AdapSafe: Adaptive and Safe-Certified Deep Reinforcement Learning-Based Frequency Control for Carbon-neutral Power Systems

Anonymous submission

Abstract

With the increasing penetration of inverter-based renewable energy resources, deep reinforcement learning (DRL) has been proposed as one of the most promising solutions to realize real-time and autonomous control for future carbon-neutral power systems. In particular, DRL-based frequency control approaches have been extensively investigated to overcome the limitations of model-based approaches, such as the computational cost and scalability for large-scale systems. Nevertheless, the real-world implementation of DRL-based frequency control methods is facing the following fundamental challenges: 1) safety guarantee during the learning and decision-making processes; 2) adaptability against the dynamic system operating conditions. To this end, this is the first work that proposes an **Adap**tive and **Safe**-Certified DRL (**AdapSafe**) algorithm for frequency control to simultaneously address the aforementioned challenges. In particular, a novel self-tuning control barrier function is designed to actively compensate the unsafe frequency control strategies under variational safety constraints and thus achieve guaranteed safety. Furthermore, the concept of meta-reinforcement learning is integrated to significantly enhance its adaptiveness in non-stationary power system environments without sacrificing the safety cost. Experiments are conducted based on GB 2030 power system, and the results demonstrate that the proposed AdapSafe exhibits superior performance in terms of its guaranteed safety in both training and test phases, as well as its considerable adaptability against the dynamics changes of system parameters.

Introduction

To achieve the carbon-neutral target, high penetration of renewable energy sources (RES) is expected in power systems (Zhang et al. 2020). Nevertheless, the wide deployment of power electronic devices for RES integration renders the decrease of system synchronous inertia (Ulbig, Borsche, and Andersson 2014), which may endanger the frequency stability or limit the penetration of RES (Teng, Trovato, and Strbac 2015). For example, the UK suffered the most significant power outage in over a decade in 2019, lasting more than 1.5 hours, causing widespread disruption to the traffic light network and affecting about 1 million people. The main driver of this blackout is the rapid frequency changes as a significant amount of new devices and distributed resources were added to the system in a very short time (Bialek 2020).

In the literature, load frequency control (LFC) has been extensively studied to ensure frequency stability and improve power quality in power systems. Traditional model-based LFC methods are mainly designed based on the PID control methods (Tan 2009; Zhang et al. 2013), and their effectiveness is highly dependent on the specific physical model and the precision of parameters. In order to achieve self-tuning of PID parameters under model uncertainty, Ranjit et al. (Singh and Ramesh 2019) proposed a PID controller with filtering for the LFC problem in a two-region interconnected grid containing PV generation, which exhibits strong robustness under parameter uncertainty. Furthermore, fuzzy control (Sahu, Panda, and Pradhan 2015; Debnath, Jena, and Sanyal 2019), BAT optimization algorithm (Abd-Elazim and Ali 2016) and model predictive control (Ersdal, Imsland, and Uhlen 2015) also show improvements to a certain extent compared to the traditional PID methods. However, it is imperative to note that most model-based methods are computationally expensive and thus significantly limit their implementation and performance for real-time LFC scenarios under uncertainty.

To this end, the data-driven approach, especially the Deep Reinforcement Learning (DRL) control scheme, is considered as a promising solution to realize the real-time LFC, especially in the carbon-neutral case, due to its capability to deal with dynamic uncertainty and sequential decision-making problems. In particular, Yin et al. (Yin et al. 2017) proposed a novel emotional reinforcement learning (ERL) with nine control strategies for designing controllers in a two-area LFC power system, where each strategy is combined with different converting functions, reward function, and learning rate of RL. Moreover, a collaborative multi-agent DRL method is proposed for multi-area power systems (Yan and Xu 2020). The reward function is set up to minimize the regional control error signals in all regions, and the performance of the proposed method is demonstrated on the New-England 39-bus system. For the existing DRL-based LFC approaches, although fully exploring the whole action space during the learning process can assist in obtaining the optimal strategy, it may also lead to unsafe actions that need to be avoided. Therefore, it is of great importance to balance the exploration and safety of a DRL-based LFC method.

To solve the aforementioned challenge, safe DRL-based control methods have been proposed in the machine learning community. Xia et al. (Xia et al. 2022) established a safety evaluation network to generate safe frequency control actions based on the historical data of off-line learning. In addition, Gupta et al. (Gupta, Pal, and Vittal 2021) designed a DDPG algorithm based on bounded exploration control (BEC), which can adapt to variational environments with uncertainty and unexpected scenarios. Nevertheless, the aforementioned safe DRL approaches lack a theoretical guarantee for safety and can not take into account the time-varying characteristics of the inherent parameters in the real physical system. Thereby, it is risky to directly implement existing approaches to real-world cyber-physical systems, such as power systems, where safety and adaptability are both crucial and fundamental requirements.

More specifically, it is of great importance to develop a DRL-based LFC approach that can jointly deal with the fundamental challenges of 1) safety guarantee during the learning and decision-making processes; 2) adaptability against the dynamic system operating conditions. Therefore, this paper aims to fill this research gap and investigate the fundamental limitations of existing approaches by proposing a novel adaptive and safe-certified DRL-based LFC method with the integration of meta-reinforcement learning and self-tuning control barrier function (CBF). To summarize, this study makes the following original contributions:

(1) To the best of the authors' knowledge, this is the first work that proposes an **Adap**tive and **Safe**-Certified DRL (**AdapSafe**) algorithm for power system frequency control to handle the dynamic system operating conditions while providing the guarantee that the executed control actions are safe during both the offline learning and online implementing procedures.

(2) Based on the proposed AdapSafe, we further improve the efficiency of the meta-training phase using transition post-processing and noise elimination measures. In addition, parameter self-tuning is proposed for our CBF-based compensator, thus enabling the adjustment of the compensation according to the safety risk level.

(3) The effectiveness of the proposed AdapSafe is demonstrated based on a GB 2030 power system, and the results demonstrate that AdapSafe can minimize the control cost without sacrificing safety.

Background and Problem Definition

Power System Frequency Control

A typical frequency trajectory after a generator outage, shown in Figure 1, can be divided into four response stages (Zhang et al. 2020). Firstly, the inertial response takes effect immediately after the accident. At this stage, due to the control deadband, the rate of frequency (RoCoF) $\Delta\dot{f}$ is solely determined by the system inertia. Then, when the frequency exceeds the dead zone, the thermal unit governors and frequency controllers of RES gradually contaminate the frequency decline and pull the frequency back to the quasi-steady state by adjusting their active power output (primary frequency control). Finally, the automatic generation control of generators is activated to reset the system frequency to its nominal value (secondary frequency control), followed by a generation redispatch to achieve optimal economic operation and get ready for the next possible emergency (tertiary frequency control). In this paper, without loss of generality, we combine inertial response and primary frequency control into phase I, the secondary and tertiary control as phase II.

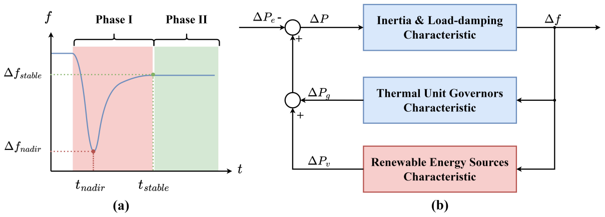

Figure 1: Power system frequency response: (a) typical system frequency trajectory after a generator outage; (b) block diagram of system frequency dynamics

.

In the end, the frequency response dynamics can be expressed by the following swing equation (Kundur, Balu, and Lauby 1994; Zhang et al. 2020):

$$2H\Delta\dot{f}(t) + D\Delta f(t) = \Delta P_g(t) + \Delta P_v(t) - \Delta P_e(t) \quad (1)$$

where H and D are the system inertia and load damping constant, respectively; $\Delta f(t)$, $\Delta P_g(t)$, $\Delta P_v(t)$ and $\Delta P_e(t)$ are the deviations of frequency, generator generation, RES output power adjustment and total power imbalance, respectively.

Frequency Control as a CMDP

In order to solve the above LFC problem subject to safety constraints, we can naturally transform it into a Constrained Markov Decision Process (CMDP) (Altman 1999). A CMDP is an MDP with constraints that limit the permissible policy set. For the model-based method, it can be set as the form of a tuple: $< \mathcal{S}, \mathcal{A}, f, g, \mathcal{R}, \mathcal{C} >$, where $\mathcal{S}$ is the continuous state set, i.e., $\mathcal{S} \subset \mathbb{R}^m$, $\mathcal{A}$ is the continuous action set, i.e., $\mathcal{A} \subset \mathbb{R}^n$, $f:\ \mathbb{R}^m \rightarrow \mathbb{R}^m$ is the unactuated dynamics, $g:\ \mathbb{R}^m \rightarrow \mathbb{R}^n$ is the actuated dynamics, and the state transitions can be defined as:

$$s_{t+1} = s_t + (f(s_t) + g(s_t)\, a_t)\Delta t \quad (2)$$

where $s_t \in \mathcal{S}$, $a_t \in \mathcal{A}$; Δt is the interaction time step and the system dynamics f and g are not immutable. In the later experiment, we set up a $\hat{f}$ and $\hat{g}$ as nominal system dynamics to simulate the difference between the nominal and the actual environments. In addition, $\mathcal{R}$ is an immediate reward after a state transition, and $\mathcal{C}$ is an additional set of safety constraint which includes state constraints and action constraints, i.e., $\mathcal{C} := \{c_1, c_2, ...c_i\},\ i \in 1, 2, ..., N_c$ and each c_i can be equality or inequality constraints. More specifically, the correspondence of the tuples of CMDP is as follows:

State $\mathcal{S}$. Since the current frequency change Δf is determined by both governor control and RES frequency control, in addition to $\Delta f(t)$, the current power output from generators $\Delta P_g(t)$ should also be a state variable.

Action $\mathcal{A}$. RES plants are generally connected to the power system through power electronic devices, and thus the output power adjustment $\Delta P_v(t)$ finally provided by the RES to the power system is considered as the action a.

System Dynamics f, g**.** Through the s-domain transformation of Eq. 1 combined with the state variables Δf and ΔP_g, the system dynamics f and g are expressed as follows:

$$f = \begin{bmatrix} \frac{-D}{2H}\Delta f + \frac{\Delta P_g - \Delta P_e}{2H} \\ \frac{1}{R_g T_g}\Delta f - \frac{\Delta P_g}{T_g} \end{bmatrix} \tag{3}$$

$$g = \begin{bmatrix} \frac{1}{2H} \\ 0 \end{bmatrix} \tag{4}$$

where a droop gain R_g and a low-pass filter with a time constant T_g is used to represent the dynamics of governor control (Chu et al. 2020).

Constraint $\mathcal{C}$**.** For frequency nadir, in phase I, our safe constraint is that Δf_{nadir} can be above the maximum dead zone limit Δf_{bound}; once the frequency is restored to a quasi-steady state, our safety constraint will become keep $\Delta \dot{f}$ within the range of $\Delta \dot{f}_{bound}$ while ensuring that the steady-state frequency is above a certain limit Δf_{stable}; at the same time, the action bound that RES can execute is also one of the constraints.

$$\mathcal{C} := \begin{cases} c_1 : \{\Delta f_{nadir} \geq \Delta f_{bound}\} \\ c_2 : \{\Delta f \geq \Delta f_{stable}\} \\ c_3 : \{|\Delta \dot{f}| \leq \Delta \dot{f}_{bound}\} \\ c_4 : \{\Delta P_{min} \leq \Delta P_e < \Delta P_{max}\} \end{cases} \tag{5}$$

where $\Delta \dot{f}$ is used to measure stability. In phase I, the controller needs to meet constraints c_1 and c_4, while in phase II, the constraints become c_2, c_3 and c_4.

Reward $\mathcal{R}$**.** As the safety constraints set $\mathcal{C}$ will change according to the phase of the state, a segmented reward function based on the frequency control action a is set to adjust the system state, as follows:

$$r = \begin{cases} -100, & \text{if } \Delta f_{nadir} < \Delta f_{bound} \\ -m_1|a| + (-1)^q m_2, & \text{else if } |\Delta \dot{f}| \leq \Delta \dot{f}_{bound} \\ -m_1|a| - m_3 t, & \text{otherwise} \end{cases} \tag{6}$$

where q indicates whether the next state satisfy $\Delta f > \Delta f_{stable}$; $m_1, m_2, m_3 \in \mathbb{R}+$ are the penalty coefficients of the performed action, system stability and time step, respectively.

AdapSafe: Adaptive and Safe-Certified DRL-based Frequency Control

Problem Formulation for AdapSafe

Let $\mathcal{S}_{safe}$ and $\mathcal{A}_{safe}$ denote the sets of safe state and safe action, respectively, indicating that if the current state s_t and action a_t are in the safe set, all the constraints in $\mathcal{C}$ can be satisfied and the corresponding next state s_{t+1} will also be safe. To guarantee that $a_{rl} \in \mathcal{A}_{safe}$ and $s_{t+1} \in \mathcal{S}_{safe}$ can be jointly satisfied, a compensation action u is introduced in this work. Considering the dynamics of the investigated non-stationary system, the safety policy π has to be adaptively adjusted in response to different environments to maximize the overall rewards. To this end, the overall objective of the optimization problem is to maximize the average episodes' rewards for multi-tasks while satisfying safety constraints with minimum compensation costs u, as expressed below:

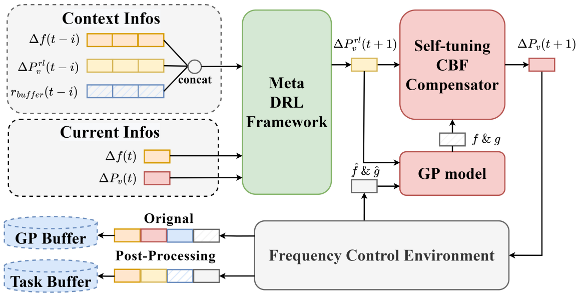

Figure 2: The control framework of the proposed AdapSafe for power system frequency control.

.

$$\max_{\pi} \frac{1}{N_{task}} \sum_{k=1}^{N_{task}} \min_{u_t} \mathbb{E}_{\tau^k \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t r^k (s_t, a_{rl} + u_t) \right] \tag{7}$$
$$\text{s.t.: } s_{t+1} \in \mathcal{S}_{safe} \subset \mathcal{S}$$
$$a_t = (a_{rl} + u_t) \in \mathcal{A}_{safe} \subseteq \mathcal{A}$$

where $\gamma \in [0, 1)$ is the discount factor and a_{rl} is the deterministic action of actor network output.

This section aims to introduce the proposed AdapSafe, which includes two main steps 1) self-tuning safety certification; and 2) adaptiveness enhancement, as presented in Figure 2.

Step 1: Self-tuning Safety Certification

To achieve the objective above, the safe state set $\mathcal{S}_{safe}$ is first defined as:

$$\begin{aligned} \mathcal{S}_{safe} &= \{s \in \mathbb{R}^n : h(s) \geq 0\} \\ \partial \mathcal{S}_{safe} &= \{s \in \mathbb{R}^n : h(s) = 0\} \\ \mathrm{Int}(\mathcal{S}_{safe}) &= \{s \in \mathbb{R}^n : h(s) > 0\} \end{aligned} \tag{8}$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, $\partial \mathcal{S}_{safe}$ and $\mathrm{Int}(\mathcal{S}_{safe})$ represent the boundary and interior of set $\mathcal{S}_{safe}$, respectively.

Inspired by using the Lyapunov-like function to prove the stability of the set without calculating the exact solution of the system (Sloth, Wisniewski, and Pappas 2012; Wisniewski and Sloth 2015), this paper employs the barrier function to ensure the forward invariance of the safe set (see **Definition 1**) and avoid the high computational complexity of calculating all state reachable sets in every time step. Compared with the Lyapunov-based methods (Berkenkamp et al. 2017; Chow et al. 2018), barrier function-based approaches (Cheng et al. 2019) relax the condition to guarantee the forward invariant property in the safe set and significantly increase the feasibility.

Definition 1. A set $\mathcal{S}$ is *forward invariant* for a dynamical system $\dot{x} = f(x)$ if $x_0 \in \mathcal{S} \to \varphi(t, x_0) \in \mathcal{S}$ for all positive t where $\varphi(t, x_0)$ is the flow of f starting at (t, x_0). In our framework, that means once the state s_0 is in the safe set $\mathcal{S}_{safe}$, during the entire episode, s_t will not cross the boundary of $\mathcal{S}_{safe}$.

To this end, we inherit the setting of barrier functions in (Ames et al. 2016), and divide the selection of barrier functions into reciprocal barrier functions (RBF) and zeroing barrier functions(ZBF). The difference between them lies in the different processing of the safe set and the corresponding forward invariance guarantee. For ZBF, $h(\cdot)$ can be directly selected as a candidate; however, the selection of RBF must have the following properties:

$$\inf_{x \in \text{Int}(\mathcal{C})} B(x) \geq 0, \quad \lim_{x \to \partial \mathcal{C}} B(x) = \infty \tag{9}$$

On this basis, the corresponding reciprocal control barrier functions (RCBF) and zeroing control barrier functions (ZCBF) can be developed and ensure the forward invariance for AdapSafe, which are defined in **Definition 2** and **Definition 3**, respectively.

Definition 2. Given a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ and a set $\mathcal{S}_{safe} \subset \mathbb{R}^n$ defined by Eq. 8, the function $h(\cdot)$ is called a *ZCBF* defined on the state space $\mathcal{S}$ with $\mathcal{S}_{safe} \subseteq \mathcal{S} \subset \mathbb{R}^n, \forall s_t \in \mathcal{S}$, if there exists a continuous function $\alpha : (-b, a) \to (-\infty, \infty)$ such that for all $s_t \in \text{Int}(\mathcal{S}_{safe})$:

$$\sup_{u_t \in \mathcal{A}} \left[L_f h(s_t) + L_g h(s_t)(a_{rl} + u_t) + \alpha(h(s_t)) \right] \geq 0 \tag{10}$$

where L_f and L_g are the lie derivative of the environment dynamics f and g, respectively. Then any feasible compensation $u_t \in U_{\text{ZCBF}}(s_t) = \{(a_{rl} + u_t) \in \mathcal{A} : L_f h(s_t) + L_g h(s_t)(a_{rl} + u) + \alpha(h(s_t)) \geq 0\}$ will render the set $\mathcal{S}_{safe}$ forward invariant in time t.

Definition 3. Given a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ and a set $\mathcal{S}_{safe} \subset \mathbb{R}^n$ defined by Eq. 8. A continuously differentiable function $B : \text{Int}(\mathcal{S}_{safe}) \to \mathbb{R}$ is called a *RCBF* if there exist three continuous functions $\beta_1, \beta_2, \beta_3 : [0, a) \to [0, \infty)$ for some $a > 0$ such that, for all $s_t \in \text{Int}(\mathcal{S}_{safe})$:

$$\begin{gathered} \frac{1}{\beta_1(h(s_t))} \leq B(s_t) \leq \frac{1}{\beta_2(h(s_t))} \\ \inf_{u_t \in \mathcal{A}} \left[L_f B(s_t) + L_g B(s_t)(a_{rl} + u_t) - \beta_3(h(s_t)) \right] \leq 0 \end{gathered} \tag{11}$$

then any feasible compensation $u_t \in U_{\text{RCBF}}(s_t) = \{(a_{rl} + u_t) \in \mathcal{A} : L_f B(s_t) + L_g B(s_t)(a_{rl} + u) - \beta_3(h(s_t)) \leq 0\}$ will render the set $\mathcal{S}_{safe}$ forward invariant in time t.

In this way, the safety state and action constraints in Eq. 7 can be replaced with RCBF or ZCBF, which express the safety guarantees in the form of inequality constraints. As the action space is limited, there may not be a feasible solution u_t that can transfer current state s_t to $\mathcal{S}_{safe}$ via only one step in some extreme cases. Therefore, a slack variable ϵ needs to be introduced to relax the computational safety constraints, and then the calculation of u_t can be transformed into an inner-level minimization problem as follows:

$$\begin{gathered} \min_{u_t} u_t^2 + \epsilon^2 \\ \text{s.t.:} L_{\hat{f}} h(s_t) + L_{\hat{g}} h(s_t)(a_{rl} + u_t) + \alpha(h(s_t)) \geq 0 \\ or\ L_{\hat{f}} B(s_t) + L_{\hat{g}} B(s_t)(a_{rl} + u_t) - \beta_3(h(s_t)) \leq 0 \\ a_{min} \leq (a_{rl} + u_t) \leq a_{max} \end{gathered} \tag{12}$$

Note that the above problem can be solved via the Quadratic Programs (QP) method, and the parameters of RCBF or ZCBF need to be determined.

For the real-world implementation, it is imperative to note that the aforementioned approach still can not guarantee safety as there is a mismatch between the nominal environmental parameters and the real-world parameters. In addition, the setting of hyperparameters α or $\beta_1, \beta_2, \beta_3$ in CBF inequality constraints will also significantly affect the compensation action. In particular, tight constraints may lead to conservative actions and thus increase the control cost. On the other hand, loose constraints may result in the violation of safety constraints and state oscillation. The following innovations in parameter adaptation are proposed to solve the above issue.

Adaptive GP regression. For the non-stationary power system environment, estimating the dynamics via the conventional Gaussian Process (GP) model can not achieve high confidence in the context of varying model parameters. On the other hand, the model may encounter new scenarios that have not been observed/learned during the training process. Under this reality, if the GP model is established for each environment, the computational cost will be greatly increased because of the huge number of training and testing tasks.

To this end, we propose to store the current imprecise model parameters in the buffer as part of each transition, that is, the transition in the GP buffer contains $(s_t, a_t, s_{t+1}, r_t, \hat{f}, \hat{g})$. During the training process, the current environmental parameters $\hat{f}, \hat{g}$, and state s_t are taken as inputs to predict the parameters' error y_t to realize the follow-up safety compensation. The specific calculation equation is expressed as follows:

$$\begin{gathered} y_t = s_{t+1} - (\hat{f}(s_t) + \hat{g}(s_t) a_t) \\ p(y \mid \mathcal{D}_{gp}, s_t, a_t, \hat{f}, \hat{g}) = \prod_{i=0}^{n_{gp}} \mathcal{N}(\text{m}(y_i), \text{cov}(y_i)) \end{gathered} \tag{13}$$

where D_{gp} is the buffer for GP update; to reduce the computational cost, we only save the most recent episode train data; n_{gp} is the number of samples in buffer, $\text{m}(\cdot)$ and $\text{cov}(\cdot)$ are the mean function and covariance matrix in GP regression, respectively.

Safety constraint adaptation. Take ZCBF as an example, the larger the value of parameter α, the looser the safety constraint is, and vice versa. When the safety risk is large, α needs to be reduced, and more conservative compensation u will be implemented to guarantee that the states can be kept in the safe area; while α needs to be increased to relax restrictions when the safety risk is small, thus reduce the oscillation caused by compensation. To improve the adaptability of the parameter, we propose to establish a nonlinear function to map the relationship between safety risk and parameter α:

$$\alpha = e_2 \, e^{-\tan(e_3 \; clip(\Delta f - \delta f_{bound}, -\frac{\pi}{2}, \frac{\pi}{2}))} \tag{14}$$

where the current risk is defined as the numerical error between Δf with the safe constraint Δf_{bound} and the error is clipped to $(-\frac{\pi}{2}, \frac{\pi}{2})$; e_2, e_3 are hyperparameters that control the safety risk and compensation intensity. Then the tangent function is used to extend it to the whole domain, and finally, α is adaptive by the mapping of the exponential function.

Step 2: Adaptiveness Enhancement

To enhance the performance of DRL under multi-tasks, Rasool et al. (Fakoor et al. 2019) take advantage of the current meta reinforcement learning in adaptive control (Sæmundsson, Hofmann, and Deisenroth 2018) and introduce one of the state-of-the-art meta-reinforcement learning algorithms Meta Q Learning (MQL). The meta-training phase of MQL is only set to maximize the average reward of all training tasks. The objective functions of meta-training and meta-adaptation are set in Eq. 16 and Eq. 17, respectively.

$$\ell^k(\theta) = \underset{s,\tau\sim\mathcal{D}_k}{\mathbb{E}} \left[Q\left(s, \pi_\theta(s), c_\theta(\tau)\right)\right] \tag{15}$$

$$\widehat{\theta}_{\text{train}} = \arg\max_\theta \frac{1}{N_{task}} \sum_{k=1}^{N_{task}} \ell^k(\theta) \tag{16}$$

$$\theta^* = \arg\max_\theta \underset{s,\tau\sim\mathcal{D}_{\text{meta}}}{\mathbb{E}} \left[\beta \, \ell^{eval}(\theta)\right] - \lambda \left\|\theta - \widehat{\theta}_{\text{train}}\right\|_2^2 \tag{17}$$

where N_{task} denotes the number of training tasks, $\ell^k(\theta)$ is the optimization target corresponding to the k training task or evaluation task; $c_\theta(\tau)$ is the context information generated by gate recurrent unit based on the data from each trajectory τ; $\beta\left(s,\ c_\theta(\tau),\ D_{\text{eval}}, \mathcal{D}_{\text{meta}}\right)$ is the propensity score calculated based on the similarity between the training tasks and the evaluation tasks; λ is based on the coefficient related to the Effective Sample Size (ESS) obtained by β to limit the update amplitude of policy parameters.

Given that the future carbon-neutral power systems exhibit the characteristics of high uncertainty and frequently varying system operation conditions, the next step for AdapSafe is to improve its adaptiveness while still meeting the safety constraints. To this end, based on meta Q-learning, we propose two improvement schemes to ensure frequency safety in the meta-training process without affecting the learning performance.

Transition post-processing. The first improvement scheme we developed to MQL is the post-processing of the transition (s_t, a_t, r_t, s_{t+1}) deposited into the train tasks' buffers. Due to the action compensation u_t, there exists a discrepancy between the output by the actor-network a_{rl} and the final executed action a_t. Since the real executed action a_t is compensated by the CBF-based method, directly depositing (s_t, a_t, r_t, s_{t+1}) into the DRL replay buffer will inevitably make all training experiences safe and high-value. Consequently, it will make the actor-network greatly dependent on the compensation term during the training process and then reduce the speed of DRL learning to the safety policy.

Therefore, we carry out the post-processing of the action and choose to store the unprocessed action a_{rl} in the replay buffer. However, this will cause a mismatch between r_t and a_{rl} as the reward in the buffer is calculated using $a_{rl}+u_t$ and thus, the reward post-processing needs to be implemented. Given that a safe strategy is required by the agent (i.e., the learned action does not need compensation u_t), the form of reward post-processing is designed as follows:

$$r_{buffer}(s_t, a_t) = r_t(s_t, a_t) - e_1 \left|u_t\right| \tag{18}$$

where e_1 is the penalty factor for CBF-based compensation.

Noise elimination. Furthermore, the second improvement is eliminating the noise about state and action during the exploration. For the LFC problem, frequency stability is judged based on the frequency derivative $\Delta\dot{f} < \Delta\dot{f}_{bound}$, which is also considered as a reference in the design of the reward function in Eq. 6. Nevertheless, the noise terms designed for actions and states may significantly reduce the time to approach the stable state (i.e., the time to enter phase II). Therefore, we eliminate the noise during the meta-train process. Although this change may reduce the exploration performance to a certain extent, it is experimentally verified that the overall learning rate is better than the MQL method with noise due to the existence of safety constraints that can effectively reduce meaningless exploration and improve the efficiency of the training process.

Overall, the complete algorithm of the proposed AdapSafe is presented in Algorithm 1.

Experiments and Analysis

This section aims to verify the proposed AdapSafe based on a GB 2030 power system. The details regarding the system description and characteristics can be found in (Badesa, Teng, and Strbac 2019). At the same time, our safety constraints also adopt the standard frequency limits set by the National Grid: $\Delta f_{bound} = 0.8$ Hz, $\Delta f_{stable} = 0.5$ Hz and $\Delta\dot{f}_{bound} = 0.01$ Hz/s. To simulate the non-stationary environment and consider the uncertainty brought by RES, we set ΔP_e, H, R_g as variables and generate random samples from a uniform distribution during the training and evaluation phases. The specific settings are provided in Appendix A.

Baselines

We compare AdapSafe with the following baselines:

VSM controller. A traditional control method (Markovic et al. 2018) is implemented as the baseline, which employs an adaptive Virtual Synchronous Machine (VSM) to maintain the effectiveness of frequency regulation and dynamically optimize the control parameters in a variational environment.

Natural TD3. An off-policy DRL algorithm TD3 (Fujimoto, Hoof, and Meger 2018) is considered as the baseline, which avoids unsafe actions via the development of a reward function rather than using a CBF-based compensator.

Algorithm 1: AdapSafe: Adaptive and Safe-Certified DRL-based Frequency Control

```
Input: Train tasks $\mathcal{T}_{train}$ and meta train buffer $\mathcal{D}^k_{meta}$; evaluation tasks $\mathcal{T}_{eval}$ and buffer $\mathcal{D}_{eval}$; GP update buffer $\mathcal{D}_{gp}$; an off-policy DRL algorithm, e.g. TD3.
Parameter: Policy parameters $\theta$; nominal dynamics $\hat{f}, \hat{g}$; safety hyperparameter $\alpha$(or $\beta_3$)$, e_1, e_2, e_3$; total training steps $N$; policy train steps $n$
Output: Policy parameters $\theta$
// Burn Up
Burning up some episodes using $\pi_\theta$ and CBF-based compensation $u$ with $\alpha$(or $\beta_3$) adapted by Eq. 14, save row trajectories to $\mathcal{D}_{gp}$
Post-processing for each transition and save burning trajectories to $\mathcal{D}^k_{meta}$ for each train task $k$ in $\mathcal{T}_{train}$
Update $\hat{f}, \hat{g}$ using buffer $\mathcal{D}_{gp}$ and current GP model
while $i < N$ do
    // Safe Meta Training
    Gather safe trajectories using $\theta$, $u$ and self-tuning $\alpha$(or $\beta_3$) for $\mathcal{T}_{train}$, save post-processing transitions to $\mathcal{D}_{meta}$ and add corresponding steps to $i$
    Sample batch trajectories $\tau$ from $\mathcal{D}^k_{meta}$ while feeding transitions through context and get $c_\theta(\tau)$
    Do meta training using Eq. 16
    update $\theta \leftarrow \theta_{meta}$
    // Safe Meta Adaption
    Gather one episode safe trajectories using $\theta$, $u$ and self-tuning $\alpha$(or $\beta_3$) for $\mathcal{T}_{eval}$, save post-processing transitions to $\mathcal{D}_{eval}$ and add corresponding steps to $i$
    Sample mini-batch trajectories $\tau$ from $\mathcal{D}_{eval}$ and calculate optimization $\ell^{eval}(\theta)$ by Eq 15
    Train logistic regression for estimating the propensity score $\beta\left(s,\ c_\theta(\tau),\ D_{\text{eval}}, \mathcal{D}_{\text{meta}}\right)$ and calculate normalized ESS to get the coefficient $\lambda$
    for $j < n$ do
        Sample mini-batch from $\mathcal{D}_{meta}$
        Calculate $\beta$ for sampled mini-batch
        Do meta adaption using Eq. 17
    end for
end while
```

CBF-TD3. The CBF-based TD3 only uses our CBF-based compensator to guarantee safe constraints without meta-training in non-stationary environments.

MQL. We also implement the MQL method that only conducts meta-training and meta-adaptation. Its treatment of ensuring safety is the same as natural TD3.

Since the natural TD3 and CBF-TD3 methods do not support multi-task training, we refer to the framework of multi-task training in MQL, sample and train each task in turn and modify the same models through parameters sharing.

Training Performance Analysis

Since the VSM controller does not require a training phase, we first compare four DRL-based frequency control schemes in terms of the training performance under the multi-tasks training setting, as shown in Figure 3.

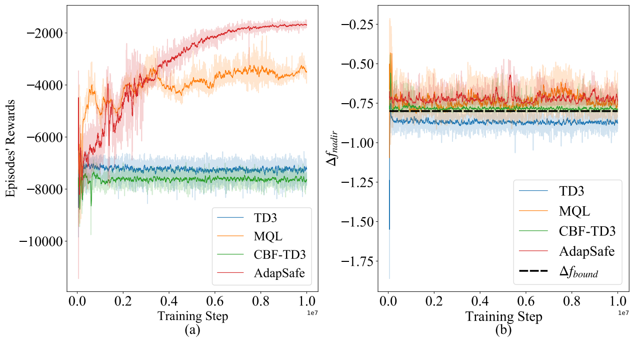

Figure 3: The performance of each algorithm in the training phase, where (a) represents the episodes' rewards of the training phase, (b) represents the frequency nadir of each episode, and the dotted line represents the standard safety constraint Δf_{bound}.

Cost. From the perspective of episodes' reward, it is evident that the proposed AdapSafe exhibits the best performance among the four tested approaches. In particular, since the natural TD3 method and CBF-TD3 method do not have any special treatment for multi-task learning, their training episodes' rewards in the training phase are always kept at a low level; while MQL and AdapSafe both introduce contextual variables and update model parameters via meta-learning, thus show higher rewards under non-stationary environments. Nevertheless, AdapSafe still obtain approximately 50% improvement than the MQL.

Safety. From the perspective of safety, the result indicates that both of the CBF-based methods (i.e., the CBF-TD3 and the AdapSafe) can satisfy that Δf is 100% guaranteed above the safety constraint (lower bound) Δf_{bound} in phase I while for TD3 algorithm without action compensation u, approximately 96.5% of the training episodes exist unsafe states. On the other hand, only about 67.7% training episodes can meet the safety constraints for the MQL.

Test Performance Analysis

In the test phase, we randomly select 50 tasks from the parameter distribution set in Appendix A and set six corresponding evaluation metrics to measure control performance. In particular, the metrics for measuring cost include the average episodes' reward $\overline{R}$ and the average episodes' control action $\overline{A}$ under all test tasks; the metrics for safety are designed based on the change of control task requirements and divided into phase I and phase II. In phase I, we define the worst frequency nadir metric $\underline{N}$ and the total times T of Δf below Δf_{bound} (i.e., unsafe) in all test tasks. In addition, given that phase II is established when the frequency has been stabilized, $\overline{T}_{stab}$ and $\overline{T}_{safe}$ are defined to represent the average convergence speed and the average number of Δf above f_{stable} in the stabilization phase. The test results of all the methods are shown in Table 1. Furthermore, Δf and RES power adjustment ΔP_v under two test environments with the boundary of parameter ΔP_e are shown in Figure 4 to visually inspect the changes in states and actions

Methods	Cost		Safety			
	$\overline{R}$	$\overline{A}$	$\underline{N}$	T	$\overline{T}_{stab}$	$\overline{T}_{safe}$
VSM	-	-0.71	-0.947	54	107.84	105.66
TD3	-7738	-0.05	-0.833	57	60.47	16.90
MQL	-6302	**-0.03**	-0.844	22	50.91	21.47
CBF-TD3	-8584	-0.63	-0.825	12	81.86	81.73
AdapSafe	**-3345**	-0.27	**-0.797**	**0**	**137.46**	**137.42**

Table 1: The performance of each algorithm in 50 test tasks.

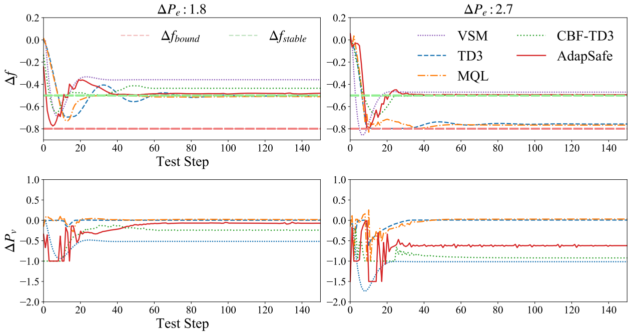

Figure 4: The curves of Δf and ΔP_v across different control methods under a variational environment, where the upper and lower bound of the environmental parameter ΔP_e distribution are selected.

of different control methods under varying environments.

Cost. As can be seen, the AdapSafe obtains the highest average reward $\overline{R}$, indicated by the approximately 56% higher value than that of the TD3. On the other hand, although the TD3 and MQL methods learned the strategies to use smaller power adjustment $\overline{A}$ to achieve acceptable performance in some simple scenarios (e.g., $\Delta P_e = 1.8$ in Figure 4), they cannot achieve phase II safe target: $\Delta f > \Delta f_{stable}$ under more difficult scenarios (e.g., $\Delta P_e = 2.7$ in Figure 4). Meanwhile, although the VSM and the CBF-TD3 can dynamically adjust ΔP_v for ensuring the safe constraints, the cost of the control action is 163% and 133% higher than ours, respectively, which leads to the worst average episodes' rewards.

Safety. From the perspective of safety, it is imperative to highlight that the proposed AdaptSafe is the only method that can realize the safety guarantee (i.e., $T = 0$, $\underline{N} > -0.8$) for both I and II under varying environmental parameters. Meanwhile, the advantage of AdapSafe in II is more prominent compared with other methods. In particular, the metric value of $\overline{T}_{stab}$ is approximately 27%, 128%, 169%, and 71% higher than those of the VSM, TD3, MQL and CBF-TD3, respectively. Furthermore, during the stabilization phase, AdapSafe also exhibits the best performance in ensuring safety. Almost all the states that enter the stable phase are safe, while only about 28% and 42% stable states of the TD3 and MQL methods can satisfy phase II safety constraint.

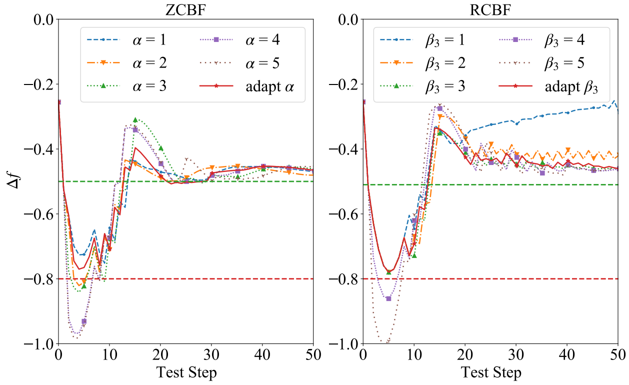

Figure 5: Ablation study of Z(R)CBF-based Adapsafe with fixed $\alpha(\beta_3)$ and the red line is our self-adapt strategy.

Ablation Study

For the proposed AdapSafe, the hyperparameters α and β_3, which control the compensation intensity, are of great importance in the experimental effect. To this end, as shown in Figure 5, we conducted the ablation experiments to compare the proposed self-tuning approach with five fixed values of α and β_3. First, it can be observed that the CBF compensation becomes less conservative and subsequently triggers unsafe Δf with the increasing value of α or β_3. The results indicate that the fixed hyperparameter does not accommodate the variational safety constraints of both phase I and phase II. Meanwhile, the proposed self-tuning method performs a relatively conservative (i.e., slightly smaller α or β_3) to meet the Δf_{bound} constraint in phase I. In phase II, α or β_3 of the proposed self-tuning method can be actively increased to meet the Δf_{stable} constraint and reduce the control cost for achieving the target of economic scheduling.

Conclusion

This paper proposes a novel DRL-based frequency control framework, AdapSafe, to simultaneously address the two crucial challenges of safety guarantee and adaptiveness enhancement in a non-stationary environment to facilitate its real-world implementation. In particular, a self-tuning CBF-based compensator is designed to realize the optimal safety compensation under different risk conditions, which greatly reduces the control cost. Furthermore, to minimize the control cost without sacrificing safety, the CBF-based safe control method is integrated with meta reinforcement learning algorithm with the innovations of transition post-procession and noise elimination scheme to achieve safety assurance during the meta-training phase for multi-task scenarios. Through a comparative study with the state-of-the-art frequency control methods based on a GB 2030 power system, the results demonstrate that AdapSafe can achieve the target of safety guarantee under a variational environment with superior control performance for both the training and testing phases. In the future, the proposed AdapSafe will be further developed to the multi-machine frequency control problem based on large-scale power systems.

References

Abd-Elazim, S.; and Ali, E. 2016. Load frequency controller design via BAT algorithm for nonlinear interconnected power system. *International Journal of Electrical Power & Energy Systems*, 77: 166–177.

Altman, E. 1999. *Constrained Markov decision processes: stochastic modeling*. Routledge.

Ames, A. D.; Xu, X.; Grizzle, J. W.; and Tabuada, P. 2016. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8): 3861–3876.

Badesa, L.; Teng, F.; and Strbac, G. 2019. Simultaneous scheduling of multiple frequency services in stochastic unit commitment. *IEEE Transactions on Power Systems*, 34(5): 3858–3868.

Berkenkamp, F.; Turchetta, M.; Schoellig, A.; and Krause, A. 2017. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30.

Bialek, J. 2020. What does the GB power outage on 9 August 2019 tell us about the current state of decarbonised power systems? *Energy Policy*, 146: 111821.

Cheng, R.; Orosz, G.; Murray, R. M.; and Burdick, J. W. 2019. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3387–3395.

Chow, Y.; Nachum, O.; Duenez-Guzman, E.; and Ghavamzadeh, M. 2018. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31.

Chu, Z.; Markovic, U.; Hug, G.; and Teng, F. 2020. Towards optimal system scheduling with synthetic inertia provision from wind turbines. *IEEE Transactions on Power Systems*, 35(5): 4056–4066.

Debnath, M. K.; Jena, T.; and Sanyal, S. K. 2019. Frequency control analysis with PID-fuzzy-PID hybrid controller tuned by modified GWO technique. *International Transactions on Electrical Energy Systems*, 29(10): e12074.

Ersdal, A. M.; Imsland, L.; and Uhlen, K. 2015. Model predictive load-frequency control. *IEEE Transactions on Power Systems*, 31(1): 777–785.

Fakoor, R.; Chaudhari, P.; Soatto, S.; and Smola, A. J. 2019. Meta-q-learning. *arXiv preprint arXiv:1910.00125*.

Fujimoto, S.; Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, 1587–1596. PMLR.

Gupta, P.; Pal, A.; and Vittal, V. 2021. Coordinated Wide-Area Damping Control Using Deep Neural Networks and Reinforcement Learning. *IEEE Transactions on Power Systems*, 37(1): 365–376.

Kundur, P.; Balu, N. J.; and Lauby, M. G. 1994. *Power system stability and control*, volume 7. McGraw-hill New York.

Markovic, U.; Chu, Z.; Aristidou, P.; and Hug, G. 2018. LQR-based adaptive virtual synchronous machine for power systems with high inverter penetration. *IEEE Transactions on Sustainable Energy*, 10(3): 1501–1512.

Sæmundsson, S.; Hofmann, K.; and Deisenroth, M. P. 2018. Meta reinforcement learning with latent variable gaussian processes. *arXiv preprint arXiv:1803.07551*.

Sahu, R. K.; Panda, S.; and Pradhan, P. C. 2015. Design and analysis of hybrid firefly algorithm-pattern search based fuzzy PID controller for LFC of multi area power systems. *International Journal of Electrical Power & Energy Systems*, 69: 200–212.

Singh, R.; and Ramesh, L. 2019. Comparison of automatic load frequency control in two area power systems using pso algorithm based pid controller and conventional pid controller. In *Journal of Physics: Conference Series*, volume 1172, 012054. IOP Publishing.

Sloth, C.; Wisniewski, R.; and Pappas, G. J. 2012. On the existence of compositional barrier certificates. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 4580–4585. IEEE.

Tan, W. 2009. Unified tuning of PID load frequency controller for power systems via IMC. *IEEE Transactions on power systems*, 25(1): 341–350.

Teng, F.; Trovato, V.; and Strbac, G. 2015. Stochastic scheduling with inertia-dependent fast frequency response requirements. *IEEE Transactions on Power Systems*, 31(2): 1557–1566.

Ulbig, A.; Borsche, T. S.; and Andersson, G. 2014. Impact of low rotational inertia on power system stability and operation. *IFAC Proceedings Volumes*, 47(3): 7290–7297.

Wisniewski, R.; and Sloth, C. 2015. Converse barrier certificate theorems. *IEEE Transactions on Automatic Control*, 61(5): 1356–1361.

Xia, Y.; Xu, Y.; Wang, Y.; Mondal, S.; Dasgupta, S.; Gupta, A. K.; and Gupta, G. 2022. A Safe Policy Learning-Based Method for Decentralized and Economic Frequency Control in Isolated Networked-Microgrid Systems. *IEEE Transactions on Sustainable Energy*.

Yan, Z.; and Xu, Y. 2020. A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system. *IEEE Transactions on Power Systems*, 35(6): 4599–4608.

Yin, L.; Yu, T.; Zhou, L.; Huang, L.; Zhang, X.; and Zheng, B. 2017. Artificial emotional reinforcement learning for automatic generation control of large-scale interconnected power grids. *IET Generation, Transmission & Distribution*, 11(9): 2305–2313.

Zhang, C.-K.; Jiang, L.; Wu, Q.; He, Y.; and Wu, M. 2013. Delay-dependent robust load frequency control for time delay power systems. *IEEE Transactions on Power Systems*, 28(3): 2192–2201.

Zhang, Z.; Du, E.; Teng, F.; Zhang, N.; and Kang, C. 2020. Modeling frequency dynamics in unit commitment with a high share of renewable energy. *IEEE Transactions on Power Systems*, 35(6): 4383–4395.

Appendix A: Case Study Setting

In this section, we describe the detailed system dynamics, as well as give the specific procedure for calculating compensation u. Moreover, we explain the parameter settings involved in the GB 2030 power system.

System Frequency Dynamics

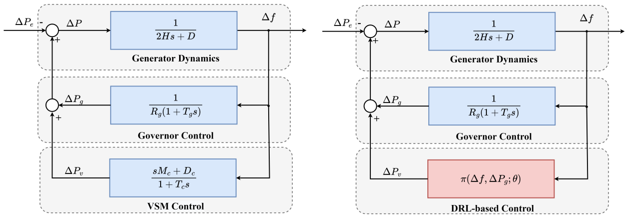

Figure 6: Simplified system frequency dynamics model.

A simplified model of system frequency dynamics is shown in Figure 6, and the low-order model proposed in (Chu et al. 2020) is used for modeling the governor dynamics. For the VSM-based control scheme proposed in (Markovic et al. 2018), T_c is the converter time constant, while M_c and D_c represent the virtual inertia and damping constant of the converter. Meanwhile, for the DRL-based control scheme, the frequency adjustment ΔP_v brought by RES will be determined by its policy network $\pi(\Delta f, \Delta P_g; \theta)$.

From Fig 6, we can get the specific algebraic expression of the dynamics for DRL-based methods as shown in Eq 19:

$$\begin{aligned} \Delta \dot{f} &= \frac{1}{2H}(-D\Delta f - \Delta P_g - \Delta P_e + \Delta P_v) \\ \Delta \dot{P}_g &= \frac{1}{RgTg}(\Delta f - R_g \Delta P_g) \\ \Delta P_v &= \pi(\Delta f, \Delta P_g; \theta) \end{aligned} \tag{19}$$

Based on the above dynamics, the control objective of our LFC problem can be divided into two phases in Figure 7, where phase I needs to satisfy the deadband frequency Δf_{nadir} above Δf_{bound}, and phase II needs to achieve stable frequency Δf_{stable} with less control cost.

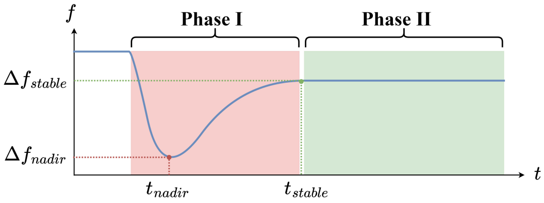

Figure 7: Two-phase control objective for the LFC problem.

CBF-based Compensator

The action compensation u based on the CBF method is realized by solving quadratic programming (QP) problems with inequality constraints. In view of our LFC problem, the specific QP equation is set as follows:

$$\begin{aligned} &\min_x \frac{1}{2} x^T P x \\ &\text{s.t. } : Gx \le h \\ &\qquad Ax = b \end{aligned} \tag{20}$$

$$\begin{aligned} \text{where} \quad P &= \begin{bmatrix} 1 & 0 \\ 0 & 10^{24} \end{bmatrix} \; x = \begin{bmatrix} u \\ \epsilon \end{bmatrix} \\ H &= \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \; F = \begin{bmatrix} -\Delta f_{bound} \\ -\Delta f_{stable} \end{bmatrix} \end{aligned} \tag{21}$$

$$G^1 = \begin{bmatrix} -H_1 g & -1 \\ -H_2 g & -1 \\ 1 & 0 \\ -1 & 0 \end{bmatrix} G^2 = \begin{bmatrix} -H_1 g & -1 \\ -H_2 g & -1 \\ 1 & 0 \\ -1 & 0 \\ g[0] & 0 \\ -g[0] & 0 \end{bmatrix} \tag{22}$$

$$\begin{aligned} h^1_{\text{ZCBF}} &= \begin{bmatrix} \alpha F_1 + H_1 f + H_1 g \; a_{rl} - \\ (1-\alpha) H_1 x - k_\delta |H_1| \sigma \\ \alpha F_1 + H_2 f + H_2 g \; a_{rl} - \\ (1-\alpha) H_1 x - k_\delta |H_2| \sigma \\ -a_{rl} + \Delta f_{bound} \\ a_{rl} - \Delta f_{bound} \end{bmatrix} \\ h^2_{\text{ZCBF}} &= \begin{bmatrix} \alpha F_2 + H_1 f + H_1 g \; a_{rl} - \\ (1-\alpha) H_1 x - k_\delta |H_1| \sigma \\ \alpha F_2 + H_2 f + H_2 g \; a_{rl} - \\ (1-\alpha) H_1 x - k_\delta |H_2| \sigma \\ -a_{rl} + \Delta f_{bound} \\ a_{rl} - \Delta f_{bound} \\ -f[0] - g[0] a_{rl} + \Delta \dot{f}_{bound} \\ f[0] + g[0] a_{rl} - \Delta \dot{f}_{bound} \end{bmatrix} \end{aligned} \tag{23}$$

$$\begin{aligned} h^1_{\text{RCBF}} &= \begin{bmatrix} \frac{\beta_3 (H_1 x + F_2 + 1)(H_1 x + F_1)}{\ln(\frac{1 + H_1 x + F_1}{H_1 x + F_1}) + (H_1 f + H_1 g \; a_{rl} - k_\delta |H|^T \sigma)} \\ \frac{\beta_3 (H_2 x + F_2 + 1)(H_2 x + F_1)}{\ln(\frac{1 + H_2 x + F_1}{H_2 x + F_1}) + (H_2 f + H_2 g \; a_{rl} - k_\delta |H|^T \sigma)} \\ -a_{rl} + \Delta f_{bound} \\ a_{rl} - \Delta f_{bound} \end{bmatrix} \\ h^2_{\text{RCBF}} &= \begin{bmatrix} \frac{\beta_3 (H_1 x + F_2 + 1)(H_1 x + F_2)}{\ln(\frac{1 + H_1 x + F_2}{H_1 x + F_2}) + (H_1 f + H_1 g \; a_{rl} - k_\delta |H|^T \sigma)} \\ \frac{\beta_3 (H_2 x + F_2 + 1)(H_2 x + F_2)}{\ln(\frac{1 + H_2 x + F_2}{H_2 x + F_2}) + (H_2 f + H_2 g \; a_{rl} - k_\delta |H|^T \sigma)} \\ -a_{rl} + \Delta f_{bound} \\ a_{rl} - \Delta f_{bound} \\ -f[0] - g[0] a_{rl} + \Delta \dot{f}_{bound} \\ f[0] + g[0] a_{rl} - \Delta \dot{f}_{bound} \end{bmatrix} \end{aligned} \tag{24}$$

where G^1 matches h^1_{ZCBF} or h^1_{RCBF} and G^2 matches h^2_{ZCBF} or h^2_{RCBF}.

Since the safety constraints $\mathcal{C}$ vary at different phases, we use $G^1 x \le h^1$as the QP constraint in phase I and $G^2 x \le h^2$ in phase II. Furthermore, two forms of CBF-based compensation are designed for the LFC problem by setting different h_{ZCBF} and h_{RCBF}.

Parameters Setting

Table 2 shows our settings for the variable parameters during the generation of the training and testing environments. It should be noted that the time scale for the change of environment dynamics is greater than the time for simulating a DRL episode (we set it to 100s). Moreover, for generation loss ΔP_e and system inertia H, we have a preference for sampling tasks. For ΔP_e , 1/3 of the training tasks are sampled from the uniform distribution $\mathcal{U}(1.8, 2.2)$, while the remaining 2/3 of the training tasks are sampled from $\mathcal{U}(2.2, 2.7)$.

Parameters	Value distribution	Preference
ΔP_e	$\mathcal{U}(1.8, 2.2) + \mathcal{U}(2.2, 2.7)$	1:2
H	$\mathcal{U}(1.5, 2.0) + \mathcal{U}(2.0, 2.5)$	2:1
R_g	$\mathcal{U}(0.2, 0.3)$	N

Table 2: Parameters value distribution and preference ratio setting, $\mathcal{U}(a, b)$ denotes a uniform distribution of parameters a and b, and preference 'N' indicates this parameter has no preference setting.

Besides, the following Table 3 shows the hyperparameter configuration of four comparison algorithms in the same set of interaction environment.

	TD-3	MQL	CBF-TD3	AdapSafe
Burning up	1e4			
Total steps	1e7			
$[m_1, m_2, m_3]$ (Eq. 6)	[50, 10, 50, 0.5]			
Exploration noise	0.3	0.3	0.3	0
Context dimension	0	30	0	20
History length	0	30	0	30
Meta-adaptation freq	-	3e4	-	3e4
Adam learning rate	8e-4	8e-4	3e-4	3e-4
e_1 (Eq. 18)	-	-	5	5
$[e_2, e_3]$ (Eq. 14)	-	-	[1,2]	[1,2]
GP kernel	-	-	k_{RBF}	k_{RBF}

Table 3: Hyperparameter for four DRL-based LFC schemes in the same non-stationary environment.

Appendix B: Training and Testing Results

In this section, we first fix a set of environmental parameters and compare the TD-3 and CBF-TD3 algorithms as a way to show that our CBF-based compensator can achieve safety guarantees in a static environment. After that, we randomly select a combination of parameter distributions from Table 2 as test environments to demonstrate the superiority of our AdapSafe method by comparing it with the other four algorithms. Finally, we also test more difficult unknown environments beyond the parameter distributions in Table 2, and present the comparison results in Table 3.

Results in Stationary Environment

As can be seen from Figure 8, the CBF-based compensator can guarantee all training episodes satisfy the safety constraint of $\Delta f_{nadir} > \Delta f_{bound}$ in static environments. In contrast, even in a simple task, such as $\Delta P_e = 1.8$, the TD3 algorithm can hardly guarantee the safety of the frequency all the time.

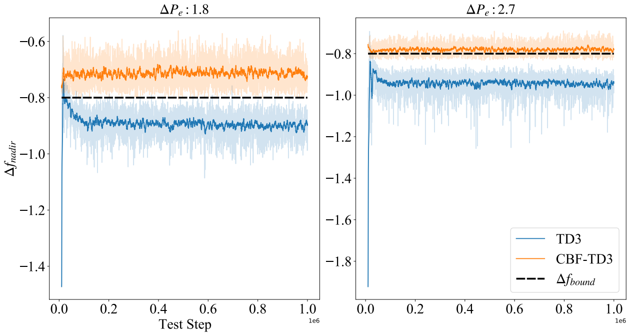

Figure 8: Comparison of the Δf_{nadir} for the TD3 method and the CBF-TD3 method with environmental parameters of $\Delta P_e = 1.8$ (or 2.7), $H = 2$ and $R_g = 0.3$.

Results in Non-Stationary Environment

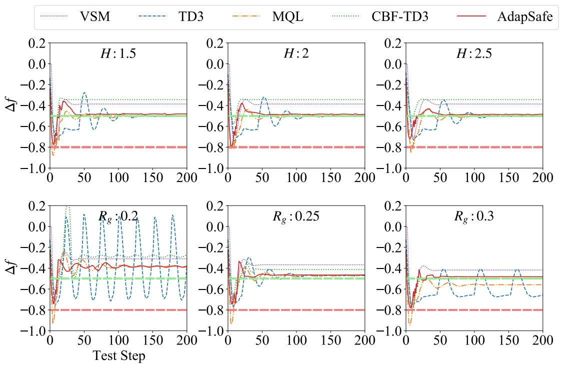

Figure 9: Variation of Δf under environmental parameters H and R_g changes.

Figure 9 illustrates the Δf variation of the five control methods for variable environmental parameters H and R_g. It can be seen that our AdapSafe method can significantly improve the adaptivity in non-stationary environment setting without sacrificing any safety cost.

Results in Out-of-Distribution Environments

Although we have set the range of environmental parameters within the parameter bounds that the actual frequency control system may reach, we also test more extreme cases, such as too small or too large parameters, and show the test performance in Tables 4 and Tabel 5.

From Tables 4 and 5, we can see that for the case of extremely small environmental parameters, our method still has the best control performance from both cost and safety perspectives; while for the case of extremely large environmental parameters, all algorithms cannot guarantee that Δf

Methods	Cost		Safety			
	$\overline{R}$	$\overline{A}$	$\underline{N}$	T	$\overline{T}_{stab}$	$\overline{T}_{safe}$
VSM	-	-0.33	-0.776	**0**	103.72	103.72
TD3	466	**-0.04**	-0.726	**0**	136.04	135.84
MQL	-1789	-0.06	-0.840	6	97.28	97.16
CBF-TD3	-6454	-0.66	**-0.477**	**0**	145.72	145.64
AdapSafe	**925**	-0.05	-0.788	**0**	**170.60**	**170.44**

Table 4: The performance and safety of each algorithm in 100 test tasks where parameters are extremely small: $\Delta P_e \sim \mathcal{U}(1.7, 1.8), R_g \sim \mathcal{U}(0.15, 0.2), H \sim \mathcal{U}(1, 1.5)$.

Methods	Cost		Safety			
	$\overline{R}$	$\overline{A}$	$\underline{N}$	T	$\overline{T}_{stab}$	$\overline{T}_{safe}$
VSM	-	-0.80	-1.119	652	45.04	25
TD3	-23335	-0.18	**-1.106**	1666	100	0
MQL	-22800	**-0.17**	-1.153	3668	36.84	0
CBF-TD3	**-20809**	-1.25	-1.110	**638**	175.84	**78.12**
AdapSafe	-21731	-1.29	-1.146	664	**177.24**	76.08

Table 5: The performance and safety of each algorithm in 100 test tasks where parameters are extremely large: $\Delta P_e \sim \mathcal{U}(2.7, 2.8), R_g \sim \mathcal{U}(0.3, 0.35), H \sim \mathcal{U}(2.5, 3)$.

will recover to safety by one step action adjustment, but the overall effect of the CBF-TD3 and AdapSafe is better than that of MQL and TD3 algorithms.