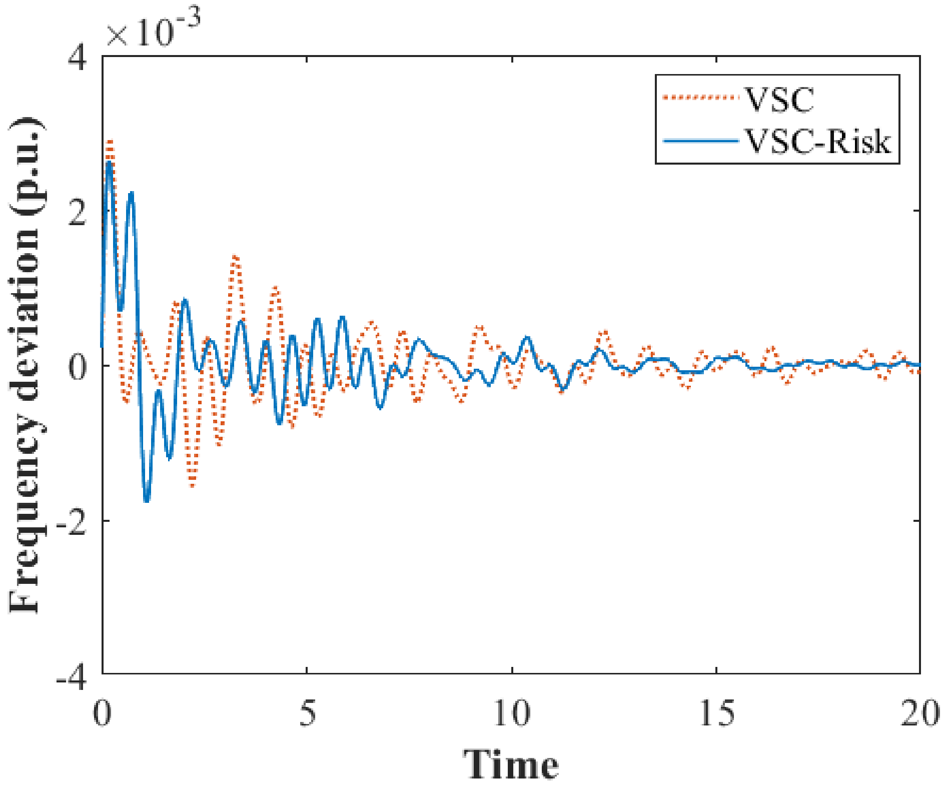


Figure 5.5: Comparison on the frequency deviation between VSC and VSC-Risk.

VSCs have provided additional actuation capabilities, leading to better damping performance.

To provide quantitative results for corroborating the risk constraint, we presents Fig. 5.7-5.9 that demonstrate the statistical information of the LQR objective values based on 100 testing scenarios. Fig. 5.7, Fig. 5.8 and Fig. 5.9 are the box plots for all WADC designs with the median value, lower/upper quartiles and minimum/maximum, for each delay setting. In general, the risk-constrained designs, both SG-Risk and VSC-Risk, have slightly increased the objective values on av-

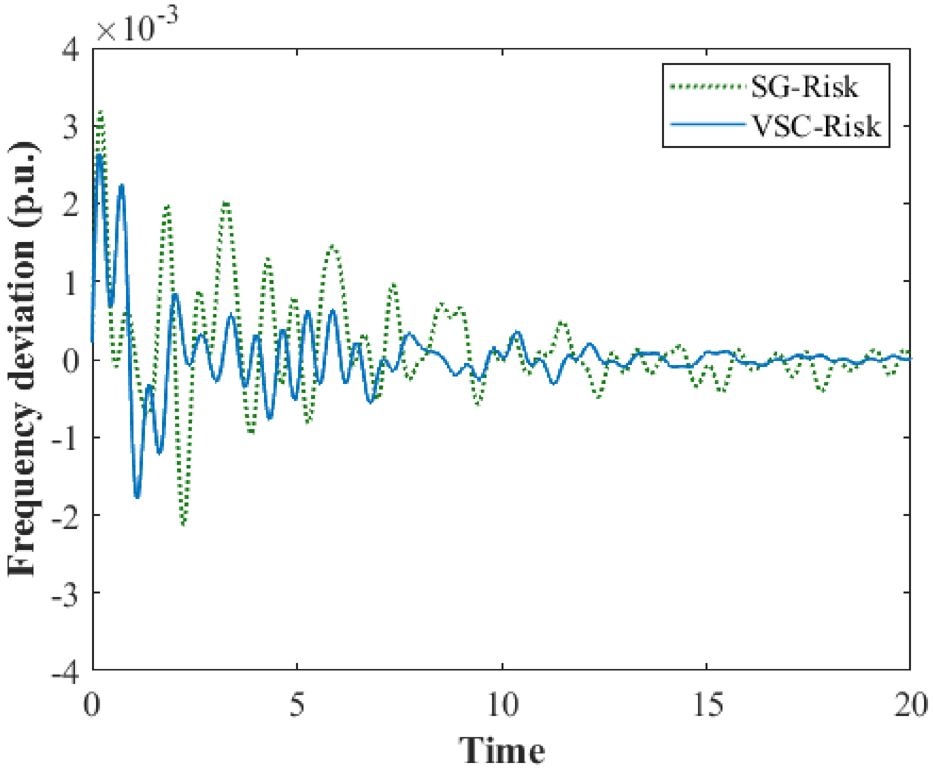


Figure 5.6: Comparison on the frequency deviation between SG-Risk and VSC-Risk.

erage, yet significantly reducing the variance and also the maximum of objective values. This result illustrates the effectiveness of using the risk constraint in mitigating the worst-case performance, thereby increasing the stability margin of power system operations. This comparison also verifies the improvements provided by the additional VSC resources, as VSC and VSC-Risk respectively outperforms SG and SG-Risk.

To further highlight the benefits of considering the risk constraint, we compare the state cost of VSC and VSC-Risk with an increasing delays, as shown in

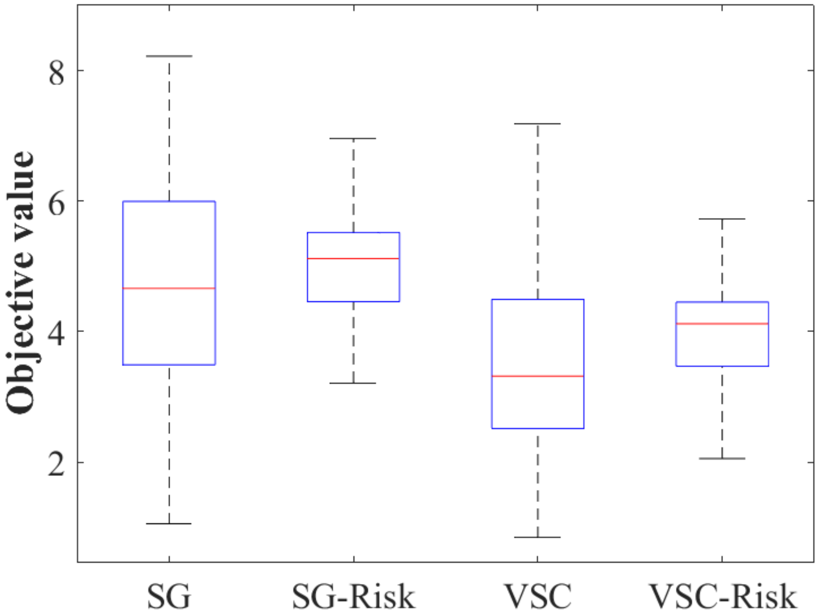


Figure 5.7: Objective values when the maximum delays is 0.02s.

Fig. 5.10. The solid line represents the average value across 100 scenarios for each delay setting, while the blue and orange shaded areas indicate the state cost variations of VSC and VSC-Risk, respectively. It is evident that incorporating the risk constraint helps to mitigate the rise in the state cost at high delays, while effectively reducing its variability with a smaller shaded area than the unconstrained ones. Therefore, the risk constraint can effectively enhance the robustness of WADC design in the presence of increasing delays in WAMS.

Last, we investigate the impact of choosing the risk tolerance parameter c with the highest delay setting. Fig. 5.11 shows the box plots for the objective values and state costs for three different levels of c, with the same maximum delays of

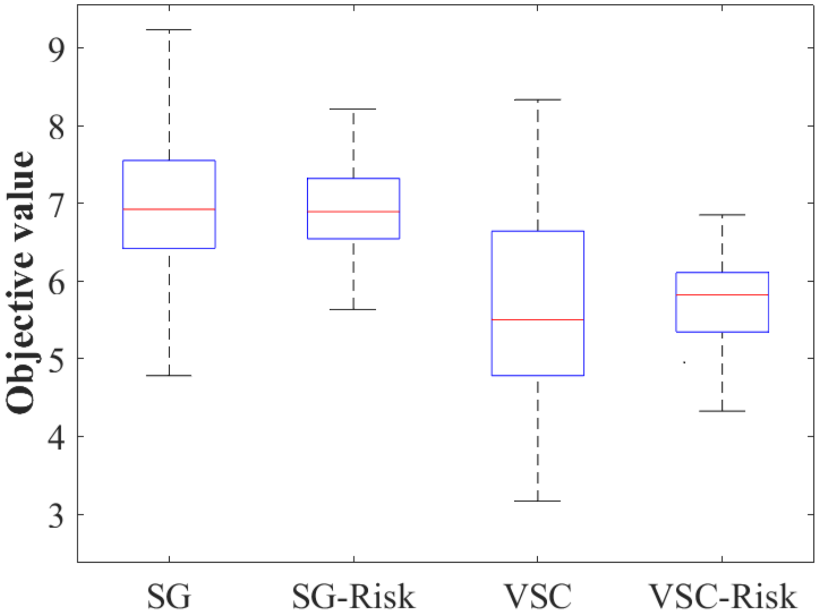


Figure 5.8: Objective values when the maximum delays is 0.06s.

0.10s. In Fig. 5.11(a), a smaller value of c, which further limits the mean-variance risk, leads to a smaller objective value but higher variance. However, Fig 5.11(b) shows that both the state cost and the variance decrease with the c value. This implies that further reducing the risk level in WADC could improve the performance in the state deviation including frequency deviation, both in terms of the average value or the variance. However, this improvement may increase the overall LQR cost at the price of needing additional control efforts.

To sum up, our numerical tests have validated the convergence of the proposed algorithm in solving the risk-constrained problem in (5.8). Integration of VSCs has shown the improved damping performance by providing additional ac-

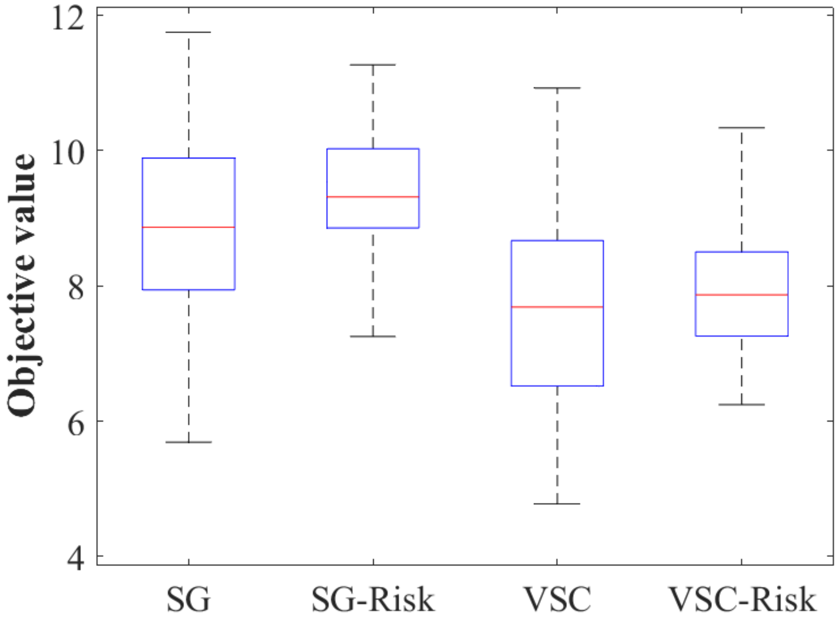


Figure 5.9: Objective values when the maximum delays is 0.10s.

tuation capabilities. Most importantly, the effectiveness of risk-constraint has been verified in the testing, especially in mitigating the worst-case performance while improving the system stability.

This work designed a risk-constrained WADC approach that aims to address the communication delays. Based on a linearized system model that incorporates both SGs and VSCs to provide damping capabilities, we cast it to minimize the LQR objective over the structured feedback gain matrix according to the communication network's connectivity. Our analysis suggested that the level of system perturbations grows with larger communication delays, leading to higher state variability. Thereby, we introduced the mean-variance risk constraint to bound the variation

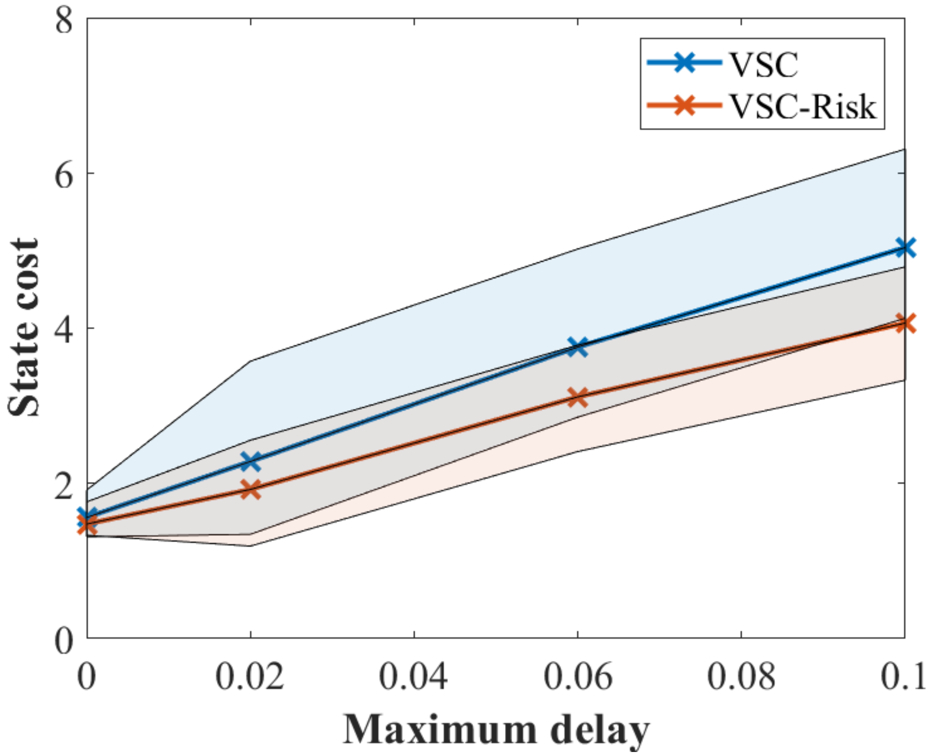
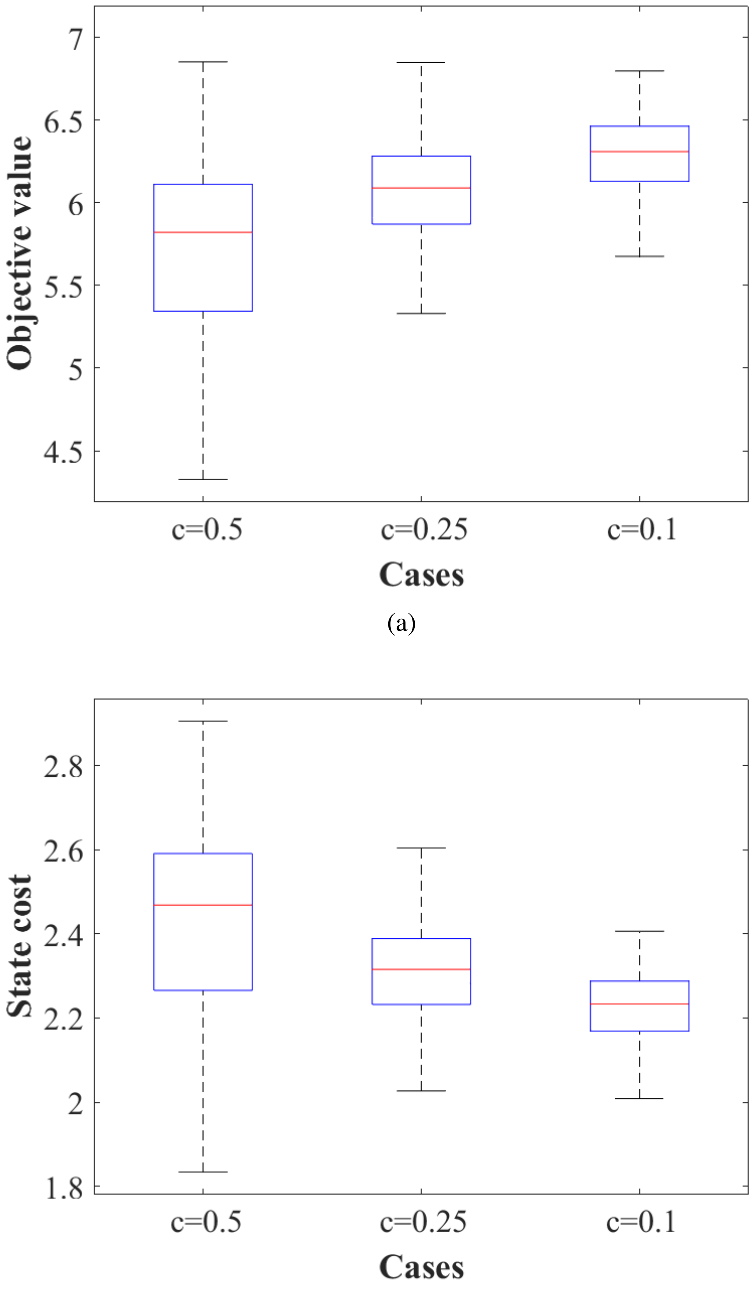


Figure 5.10: State costs of VSC and VSC-Risk.

of the state cost, in order to reduce the delay-induced variability and improve the worst-case performance. By reformulating the constraint into a tractable quadratic form, we solve the dual maximin problem by developing a RL-based SGDmax algorithm. The latter works by iteratively searching for a policy using efficient ZOPG-based gradient updates, which can provably attain the SP with high probability. Numerical tests on the IEEE 68-bus system demonstrated the effectiveness and advantages of the proposed risk-aware WADC design in reducing the variability of the total LQR cost, thereby improving the stability performance especially in worst-case scenarios of oscillations and delays. Future research directions include expanding the types of risk measures to e.g., conditional value at risk (CVaR), investigating large-scale implementations of our proposed design, as well as considering generalized grid control tasks for new resources and renewables.



(a)

(b)

Figure 5.11: (a) Objective values and (b) state costs for different risk tolerance parameters.

Chapter 6

RL-based Grid-forming Inverter Control

This chapter presents a risk-aware grid-forming inverter controller considering high load perturbations. Section 6.1 introduces a grid-forming inverter (GFM) problem and the motivation of this work. Section 6.2 formulates the system model considering the dynamics of both synchronous generators (SGs) and GFMs, which are combined through network coupling. In Section 6.3, we design a risk-constrained GFM problem with a mean-variance risk constraint to mitigate frequency oscillations from high load perturbations. Section 6.4 showcases the numerical tests result using the modified IEEE 68-bus system to demonstrate the impact of considering the risk constraint.

This chapter is based on the following submissions and preparations:

K. Kwon, S. Mukherjee, T. L. Vu, and H. Zhu, "Risk-Constrained Reinforcement Learning Control of Inverter-Dominated Power Systems." submitted to *IEEE Control System Letters*, 2024. (Under revision)

K. Kwon, S. Mukherjee, T. L. Vu and H. Zhu, "Risk-Constrained Reinforcement Learning for Inverter-Dominated Power System Controls Under Modeling Errors." to be submitted to *IEEE Transactions on Smart Grid*, 2023.

6.1 Grid-forming Inverter Control Problem

Grid-forming inverters (GFMs) are increasingly important for establishing grid voltage and frequency in next-generation power systems with high penetration of low-carbon energy resources [116]. Photovoltaics, wind generators, and energy storage devices lack in conventional primary and secondary controls as synchronous generators (SGs), and thus their integration greatly challenges grid stability. Advanced GFM technology can address this issue as they operate as independent voltage sources to support grid stability by controlling the voltage and frequency at the interfaces of new resources [9].

The existing GFM control strategies consist of three main categories: droop control [117–119], virtual synchronous generators [120, 121], and virtual oscillator control [122]. Especially, droop control is a well-established method to mitigate voltage and frequency fluctuations by following P-ω and Q-V droop curves. By observing active and reactive powers from the network as inputs, it can vary the terminal frequency and voltage depending on the internal voltage/power set-points [123]. Thus, changing these set-points can affect the overall grid dynamics to quickly attain the steady-state operations after huge external perturbations due to, i.e., sudden changes of the load/renewable. As a local control design, it is known that the overall performance of multiple droop-controllers could degrade for reducing inter-area oscillations in large-scale interconnection [117]. A decentralized control design among all GFMs can address this issue, with state information exchange among GFMs as enabled by communication network [124]. The number of communication links are typically limited, and thus a structured feedback design per the information-exchange graph among GFMs will be adopted later on.

Recent advances in data-driven methods, including both model-based and model-free ones, have provided significant advantages for solving optimal control problems. To design the decentralized GFM controller, there have been several data-driven techniques based on reinforcement learning (RL) [72,125], adaptive dynamic programming [101], regression trees [126] and neural networks [127, 128]. While these model-free approaches do not require to know the system's mathematical model, they are known to suffer from the sample complexity issue which needs extensive data samples and large training time [109]. Thus, we develop a model-based approach by simulating the underlying system dynamics, which will greatly accelerate the policy search in practice.

However, the effectiveness of a policy developed through a model-based approach may diminish when confronted with an inaccurate system model due to parameter mismatches between the model and the actual system [129]. Since the model relies on estimated parameters of the components e.g. SGs and GFMs, any disparities in these model parameters can introduce errors in the system dynamics, leading to increased variability in the system trajectory. Therefore, relying solely on a linear quadratic regulator (LQR) objective to minimize expected state deviations and control costs in the GFM control problem can result in significant performance degradation, especially in the worst-case oscillations. While there is some research that tackles parametric mismatch in GFM control [15, 16, 130], most of the work is limited to the model predictive control (MPC) approach. As MPC depends on a predefined model, its performance can deteriorate as model uncertainty increases. This motivates us to adopt an RL-based method since RL can adapt to model uncertainty and enhance control performance by learning from data.

To this end, we develop a risk-aware RL approach for the GFM control design problem while aiming to address model parameter mismatch when employing a model-based approach. We formulate it as a constrained LQR problem with the so-termed mean-variance risk constraint. The latter is imposed on the overall deviations of state cost from its expectation, which can reduce the level of high system variability as a result of significant disturbances to enhance the worst-case performance. To solve this problem, we implement an RL-based algorithm termed as stochastic gradient-descent with max-oracle (SGDmax), which utilizes zero-order policy gradient (ZOPG) as estimated gradients for reduced computational complexity.

Our main contributions are three-fold. First, we represent the GFM control problem as a risk-constrained LQR problem by developing a linearized system model that incorporates both SGs and GFMs. Second, our risk-aware GFM controller incorporates the mean-variance risk constraint and solves it using the RL-based SGDmax algorithm. Last, we demonstrate the effectiveness of the proposed method through numerical tests in the presence of model parameter mismatch. Most importantly, we validate that introducing the risk constraint can reduce the variability of frequency deviations, thereby improving the worst-case performance.

6.2 System Modeling

We consider a power system consisting of N_a areas with a total of N_g synchronous generators (SGs) and N_f grid-forming inverters (GFMs), as illustrated in Fig. 6.1. We denote SGs and GFMs as indexed by the sets $\mathcal{G} = \{1, 2, \ldots, N_g\}$ and $\mathcal{F} = \{N_g + 1, N_g + 2, \ldots, N_g + N_f\}$, respectively. Without loss of generality (Wlog), every load bus is assumed to be connected to one GFM, as other load buses can be reduced.

To model the overall system, we first present the dynamics of each SG and GFM. The state of SG $i \in \mathcal{G}$ is represented as $\mathbf{x}_i = [\delta_i, \omega_i, E_i, E_i^{fd}]$. The electromechanical (EM) states δ_i and ω_i indicate the deviation of the internal rotor angel and speed from the operating point, while the non-EM states E_i and E_i^{fd} denote the generator internal and excitation voltage, respectively. The dynamics of SG $i \in \mathcal{G}$ are described by the following the fourth-order model:

$$\dot{\delta}_i = \omega_i - \omega_0, \tag{6.1a}$$

$$\dot{\omega}_i = \frac{1}{M_i}\big[D_i(\omega_0 - \omega_i) + P_i - P_i^n\big] \tag{6.1b}$$

$$\dot{E}_i = \frac{1}{\tau_i^{d'}}\left[-\frac{x_i^d}{x_i^{d'}}E_i + (x_i^d - x_i^{d'})I_i^d + E_i^{fd}\right], \tag{6.1c}$$

$$\dot{E}_i^{fd} = \frac{1}{\tau_i^a}[-E_i^{fd} - K_i^a(E_i - x_i^{d'}I_i^d - E_i^s)]. \tag{6.1d}$$

Note that the parameters $\{M_i, D_i\}$ represent the inertia and damping coefficients, while $\{\tau_i^{d'}, \tau_i^a, x_i^d, x_i^{d'}, K_i^a\}$ are fixed parameters of the excitation component. In addition, the power delivered to the network P_i^n and the d-axis current I_i^d are algebraic variables that depend on the nonlinear power flow, which will be linearized

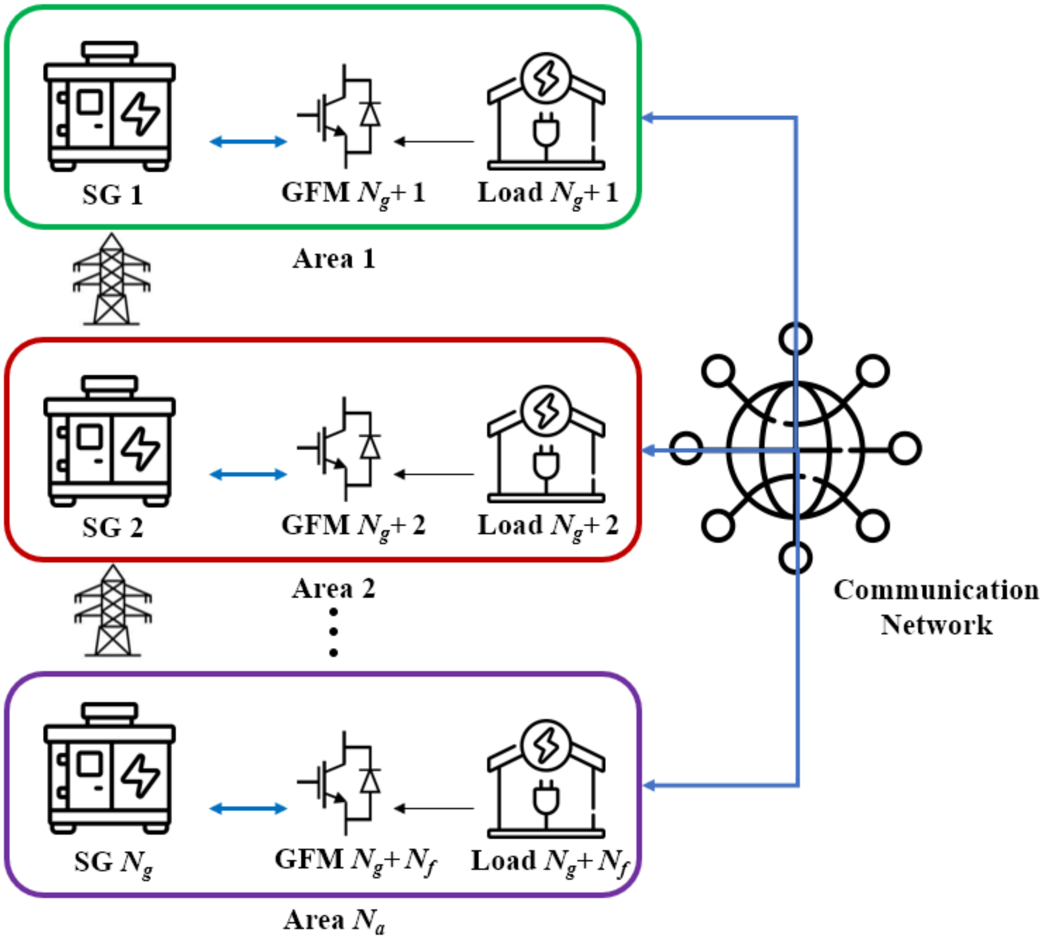


Figure 6.1: An illustration of a power system with synchronous generators (SGs) and grid-forming inverters (GFMs).

shortly. By controlling the damping control signal $\mathbf{u}_i := [E_i^s]$ in (6.1d), we adjust E_i^{fd}, which improves the damping performance by affecting the other SG states.

Each of the N_f GFMs acts as a controllable voltage source at the DER-connected load bus [118]. As illustrated in Fig. 6.2, the internal dynamics utilizes the P-ω and Q-V droop control curves as depicted in Fig. 6.3. The voltage and

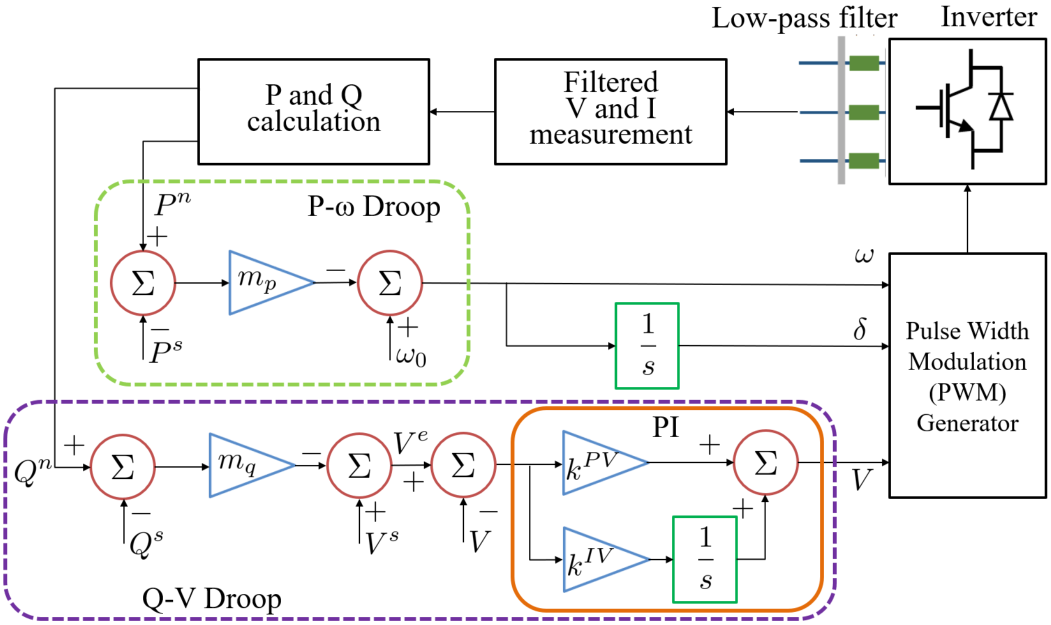


Figure 6.2: GFM dynamics based on droop controls.

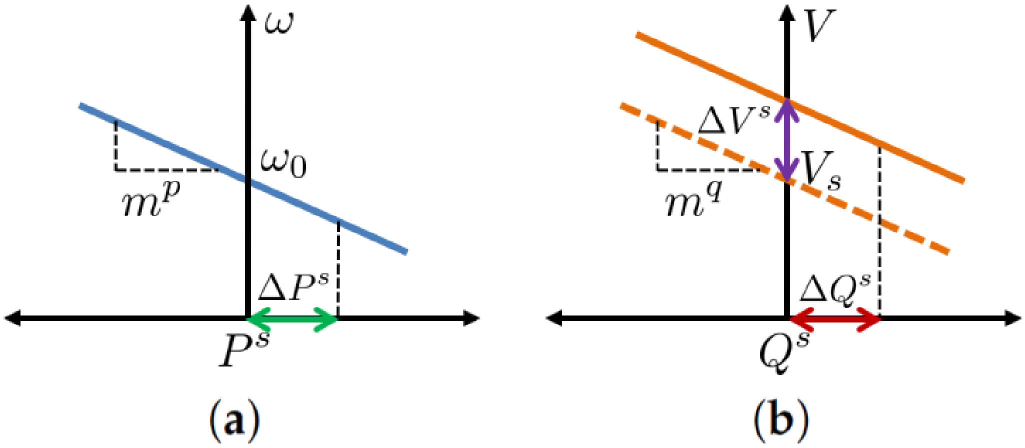


Figure 6.3: (a) P-ω and (b) Q-V droop characteristics.

current measurements of a GFM node go through a low-pass filter to eliminate possible oscillations due to harmonics or measurement error. Per GFM $j \in \mathcal{F}$, by calculating the active and reactive powers delivered by the network, namely P_j^n and Q_j^n, using the terminal voltage/current measurements, each droop controller uses the difference from the corresponding power set-point to determine the actuation signal. For example, the difference between P_j^n and the active power set-point P_j^s is multiplied by the droop gain m_j^p to determine the signal ω_j; and similarly for the voltage error signal V_j^e. Note that the Q-V droop is also followed by a proportional-integral (PI) controller to further regulate the deviations of the voltage error V_f^e, with k_j^{pv} and k_j^{iv} as the proportional and integral gains, respectively. Last, ω, δ and V are sent to the pulse width modulation (PWM) generator and the frequency and voltage of the node are set accordingly by the inverter. Thus, the dynamic model of GFM $j \in \mathcal{F}$ can be expressed as follows:

$$\dot{\theta}_j = \omega_j - \omega_0, \tag{6.2a}$$

$$\dot{\omega}_j = \frac{1}{\tau_j}\big[\omega_0 - \omega_j + m_j^p(P_j^s - P_j^n)\big], \tag{6.2b}$$

$$\dot{V}_j^e = \frac{1}{\tau_j}\big[V_j^s - V_j^e - V_j + m_j^q(Q_j^s - Q_j^n)\big], \tag{6.2c}$$

$$\dot{V}_j = k_j^{pv}\dot{V}_j^e + k_j^{iv}V_j^e \tag{6.2d}$$

where τ_j is a pre-determined droop time constant. The state vector per GFM j becomes $\mathbf{x}_j := [\theta_j, \omega_j, V_j^e, V_j]$. Thus, the GFM works by controlling its V_j and ω_j via adjusting the voltage and power set-points, which are included by the vector $\mathbf{u}_j := [V_j^s, P_j^s.Q_j^s]$.

Based on (6.1) and (6.2), the dynamics of SGs and GFMs are coupled through the network power flow (PF), which determines $\{P_\ell^n\}$ and $\{Q_\ell^n\}$ for $\ell \in$

$\mathcal{G} \cup \mathcal{F}$. As a result, we can establish the overall system dynamics through steady-state PF analysis. By employing Kron reduction [111], we initially eliminate all other buses, leaving only the SG and GFM buses. This allows us to consider the PF among $\{E_i, \delta_i\}_{i\in\mathcal{G}}$ and $\{V_j, \theta_j\}_{j\in\mathcal{F}}$ as follows:

$$\begin{aligned}
P_i^n &= \textstyle\sum_{\ell=1}^{N_g} E_i E_\ell (G_{i\ell}\cos(\delta_i - \delta_\ell) + B_{i\ell}\sin(\delta_i - \delta_\ell)) \\
&\quad + \textstyle\sum_{\ell=N_g+1}^{N_g+N_f} E_i V_\ell (G_{i\ell}\cos(\delta_i - \theta_\ell) + B_{i\ell}\sin(\delta_i - \theta_\ell)), \\
Q_i^n &= \textstyle\sum_{\ell=1}^{N_g} E_i E_\ell (G_{i\ell}\sin(\delta_i - \delta_\ell) - B_{i\ell}\cos(\delta_i - \delta_\ell)) \\
&\quad + \textstyle\sum_{\ell=N_g+1}^{N_g+N_f} E_i V_\ell (G_{i\ell}\sin(\delta_i - \theta_\ell) - B_{i\ell}\cos(\delta_i - \theta_\ell)), \\
P_j^n &= \textstyle\sum_{\ell=1}^{N_g} V_j E_\ell (G_{j\ell}\cos(\theta_j - \delta_\ell) + B_{j\ell}\sin(\theta_j - \delta_\ell)) \\
&\quad + \textstyle\sum_{\ell=N_g+1}^{N_g+N_f} V_j V_\ell (G_{j\ell}\cos(\theta_j - \theta_\ell) + B_{j\ell}\sin(\theta_j - \theta_\ell)), \\
Q_j^n &= \textstyle\sum_{\ell=1}^{N_g} V_j E_\ell (G_{j\ell}\sin(\theta_j - \delta_\ell) - B_{j\ell}\cos(\theta_j - \delta_\ell)) \\
&\quad + \textstyle\sum_{\ell=N_g+1}^{N_g+N_f} V_j V_\ell (G_{j\ell}\sin(\theta_j - \theta_\ell) - B_{j\ell}\cos(\theta_j - \theta_\ell)).
\end{aligned}$$

Here, we can derive the expression for the d-axis current as $I_i^d = Q_i^n / E_i$, which can be expressed as:

$$\begin{aligned}
I_i^d &= \textstyle\sum_{\ell=1}^{N_g} E_\ell (G_{i\ell}\sin(\delta_i - \delta_\ell) - B_{i\ell}\cos(\delta_i - \delta_\ell)) \\
&\quad + \textstyle\sum_{\ell=N_g+1}^{N_g+N_f} V_\ell (G_{i\ell}\sin(\delta_i - \theta_\ell) - B_{i\ell}\cos(\delta_i - \theta_\ell))
\end{aligned}$$

To simplify these PF equations, we can linearize them around the steady-state operating point using a Jacobian matrix, resulting in

$$\begin{bmatrix} \Delta\mathbf{P}^g \\ \Delta\mathbf{I}^d \\ \Delta\mathbf{P}^f \\ \Delta\mathbf{Q}^f \end{bmatrix} = \begin{bmatrix} \frac{\partial\mathbf{P}^g}{\partial\boldsymbol{\delta}} & \frac{\partial\mathbf{P}^g}{\partial\mathbf{E}} & \frac{\partial\mathbf{P}^g}{\partial\boldsymbol{\theta}} & \frac{\partial\mathbf{P}^g}{\partial\mathbf{V}} \\ \frac{\partial\mathbf{I}^d}{\partial\boldsymbol{\delta}} & \frac{\partial\mathbf{I}^d}{\partial\mathbf{E}} & \frac{\partial\mathbf{I}^d}{\partial\boldsymbol{\theta}} & \frac{\partial\mathbf{I}^d}{\partial\mathbf{V}} \\ \frac{\partial\mathbf{P}^f}{\partial\boldsymbol{\delta}} & \frac{\partial\mathbf{P}^f}{\partial\mathbf{E}} & \frac{\partial\mathbf{P}^f}{\partial\boldsymbol{\theta}} & \frac{\partial\mathbf{P}^f}{\partial\mathbf{V}} \\ \frac{\partial\mathbf{Q}^f}{\partial\boldsymbol{\delta}} & \frac{\partial\mathbf{Q}^f}{\partial\mathbf{E}} & \frac{\partial\mathbf{Q}^f}{\partial\boldsymbol{\theta}} & \frac{\partial\mathbf{Q}^f}{\partial\mathbf{V}} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\delta} \\ \Delta\mathbf{E} \\ \Delta\boldsymbol{\theta} \\ \Delta\mathbf{V} \end{bmatrix} \tag{6.3}$$

where bold notation indicates the vector form obtained by concatenating all corresponding scalar variables, with g and f indicating the SG and GFM components, respectively. By substituting (6.3) into (6.1) and (6.2), we can formulate the overall dynamics in continuous-time as

$$\dot{\mathbf{x}} = \mathbf{A}_c\mathbf{x} + \mathbf{B}_c\mathbf{u} + \boldsymbol{\xi}. \tag{6.4}$$

where $\mathbf{x} := [\Delta\boldsymbol{\delta}_g, \Delta\boldsymbol{\omega}_g, \Delta\mathbf{E}_g, \Delta\mathbf{E}_g^{fd}, \Delta\boldsymbol{\delta}_f, \Delta\boldsymbol{\omega}_f, \Delta\mathbf{V}_f^e, \Delta\mathbf{V}_f]^\intercal \in \mathbb{R}^{4N_g+4N_f}$ and $\mathbf{u} := [\Delta\mathbf{E}_g^s, \Delta\mathbf{V}_f^s, \Delta\mathbf{P}_f^s, \Delta\mathbf{Q}_f^s]^\intercal \in \mathbb{R}^{N_g+3N_f}$. Due to the linearization, all the variables in $\mathbf{x}$ and $\mathbf{u}$ now represent the deviations from the corresponding steady-state values. Note that we add $\boldsymbol{\xi}$ which denotes random perturbations to system states, such as external disturbance or imperfect system modeling. By considering the GFM control time Δt, we represent the discrete-time dynamics based on (6.4) as

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \boldsymbol{\xi}_t, t = 0, 1, \ldots \tag{6.5}$$

where $\mathbf{A} = \mathbf{I} + \Delta t \cdot \mathbf{A}_c$ and $\mathbf{B} = \Delta t \cdot \mathbf{B}_c$with $\mathbf{I}$ and Δt denoting the identity matrix and a small enough time step, respectively.

6.3 Risk-constrained GFM Problem

Under the system dynamics in (6.5), we can formulate the GFM control problem as an optimal control one with the linear quadratic regulator (LQR) objective, by minimizing the total cost of state and control, as

$$\min_{\mathbf{K}\in\mathcal{K}} \; R_0(\mathbf{K}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}\left[\mathbf{x}_t^\intercal\mathbf{Q}\mathbf{x}_t + \mathbf{u}_t^\intercal\mathbf{R}\mathbf{u}_t\right] \tag{6.6}$$

where matrices $\{\mathbf{Q}, \mathbf{R}\}$ are positive (semi-)definite matrices used to weight the state and control variables into a single cost. Our goal is to find the best structured controller gain matrix $\mathbf{K} \in \mathbb{R}^{3N_f \times (2N_g + 4N_f)}$ that linearly maps from $\mathbf{x}_t$ to $\mathbf{u}_t$, namely $\mathbf{u}_t = -\mathbf{K}\mathbf{x}_t$. Here, $\mathcal{K}$ indicates the structured feedback set defined by the information-exchange graph. Specifically, for any GFM or SG node $\ell \in \mathcal{G} \cup \mathcal{F}$ and GFM node $j \in \mathcal{F}$, the structured set $\mathcal{K}$ is defined as

$$\mathcal{K} = \{\mathbf{K} : \mathbf{K}_{j,\ell} = 0 \text{ if and only if } j \nleftrightarrow \ell)\}$$

with $j \nleftrightarrow \ell$ implying that nodes ℓ and j are not connected through a communication link. Note that the size of $\mathbf{K}_{j,\ell}$ is different according to ℓ, i.e. $\mathbf{K}_{j,\ell} \in \mathbb{R}^{3\times 2}$ if $\ell \in \mathcal{G}$ and $\mathbf{K}_{j,\ell} \in \mathbb{R}^{3\times 4}$ if $\ell \in \mathcal{F}$. While the sparsity of $\mathbf{K}$ presents challenges in the analysis of the feasible region [84], it will not affect the implementation of our proposed algorithm.

Although the LQR objective in (5.6) effectively reduces oscillations on average, focusing solely on the average trajectory performance cannot account for the substantial system variability. Specifically, model parameters mismatch arising from imperfect modeling results in errors $\mathbf{A}_e$ and $\mathbf{B}_e$ in the matrices $\mathbf{A}$ and $\mathbf{B}$ in (6.5), respectively. Consequently, this mismatch introduces additional uncertainty and perturbations to the system dynamics, represented as

$$\begin{aligned}\mathbf{x}_{t+1} &= (\mathbf{A} + \mathbf{A}_e)\mathbf{x}_t + (\mathbf{B} + \mathbf{B}_e)\mathbf{u}_t + \boldsymbol{\xi}_t \\ &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + (\boldsymbol{\xi}_t + \mathbf{A}_e\mathbf{x}_t + \mathbf{B}_e\mathbf{u}_t) \\ &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \boldsymbol{\xi}'_t. \qquad (6.7)\end{aligned}$$

Notably, we modify the perturbation term in (6.5) to $\boldsymbol{\xi}'_t$, which encompasses both the original random noise $\boldsymbol{\xi}_t$ and an additional perturbation due to model parameter mismatch. As the degree of mismatch increases, characterized by higher values in $\mathbf{A}_e$ and $\mathbf{B}_e$, there is a corresponding increase in the perturbation $\boldsymbol{\xi}'_t$. This escalation of perturbations leads to a higher level of variability in the system trajectory. This increased variability presents a significant challenge to LQR-based design, especially in the context of interconnected grids with inter-area oscillations and thereby diminishing the worst-case damping performance of the controller.

To tackle this issue, we put forth a risk-constrained LQR formulation by limiting the so-termed mean-variance risk measure, as

$$\begin{aligned}\min_{\mathbf{K}\in\mathcal{K}}\ R_0(\mathbf{K})&=\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}\left[\mathbf{x}_t^\intercal\mathbf{Q}\mathbf{x}_t+\mathbf{u}_t^\intercal\mathbf{R}\mathbf{u}_t\right] \qquad (6.8)\\ \text{s.t. } R_c(\mathbf{K})&=\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\sum_{t=0}^{T-1}\left(\mathbf{x}_t^\intercal\mathbf{Q}\mathbf{x}_t-\mathbb{E}\left[\mathbf{x}_t^\intercal\mathbf{Q}\mathbf{x}_t|\mathcal{H}_t\right]\right)^2\leq c.\end{aligned}$$

We denote $\mathcal{H}_t=[\mathbf{x}_0,\mathbf{u}_0,\ldots,\mathbf{x}_{t-1},\mathbf{u}_{t-1}]$ as the system state and control trajectory up to time t, and c as a risk tolerance parameter. Note that the risk constraint bounds the deviations of the realized state cost $\mathbf{x}_t^\intercal\mathbf{Q}\mathbf{x}_t$ from its expected value. This constraint enables us to mitigate the worst-case scenarios of very high system variability as caused by external load disturbances and imperfect modeling.

The risk-constrained optimization problem in (6.8) is solved by utilizing SGDmax with ZOPG in Algorithm 5 and 6, as introduced in Section 4.5.

6.4 Numerical Tests

To demonstrate the effectiveness of our risk-aware GFM control, we conduct numerical tests on a modified version of the IEEE 68-bus system [115]. As shown in Fig. 5.1, this system consists of 16 SGs across five ares. We have added a total of 10 GFMs to selected load buses. The GFM parameters follow from [123] and are set to: $\tau = 0.01$ s, $m^p = 0.01$ pu, $m^q = 0.05$ pu, $k^{pv} = 1$ pu, $k^{iv} = 5.86$ pu/s. We assume that only SGs and GFMs in neighboring areas can exchange data, forming a structured feedback following the information-exchange graph. For example, SG1 at bus 1 can exchange data with the GFM located at bus 49 since Area 1 and 2 are connected via a communication line. However, SG1 cannot exchange data with the GFM at bus 18, as Area 1 and 5 are not adjacent to each other, and thus there is no communication line connecting them.

We consider two cases, namely *GFM* and *GFM-Risk*. *GFM* represents the policy trained using Algorithm 6 to solve (5.6) with (6.4) without considering the risk constraint, while *GFM-Risk* is the solution of our risk-constrained problem(6.8) obtained using Algorithm 6. Both *GFM* and *GFM-Risk* employ Algorithm 6 with estimated gradients obtained from Algorithm 5. The parameters used in the simulation are as follows: $r = 0.1, M = 50, \eta = 10^{-4}$ and $c = 0.25$. The control time step is set to $\Delta t = 0.01$s and the observation time window is set to be a total of 8s. As for perturbations, we introduce a line-to-ground fault at a random location on a random line at $t = 0$ for each episode during both training and testing [110].

Using the two different policies *GFM* and *GFM-Risk* obtained from the training, we conduct tests involving 100 new scenarios, each featuring a fault on a

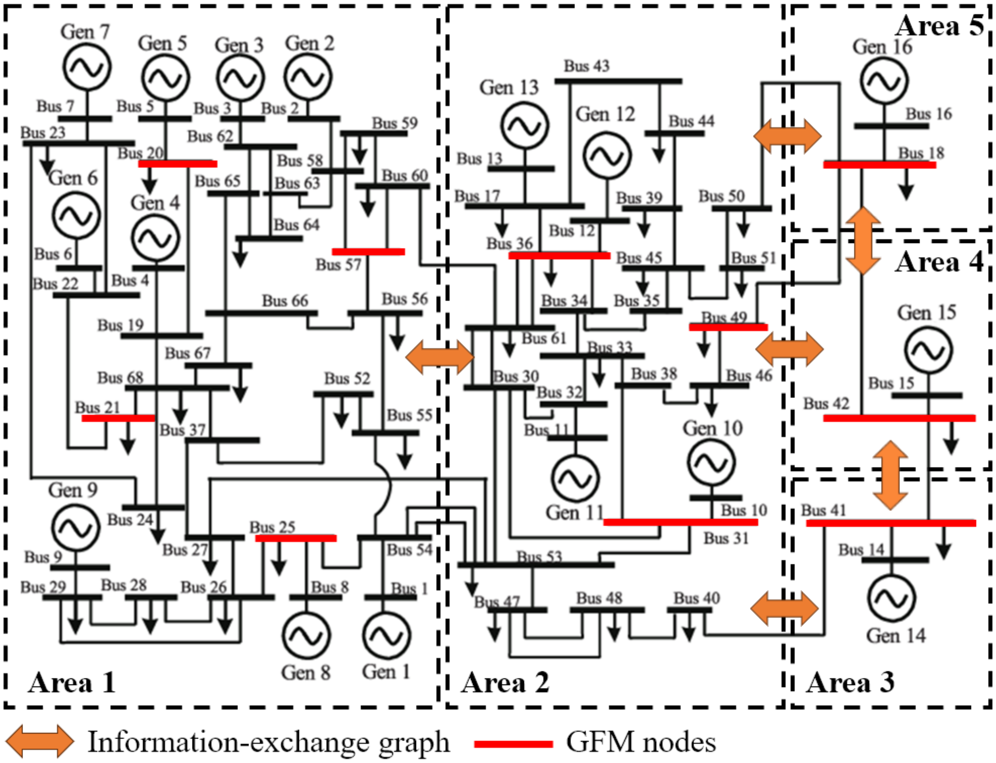


Figure 6.4: Modified IEEE 68-bus system with 16 SGs and 10 GFMs.

random line. To demonstrate the effectiveness of the risk constraint in the presence of model parameter mismatch, we consider three mismatch settings: No mismatch, 20% mismatch and 40% mismatch. For the 20% and 40% mismatch settings, we introduce random error with the maximum bounds of 20% and 40% on the SG parameters $[M, D, \tau^{d'}, x^d, x^{d'}, \tau^a, K^a]$ and GFM parameters $[\tau, m^p, m^q, k^{pv}, k^{iv}]$ for each SG and GFM, respectively.

To illustrate the impact of model parameter mismatch, Fig. 6.5 shows the frequency deviation for an extreme scenario at different mismatch levels when using the GFM-Risk design. In this extreme scenario, we consider a fault occurring in the

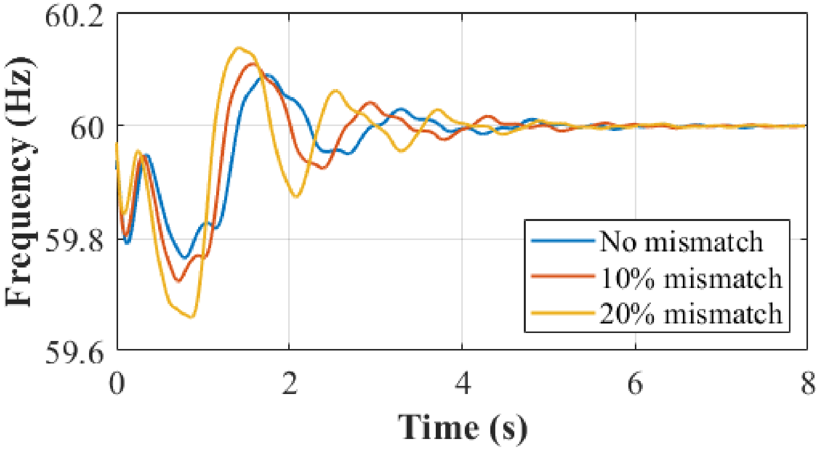
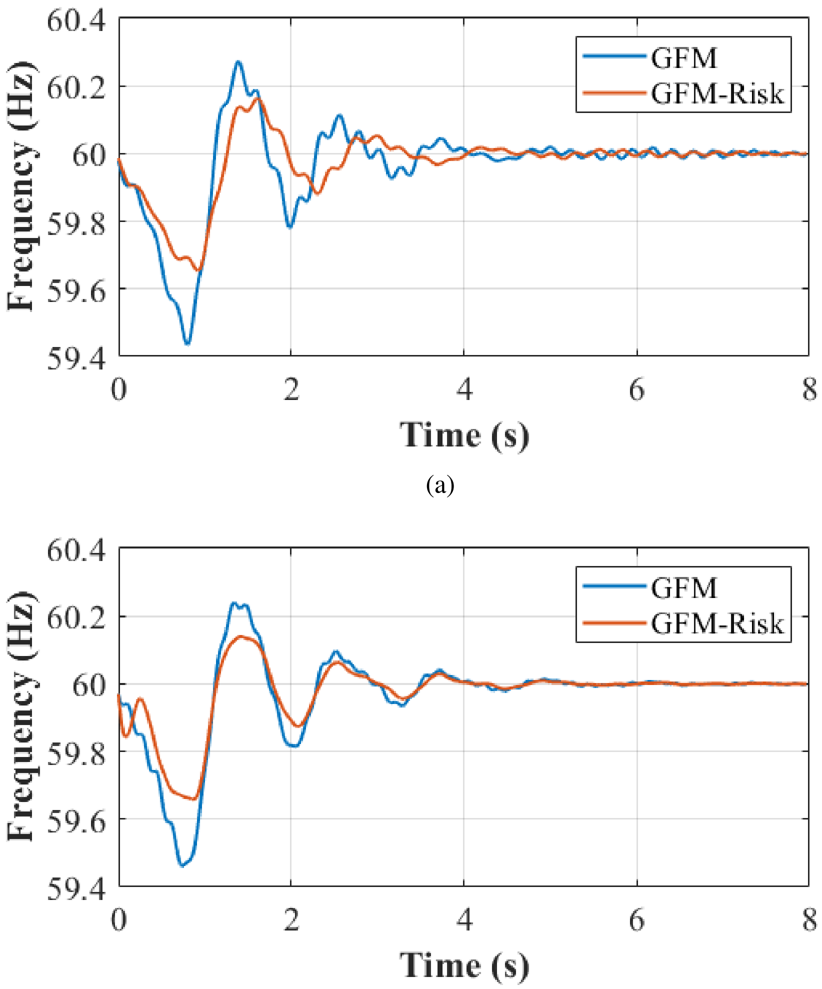


Figure 6.5: Comparison on the frequency deviation with different model parameter mismatch levels.

line between bus 17 and bus 36. Evidently, as the level of mismatch increases, the damping performance gradually deteriorates, requiring more time to reach a steady-state. This implies that modeling errors lead to an increase in system variability, thereby diminishing the damping performance of the controller.

To better analyze the effectiveness of incorporating the risk constraint, we compare the frequency deviations on buses near the fault and far from the fault with a 40% mismatch setting, as depicted in Fig. 6.6. We select bus 45 and 24 as representative buses near and far from the fault. Clearly, *GFM-Risk* exhibits smaller deviations and reaches steady-state faster than *GFM* in both buses. In Fig 6.6(a), there are more fluctuations than in Fig 6.6(b) in both cases when approaching the steady-state, which is reasonable as bus 45 is closer to the line with the fault. Furthermore, *GFM* experiences small, unstable fluctuations, particularly at bus 45, primarily due to model parameter mismatch. On the other hand, *GFM-Risk* effectively mitigates



(a)

(b)

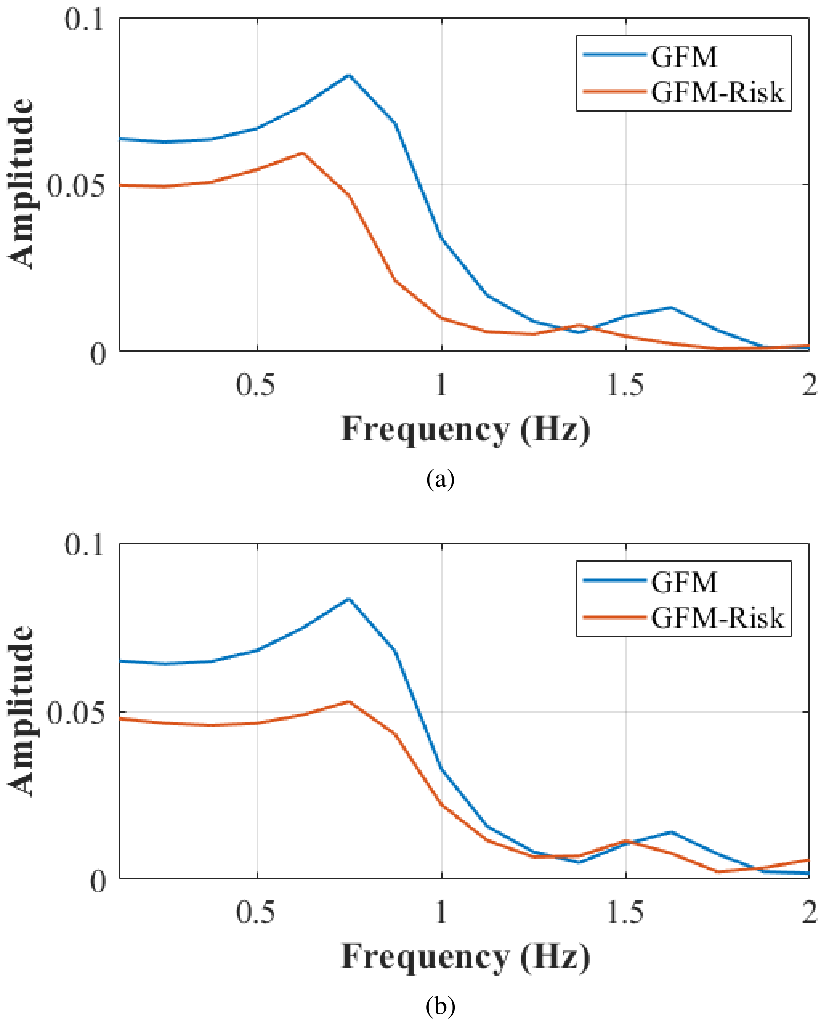
Figure 6.6: Comparison on the frequency deviation at (a) bus 45 near the fault and (b) bus 24 far from the fault.

this unstable behavior in both cases. This indicates that *GFM-Risk* provides more damping in extreme cases compared to *GFM* by considering the risk constraint.

Next, we conducted spectrum analysis using fast Fourier transform (FFT) on

Fig. 6.6 to demonstrate the effectiveness of incorporating a risk constraint in reducing wide-area oscillations. Fig. 6.7 presents the FFT results for buses located near and far from the fault location, specifically, bus 45 and bus 27. Since our primary focus is on wide-area oscillations, we concentrated on the low-frequency range between 0.1 and 2 Hz. It is evident that both buses exhibit significant frequency components around 0.75 Hz, which arise due to wide-area oscillations. Consistent with the trends observed in Fig. 6.6, *GFM-Risk* demonstrates better damping performance than *GFM* by reducing the level of oscillation amplitude. It is worth highlighting that significant improvements in damping performance are particularly noticeable at the bus located near the fault, with a more significant decrease in the peak frequency. These enhancements make a substantial contribution to improving system stability, especially since buses near the fault location experience a higher level of oscillation than the rest of the system.

To further quantify the benefits of including the risk constraint, we compare the statistics of the LQR objective costs over 100 testing scenarios in Fig. 6.8-6.10. Each subplot corresponds to a certain parameter mismatch level of 0%, 20%, or 40%. The red lines represent the median values, while the lower and upper quartiles are indicated by the blue boxes. The maximum and minimum values are depicted with black lines. First, by comparing across the subplots, we observe that both *GFM* and *GFM-Risk* experience higher costs with increasing mismatch levels, in terms of both the median and variance values. This is expected as an increasing mismatch level introduces higher system variability and frequency oscillations [cf. (6.7)]. Second, compared with *GFM*, our proposed *GFM-Risk* slightly increases the median



(a)

(b)

Figure 6.7: Fast Fourier transform (FFT) results of (a) bus 45 near the fault and (b) bus 24 far from the fault.

value in all subplots, as a result of using the risk constraint. However, *GFM-Risk* leads to significantly smaller variances, especially lowering the maximum value corresponding to the highest oscillation level. This highlights the effectiveness of using the risk constraint in mitigating the worst-case performance and thus enhanc-

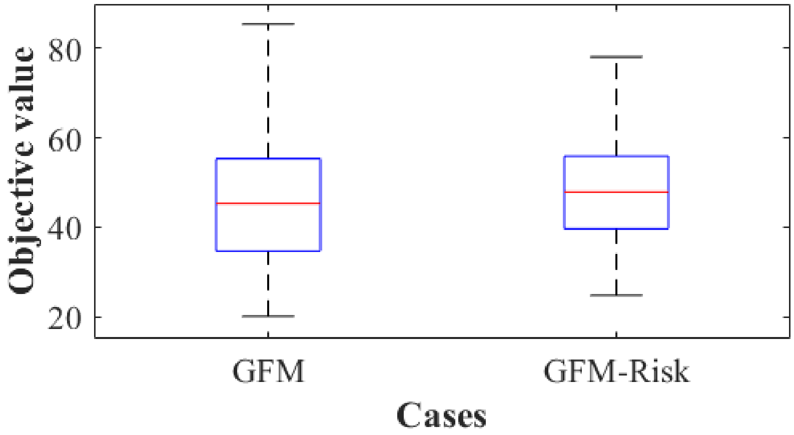


Figure 6.8: Comparison on the LQR objective values in no mismatch case.

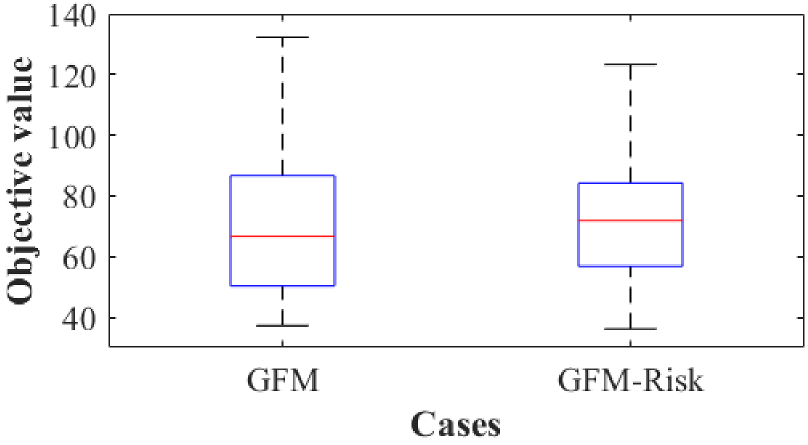


Figure 6.9: Comparison on the LQR objective values in 20% mismatch case.

ing the overall system stability. We also observe that the decrease in the variance of *GFM-Risk* is more evident at higher mismatch level, corroborating the applicability of our proposed approach to practical implementations where large model mismatches or system uncertainty in general tend to be present.

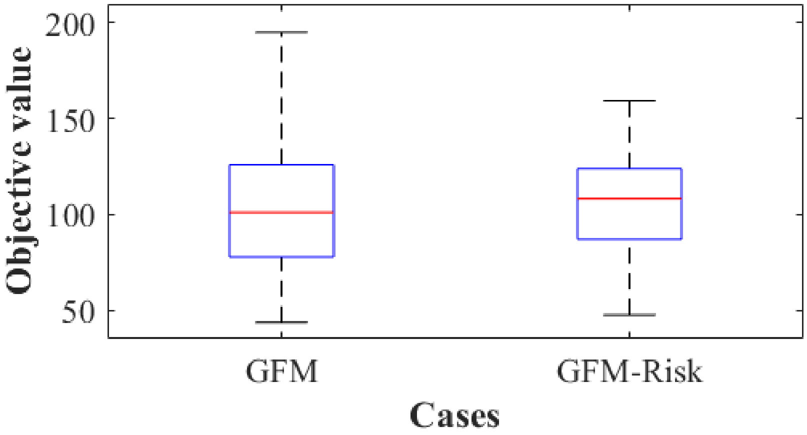


Figure 6.10: Comparison on the LQR objective values in 40% mismatch case.

In conclusion, our numerical tests have validated the convergence of the proposed algorithm for solving the GFM problem considering a mean-variance risk constraint. The test results have confirmed that the constraint contributes to enhancing damping in extreme cases and reducing the variance in LQR objective values. Most importantly, the worst-case performance improves as the model parameter mismatch level increases, demonstrating the efficacy of the proposed method in the presence of significant disturbance.

This work designed a risk-aware controller for GFMs that aims to address the frequency oscillations in the presence of model parameter mismatch. Based on the linearized system model that incorporates both SGs and GFMs, we formulated the problem to minimize the LQR objective over the structured feedback gain matrix according to the connectivity of communication network. As we adopt model-based RL method, the errors on the SG and GFM parameters lead to the increase in state variability, which results in high frequency deviations. To tackle this issue, we introduced the mean-variance risk constraint to limit the state cost variations, thereby reducing the system variability caused by model parameter mismatch and enhancing the worst-case performance. By reformulating this constraint into a tractable quadratic form, we solved the dual problem, represented as a maximin problem, using an RL-based SGDmax algorithm. This method searched for a policy through GD iterations by leveraging efficient ZOPG for gradient estimation. Numerical tests on the modified IEEE 68-bus system highlighted the effectiveness of the proposed risk-aware GFM controller in reducing the variability of total LQR cost, thus improving the performance in worst-case scenarios involving the high level of parametric mismatch.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This work first focuses on developing RL-based control policies for power system resources, encompassing batteries and electric vehicle charging stations (EVCS) while addressing the issues from complex and high-dimensional MDP representation. Second, it develops risk-aware RL-based control policies for power system stability to mitigate the high variability from the uncertainty factors, such as high load perturbation in load frequency control (LFC), communication delays in wide-area damping control (WADC) and modeling errors in grid-forming inverter (GFM) control.

Chapter 2 presents an accurate model of cycle-based degradation cost in order to allow for efficient battery control designs using RL. In order to model the degradation which depends on the full cycle, we introduce additional state variables to judiciously keep track of important switching points of SoC trajectory for effectively identifying (dis)charging cycles. This way, the actual degradation cost is separated into instantaneous terms along with other operation costs such as the net cost for electricity usage and FR penalty, such that powerful DQN based RL algorithms are readily applicable. Numerical tests confirm the effectiveness of proposed cycle-based degradation model and demonstrate the performance improvements in effectively mitigating battery degradation over existing linearized approximation approach.

Chapter 3 has developed a practical modeling approach for the optimal EV charging station operation problem, allowing for efficient solutions using RL. To deal with the high and variable dimensions of states/actions, we propose to design efficient aggregation schemes by utilizing the EV's laxity that measures the emergency level of its charging need. First, the LLF rule has made it possible to consider only the total charging action across the EVCS, which is shown to recover feasible individual EV charging schedules if existing. Second, we propose aggregating the state into the number of EVs in each laxity group, which satisfies reward and dynamic homogeneities and thus leads to equivalent policy search. We have developed the policy gradient method based on the proposed MDP representation to find the optimal parameters for the linear Gaussian policy. Case studies based on real-world data have demonstrated the performance improvement of the proposed MDP representation over the earlier approximation-based approach for the EVCS problem. The RL parameter results imply that further state aggregation can deal with many laxity levels in practical systems at a minimal loss of optimality.

Chapter 4 presents a learning-based method to solve the LFC problem in the networked microgrids while considering the structured feedback and the mean-variance risk constraint. To solve this problem, we consider the minimax reformulation of the dual problem and leverage the stochastic (S)GDmax algorithms to approach the stationary points (SPs). Specifically, the SGDmax algorithm relies on the ZOPG-based updates, making it suitable for model-free learning. Using the recent results on the local Lipschitz and smoothness of LQR cost, convergence of the (S)GDmax algorithms can be established by properly bounding the related constants for choosing the stepsize. Notably, for SGDmax the convergence can only be shown with a high probability, due to the additional noise in the gradient estimate. Numerical tests on a simple networked microgrids system have validated the convergence of our proposed algorithms while demonstrating the impact of risk and structured constraints for the LQR problem.

Chapter 5 designs a risk-constrained WADC approach that aims to address the communication delays. Based on a linearized system model that incorporates both SGs and VSCs to provide damping capabilities, we cast it to minimize the LQR objective over the structured feedback gain matrix according to the communication network's connectivity. Our analysis suggest that the level of system perturbations grows with larger communication delays, leading to higher state variability. Thereby, we introduce the mean-variance risk constraint to bound the variation of the state cost, in order to reduce the delay-induced variability and improve the worst-case performance. By reformulating the constraint into a tractable quadratic form, we solve the dual maximin problem by developing a RL-based SGDmax algorithm. The latter works by iteratively searching for a policy using efficient ZOPG-based gradient updates, which can provably attain the SP with high probability. Numerical tests on the IEEE 68-bus system demonstrates the effectiveness and advantages of the proposed risk-aware WADC design in reducing the variability of the total LQR cost, thereby improving the stability performance especially in worst-case scenarios of oscillations and delays.

Chapter 6 designs a risk-aware controller for GFMs that aims to address the frequency oscillations resulting from parameter mismatch in the system model. We consider GFMs along with SGs and constructed the overall system dynamics by accounting for network coupling. Our GFM control policy design involves a constrained optimization problem with a standard LQR objective and a mean-variance risk constraint, which latter aimed to bound time-averaged metrics related to state deviations. To solve this problem, we reformulate the constraint in quadratic form and formulated the dual problem as a maximin problem. The RL-based algorithm termed as SGDmax is adopted to find the best policy considering the structured feedback, with the ZOPG utilized as estimated gradients. Numerical tests conducted on the modified IEEE 68-bus system validates the effectiveness of our proposed approach. The training results confirms the convergence of our algorithm. The test results based on 100 scenarios with different model parameter mismatch levels highlights the benefits of the constraint in mitigating frequency deviations in extreme cases. Notably, it reduces the variance in objective values and improved worst-case performance, particularly in scenarios involving significant parametric mismatch.

7.2 Future Work

The RL-based EVCS operation problem in Chapter 2 has limitations in satisfying all EV demands and the constant EV charging rate in action. In this context, exciting future research directions open up regarding more general EVCS problem set-ups such as penalizing non-fully charged EVs at departure, as well as variable EV charging rate and action. The former makes it relevant to consider a constrained RL formulation that limits the number (or total demand) of unsatisfied EVs at departure or the corresponding statistical risk, following from the safe RL framework [131]. As for the variable charging power, it would be interesting to pursue the connection to recent work [55] that uses a smoothed LLF approach to deal with different charging rates.

The RL-based WADC and GFM comtrol in Chapter 5 and Chapter 6 can be extended to include the different types of risk measures to e.g., conditional value at risk (CVaR). In addition, we can investigate large-scale implementations of our proposed design, as well as considering generalized grid control tasks for new resources and renewables.

Appendices

Appendix A

Proof of Theorem 1

The key step is to ensure that the iterates stay within the sublevel set $\mathcal{G}^0$ defined in Section 4.4. To this end, consider the function $\Phi(\cdot)$ in (4.17) with its *Moreau envelope* $\Phi_\mu(\cdot)$ defined as (4.18). Based on Lemma 3, the problem becomes to show the convergence of $\Phi_{\mu_0}(\cdot)$ instead. To bound the iterative change in $\Phi_{\mu_0}(\cdot)$, one can use L_0-Lipschitz and ℓ_0-weakly convex properties of $\Phi(\cdot)$ to analyze the update $K^{j+1} \leftarrow K^j - \eta\nabla\Phi(K^j)$ and obtain [80, Lemma D.3]

$$\Phi_{\mu_0}(K^{j+1}) \leq \Phi_{\mu_0}(K^j) - \frac{\eta}{4}\left\|\nabla\Phi_{\mu_0}(K^j)\right\|^2 + \eta^2\ell_0 L_0^2. \tag{A.1}$$

Note that $\eta \leq \rho_0$ is needed to apply the constants L_0 and ℓ_0. Furthermore, by setting $\eta \leq \epsilon^2/(4\ell_0 L_0^2)$, the last term is upper bounded by $\epsilon^4/(16\ell_0 L_0^2)$, while the second term is lower bounded by the same value as $\|\nabla\Phi_{\mu_0}(K^j)\| > \epsilon$ holds before reaching K_ϵ. Therefore, we can guarantee that $\Phi_{\mu_0}(K^j)$ is non-increasing and $K^j \in \mathcal{G}^0\ \forall j$. As a result, L_0-Lipschitz and ℓ_0-smoothness properties hold throughout the iterative updates.

To verify the SP condition in Lemma 3, summing up (A.1) over $j = 0, 1, \ldots,$ $J-1$ yields

$$\frac{1}{J}\sum_{j=0}^{J-1}\left\|\nabla\Phi_{\mu_0}(K^j)\right\|^2 \leq \frac{4\big[\Phi_{\mu_0}(K^0)-\Phi_{\mu_0}(K^J)\big]}{J\eta}+4\eta\ell_0 L_0^2$$

$$\leq \frac{4\ell_0 L_0^2 \Phi_{\mu_0}(K^0)}{J\epsilon^2} + \epsilon^2$$

where the second step uses the choice of stepsize in (4.19). As K^0 is stable, the value $\Phi_{\mu_0}(K^0)$ is finite and thus the first term is in the order of ϵ^2 with $J = O(\ell_0 L_2^2 \Phi_{\mu_0}(K^0)/\epsilon^4)$ iterations. As a result, the gradient norm $\|\nabla \Phi_{\mu_0}(K^j)\|$ eventually approaches ϵ, satisfying the ϵ-SP condition. □

Appendix B

Proof of Theorem 2

Similar to Appendix A, the key lies in the iterative analysis of function $\Phi_{\mu_0}(K)$, or in this case its expectation. First, due to the noisy gradient of ZOPG, one can obtain the following inequality similar to (A.1) [80, Lemma D.4]:

$$\begin{aligned}\mathbb{E}\left[\Phi_{\mu_0}(K^{j+1})\right] \leq{}& \mathbb{E}\left[\Phi_{\mu_0}(K^{j})\right] \\ &- \frac{\eta}{4}\mathbb{E}\|\nabla\Phi_{\mu_0}(K^j)\|^2 + \eta^2\ell_0\left(L_0^2 + \ell_0^2 r^2/M\right)\end{aligned} \tag{B.1}$$

where the last term is because the noise variance of each ZO gradient sample with a smoothing radius r is bounded by $\ell_0^2 r^2$ as shown in [72, 82], while M is the total number of samples. Note that by choosing the smoothing radius r as in Theorem 2, we prevent the overall noise variance $(\ell_0^2 r^2/M)$ to be dominant in the last term. Moreover, r needs to be smaller than ρ_0 to ensure that each ZOPG iteration can use the local Lipschitz and smoothness constants.

Summing up (B.1) over iterations $j = 0, \ldots, J-1$ yields

$$\begin{aligned}&\frac{1}{J}\sum_{j=0}^{J-1}\mathbb{E}\left[\|\nabla\Phi_{\mu_0}(K^j)\|^2\right] \\ \leq&\frac{4\left[\Phi_{\mu_0}(K^0)-\mathbb{E}[\Phi_{\mu_0}(K^J)]\right]}{J\eta} + 4\eta\ell_0(L_0^2 + \ell_0^2 r^2/M).\end{aligned}$$

With $\eta = O(\epsilon^2)$ and J inversely proportional to $\eta\epsilon^2$ given in Theorem 2, this upper bound is in the order of ϵ^2, as detailed soon. To eliminate the expectation therein, one can analyze the probability of exceeding ϵ^2 by considering whether $\{K^j\}$ exceeds $\mathcal{G}^1$ within J iterations, as given by

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{J}\sum_{j=0}^{J-1}\|\nabla\Phi_{\mu_0}(K^j)\|^2 \geq \epsilon^2\right)\\
=&\mathbb{P}\left(\frac{1}{J}\sum_{j=0}^{J-1}\|\nabla\Phi_{\mu_0}(K^j)\|^2 \geq \epsilon^2, \tau > J\right)\\
+&\mathbb{P}\left(\frac{1}{J}\sum_{j=0}^{J-1}\|\nabla\Phi_{\mu_0}(K^j)\|^2 \geq \epsilon^2, \tau \leq J\right),
\end{aligned} \tag{B.2}$$

where $\tau := \min\{j \geq 0 : K^j \notin \mathcal{G}^1\}$. The first term of (B.2) can be bounded by

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{J}\sum_{j=0}^{J-1}\|\nabla\Phi_{\mu_0}(K^j)\|^2 \geq \epsilon^2, \tau > J\right)\\
\leq&\frac{1}{\epsilon^2}\mathbb{E}\left[\frac{1}{J}\sum_{j=0}^{J-1}\|\nabla\Phi_{\mu_0}(K^j)\|^2\right]\\
\leq&\frac{1}{\epsilon^2}\left\{\frac{4\left[\Phi_{\mu_0}(K^0)-\mathbb{E}[\Phi_{\mu_0}(K^J)]\right]}{J\eta} + 4\eta\ell_0(L_0^2+\ell_0^2r^2/M)\right\}\\
\leq&\frac{4\Phi_{\mu_0}(K^0)}{J\eta\epsilon^2}+\frac{4}{\alpha}=\frac{4}{\beta}+\frac{4}{\alpha}
\end{aligned} \tag{B.3}$$

where the first step follows from the Markov's inequality, while the last one uses the parameter settings in (4.22) with β simplifying the first fractional term to be determined soon.

In addition, the second term can be bounded by recognizing that the sequence $Y^j := \Phi_{\mu_0}(K^{\min(j,\tau)}) + (J-j)\eta\ell_0(L_0^2+\ell_0^2r^2/M)$ is a supermartingale, as shown in [82]. Thus, using the Doob's maximal inequality for supermartingales, one can bound the second term of (B.2) as

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{J}\sum_{j=0}^{J-1}\|\nabla\Phi_{\mu_0}(K^j)\|^2 \geq \epsilon^2, \tau \leq J\right)\\
\leq&\mathbb{P}(\tau \leq J)\\
\leq&\frac{\Phi_{\mu_0}(K^0) + J\eta^2\ell_0(L_0^2 + \ell_0^2 r^2/M)}{10\Phi_{\mu_0}(K^0)}\\
\leq&\frac{1}{10} + \frac{J\eta\epsilon^2}{10\alpha\Phi_{\mu_0}(K^0)} = \frac{1}{10} + \frac{\beta}{10\alpha}
\end{aligned} \tag{B.4}$$

where the first step relaxes the probability, the second step follows from Doob's maximal inequality, while the last one again uses the parameter settings in (4.22). Therefore, the probability that $\{K^j\}$ exceeds $\mathcal{G}^1$ before J is bounded, and we can ensure that $\{K^j\}$ is within $\mathcal{G}^1$ with a high probability. This is why the compact sublevel set $\mathcal{G}^1$ can be used to bound the Lipschitz and smoothness constants for $\Phi(K)$.

By substituting (B.3)-(B.4), the overall probability becomes

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{J}\sum_{j=0}^{J-1}\|\nabla\Phi_{\mu_0}(K^j)\|^2 \geq \epsilon^2\right) &\leq \frac{1}{10} + \frac{4}{\beta} + \frac{\beta}{10\alpha} + \frac{4}{\alpha}\\
&\leq \frac{1}{10} + \frac{4}{\alpha} + \frac{4}{\sqrt{10\alpha}}
\end{aligned}$$

where the last step uses the best choice of $\beta = 2\sqrt{10\alpha}$. As a result, with probability $(0.9 - \frac{4}{\alpha} - \frac{4}{\sqrt{10\alpha}})$, the ϵ-SP can be attained by the iterations $\{K^j\}$ within J iterations. Note that this probability increases with α, but a large α also reduces the stepsize which potentially slows down the convergence. Therefore, the choice of α is very important for the algorithm implementation. □

Appendix C

Proof of Eq. (5.2)

To formulate (5.3a) and (5.3b), we perform the algebraic analysis of (5.2) that is similar to [102, Appendix]. First, to simplify the notation, let us denote the Jacobian matrix in (5.2) by

$$\begin{bmatrix} \frac{\partial \mathbf{P}^e}{\partial \boldsymbol{\delta}} & \frac{\partial \mathbf{P}^e}{\partial \mathbf{E}} & \frac{\partial \mathbf{P}^e}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{P}^e}{\partial \mathbf{V}} \\ \frac{\partial \mathbf{I}^q}{\partial \boldsymbol{\delta}} & \frac{\partial \mathbf{I}^q}{\partial \mathbf{E}} & \frac{\partial \mathbf{I}^q}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{I}^q}{\partial \mathbf{V}} \\ \frac{\partial \mathbf{P}^v}{\partial \boldsymbol{\delta}} & \frac{\partial \mathbf{P}^v}{\partial \mathbf{E}} & \frac{\partial \mathbf{P}^v}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{P}^v}{\partial \mathbf{V}} \\ \frac{\partial \mathbf{Q}^v}{\partial \boldsymbol{\delta}} & \frac{\partial \mathbf{Q}^v}{\partial \mathbf{E}} & \frac{\partial \mathbf{Q}^v}{\partial \boldsymbol{\theta}} & \frac{\partial \mathbf{Q}^v}{\partial \mathbf{V}} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} & \mathbf{A}_{14} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & \mathbf{A}_{24} \\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} & \mathbf{A}_{34} \\ \mathbf{A}_{41} & \mathbf{A}_{42} & \mathbf{A}_{43} & \mathbf{A}_{44} \end{bmatrix}.$$

By combining the third and fourth rows of (5.2), we have

$$\begin{aligned} \Delta\boldsymbol{\theta} &= \mathbf{F}_1\mathbf{A}_{34}^{-1}(\Delta\mathbf{P}^v - \mathbf{A}_{31}\Delta\boldsymbol{\delta} - \mathbf{A}_{32}\Delta\mathbf{E}) \\ &\quad - \mathbf{F}_1\mathbf{A}_{44}^{-1}(\Delta\mathbf{Q}^v - \mathbf{A}_{41}\Delta\boldsymbol{\delta} - \mathbf{A}_{42}\Delta\mathbf{E}), \\ \Delta\mathbf{V} &= \mathbf{F}_2\mathbf{A}_{43}^{-1}(\Delta\mathbf{Q}^v - \mathbf{A}_{41}\Delta\boldsymbol{\delta} - \mathbf{A}_{42}\Delta\mathbf{E}) \\ &\quad - \mathbf{F}_2\mathbf{A}_{33}^{-1}(\Delta\mathbf{P}^v - \mathbf{A}_{31}\Delta\boldsymbol{\delta} - \mathbf{A}_{32}\Delta\mathbf{E}), \end{aligned}$$

where the two new matrices are

$$\mathbf{F}_1 = (\mathbf{A}_{34}^{-1}\mathbf{A}_{33} - \mathbf{A}_{44}^{-1}\mathbf{A}_{43})^{-1},\ \mathbf{F}_2 = (\mathbf{A}_{43}^{-1}\mathbf{A}_{44} - \mathbf{A}_{33}^{-1}\mathbf{A}_{34})^{-1}.$$

Substituting them into the first and second rows of (5.2) yields

$$\Delta\mathbf{P}^e = \mathbf{A}_1^P\Delta\boldsymbol{\delta} + \mathbf{A}_2^P\Delta\mathbf{E} + \mathbf{A}_3^P\Delta\mathbf{P}^v + \mathbf{A}_4^P\Delta\mathbf{Q}^v,$$

$$\Delta \mathbf{I}^q = \mathbf{A}_1^I \Delta \boldsymbol{\delta} + \mathbf{A}_2^I \Delta \mathbf{E} + \mathbf{A}_3^I \Delta \mathbf{P}^v + \mathbf{A}_4^I \Delta \mathbf{Q}^v.$$

with the coefficient matrices given by

$$\begin{aligned}
\mathbf{A}_1^P &= \mathbf{A}_{11} + \mathbf{A}_{13}\mathbf{F}_1(\mathbf{A}_{44}^{-1}\mathbf{A}_{41} - \mathbf{A}_{34}^{-1}\mathbf{A}_{31}) \\
&\quad + \mathbf{A}_{14}\mathbf{F}_2(\mathbf{A}_{33}^{-1}\mathbf{A}_{31} - \mathbf{A}_{43}^{-1}\mathbf{A}_{41}), \\
\mathbf{A}_2^P &= \mathbf{A}_{12} + \mathbf{A}_{13}\mathbf{F}_1(\mathbf{A}_{44}^{-1}\mathbf{A}_{42} - \mathbf{A}_{34}^{-1}\mathbf{A}_{32}) \\
&\quad + \mathbf{A}_{14}\mathbf{F}_2(\mathbf{A}_{33}^{-1}\mathbf{A}_{32} - \mathbf{A}_{43}^{-1}\mathbf{A}_{42}), \\
\mathbf{A}_3^P &= \mathbf{A}_{13}\mathbf{F}_1\mathbf{A}_{34}^{-1} - \mathbf{A}_{14}\mathbf{F}_2\mathbf{A}_{33}^{-1}, \\
\mathbf{A}_4^P &= \mathbf{A}_{14}\mathbf{F}_2\mathbf{A}_{43}^{-1} - \mathbf{A}_{13}\mathbf{F}_1\mathbf{A}_{44}^{-1}, \\
\mathbf{A}_1^I &= \mathbf{A}_{21} + \mathbf{A}_{23}\mathbf{F}_1(\mathbf{A}_{44}^{-1}\mathbf{A}_{41} - \mathbf{A}_{34}^{-1}\mathbf{A}_{31}) \\
&\quad + \mathbf{A}_{24}\mathbf{F}_2(\mathbf{A}_{33}^{-1}\mathbf{A}_{31} - \mathbf{A}_{43}^{-1}\mathbf{A}_{41}), \\
\mathbf{A}_2^I &= \mathbf{A}_{22} + \mathbf{A}_{23}\mathbf{F}_1(\mathbf{A}_{44}^{-1}\mathbf{A}_{42} - \mathbf{A}_{34}^{-1}\mathbf{A}_{32}) \\
&\quad + \mathbf{A}_{24}\mathbf{F}_2(\mathbf{A}_{33}^{-1}\mathbf{A}_{32} - \mathbf{A}_{43}^{-1}\mathbf{A}_{42}), \\
\mathbf{A}_3^I &= \mathbf{A}_{23}\mathbf{F}_1\mathbf{A}_{34}^{-1} - \mathbf{A}_{24}\mathbf{F}_2\mathbf{A}_{33}^{-1}, \\
\mathbf{A}_4^I &= \mathbf{A}_{24}\mathbf{F}_2\mathbf{A}_{43}^{-1} - \mathbf{A}_{23}\mathbf{F}_1\mathbf{A}_{44}^{-1}.
\end{aligned}$$

This completes the proof of (5.3). □

Bibliography

[1] B. Recht, “A tour of reinforcement learning: The view from continuous control,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, no. 1, pp. 253–279, 2019.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, second ed., 2018.

[3] Y. Shi, B. Xu, Y. Tan, D. Kirschen, and B. Zhang, “Optimal battery control under cycle aging mechanisms in pay for performance settings,” *IEEE Transactions on Automatic Control*, vol. 64, no. 6, pp. 2324–2339, 2019.

[4] B. Xu, Y. Shi, D. S. Kirschen, and B. Zhang, “Optimal battery participation in frequency regulation markets,” *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6715–6725, 2018.

[5] Z. Ma, D. Callaway, and I. Hiskens, “Optimal charging control for plug-in electric vehicles,” *Control and Optimization Methods for Electric Smart Grids*, pp. 259–273, 2012.

[6] Y. Xu and F. Pan, “Scheduling for charging plug-in hybrid electric vehicles,” in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 2495–2501, IEEE, 2012.

[7] X. Fu, J. Sun, M. Huang, Z. Tian, H. Yan, H. H.-C. Iu, P. Hu, and X. Zha, “Large-Signal Stability of Grid-Forming and Grid-Following Controls in Voltage Source Converter: A Comparative Study,” *IEEE Transactions on Power Electronics*, vol. 36, no. 7, pp. 7832–7840, 2021.

[8] L. Xiong, X. Liu, Y. Liu, and F. Zhuo, “Modeling and Stability Issues of Voltage-source Converter-dominated Power Systems: A Review,” *CSEE Journal of Power and Energy Systems*, vol. 8, no. 6, pp. 1530–1549, 2022.

[9] A. Singhal, T. L. Vu, and W. Du, “Consensus Control for Coordinating Grid-Forming and Grid-Following Inverters in Microgrids,” *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 4123–4133, 2022.

[10] B. Xu, A. Oudalov, A. Ulbig, G. Andersson, and D. S. Kirschen, “Modeling of lithium-ion battery degradation for cell life assessment,” *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1131–1140, 2018.

[11] Q. Huang, Q.-S. Jia, and X. Guan, “Robust Scheduling of EV Charging Load With Uncertain Wind Power Integration,” *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1043–1054, 2018.

[12] C. Luo, Y.-F. Huang, and V. Gupta, “Stochastic Dynamic Pricing for EV Charging Stations With Renewable Integration and Energy Storage,” *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1494–1505, 2018.

[13] E. E. Vlahakis, L. D. Dritsas, and G. D. Halikias, “Distributed LQR design for identical dynamically coupled systems: Application to Load Frequency

Control of multi-area Power Grid," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 4471–4476, IEEE, 2019.

[14] J. W. Stahlhut, T. J. Browne, G. T. Heydt, and V. Vittal, "Latency Viewed as a Stochastic Process and its Impact on Wide Area Power System Control Signals," *IEEE Transactions on Power Systems*, vol. 23, no. 1, pp. 84–91, 2008.

[15] R. Heydari, H. Young, F. Flores-Bahamonde, S. Vaez-Zadeh, C. González-Castaño, S. Sabzevari, and J. Rodríguez, "Model-free predictive control of grid-forming inverters with lcl filters," *IEEE Transactions on Power Electronics*, vol. 37, no. 8, pp. 9200–9211, 2022.

[16] H. A. Young, V. A. Marin, C. Pesce, and J. Rodriguez, "Simple finite-control-set model predictive control of grid-forming inverters with lcl filters," *IEEE Access*, vol. 8, pp. 81246–81256, 2020.

[17] M. Arbabzadeh, R. Sioshansi, J. X. Johnson, and G. A. Keoleian, "The role of energy storage in deep decarbonization of electricity production," *Nature Communications*, vol. 10, p. 3413, Jul 2019.

[18] P. Denholm, E. Ela, B. Kirby, and M. Milligan, "Role of energy storage with renewable electricity generation," tech. rep., National Renewable Energy Lab (NREL), 2010.

[19] D. Krishnamurthy, C. Uckun, Z. Zhou, P. R. Thimmapuram, and A. Botterud, "Energy storage arbitrage under day-ahead and real-time price uncertainty,"

IEEE Transactions on Power Systems, vol. 33, no. 1, pp. 84–93, 2018.

[20] N. Padmanabhan, M. Ahmed, and K. Bhattacharya, "Battery energy storage systems in energy and reserve markets," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 215–226, 2020.

[21] Y. Shi, B. Xu, D. Wang, and B. Zhang, "Using battery storage for peak shaving and frequency regulation: Joint optimization for superlinear gains," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 2882–2894, 2018.

[22] S. Gupta, V. Kekatos, and W. Saad, "Optimal real-time coordination of energy storage units as a voltage-constrained game," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3883–3894, 2019.

[23] T. Morstyn, B. Hredzak, R. P. Aguilera, and V. G. Agelidis, "Model predictive control for distributed microgrid battery energy storage systems," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 3, pp. 1107–1114, 2017.

[24] A. Oshnoei, M. Kheradmandi, and S. M. Muyeen, "Robust control scheme for distributed battery energy storage systems in load frequency control," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4781–4791, 2020.

[25] K. Meng, Z. Y. Dong, Z. Xu, and S. R. Weller, "Cooperation-driven distributed model predictive control for energy storage systems," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2583–2585, 2015.

[26] B. J. Kirby, "Frequency regulation basics and trends," 5 2005.

[27] C. Zou, C. Manzie, and D. Nešić, “Model predictive control for lithium-ion battery optimal charging,” *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 2, pp. 947–957, 2018.

[28] D. M. Rosewater, D. A. Copp, T. A. Nguyen, R. H. Byrne, and S. Santoso, “Battery energy storage models for optimal control,” *IEEE Access*, vol. 7, pp. 178357–178391, 2019.

[29] R. Spotnitz, “Simulation of capacity fade in lithium-ion batteries,” *Journal of Power Sources*, vol. 113, no. 1, pp. 72–80, 2003.

[30] J. Cao, D. Harrold, Z. Fan, T. Morstyn, D. Healey, and K. Li, “Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model,” *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4513–4521, 2020.

[31] Y. Shi, B. Xu, Y. Tan, and B. Zhang, “A convex cycle-based degradation model for battery energy storage planning and operation,” in *2018 Annual American Control Conference (ACC)*, pp. 4590–4596, 2018.

[32] T. Jónsson, *Forecasting and decision-making in electricity markets with focus on wind energy*. PhD thesis, 2012.

[33] D. Zhu and Y.-J. A. Zhang, “Optimal coordinated control of multiple battery energy storage systems for primary frequency regulation,” *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 555–565, 2019.

[34] “PJM Manual 12: Balancing Operations.” https://pjm.com/ /media/documents /manuals/m12-redline.ashx. Accessed: 2022-03-01.

[35] J. Wang, P. Liu, J. Hicks-Garner, E. Sherman, S. Soukiazian, M. Verbrugge, H. Tataria, J. Musser, and P. Finamore, “Cycle-life model for graphite-LiFePO4 cells,” *Journal of power sources*, vol. 196, no. 8, pp. 3942–3948, 2011.

[36] B. Foggo and N. Yu, “Improved battery storage valuation through degradation reduction,” *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5721–5732, 2017.

[37] L.-J. Lin, “Self-improving reactive agents based on reinforcement learning, planning and teaching,” *Machine Learning*, vol. 8, pp. 293–321, May 1992.

[38] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, pp. 529–533, Feb 2015.

[39] “Market Prices - ERCOT.” http://www.ercot.com/mktinfo/prices. Accessed: 2021-06-07.

[40] “*PJM Ancillary Services.* [Online].” Available: https://www.pjm.com/markets-and-operations/ancillary-services.aspx.

[41] “*Keras: The Python Deep Learning Library.*.” Available: https://keras.io.

[42] I. E. Agency, “Global EV Outlook 2020,” tech. rep., International Energy Agency, June 2020.

[43] Z. Stevic and I. Radovanovic, “Energy Efficiency of Electric Vehicles,” *New Generation of Electric Vehicles, edited by Z. Stevic (Intech, Rijeka, 2012)*, 2012.

[44] X. Hu, N. Chen, N. Wu, and B. Yin, “The Potential Impacts of Electric Vehicles on Urban Air Quality in Shanghai City,” *Sustainability*, vol. 13, no. 2, 2021.

[45] W. Tang, S. Bi, and Y. J. Zhang, “Online coordinated charging decision algorithm for electric vehicles without future information,” *IEEE Transactions on Smart Grid*, vol. 5, no. 6, pp. 2810–2824, 2014.

[46] H. Zhang, Z. Hu, Z. Xu, and Y. Song, “Optimal Planning of PEV Charging Station With Single Output Multiple Cables Charging Spots,” *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2119–2128, 2017.

[47] Q. Yan, B. Zhang, and M. Kezunovic, “Optimized Operational Cost Reduction for an EV Charging Station Integrated With Battery Energy Storage and PV Generation,” *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2096–2106, 2019.

[48] Q. Chen, N. Liu, C. Hu, L. Wang, and J. Zhang, “Autonomous Energy Management Strategy for Solid-State Transformer to Integrate PV-Assisted EV

Charging Station Participating in Ancillary Service," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 1, pp. 258–269, 2017.

[49] M. Alizadeh, H.-T. Wai, M. Chowdhury, A. Goldsmith, A. Scaglione, and T. Javidi, "Optimal pricing to manage electric vehicles in coupled power and transportation networks," *IEEE Transactions on control of network systems*, vol. 4, no. 4, pp. 863–875, 2016.

[50] F. He, D. Wu, Y. Yin, and Y. Guan, "Optimal deployment of public charging stations for plug-in hybrid electric vehicles," *Transportation Research Part B: Methodological*, vol. 47, pp. 87–101, 2013.

[51] K. Zhang, L. Lu, C. Lei, H. Zhu, and Y. Ouyang, "Dynamic operations and pricing of electric unmanned aerial vehicle systems and power networks," *Transportation Research Part C: Emerging Technologies*, vol. 92, pp. 472–485, 2018.

[52] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5246–5257, 2019.

[53] 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, 2020.

[54] S. Wang, S. Bi, and Y. A. Zhang, "Reinforcement Learning for Real-Time

Pricing and Scheduling Control in EV Charging Stations," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 849–859, 2021.

[55] N. Chen, C. Kurniawan, Y. Nakahira, L. Chen, and S. H. Low, "Smoothed least-laxity-first algorithm for EV charging," *CoRR*, vol. abs/2102.08610, 2021.

[56] R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation*, vol. 12. MIT Press, 2000.

[57] R. Givan, T. Dean, and M. Greig, "Equivalence notions and model minimization in Markov decision processes," *Artificial intelligence*, vol. 147, no. 1, pp. 163–223, 2003.

[58] K. Doya, "Reinforcement Learning in Continuous Time and Space," *Neural Comput.*, vol. 12, p. 219–245, Jan. 2000.

[59] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High Dimensional Continuous Control Using Generalized Advantage Estimation," 2018.

[60] "UC Davis. Richards ave station arrivals, 2019." http://anson.ucdavis.edu/clarkf/richards.csv.gz. Accessed: 2021-06-07.

[61] S. K. Pandey, S. R. Mohanty, and N. Kishor, "A literature survey on load frequency control for conventional and distribution generation power systems," *Renewable and Sustainable Energy Reviews*, vol. 25, pp. 318–334, 2013.

[62] S. Wen, X. Yu, Z. Zeng, and J. Wang, "Event-triggering load frequency control for multiarea power systems with communication delays," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 2, pp. 1308–1317, 2016.

[63] S. Zhu, Y. Zhang, and P. Chang, "Load frequency control of multi-area interconnected power system with renewable energy," in *2021 IEEE Sustainable Power and Energy Conference (iSPEC)*, pp. 1814–1817, 2021.

[64] S. Falahati, S. A. Taher, and M. Shahidehpour, "Grid secondary frequency control by optimized fuzzy control of electric vehicles," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5613–5621, 2018.

[65] P. Kundur and N. Balu, *Power System Stability and Control*. EPRI power system engineering series, McGraw-Hill, 1994.

[66] T. Yang, Y. Zhang, W. Li, and A. Y. Zomaya, "Decentralized networked load frequency control in interconnected power systems based on stochastic jump system theory," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4427–4439, 2020.

[67] F. Zhao, K. You, and T. Başar, "Global Convergence of Policy Gradient Primal-dual Methods for Risk-constrained LQRs," *arXiv preprint arXiv:2104.04901*, 2021.

[68] D. M. Andrade, S. Gamboa, and J. A. Torres, "Distributed load-frequency control in power systems," in *2020 IEEE ANDESCON*, pp. 1–6, 2020.

[69] H. Bevrani, *Robust Power System Frequency Control.* Power Electronics and Power Systems, Springer US, 2008.

[70] V. Kučera, *Algebraic Riccati Equation: Hermitian and Definite Solutions*, pp. 53–88. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991.

[71] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," *arXiv preprint arXiv:1907.08921*, 2019.

[72] Y. Li, Y. Tang, R. Zhang, and N. Li, "Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach," *IEEE Transactions on Automatic Control*, 2021.

[73] S. Mukherjee, H. Bai, and A. Chakrabortty, "Reduced-dimensional reinforcement learning control using singular perturbation approximations," *Automatica*, vol. 126, p. 109451, 2021.

[74] S. Mukherjee and T. L. Vu, "Reinforcement learning of structured stabilizing control for linear systems with unknown state matrix," *IEEE Transactions on Automatic Control*, pp. 1–1, 2022.

[75] K.-b. Kwon, L. Ye, V. Gupta, and H. Zhu, "Model-free Learning for Risk-constrained Linear Quadratic Regulator with Structured Feedback in Networked Systems," *arXiv preprint arXiv:2204.01779*, 2022.

[76] Z. Yan and Y. Xu, "A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4599–4608, 2020.

[77] W. Cui, Y. Jiang, and B. Zhang, “Reinforcement learning for optimal primary frequency control: A lyapunov approach,” 2020.

[78] A. Tsiamis, D. S. Kalogerias, L. F. O. Chamon, A. Ribeiro, and G. J. Pappas, “Risk-Constrained Linear-Quadratic Regulators,” in *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 3040–3047, 2020.

[79] A. Tsiamis, D. S. Kalogerias, A. Ribeiro, and G. J. Pappas, “Linear Quadratic Control with Risk Constraints,” *arXiv preprint arXiv:2112.07564*, 2021.

[80] T. Lin, C. Jin, and M. Jordan, “On gradient descent ascent for nonconvex-concave minimax problems,” in *International Conference on Machine Learning*, pp. 6083–6093, PMLR, 2020.

[81] J. C. Spall, “A one-measurement form of simultaneous perturbation stochastic approximation,” *Automatica*, vol. 33, no. 1, pp. 109–112, 1997.

[82] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, “Derivative-free methods for policy optimization: Guarantees for linear quadratic systems,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2916–2925, PMLR, 2019.

[83] E. Vlahakis, L. Dritsas, and G. Halikias, “Distributed LQR design for a class of large-scale multi-area power systems,” *Energies*, vol. 12, no. 14, p. 2664, 2019.

[84] H. Feng and J. Lavaei, “On the Exponential Number of Connected Components for the Feasible Set of Optimal Decentralized Control Problems,” in *2019 American Control Conference (ACC)*, pp. 1430–1437, 2019.

[85] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[86] C. Jin, P. Netrapalli, and M. Jordan, “What is local optimality in nonconvex-nonconcave minimax optimization?,” in *International Conference on Machine Learning*, pp. 4880–4889, PMLR, 2020.

[87] J. Lian, S. Wang, M. A. Elizondo, J. Hansen, R. Huang, R. Fan, H. Kirkham, L. D. Marinovici, D. Schoenwald, and F. Wilches-Bernal, “Universal Wide-Area Damping Control for Mitigating Inter-Area Oscillations in Power Systems,” tech. rep., Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2017.

[88] Z. Shi, J. Li, H. I. Nurdin, and J. E. Fletcher, “Comparison of Virtual Oscillator and Droop Controlled Islanded Three-Phase Microgrids,” *IEEE Transactions on Energy Conversion*, vol. 34, no. 4, pp. 1769–1780, 2019.

[89] A. Kumar and M. Bhadu, “Wide-Area Damping Control System for Large Wind Generation with Multiple Operational Uncertainty,” *Electric Power Systems Research*, vol. 213, p. 108755, 2022.

[90] L. Zacharia, L. Hadjidemetriou, and E. Kyriakides, “Integration of Renewables Into the Wide Area Control Scheme for Damping Power Oscillations,”

IEEE Transactions on Power Systems, vol. 33, no. 5, pp. 5778–5786, 2018.

[91] C. Liu, G. Cai, W. Ge, D. Yang, C. Liu, and Z. Sun, “Oscillation Analysis and Wide-Area Damping Control of DFIGs for Renewable Energy Power Systems Using Line Modal Potential Energy,” *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3460–3471, 2018.

[92] K. Mahapatra, M. Ashour, N. R. Chaudhuri, and C. M. Lagoa, “Malicious Corruption Resilience in PMU Data and Wide-Area Damping Control,” *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 958–967, 2020.

[93] A. Patel, S. Roy, and S. Baldi, “Wide-Area Damping Control Resilience Towards Cyber-Attacks: A Dynamic Loop Approach,” *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 3438–3447, 2021.

[94] K. Sun, W. Qiu, Y. Dong, C. Zhang, H. Yin, W. Yao, and Y. Liu, “WAMS-Based HVDC Damping Control for Cyber Attack Defense,” *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 702–713, 2023.

[95] M. E. C. Bento, “A Hybrid Particle Swarm Optimization Algorithm for the Wide-Area Damping Control Design,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 1, pp. 592–599, 2022.

[96] Y. Zhou, J. Liu, Y. Li, C. Gan, H. Li, and Y. Liu, “A Gain Scheduling Wide-Area Damping Controller for the Efficient Integration of Photovoltaic Plant,” *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 1703–1715, 2019.

[97] M. Li and Y. Chen, “A Wide-Area Dynamic Damping Controller Based on Robust H∞ Control for Wide-Area Power Systems With Random Delay and Packet Dropout,” *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4026–4037, 2018.

[98] T. Surinkaew and I. Ngamroo, “Inter-Area Oscillation Damping Control Design Considering Impact of Variable Latencies,” *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 481–493, 2019.

[99] R. K. Pandey and D. K. Gupta, “Integrated Multi-Stage LQR Power Oscillation Damping FACTS Controller,” *CSEE Journal of Power and Energy Systems*, vol. 4, no. 1, pp. 83–91, 2018.

[100] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, “Sparse and Optimal Wide-Area Damping Control in Power Networks,” in *2013 American Control Conference*, pp. 4289–4294, 2013.

[101] S. Mukherjee, A. Chakrabortty, H. Bai, A. Darvishi, and B. Fardanesh, “Scalable Designs for Reinforcement Learning-Based Wide-Area Damping Control,” *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2389–2401, 2021.

[102] J. Guo, I. Zenelis, X. Wang, and B.-T. Ooi, “WAMS-Based Model-Free Wide-Area Damping Control by Voltage Source Converters,” *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 1317–1327, 2021.

[103] B. Pang, H. Nian, C. Wu, and F. Blaabjerg, “Damping Control of High-Frequency Resonance based on Voltage Feedforward for Voltage Source Con-

verter under a Parallel Compensated Grid," *IET Power Electronics*, vol. 13, no. 13, pp. 2682–2691, 2020.

[104] A. Thakallapelli and S. Kamalasadan, "Measurement-based Wide-Area Damping of Inter-Area Oscillations based on MIMO Identification," *IET Generation, Transmission & Distribution*, vol. 14, no. 13, pp. 2464–2475, 2020.

[105] J. Duan, H. Xu, and W. Liu, "Q-Learning-Based Damping Control of Wide-Area Power Systems Under Cyber Uncertainties," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6408–6418, 2018.

[106] P. Gupta, A. Pal, and V. Vittal, "Coordinated Wide-Area Damping Control Using Deep Neural Networks and Reinforcement Learning," *IEEE Transactions on Power Systems*, vol. 37, no. 1, pp. 365–376, 2022.

[107] Y. Hashmy, Z. Yu, D. Shi, and Y. Weng, "Wide Area Measurement System-based Low Frequency Oscillation Damping Control through Reinforcement Learning," 2020.

[108] S. Tu and B. Recht, "The Gap between Model-based and Model-Free Methods on the Linear Quadratic Regulator: An Asymptotic Viewpoint," in *Conference on Learning Theory*, pp. 3036–3083, PMLR, 2019.

[109] L. Ye, H. Zhu, and V. Gupta, "On the Sample Complexity of Decentralized Linear Quadratic Regulator With Partially Nested Information Structure," *IEEE Transactions on Automatic Control*, vol. 68, no. 8, pp. 4841–4856, 2023.

[110] X. Fan, J. Shu, and B. Zhang, "Coordinated Control of DC Grid and Offshore Wind Farms to Improve Rotor-Angle Stability," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4625–4633, 2018.

[111] V. Vittal, J. D. McCalley, P. M. Anderson, and A. Fouad, *Power System Control and Stability*. John Wiley & Sons, 2019.

[112] R. Preece, J. V. Milanović, A. M. Almutairi, and O. Marjanovic, "Damping of Inter-Area Oscillations in Mixed AC/DC Networks using WAMS based Supplementary Controller," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1160–1169, 2013.

[113] N. T. Trinh, I. Erlich, and S. P. Teeuwsen, "Methods for utilization of MMC-VSC- HVDC for power oscillation damping," in *2014 IEEE PES General Meeting | Conference & Exposition*, pp. 1–5, 2014.

[114] T. Weckesser, H. Jóhannsson, and J. Østergaard, "Impact of Model Detail of Synchronous Machines on Real-Time Transient Stability Assessment," in *2013 IREP Symposium Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid*, pp. 1–9, 2013.

[115] B. Pal and B. Chaudhuri, *Robust Control in Power Systems*. Springer Science & Business Media, 2006.

[116] S. Anttila, J. S. Döhler, J. G. Oliveira, and C. Boström, "Grid Forming Inverters: A Review of the State of the Art of Key Elements for Microgrid Operation," *Energies*, vol. 15, no. 15, 2022.

[117] D. B. Rathnayake, M. Akrami, C. Phurailatpam, S. P. Me, S. Hadavi, G. Jayasinghe, S. Zabihi, and B. Bahrani, “Grid Forming Inverter Modeling, Control, and Applications,” *IEEE Access*, vol. 9, pp. 114781–114807, 2021.

[118] W. Du, Z. Chen, K. P. Schneider, R. H. Lasseter, S. P. Nandanoori, F. K. Tuffner, and S. Kundu, “A Comparative Study of Two Widely Used Grid-forming Droop Controls on Microgrid Small-Signal Stability,” *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 8, no. 2, pp. 963–975, 2019.

[119] S. Peyghami, P. Davari, H. Mokhtari, and F. Blaabjerg, “Decentralized Droop Control in DC Microgrids based on a Frequency Injection Approach,” *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6782–6791, 2019.

[120] M. Ebrahimi, S. A. Khajehoddin, and M. Karimi-Ghartemani, “An Improved Damping Method for Virtual Synchronous Machines,” *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1491–1500, 2019.

[121] I. Serban and C. P. Ion, “Microgrid control based on a grid-forming inverter operating as virtual synchronous generator with enhanced dynamic response capability,” *International Journal of Electrical Power & Energy Systems*, vol. 89, pp. 94–105, 2017.

[122] D. Groß, M. Colombino, J.-S. Brouillon, and F. Dörfler, “The Effect of Transmission-Line Dynamics on Grid-Forming Dispatchable Virtual Oscillator Control,” *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 1148–1160, 2019.

[123] W. Du, Y. Liu, R. Huang, F. K. Tuffner, J. Xie, and Z. Huang, "Positive-Sequence Phasor Modeling of Droop-Controlled, Grid-Forming Inverters with Fault Current Limiting Function," in *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference*, pp. 1–5, 2022.

[124] B. Mirafzal and A. Adib, "On Grid-Interactive Smart Inverters: Features and Advancements," *IEEE Access*, vol. 8, pp. 160526–160536, 2020.

[125] M. Eskandari, A. V. Savkin, and J. Fletcher, "A Deep Reinforcement Learning-based Intelligent Grid-Forming Inverter for Inertia Synthesis by Impedance Emulation," *IEEE Transactions on Power Systems*, 2023.

[126] H. O. Omotoso, A. A. Al-Shamma'a, M. Alharbi, H. M. H. Farh, A. Alkuhayli, A. M. Abdurraqeeb, F. Alsaif, U. Bawah, and K. E. Addoweesh, "Machine Learning Supervisory Control of Grid-Forming Inverters in Islanded Mode," *Sustainability*, vol. 15, no. 10, p. 8018, 2023.

[127] H. Issa, V. Debusschere, L. Garbuio, P. Lalanda, and N. Hadjsaid, "Artificial Intelligence-Based Controller for Grid-Forming Inverter-Based Generators," in *IEEE PES Innovative Smart Grid Technologies Conference Europe*, pp. 1–6, IEEE, 2022.

[128] S. Li, M. Fairbank, C. Johnson, D. C. Wunsch, E. Alonso, and J. L. Proao, "Artificial Neural Networks for Control of a Grid-Connected Rectifier/Inverter under Disturbance, Dynamic and Power Converter Switching Conditions," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 4, pp. 738–750, 2013.

[129] M. Mehrasa, M. Babaie, M. Sharifzadeh, S. Bacha, and K. Al-Haddad, "An intelligent linearization control method for grid-tied packed e-cell inverter under load variations and parameters mismatch," in *2021 22nd IEEE International Conference on Industrial Technology (ICIT)*, vol. 1, pp. 310–315, 2021.

[130] O. Babayomi, Z. Zhang, Y. Li, and R. Kennel, "Adaptive predictive control with neuro-fuzzy parameter estimation for microgrid grid-forming converters," *Sustainability*, vol. 13, no. 13, 2021.

[131] J. Garcia and F. Fernández., "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, p. 1437–1480, 2015.

[132] D. Ernst, M. Glavic, F. Capitanescu, and L. Wehenkel, "Reinforcement learning versus model predictive control: A comparison on a power system problem," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 517–529, 2009.

[133] Y. Lin, J. McPhee, and N. L. Azad, "Comparison of deep reinforcement learning and model predictive control for adaptive cruise control," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 2, pp. 221–231, 2021.

[134] I. Saeed, T. Alpcan, S. M. Erfani, and M. B. Yilmaz, "Distributed nonlinear model predictive control and reinforcement learning," in *2019 Australian New Zealand Control Conference (ANZCC)*, pp. 1–3, 2019.

[135] P. Hildalgo-Gonzalez, D. Kammen, J. Szinai, S. Melissa, and N. Caroline, "The role of storage in the path to net zero," tech. rep., Renewable and Appropriate Energy Laboratory (RAEL), 2021.

[136] F. Altaf, B. Egardt, and L. Johannesson Mårdh, "Load management of modular battery using model predictive control: Thermal and state-of-charge balancing," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 1, pp. 47–62, 2017.

[137] K. Ojand and H. Dagdougui, "Q-learning-based model predictive control for energy management in residential aggregator," *IEEE Transactions on Automation Science and Engineering*, pp. 1–12, 2021.

[138] X. Fang, B.-M. Hodge, L. Bai, H. Cui, and F. Li, "Mean-variance optimization-based energy storage scheduling considering day-ahead and real-time lmp uncertainties," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7292–7295, 2018.

[139] Z. Guo, W. Wei, L. Chen, Z. Y. Dong, and S. Mei, "Impact of energy storage on renewable energy utilization: A geometric description," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 2, pp. 874–885, 2021.

[140] M. Vašak and G. Kujundžić, "A battery management system for efficient adherence to energy exchange commands under longevity constraints," *IEEE Transactions on Industry Applications*, vol. 54, no. 4, pp. 3019–3033, 2018.

[141] M. Islam, F. Yang, J. Hossain, C. Ekanayeke, and U. B. Tayab, "Battery energy management to minimize the grid fluctuation in residential microgrids," in *2018 Australasian Universities Power Engineering Conference (AUPEC)*, pp. 1–4, 2018.

[142] W. B. Powell and S. Meisel, "Tutorial on stochastic optimization in energy—part ii: An energy storage illustration," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1468–1475, 2016.

[143] A. S. Zamzam, B. Yang, and N. D. Sidiropoulos, "Energy storage management via deep q-networks," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, IEEE, 2019.

[144] J. Liu, H. Tang, M. Matsui, M. Takanokura, L. Zhou, and X. Gao, "Optimal management of energy storage system based on reinforcement learning," in *Proceedings of the 33rd Chinese Control Conference*, pp. 8216–8221, 2014.

[145] B. V. Mbuwir, F. Ruelens, F. Spiessens, and G. Deconinck, "Battery energy management in a microgrid using batch reinforcement learning," *Energies*, vol. 10, no. 11, 2017.

[146] S. Kim and H. Lim, "Reinforcement learning based energy management algorithm for smart energy buildings," *Energies*, vol. 11, no. 8, 2018.

[147] Y. Shi, B. Xu, B. Zhang, and D. Wang, "Leveraging energy storage to optimize data center electricity cost in emerging power markets," in *Proceedings*

of the Seventh International Conference on Future Energy Systems, e-Energy '16, (New York, NY, USA), Association for Computing Machinery, 2016.

[148] M. Ecker, N. Nieto, S. Käbitz, J. Schmalstieg, H. Blanke, A. Warnecke, and D. U. Sauer, "Calendar and cycle life study of li(nimnco)o2-based 18650 lithium-ion batteries," *Journal of Power Sources*, vol. 248, pp. 839–851, 2014.

[149] E. Even-Dar, S. M. Kakade, and Y. Mansour, "Online markov decision processes," *Mathematics of Operations Research*, vol. 34, no. 3, pp. 726–736, 2009.

[150] B. Xu, J. Zhao, T. Zheng, E. Litvinov, and D. S. Kirschen, "Factoring the cycle aging cost of batteries participating in electricity markets," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2248–2259, 2018.

[151] T. Dragičević, H. Pandžić, D. Škrlec, I. Kuzle, J. M. Guerrero, and D. S. Kirschen, "Capacity optimization of renewable energy sources and battery storage in an autonomous telecommunication facility," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 4, pp. 1367–1378, 2014.

[152] M. Musallam and C. M. Johnson, "An efficient implementation of the rainflow counting algorithm for life consumption estimation," *IEEE Transactions on Reliability*, vol. 61, no. 4, pp. 978–986, 2012.

[153] A. K. Dixit, *Optimization in economic theory / by Avinash K. Dixit*. Oxford University Press Oxford, 2nd ed. ed., 1990.

[154] D. Rosewater, R. Baldick, and S. Santoso, "Risk-averse model predictive control design for battery energy storage systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2014–2022, 2020.

[155] A. Oshnoei, M. Kheradmandi, S. M. Muyeen, and N. D. Hatziargyriou, "Disturbance observer and tube-based model predictive controlled electric vehicles for frequency regulation of an isolated power grid," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 4351–4362, 2021.

[156] V. Gupta, S. R. Konda, R. Kumar, and B. K. Panigrahi, "Multiaggregator Collaborative Electric Vehicle Charge Scheduling Under Variable Energy Purchase and EV Cancelation Events," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 2894–2902, 2018.

[157] Q. Chen, F. Wang, B.-M. Hodge, J. Zhang, Z. Li, M. Shafie-Khah, and J. P. S. Catalão, "Dynamic Price Vector Formation Model-Based Automatic Demand Response Strategy for PV-Assisted EV Charging Stations," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2903–2915, 2017.

[158] E. Akhavan-Rezai, M. F. Shaaban, E. F. El-Saadany, and F. Karray, "New EMS to Incorporate Smart Parking Lots Into Demand Response," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1376–1386, 2018.

[159] Y. Zhang and L. Cai, "Dynamic Charging Scheduling for EV Parking Lots With Photovoltaic Power System," *IEEE Access*, vol. 6, pp. 56995–57005, 2018.

[160] K. Chaudhari, A. Ukil, K. N. Kumar, U. Manandhar, and S. K. Kollimalla, “Hybrid Optimization for Economic Deployment of ESS in PV-Integrated EV Charging Stations,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 1, pp. 106–116, 2018.

[161] H. Mehrjerdi and R. Hemmati, “Stochastic model for electric vehicle charging station integrated with wind energy,” *Sustainable Energy Technologies and Assessments*, vol. 37, p. 100577, 2020.

[162] H. Fathabadi, “Novel wind powered electric vehicle charging station with vehicle-to-grid (V2G) connection capability,” *Energy Conversion and Management*, vol. 136, pp. 229–239, 2017.

[163] J. Peters and J. A. Bagnell, *"Policy Gradient Methods"*, pp. 774–776. Boston, MA: Springer US, 2010.

[164] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, “Constrained reinforcement learning has zero duality gap,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[165] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, “Natural policy gradient primal-dual method for constrained markov decision processes,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 8378–8390, 2020.

[166] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, “Sparse and optimal wide-area damping control in power networks,” in *2013 American Control Conference*, pp. 4289–4294, IEEE, 2013.

[167] A. Chakrabortty, “Wide-area control of power systems: Employing data-driven, hierarchical reinforcement learning,” *IEEE Electrification Magazine*, vol. 9, no. 1, pp. 45–52, 2021.

[168] P. Shah and P. A. Parrilo, “H2-optimal decentralized control over posets: A state-space solution for state-feedback,” *IEEE Transactions on Automatic Control*, vol. 58, no. 12, pp. 3084–3096, 2013.

[169] L. Furieri, Y. Zheng, and M. Kamgarpour, “Learning the Globally Optimal Distributed LQ Regulator,” in *L4DC*, 2020.

[170] E. Hazan, “Introduction to Online Convex Optimization,” 2021.

[171] R. Chen, “Solution of minimax problems using equivalent differentiable functions,” *Computers & mathematics with applications*, vol. 11, no. 12, pp. 1165–1169, 1985.

[172] J. Diakonikolas, C. Daskalakis, and M. I. Jordan, “Efficient methods for structured nonconvex-nonconcave min-max optimization,” in *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754, PMLR, 2021.

[173] L. Li, Y. Sun, Z. Liu, X. Hou, G. Shi, and M. Su, “A decentralized control with unique equilibrium point for cascaded-type microgrid,” *IEEE Transactions on Sustainable Energy*, vol. 10, no. 1, pp. 324–326, 2018.

[174] C. Qi, K. Wang, Q. Yang, G. Li, X. Huang, J. Wu, and M. L. Crow, “Decentralized DC voltage and power sharing control of the parallel grid converters

in multi-terminal DC power integration system," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 4, pp. 1971–1980, 2018.

[175] M. Liu, P. K. Phanivong, Y. Shi, and D. S. Callaway, "Decentralized Charging Control of Electric Vehicles in Residential Distribution Networks," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 1, pp. 266–281, 2019.

[176] C. Wang, J. Duan, B. Fan, Q. Yang, and W. Liu, "Decentralized High-Performance Control of DC Microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3355–3363, 2019.

[177] R. E. Kalman *et al.*, "Contributions to the theory of optimal control," *Bol. soc. mat. mexicana*, vol. 5, no. 2, pp. 102–119, 1960.

[178] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. USA: Prentice-Hall, Inc., 1990.

[179] P. Lancaster and L. Rodman, *Algebraic riccati equations*. Clarendon press, 1995.

[180] M. Rotkowitz and S. Lall, "A Characterization of Convex Problems in Decentralized Control," *IEEE Transactions on Automatic Control*, vol. 51, no. 2, pp. 274–286, 2006.

[181] D. Tabas and B. Zhang, "Computationally Efficient Safe Reinforcement Learning for Power Systems," *arXiv preprint arXiv:2110.10333*, 2021.

[182] K. Zhang, B. Hu, and T. Basar, "Policy Optimization for $\mathcal{H}_2$ Linear Control with $\mathcal{H}_\infty$ Robustness Guarantee: Implicit Regularization and Global Convergence," in *Learning for Dynamics and Control*, pp. 179–190, PMLR, 2020.

[183] J. D. Watson, Y. Ojo, K. Laib, and I. Lestas, "A Scalable Control Design for Grid-Forming Inverters in Microgrids," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 4726–4739, 2021.

[184] H. Kikusato, D. Orihara, J. Hashimoto, T. Takamatsu, T. Oozeki, T. Matsuura, S. Miyazaki, H. Hamada, and T. Miyazaki, "Performance Evaluation of Grid-Following and Grid-Forming Inverters on Frequency Stability in Low-Inertia Power Systems by Power Hardware-In-The-Loop Testing," *Energy Reports*, vol. 9, pp. 381–392, 2023.

[185] K.-b. Kwon, L. Ye, V. Gupta, and H. Zhu, "Model-free learning for risk-constrained linear quadratic regulator with structured feedback in networked systems," in *IEEE 61st Conference on Decision and Control (CDC)*, pp. 7260–7265, 2022.

[186] M. N. Ambia, K. Meng, W. Xiao, A. Al-Durra, and Z. Y. Dong, "Interactive Grid Synchronization-Based Virtual Synchronous Generator Control Scheme on Weak Grid Integration," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 4057–4071, 2022.

[187] M. Chen, D. Zhou, A. Tayyebi, E. Prieto-Araujo, F. Dörfler, and F. Blaabjerg, "Generalized Multivariable Grid-Forming Control Design for Power Con-

verters," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2873–2885, 2022.

[188] R. Ghosh, N. R. Tummuru, and B. S. Rajpurohit, "A New Virtual Oscillator-Based Grid-Forming Controller with Decoupled Control Over Individual Phases and Improved Performance of Unbalanced Fault Ride-Through," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 12, pp. 12465–12474, 2023.

[189] G. Ravikumar and M. Govindarasu, "Anomaly Detection and Mitigation for Wide-Area Damping Control using Machine Learning," *IEEE Transactions on Smart Grid*, pp. 1–1, 2020.

[190] I. Zenelis and X. Wang, "Wide-Area Damping Control for Interarea Oscillations in Power Grids Based on PMU Measurements," *IEEE Control Systems Letters*, vol. 2, no. 4, pp. 719–724, 2018.

[191] W. Yao, "Resilient Wide-Area Damping Control Using GrHDP to Tolerate Communication Failures," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–1, 2019.

[192] X. Shi, Y. Cao, M. Shahidehpour, Y. Li, X. Wu, and Z. Li, "Data-Driven Wide-Area Model-Free Adaptive Damping Control With Communication Delays for Wind Farm," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5062–5071, 2020.

[193] A. H. Mohamed, A. E. Hussein, and M. A. Abido, "Delay-Independent

Wide-Area Damping Control Using Scattering Transformation," *Arabian Journal for Science and Engineering*, vol. 46, pp. 9465–9474, Oct 2021.

[194] I. Nacef, K. B. Kilani, and M. Elleuch, "Understanding Interarea Oscillations in Power Systems Integrating Wind Power," in *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, pp. 1–6, 2018.

[195] Y. Zhao, W. Yao, C.-K. Zhang, X.-C. Shangguan, L. Jiang, and J. Wen, "Quantifying Resilience of Wide-Area Damping Control Against Cyber Attack Based on Switching System Theory," *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 2331–2343, 2022.

[196] S. Roy, A. Patel, and I. N. Kar, "Analysis and Design of a Wide-Area Damping Controller for Inter-Area Oscillation With Artificially Induced Time Delay," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3654–3663, 2019.

[197] T. Surinkaew and I. Ngamroo, "Impact of Variable Time Delay on Oscillatory Stability in Power System with Wind and Solar Farms using Wide Area Damping Control," *IFAC-PapersOnLine*, vol. 52, no. 4, pp. 81–86, 2019.

[198] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, Y. Yang, and A. Knoll, "A Review of Safe Reinforcement Learning: Methods, Theory and Applications," 2022.

[199] A. Yazdani and R. Iravani, *Voltage-sourced Converters in Power Systems: Modeling, Control, and Applications*. John Wiley & Sons, 2010.

[200] R. Ghosh, N. R. Tummuru, B. S. Rajpurohit, and A. Monti, “Virtual Inertia from Renewable Energy Sources: Mathematical Representation and Control Strategy,” in *2020 IEEE International Conference on Power Electronics, Smart Grid and Renewable Energy (PESGRE2020)*, pp. 1–6, 2020.

[201] J. Arif, S. Ray, and B. Chaudhuri, “Multivariable Self-Tuning Feedback Linearization Controller for Power Oscillation Damping,” *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1519–1526, 2014.

[202] Y. Xu, Z. Qu, R. Harvey, and T. Namerikawa, “Data-Driven Wide-Area Control Design of Power System Using the Passivity Shortage Framework,” *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 830–841, 2021.

[203] R. Xie, I. Kamwa, and C. Y. Chung, “A Novel Wide-Area Control Strategy for Damping of Critical Frequency Oscillations via Modulation of Active Power Injections,” *IEEE Transactions on Power Systems*, vol. 36, no. 1, pp. 485–494, 2021.

[204] Y. Li, C. Rehtanz, S. Ruberg, L. Luo, and Y. Cao, “Wide-Area Robust Coordination Approach of HVDC and FACTS Controllers for Damping Multiple Interarea Oscillations,” *IEEE Transactions on Power Delivery*, vol. 27, no. 3, pp. 1096–1105, 2012.