# Dynamic Nonlinear Droop-based Fast Frequency Regulation for Power Systems with Flexible Resources Using Meta-reinforcement Learning Approach

Yuxin Ma, *Student Member*, *IEEE*, Zechun Hu, *Senior Member*, *IEEE*, and Yonghua Song, *Fellow*, *IEEE*

***Abstract*—The increasing penetration of renewable energy resources and reduced system inertia pose risks to frequency security of power systems, necessitating the development of fast frequency regulation (FFR) methods using flexible resources. However, developing effective FFR policies is challenging because different power system operating conditions require distinct regulation logics. Traditional fixed-coefficient linear droop-based control methods are suboptimal for managing the diverse conditions encountered. This paper proposes a dynamic nonlinear *P-f* droop-based FFR method using a newly established meta-reinforcement learning (meta-RL) approach to enhance control adaptability while ensuring grid stability. First, we model the optimal FFR problem under various operating conditions as a set of Markov decision processes and accordingly formulate the frequency stability-constrained meta-RL problem. To address this, we then construct a novel hierarchical neural network (HNN) structure that incorporates a theoretical frequency stability guarantee, thereby converting the constrained meta-RL problem into a more tractable form. Finally, we propose a two-stage algorithm that leverages the inherent characteristics of the problem, achieving enhanced optimality in solving the HNN-based meta-RL problem. Simulations validate that the proposed FFR method shows superior adaptability across different operating conditions, and achieves better trade-offs between regulation performance and cost than benchmarks.**

***Index Terms*—Power system, fast frequency regulation, flexible resource, meta-reinforcement learning, hierarchical neural network.**

## I. Introduction

With the rapid advancement of the global power system transformation, the traditional synchronous generators in power systems are gradually being replaced by renewable energy resources such as solar and wind energy. This shift results in lower system inertia and reduced primary frequency regulation (PFR) reserves, which threaten power system frequency security [1]. Additionally, the intermittency and uncertainty associated with wind and solar generation further enhanced the difficulties of frequency control. Traditional frequency support methods, which rely solely on traditional frequency regulation resources, are insufficient for ensuring the safe operation of the power system with high penetration of renewable energy resources. Consequently, it becomes essential to utilize emerging flexible resources such as wind and solar energy resources [2], battery energy storage [3], hybrid energy storage [4], and electric vehicle aggregators [5] to enhance the frequency support and improve the transient frequency dynamics of power systems.

Due to their mechanical characteristics, synchronous generators primarily achieve PFR through fixed-coefficient linear droop control. In contrast, flexible resources, connected to the grid via inverters, offer faster and more precise frequency response [6]. This enhanced control flexibility enables the development of customized frequency regulation standards for these resources. As a result, many transmission system operators have designed fast frequency regulation (FFR) services that utilize flexible resources to deliver rapid proportional or step frequency responses [7]. For instance, the enhanced frequency response service in UK requires the providers, predominantly storage assets, to respond proportionally to the system frequency in 1 s or less after the frequency falls out of the deadband, while the response time of the traditional PFR resources is around 10 s [8]. In the Texas power system, FFR resources provide step responses within 0.25 s once the frequency falls below 59.85 Hz [9]. In addition, the existing research has developed modified *P-f* droop-based control methods for flexible resource-based FFR. For instance, the variable *P-f* droop-based control is proposed in [10], which consists of two fixed droop coefficients activated at different frequency levels. In [11], the linear *P-f* droop-based FFR signals are decomposed into low- and high-frequency components and delivered to different

Manuscript received: January 17, 2024; revised: May 17, 2024; accepted: September 18, 2024. Date of CrossCheck: September 18, 2024. Date of online publication: October 7, 2024.

This work was supported by the Key Research and Development Program of Inner Mongolia, China (No. 2021ZD0039).



Y. Ma and Z. Hu (corresponding author) are with the Department of Electrical Engineering, Tsinghua University, Beijing 100084, China (e-mail: mayuxin21@mails.tsinghua.edu.cn; zechhu@tsinghua.edu.cn).

Y. Song is with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China (e-mail: yhsong@um.edu.mo).

DOI: 10.35833/MPCE.2024.000062

flexible resources. In addition to linear and piece-wise linear control methods, some nonlinear FFR strategies have been designed for flexible resources in [12]-[14] to achieve improved control performance.

The above-mentioned FFR services all adopt static control laws with fixed droop curves, which lack adaptability to varying operating conditions. Considering the superior control flexibility of new resources, some dynamic FFR strategies have been proposed to enhance transient frequency dynamics and improve the cost-efficiency of frequency regulation. An asymmetric droop coefficient optimization method is proposed in [15] to realize robust and cost-efficient FFR provided by wind turbines and demand response resources. The droop coefficients can be dynamically updated in a centralized manner but at a limited rate due to heavy communication and computational burdens. Hierarchical FFR schemes proposed in [16]-[18] also require high-quality communication and online optimization.

Some existing studies leverage reinforcement learning (RL) methods to develop dynamic FFR policies for flexible resources. Well-trained RL controllers can avoid online optimization and reduce the computational burden during practical implementation. Reference [19] proposed an RL-based distributed update policy for adjusting the inertia and droop coefficients of multiple virtual synchronous generators to suppress power oscillations under various disturbance sizes. However, this policy still requires communication with adjacent nodes. Reference [20] proposed an RL-based FFR controller for battery energy storage systems that relies solely on local frequency measurements. Although the methods in [19] and [20] enhance control flexibility, they cannot guarantee system stability, which is a common challenge in applying RL methods in power system control problems. Reference [12] developed an RL-based static FFR method that ensures the frequency stability through a single-input-single-output neural network structure. However, over-strict network structure constraints, such as the single-layer requirement and the single-input limit, restrict the generalization of this static method to a dynamic type.

Existing RL-based FFR methods typically assume that system frequency dynamics can be modeled as a single Markov decision process (MDP). However, these dynamics actually vary significantly with the size of load disturbances. Given the randomness and diversity of load disturbances in actual power systems, it is more appropriate to consider the optimal FFR problem as achieving fast adaption to any MDP sampled from a distribution. To date, traditional RL algorithms often solve each MDP independently and can hardly realize the rapid adaption required in the FFR context. Meta-reinforcement learning (meta-RL) is a promising method to solve this problem, whose core idea is to learn data-efficient RL algorithms capable of producing policies that adapt well to various MDPs with minimal data [21]. Various meta-RL algorithms [22], [23] have been proposed and applied across different domains, including power system operation and control. For instance, [24] proposed an optimal load frequency control method for interconnected microgrid using a meta-RL framework, and [25] focused on meta-RL-based grid voltage emergency control. However, these methods often lack theoretical guarantees for frequency or voltage stability. Applying meta-RL to the optimal FFR problem requires careful considerations to ensure frequency stability.

In summary, research gaps can be summarized as follows. Firstly, existing FFR methods are predominantly based on linear static droop control schemes or dynamic approaches burdened by heavy computation or communication demands. These methods fail to fully utilize the potential of flexible resources and lack adaptability to varying sizes of random load disturbances. Secondly, while RL methods offer potential for adaptive FFR with low computational burden during implementation, their effectiveness is limited by imperfect problem formulations in existing literature and concerns about stability guarantees. To address these gaps, this paper develops a dynamic nonlinear *P-f* droop-based FFR method using a newly established meta-RL approach to ensure both adaptability and stability. The proposed FFR method is applicable to various flexible resources integrated into power systems through power electronic inverters, presenting a possible solution for enhancing frequency stability in future power systems with high penetration of inverter-based generation. The main contributions can be summarized as follows.

1) The dynamic nonlinear FFR optimization problem is formulated as a frequency stability-constrained meta-RL problem, which leverages flexible resources to achieve stable FFR with fast adaptation to randomly varying load disturbances.

2) A hierarchical neural network (HNN) structure is proposed to parameterize dynamic nonlinear droop-based FFR policies with a theoretical frequency stability guarantee, converting the proposed meta-RL problem into a more tractable form.

3) A two-stage algorithm is specifically designed to solve the HNN-based meta-RL problem with enhanced optimality.

4) Simulations demonstrate that the proposed method provides FFR policies with superior adaptability, achieving a better balance between frequency quality and regulation cost compared with benchmark methods.

The rest of this paper is organized as follows. Section II describes the system model for controller optimization and simulation and the system model for theoretical analysis. Section III first models the optimal FFR as a stochastic optimization and then reformulates it into a constrained meta-RL problem. The HNN architecture is proposed in Section IV, and Section V presents the two-stage algorithm to solve the HNN-based meta-RL problem. Numerical simulation results are presented in Section VI. Finally, conclusions are drawn in Section VII.

## II. System Model

### A. System Model for Controller Optimization and Simulation

Considering that a control area may contain numerous flexible resources, this paper adopts the centralized optimization and distributed execution scheme for convenience of application and supervision in practical power systems. During the optimization stage, we design an aggregated FFR controller, denoted as $u$, based on the system frequency response

(SFR) model of the target control area, as illustrated in Fig. 1, where synchronous generators and flexible resources in the target control area are aggregated into equivalent blocks, respectively. The analytical approach for the model aggregation can be found in [26].

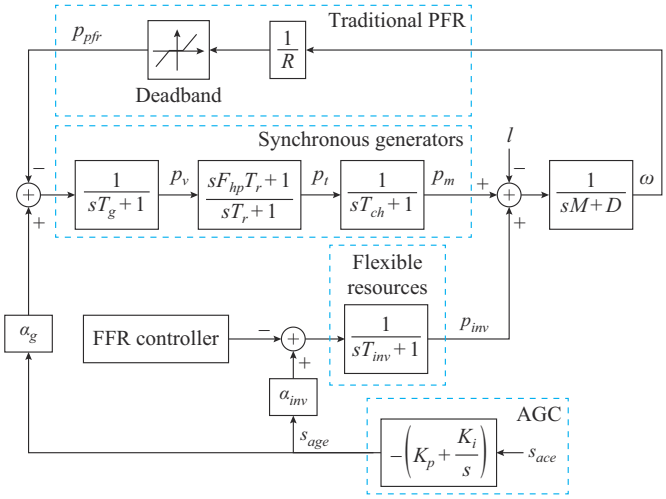


Fig. 1. Block diagram of target control area.

All variables in Fig. 1 represent deviations. $\omega$ denotes the center-of-inertia (CoI) frequency. $p_v$, $p_t$, $p_m$, and $p_{inv}$ denote the governor valve displacement, power deviation during steam reheat, mechanical output of generators, and flexible resource output, respectively. $p_{pfr}$ denotes the PFR output of synchronous generators. The control flexibility of flexible resources enables the design of a sophisticated logic for $u$ to achieve desired control performance. $l$ denotes the net load disturbance consisting of renewable power generation fluctuations, load variations, and tie-line power deviations. $T_g$, $T_r$, $T_{ch}$, and $T_{inv}$ denote the time constants of the equivalent governor, reheater, turbine, and inverter, respectively. $F_{hp}$ is the fraction of total turbine power. $M$ and $D$ denote the system inertia and load-damping coefficient, respectively. Synchronous generators are required to perform traditional PFR with a fixed linear droop coefficient $1/R$. In addition, a proportional-integral (PI) type automatic generation controller (AGC) is considered, with integral gain $K_i$ and proportional gain $K_p$. The AGC operates in flat frequency control mode, with the area control error (ACE) calculated as $s_{ace}=\beta\omega$, where $\beta$ denotes the frequency bias parameter. The command generated by AGC is denoted as $s_{agc}$, which is allocated to generators and flexible resources according to their participation factors $\alpha_g$ and $\alpha_{inv}$.

The system dynamics can be represented as a set of state-space functions as:

$$\boldsymbol{x}=\left[p_v,p_t,p_m,p_{inv},\omega,\int\omega\,\mathrm{d}t\right] \tag{1a}$$

$$\begin{cases}\int\dot{\omega}\,\mathrm{d}t=\omega\\ \dot{\omega}=\dfrac{1}{M}\left(p_m+p_{inv}-D\omega-l\right)\end{cases} \tag{1b}$$

$$\dot{p}_t=\frac{F_{hp}}{T_g}\left(p_{pfr}+\alpha_g s_{agc}\right)+\frac{T_g-F_{hp}T_r}{T_rT_g}p_v-\frac{1}{T_r}p_t \tag{1c}$$

$$\dot{p}_{inv}=\frac{1}{T_{inv}}\left(-u-p_{inv}+\alpha_{inv}s_{agc}\right) \tag{1d}$$

$$\dot{p}_v=\frac{1}{T_g}\left(-p_{pfr}+\alpha_g s_{agc}-p_v\right) \tag{1e}$$

$$\begin{cases}\dot{p}_m=\dfrac{1}{T_{ch}}\left(p_t-p_m\right)\\ s_{agc}=-K_p\beta\omega-K_i\beta\displaystyle\int\omega\,\mathrm{d}t\end{cases} \tag{1f}$$

$$p_{pfr}=\frac{1}{R}\left(\max\left(\omega-\omega_{db},0\right)+\min\left(\omega+\omega_{db},0\right)\right) \tag{1g}$$

where $\boldsymbol{x}$ is the state vector; and $\omega_{db}$ is the deadband width for generators.

### *B. System Model for Theoretical Analysis*

In this paper, the aggregated FFR controller designed in subsequent sections takes only local available information as inputs. During the application, the aggregated controller is decomposed into distributed controllers by multiplying different participation factors depending on the regulation capacity of each flexible resource. Distributed controllers work with the locally measured frequency, which can be different with the CoI frequency considered in the SFR model. Consequently, the transient frequency stability analysis should consider the specific network structure and frequency differences across the target control area, such that the frequency stability is guaranteed during the practical operation.

We denote the target control area by an undirected connected graph $(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is the set of lossless buses indexed by $i$ or $j\in\{1,2,\ldots,n\}$, and $\mathcal{E}$ is the set of transmission lines indexed by $(i,j)\in\{(i,j)|i,j\in\mathcal{V},i\neq j\}$. Each bus is equipped with an equivalent generator and an equivalent flexible resource unit aggregated from the connected resources. System dynamics model in [12] is used for theoretical stability analysis, which can be formulated as the following state-space functions:

$$\dot{\theta}_i=\omega_i \tag{2a}$$

$$\dot{\omega}_i=\frac{1}{M_i}\left[-l_i-\left(D_i+\frac{1}{R_i}\right)\omega_i-u_i-\sum_{j=1}^{n}B_{ij}\sin\left(\theta_i-\theta_j\right)\right] \tag{2b}$$

where $\omega_i$, $\theta_i$, $u_i$, $l_i$, $M_i$, $D_i$, and $R_i$ are the local frequency, phase angle, distributed FFR control signal, net load disturbance, system inertia, load-damping coefficient, and droop coefficient of synchronous generator of bus $i$, respectively; and $B_{ij}$ is the susceptance of line $(i,j)$. All variables in (2) represent deviations from their nominal values. Note that the AGC is omitted in (2) because it operates at a slower pace in practical power systems and therefore has limited effect on the transient frequency stability. The generator dynamics are simplified as a classical second-order model widely used in existing literature. The inverter dynamics are omitted for its much smaller time constant than the generator.

A static droop controller for flexible resources without lin-

earity requirement can be denoted as $u_i(\omega_i)$, taking only local frequency measurement as input. Theorem 1 gives a sufficient condition for the frequency stability of system (2) under $u_i(\omega_i)$, which will be applied in the subsequent dynamic controller optimization.

**Theorem 1** [12] Suppose the controller $u_i(\omega_i)$, $\forall i \in \{1, 2, \ldots, n\}$, is monotonically increasing with respect to the local frequency $\omega_i$, and the phase angles at the equilibrium satisfy $\left|\theta_i^* - \theta_j^*\right| \in [0, \pi/2)$ for all buses $i$ connected to $j$, then the system (2) exists a unique equilibrium that is locally exponentially stable.

Proofs can be found in [12]. According to [12], the phase angle constraint $\left|\theta_i^* - \theta_j^*\right| \in [0, \pi/2)$ is satisfied under most of the practical operating conditions. Therefore, the monotonicity of all flexible resource controllers can be considered as a sufficient condition for the system frequency stability, regardless of the power network topology. This topology-independent sufficient condition indicates that it is a practical and scalable method to first optimize an aggregated FFR droop curve based on the SFR model (1), and then decompose the curve by multiplying different positive participation factors. The distributed execution of these decomposed controllers will guarantee the system frequency stability as long as the aggregated FFR droop curve is monotonic w.r.t. the system frequency.

## III. Optimal Control Problem Formulation

In this section, we first describe the optimal FFR problem under random load disturbances from the perspective of stochastic optimization in Section III-A. Then, we show that this classical formulation can be tricky to solve if the control logic is complex. To address this, we reformulate the problem as a set of MDPs in Section III-B. Finally, in Section III-C, we formulate a frequency stability-constrained meta-RL problem to solve these MDPs.

### A. Stochastic Optimization of FFR Controller

In this subsection, we formulate the optimal FFR problem as a stochastic optimization. To be specific, the frequency quality and regulation cost are balanced through a weighted sum type objective function, and the controller $u$ is defined as a function of local measurements, including the system frequency, to facilitate distributed execution:

$$\begin{cases} \max\limits_{u} \mathbb{E}_{l\sim\mathcal{L}}\left[J = -j_1 - j_2 - j_3\right] \\ \text{s.t. } j_1 = q_1 \sum\limits_{t=0}^{T} \left|u_t\right| \\ \quad j_2 = q_2 \sum\limits_{t=0}^{T} \omega_t^2 \\ \quad j_3 = q_3 \max\limits_{t\in\{1,2,\ldots,T\}} \omega_t^2 \\ \quad \underline{u} \le u \le \bar{u} \\ \quad \text{system dynamics (1)} \\ \quad \text{frequency stability guarantee} \end{cases} \tag{3}$$

where $J$ is the objective consisting of three terms $j_1$, $j_2$, and $j_3$, which denote the control cost, the summed square error of CoI frequency deviations, and the CoI frequency nadir (or peak), respectively; $q_1$, $q_2$, and $q_3$ are the weight coefficients; $\mathbb{E}_{l\sim\mathcal{L}}[\cdot]$ is the expectation taken with respect to the random variable $l$, and $l$ follows a distribution $\mathcal{L}$; $T$ is the duration when the frequency is outside the frequency deadband after each disturbance; $t$ is the index of timesteps with small intervals such as 0.1 s; and $\underline{u}$ and $\bar{u}$ are the total upward and downward regulation capacities of flexible resources in the target control area, respectively.

This optimization formulation casts the optimal FFR problem as an infinite-dimensional optimization, making it challenging to solve. Traditional linear droop control methods simplify the problem by assuming that $u$ is a linear function of the system frequency, i.e., $u = k\omega$, where a single coefficient $k$ is tuned to handle all scenarios. This reduction transforms the infinite-dimensional problem into a one-dimensional problem. However, this simplification leads to suboptimal performance for the following reasons. First, the linearity specification restricts the control flexibility. Flexible resources can provide nonlinear frequency responses, which have been shown in [12] to outperform linear approaches. Second, using a static $k$ to handle all scenarios may be insufficient for balancing frequency deviation and regulation cost across different operating conditions. Intuitively, a gentler droop curve is preferable for small load disturbances to avoid unnecessary power output adjustments of flexible resources, thus keeping frequency deviations within an acceptable range at a low cost. When large disturbances occur, however, steeper droop curves are needed to quickly arrest the frequency and ensure system frequency stability. A static control law represents a compromise for all possible scenarios, aiming for high performance on average. However, it may not be optimal for every specific situation, leaving significant room for improvement.

### B. MDP Formulation

To address the above concerns, this paper removes the static linear type restriction and instead optimizes dynamic nonlinear controllers that can adapt rapidly to each specific disturbance event encountered during operation, although the disturbance sizes cannot be directly observed. To manage the infinite-dimensional challenge, we first reformulate the FFR optimization as a set of MDPs.

For any fixed load disturbance $l$, the FFR process can be formulated as an MDP denoted as a 5-tuple $\langle \mathcal{S}, \mathcal{A}, r, \mathbb{P}, \gamma \rangle$ [27]. $\mathcal{S}$ is the continuous state space. The state vector at timestep $t$ can be denoted as $\boldsymbol{s}_t = \left[\omega_t, \omega_{t-1}, \int \omega \, \mathrm{d}t, p_{m,t}, p_{v,t}, p_{t,t}, p_{inv,t}\right]$. $\mathcal{A}$ is the continuous action space. In this problem, the action $a_t \in \mathcal{A}$ taken at timestep $t$ is the FFR signal $u_t \in [\underline{u}, \bar{u}]$. $r: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function as shown in (4), which maps a state-action pair to a real number. $\mathbb{P}: \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ is the transition kernel, i.e., the system dynamics represented as (1), which maps a state-action pair to a probability distribution over the state space $\Delta^{\mathcal{S}}$. $\gamma \in [0, 1]$ is a discount factor.

$$r_t = -q_1\left|u_t\right| - q_2\omega_t^2 - \max\left(0, \omega_t^2 - \omega_{t-1}^2\right) \tag{4}$$

The FFR controller can be denoted as a policy $u(a|s): \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$, which maps states to action probabilities. We con-

sider policies $u_\phi$ parameterized by neural network parameters $\phi$. A policy can interact with the MDP and collect episodes $\tau=\{\boldsymbol{s}_t,a_t,r_t\}_{t=0}^{T}$ of length $T$. This paper defines an episode as a duration that starts when a load disturbance $l$ occurs and the system frequency deviates from a specific deadband, i.e., 0.015 Hz, and ends when the frequency is restored within the deadband.

Considering the stochastic load disturbances, the FFR optimization problem is actually a set of MDPs. Assume that the load disturbance $l$ occurring in different episodes follows a distribution $\mathcal{L}$. Then, during each episode, the controller encounters an MDP $M$ sampled from a distribution $\mathcal{M}$ with shared $(\mathcal{S},\mathcal{A},r,\gamma)$, but with different dynamics $\mathbb{P}$.

RL algorithms are widely used to find an optimal policy $u$ for an MDP, which maximizes the expected accumulated return within an episode $\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t r_t\right]$ based on the collected episodes. An RL algorithm can be defined as a function (5) [21], which maps the dataset $\mathcal{D}=\{\tau^h\}^H$ consisting of $H$ episodes of the target MDP to policy parameters $\phi\in\Phi$.

$$f(\mathcal{D}):\left((\mathcal{S}\times\mathcal{A}\times\mathbb{R})^T\right)^H\rightarrow\Phi \tag{5}$$

In traditional RL algorithms, $f$ is typically chosen as classical RL algorithms, such as deep $Q$-learning (DQN) [28], deep deterministic policy gradient (DDPG) [29], and proximal policy optimization (PPO) [30], to learn the optimal policy parameters $\phi$. These algorithms solve each MDP independently, requiring the controller to go through numerous episodes with the same $l$ to collect sufficient training data. However, in practical power systems, $l$ is random and non-repetitive, necessitating rapid adaption within each single episode, which is a capability that traditional RL algorithms struggle to achieve.

### C. Frequency Stability-constrained Meta-RL Problem

To achieve fast adaption to each disturbance event without destabilizing the system, we formulate a frequency stability-constrained meta-RL problem. Instead of a static policy $u_\phi$, we optimize a parameterized RL algorithm that can quickly learn the optimal $u_\phi$ for each MDP sampled from the distribution $\mathcal{M}$, which lasts for only one episode. With the objective to maximize the expected return during the whole life of the dynamic policy $u_\phi$, the stability-constrained meta-RL model can be formulated as (6), which includes two simultaneous learning loops.

$$\begin{cases}\max\limits_{\theta}\ \mathbb{E}_{M\sim\mathcal{M}}\left[\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t r_t|f_\theta,u_\phi,M\right]\right]\\ \text{s.t. stability guarantee}\end{cases} \tag{6}$$

where $\mathbb{E}_{M\sim\mathcal{M}}[\cdot]$ denotes the expectation taken with respect to $M$; and $f_\theta$ is an RL algorithm parameterized by $\theta$. The outer loop learns $f_\theta$, while the inner loop, which shares a similar mechanism with traditional RL algorithms, applies the algorithm $f_\theta$ to dynamically update the control policy $u_\phi$ based on the interacting experience with MDPs. An update at timestep $t$ of an episode can be expressed as:

$$\phi\leftarrow f_\theta\left(\mathcal{D}=\{\boldsymbol{s}_i,a_i,r_i\}_{i=0}^{t}\right) \tag{7}$$

where the dataset $\mathcal{D}$ is collected within the current episode under $M$, and it is reset at the beginning of a new episode. An ideal $f_\theta$ must be data-efficient to enable effective adaption within each episode.

Based on this meta-RL framework, we introduce non-linearity through neural network-based inner-loop policy $u_\phi$ and achieve dynamic control logic adjustment with the outer-loop RL algorithm $f_\theta$, which is capable of rapid adaption.

## IV. HNN Architecture

Due to the frequency stability constraint in the stability-constrained meta-RL model (6), existing approaches, such as those in [22] and [23], which are aimed at general unconstrained meta-RL problems, are not directly applicable. Representing hard constraints in a form compatible with the RL framework can be challenging. These constraints are often addressed using penalty terms in the reward function, which may not always ensure strict compliance. In this section, we construct an HNN to parameterize $f_\theta$ and $u_\phi$ in (6) as an event-triggered RL algorithm and a nonlinear droop-based control policy, respectively. This construction ensures that a sufficient condition for system frequency stability is always satisfied. By reformulating the frequency stability constraint in (6) as a network constraint and a trigger condition, (6) is made tractable.

### A. HNN Structure

In (6), each MDP $M$ differs in load disturbance $l$, leading to different dynamics $\mathbb{P}$. However, different dynamics $\mathbb{P}$ also share many similarities such as the generator and inverter dynamics, indicating that optimal policies of different $M$ may also share common features. Accordingly, we divide the policy parameters $\phi$ into fixed network parameters $\phi^f$ and variable external parameters $\phi^v$. Specifically, we model the common parts of different policies with the bottom neural network parameterized by $\phi^f$, and represent an RL algorithm $f_\theta$ with another top neural network, which adapts $\phi^v$ as a variable input of the policy network. The two parts form an HNN structure, as illustrated in Fig. 2.

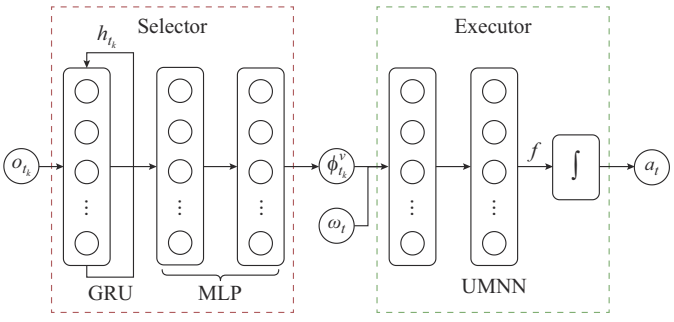


Fig. 2. HNN structure with stability guarantee.

The bottom neural network named executor can be expressed as $u(\omega;\phi)$, which takes the frequency $\omega$ as input and produces the aggregated FFR signal $u$. As common parameters of all policies, $\phi^f$ is optimized during training and

then fixed during implementation, while $\phi^v$ is always updated by the top neural network $f_\theta$ during both stages. The executor $u(\omega;\phi)$ is designed as an unconstrained monotonic neural network (UMNN) [31] to introduce monotonicity, which can be expressed as:

$$\begin{cases} f(\omega;\phi) = \dfrac{\partial u(\omega;\phi)}{\partial \omega} > 0 \\ u(\omega;\phi) = \displaystyle\int_0^{\omega} f(x;\phi)\,\mathrm{d}x \end{cases} \tag{8}$$

where $f(\omega;\phi)$ is a neural network with the input $\omega$ and parameters $\phi$.

First, the partial derivative of $u$ w.r.t. $\omega$, which is a scalar function, is parameterized as the neural network $f(\omega;\phi)$, whose output is forced to be positive through the exponential linear unit (ELU) increased by 1. The output control signal $u$ is then calculated as the integral of the positive partial derivative. In this way, the parameterized policy $u(\omega;\phi)$ is always monotonically increasing w.r.t. the system frequency $\omega$. Namely, the executor can be considered as a cluster of monotonic droop controllers indexed by $\phi^v$ with zero output at $\omega = 0$. Note that the network constraint (8) poses no limitation on the structure of the bottom neural network with parameters $\phi^f$, which can be arbitrarily complex, as long as we set a positive activation function for the final layer and add an integral layer after that.

Once the top neural network updates the output, the bottom neural network executes a different monotonic droop curve indexed by the new $\phi^v$. Therefore, the top neural network is named as the selector. While the executor updates the output at each timestep $t$, the selector works in an event-triggered mode, with the timestep of the $k^{\text{th}}$ trigger denoted as $t_k$. The detailed explanation is deferred to Section IV-B. The input $o_{t_k}$ of the selector is an observation of the system states at timestep $t_k$, which is chosen as $\left[\omega_{t_k}, \omega_{t_k-1}, \omega_{t_k} - \omega_{t_k-1}, \max_{0\le\tau\le t_k}|\omega_\tau|, \phi^v_{t_k-1}\right]$. The top neural network is designed as a recurrent neural network (RNN). The first layer comprises gate recurrent units (GRUs) [32], which introduces recurrency to store historical observation and action information in the hidden state $h_{t_k}$. $h_0$ is initialized as zeros at the beginning of each episode. The following multi-layer perceptron (MLP) learns valuable features from the historical information and produces $\phi^v_{t_k}$ accordingly, selecting the droop curve that best adapts the current operating conditions. It is worth noting that the GRU and MLP structures presented here are empirically proven to perform well in our case, but are not mandatory. The top neural network can be structured arbitrarily without constraints.

### *B. Unrolled Structure and Decision Process*

Constrained by (8), if we fix the output $\phi^v$ of the top neural network, the proposed HNN degenerates to a static monotonic controller. Based on this characteristic, we set the selector to work in an event-triggered mode with the following triggering condition:

$$t_{k+1} = \min_{t\in\{t_k+1, t_k+2, \ldots\}} |\omega_t| > |\omega_{t_k}| \tag{9}$$

That is to say, the selector is triggered if and only if the frequency deviation gets worse.

Under the triggering condition (9), the selector dynamically adjusts the droop curve selection according to its observations during the frequency arrest stage. Then, the bottom neural network keeps executing the selected static droop curve until the frequency is settled and recovered, or another disturbance occurs, inducing a larger frequency deviation and triggering the selector to update $\phi^v$. In any case, the whole network stays static and monotonic after the system frequency reaches the nadir or peak, which satisfies the sufficient condition for frequency stability described in Theorem 1.

The unrolled structure of the proposed HNN is given in Fig. 3 to illustrate the decision process of the top neural network in the event-triggered mode.

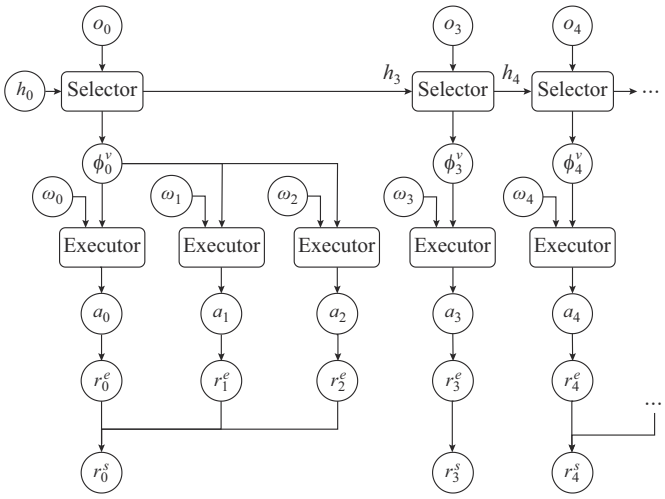


Fig. 3. Unrolled structure of proposed HNN.

At each evenly-spaced timestep $t$, $\omega_t$ is measured, and the action $a_t$, i.e., the control signal $u_t$, is updated by the executor based on $\phi^v_t$ provided by the selector. A reward $r^e_t$ for the single timestep $t$ is then obtained from the environment.

As for the selector, Fig. 3 shows the situation where the selector is triggered at $t_0 = 0$ and $t_1 = 3$. The reward for each trigger $r^s$ is defined as the accumulated individual rewards $r^e$ until the next trigger. For example, the first trigger generates a selection $\phi^v_0$ lasting for three timesteps, so the corresponding reward is calculated as $r^s_0 = \sum_{t=0}^{2} \gamma^t r^e_t$. Limited by space, only five timesteps of a certain episode are presented in Fig. 3. In the subsequent time, the selector will still be triggered whenever the frequency deteriorates.

Figure 4 shows the control logic comparison of the proposed method with two benchmark FFR methods, i.e., static linear droop control method (denoted as method 1) and static nonlinear droop control method (denoted as method 2). In Fig. 4(c), the dashed curves in different colors visualize the control logics of the executor under three different $\phi^v$. The black and blue curves with arrows show two possible dynam-

ic control logics during load disturbance events with different sizes and directions.

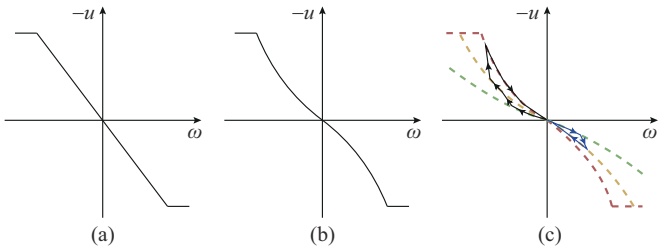


Fig. 4. Control logic comparison of different methods. (a) Method 1. (b) Method 2. (c) Proposed method.

The former analysis indicates that the network constraint (8) and the trigger condition (9) constitute a sufficient but not necessary condition for frequency stability. Consequently, the stability-constrained meta-RL problem (6) can be conservatively reformulated as follows.

$$\max_{\theta} \mathbb{E}_{M\sim\mathcal{M}}\left[\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t r_t | f_\theta, u_\phi, M\right]\right] \tag{10a}$$

$$\text{s.t. (8), (9)} \tag{10b}$$

Compared with (6), the stability constraint is replaced by network shape and trigger condition constraints that are much easier to handle.

## V. Solution Algorithm

The HNN-based meta-RL model (10) enables the optimization of a dynamic droop-based controller with a stability guarantee. Next, the goal is to solve the proposed HNN-based meta-RL problem. Inspired by [33], this section proposes an effective two-stage algorithm to solve (10) through any classical RL algorithm. Unlike the algorithm in [33], which targets adaptation over many episodes (e. g., tens of episodes), the proposed algorithm focuses on achieving much faster adaptation within every single episode.

We view the interaction process from different perspectives and reuse the experience collected by the HNN-based controller. From the view of the selector $f_\theta$, the executor actions $a_t$ and rewards $r_t^e$ can be considered as a part of the environment dynamics. The training data collected during an episode for updating $\theta$ include the selector's observation, action, and the reward for each trigger $k$, which can be denoted as $\mathcal{D}_s = \left\{o_{t_k}, \phi_{t_k}^v, r_{t_k}^s\right\}_{k=1}^{K}$, where $K$ is the total trigger number of the selector within an episode. Then, from the view of the executor, the decision process of the selector can be treated as environment transitions. The system frequency and the selector's action constitute the executor's observation $\sigma_t = \left\{\omega_t, \phi_{t_k}^v\right\}$. The training data for the executor can be expressed as $\mathcal{D}_e = \left\{\sigma_t, a_t, r_t\right\}_{t=0}^{T}$. After collecting the interaction experience of multiple episodes, any off-the-shelf RL algorithms can be used to train the network by mapping the experience buffers $\mathcal{D}_s$ and $\mathcal{D}_e$ to new parameters $\theta$ and $\phi^f$, respectively. However, we observed that simultaneous training of both selector and executor from randomly initialized $\theta$ and $\phi^f$ leads to poor performance.

To optimize the training process and achieve high performance, we propose a two-stage algorithm, which is summarized in Algorithm 1, along with the implementation process. Hyper-parameters $i$ and $j$ are the indices for the neural network updates and episodes, respectively, with a total number of $I$ and $J$. Their superscripts $e$ and $u$ distinguish the executor and united training stages.

**Algorithm 1**: HNN-based meta-RL for optimal FFR

**Initialize**: $\theta$, $\phi^f$
**Executor training**:
 **for** $i^e = \left\{0, 1, \ldots, I^e\right\}$ **do**
  Initialize an empty executor experience buffer $\mathcal{D}_e$
  **for** $j^e = \left\{0, 1, \ldots, J^e\right\}$ **do**
   Sample an MDP $M_l \sim \mathcal{M}$, and fix $\phi^v = l$
   Collect $T$ timesteps of experience using $u_\phi$
  **end for**
  Update $\phi^f$ based on $\mathcal{D}_e$
 **end for**
**United training**:
 **for** $i^u = \left\{1, 2, \ldots, I^u\right\}$ **do**
  Initialize an empty executor experience buffer $\mathcal{D}_e$
  Initialize an empty selector experience buffer $\mathcal{D}_s$
  **for** $j^u = \left\{1, 2, \ldots, J^u\right\}$ **do**
   Sample an MDP $M_l \sim \mathcal{M}$
   Collect $T$ timesteps of experience using $f_\theta$ and $u_\phi$
  **end for**
  Update $\phi^f$ based on $\mathcal{D}_e$, and update $\theta$ based on $\mathcal{D}_s$
 **end for**
**Implementation**:
 **if** $|\omega| > |\omega_{db}|$ **then**
  Begin an FFR episode, and initialize $\phi^v = 0$ and $h_0 = 0$
  **for** timestep $t = 0, 1, \ldots$ **do**
   Get an observation $o$
   **if** $|\omega| < |\omega_{db}|$ **then**
    **Break**
   **else**
    **if** condition (9) is satisfied **then**
     Select $\left(\phi^v, h\right) \leftarrow f_\theta(o, h)$
    **end if**
     Execute $a = u\left(\omega; \left(\phi^v, \phi^f\right)\right)$
   **end if**
  **end for**
 **end if**

1) Executor training stage

At the first stage, only the executor is trained to get a cluster of diversified droop curves. Since the load disturbance $l$ is a key parameter for distinguishing different MDPs, we block the selector and set the selection $\phi^v$ to be $l$. Note that although the disturbance $l$ cannot be measured during the application, it is available during training and is ex-

clusively used at the executor training stage. Only executor experience $\mathcal{D}_e$ is collected at this stage, based on which $\phi^f$ is iteratively updated.

2) United training stage

The selector network $f_\theta$ is activated at this stage, generating $\phi^v$ as the input of the executor trained at the first stage. The whole HNN interacts with the environment. The experience collected at this stage is reused to generate both $\mathcal{D}_s$ and $\mathcal{D}_e$, and parameters $\theta$ and $\phi^f$ are simultaneously updated.

3) Implementation

The implementation part in Algorithm 1 serves as a summary of the controller decision process introduced in Section IV-B. It's worth noting that, although the two training stages take hours, the time required for control signal calculation during the implementation is only a matter of milliseconds. This makes it highly suitable for practical online applications in the context of FFR. Detailed time consumption data can be found in Section VI.

The executor training state before the united training has been empirically validated to improve the final performance significantly. Through Algorithm 1, we learn a parameterized RL algorithm $f_\theta$ capable of fast adaption through classical RL algorithms. Detailed simulation results are provided in Section VI to show the effectiveness of Algorithm 1.

## VI. Case Studies

### A. Simulation Settings

The effectiveness of the proposed HNN-based meta-RL model and the solution algorithm is validated via numerical simulations. The block diagram of the simulation system is shown in Fig. 1. The simulation system is constructed on the Python platform using the OpenAI Gym framework. The system parameters are listed in Table I.

TABLE I
System Parameters

| Parameter | Value | Parameter | Value | Parameter | Value |
|---|---|---|---|---|---|
| $M$ | 9.2 s | $D$ | 2.0 p.u. | $T_g$ | 0.1 |
| $T_r$ | 12 s | $T_{ch}$ | 0.3 s | $F_{hp}$ | 0.2 |
| $R$ | 0.07 | $T_{inv}$ | 0.2 s | $K_p$ | 0.15 |
| $K_i$ | 0.015 | $\alpha_g$ | 0.5 | $\alpha_{inv}$ | 0.5 |
| $\omega_{db}$ | 0.03 | $\beta$ | 24 | | |

The control interval of the optimized FFR controller is set to be 0.1 s. For more realistic simulations of practical systems, AGC in Fig. 1 is set to update the control signal every 4 s with a transmission delay of 1.5 s. The frequency deadband for flexible resource-based FFR is set to be ±0.015 Hz. The selector in Fig. 2 is designed as a 16-unit GRU layer and an MLP composed of two fully connected 32-unit layers. The executor is designed as two fully connected 16-unit layers before the integral layer. The parameters required in Algorithm 1 are set to be $I^e=500$, $J^e=15$, $I^u=3000$, $J^u=15$, and $T=2400$. The widely used PPO algorithm [30] is leveraged to update the network parameters. Discount factor $\gamma$ in (6) is set to be 0.999. The disturbance $l$ of different MDPs is set to uniformly distributed within the range $[0.01, 0.1]$. The total FFR capacity of flexible resources is ±0.08 p. u., and the total PFR capacity of generators is ±0.07 p. u.. The weight coefficients (3) are chosen as $q_1=0.1$, $q_2=0.125$, and $q_3=5$. A single NVIDIA Quadro P2200 GPU with 5 GB memory is used to train the HNN.

### B. Result Analysis

The time required for the executor training and united training stages is 2 hours and 10 hours on average, respectively. During the implementation stage, the calculation time for the selector and the executor is 0.3 ms and 0.7 ms on average, respectively, which is fast enough for practical online applications.

Time-domain simulations on the system illustrated in Fig. 1 are performed using the well-trained HNN-based controller. The dynamics of FFR signals $u$ and frequencies $\omega$ under step load disturbances $l$ of sizes 0.01 p.u., 0.04 p.u., 0.07 p.u., and 0.1 p.u. are shown in Fig. 5.

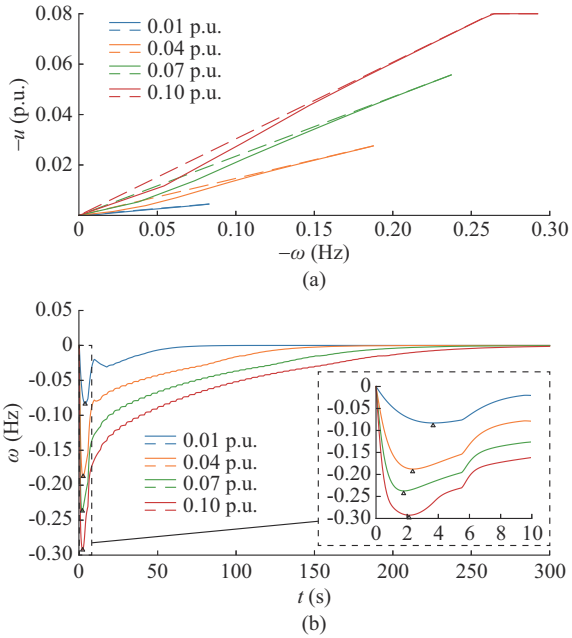


Fig. 5. Dynamics of FFR signals and frequencies under step load disturbances of different sizes. (a) FFR signals. (b) Frequencies.

Figure 5(a) shows the dynamics of FFR signals $u$ for flexible resources w. r. t. system frequency. For each disturbance size, the solid line shows the trajectories of $u$ during the frequency arrest period before the system frequency $\omega$ reaches the nadir. The dashed line illustrates the droop curve during the frequency rebound and recovery periods. The frequency nadir is marked by the triangle in Fig. 5(b). Note that the deadband of FFR is not shown in Fig. 5(a) for simplicity and clarity, but considered during simulation by resetting $u$ as 0 when $|\omega|<0.015$ Hz. The trajectories of $u$ validate the adaptability of the proposed method. To balance the control cost and frequency deviations, the proposed method executes steeper curves under larger disturbances to arrest the system frequency and avoid a catastrophic frequency nadir. In contrast, gentler curves are applied during relatively minor dis-

turbance event to suppress frequency deviation within an acceptable range at a moderate control cost. Figure 5(b) shows that the system frequency is quickly arrested within 1 to 4 s and then recovered to the nominal value under the joint action of both primary and secondary frequency regulations.

To further show the adaptability of the proposed method, it is tested under consecutive step disturbances. Specifically, a 0.04 p.u. load disturbance and a 0.06 p.u. load disturbance occur at $t=0$ and $t=30$ s, respectively. The dynamics of FFR signals and frequencies under the consecutive step disturbances are shown in Fig. 6.

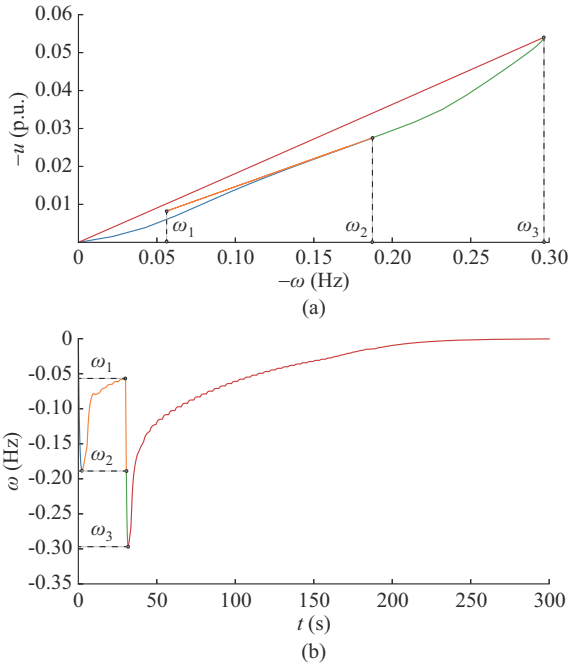


Fig. 6. Dynamics of FFR signals and frequencies under consecutive step disturbances. (a) FFR signals. (b) Frequencies.

The curves in Fig. 6 are divided into four pieces in different colors. The blue piece depicts the dynamics from the beginning of the first disturbance to the first frequency nadir $\omega_2$. During this period, the selector and the executor are both actuated. Then, the orange piece shows the dynamics during the period when the frequency rebounds to $\omega_1$ at $t=$ 30 s and falls again to $\omega_2$ after the occurrence of the second disturbance. According to the triggering condition (9), the selector is deactivated during this period because the frequency has not deteriorated. A fixed nonlinear droop curve is executed as shown in Fig. 6(a). The green piece denotes the frequency arrest period from $\omega_2$ to $\omega_3$. Here, the selector is actuated again to choose steeper droop curves that can better adapt to the frequency dynamics after the occurrence of the second disturbance. Then, the newly chosen droop curve in red is executed until the frequency is recovered to the nominal value. The piece-wise dynamics in Fig. 6 show that the proposed method can switch working states reasonably based on the triggering condition (9). This switching mode not only ensures the transient frequency stability of the system but also enables the controller to adapt to a wider range of operating conditions.

### C. Method Comparison

This subsection compares the performance of the proposed method with the two benchmark FFR methods. Method 1 is static linear droop control with a typical droop value of 1%, whose droop curve is shown in Fig. 7(a). Method 2 is static nonlinear droop control trained by the standard RL algorithm PPO without incorporating meta-learning techniques. It is parameterized by a UMNN network that is the same as the selector of the proposed HNN to ensure the frequency stability. The same reward function (4) is employed for training. This control method takes frequency $\omega$ as the single input, resulting in a static nonlinear droop curve, as depicted in Fig. 7(b).

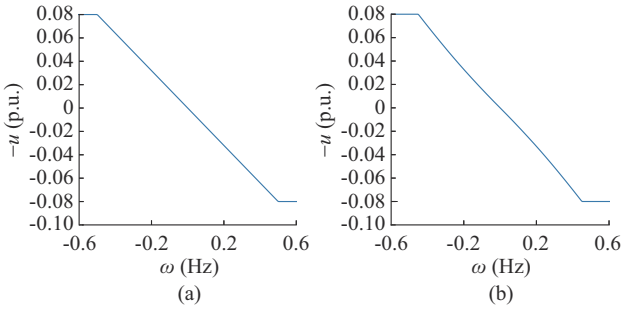


Fig. 7. Droop curves of two benchmark FFR methods for flexible resources. (a) Droop curve of method 1. (b) Droop curve of method 2.

The optimal control objective value $J$ in (3) and the proportion of the control cost term $j_1$ under various step load disturbances are listed in Table II. The objective value $J$ is largely affected by the disturbance size $l$. To better show the relative performance of different methods, we define a performance metric as:

$$P=\left(J-J_{m1}\right)\Big/\left|J_{m1}\right| \tag{11}$$

where $J_{m1}$ is the objective value of method 1. The numerator is an absolute value because the objective values are all negative. The performance of different methods under various load disturbances is plotted in Fig. 8.

TABLE II
PERFORMANCE AND CONTROL COST COMPARISONS OF DIFFERENT METHODS

| $l$ (p.u.) | Method 1 | | Method 2 | | Proposed | |
|---|---|---|---|---|---|---|
| | $J$ | $j_1$ (%) | $J$ | $j_1$ (%) | $J$ | $j_1$ (%) |
| 0.01 | –0.22 | 78 | –0.22 | 78 | –0.15 | 40 |
| 0.02 | –0.49 | 68 | –0.48 | 67 | –0.42 | 39 |
| 0.03 | –0.80 | 60 | –0.80 | 60 | –0.76 | 41 |
| 0.04 | –1.16 | 53 | –1.16 | 53 | –1.15 | 43 |
| 0.05 | –1.58 | 48 | –1.57 | 48 | –1.58 | 45 |
| 0.06 | –2.06 | 44 | –2.04 | 44 | –2.04 | 46 |
| 0.07 | –2.61 | 40 | –2.56 | 41 | –2.54 | 46 |
| 0.08 | –3.25 | 36 | –3.16 | 38 | –3.08 | 46 |
| 0.09 | –3.98 | 33 | –3.83 | 35 | –3.67 | 45 |
| 0.10 | –4.80 | 31 | –4.58 | 33 | –4.31 | 44 |

From Fig. 8, method 2 and the proposed method perform better than method 1 in all cases. As shown in Fig. 7(b), the

droop curve of method 2 becomes steeper as the frequency deviations get larger, which can be considered as a generalization of the piece-wise linear droop control method in [10]. However, such bending in the droop curve has limited improvement in the performance due to its static feature. The proposed method can dynamically modify the droop curve to realize adaptability to a greater extent. As shown in Fig. 5(a), the dynamics of FFR signals in different cases can be different even at a same frequency deviation level. After a larger disturbance, the frequency response is faster from the beginning of the event instead of accelerating after the frequency deviation reaches a high level. Consequently, the proposed method achieves the best performance in almost all cases.

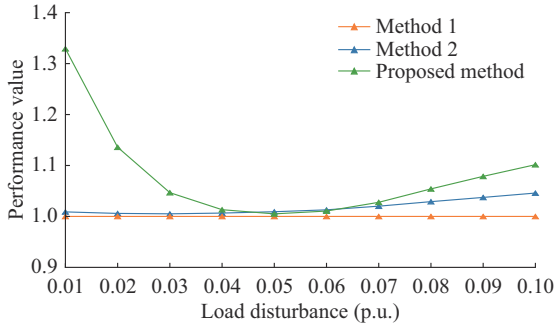


Fig. 8. Performance of different methods under various load disturbances.

Compared with other methods, the proportion of $j_1$ obtained by the proposed method is higher under larger disturbances and lower under smaller disturbances, as shown in Table II. Such results indicate that the proposed method can reasonably balance the control cost and frequency deviations case by case to achieve higher control performance.

### D. Algorithm Comparison

The proposed algorithm has an executor training stage before the united training. To validate the effectiveness of the proposed algorithm, this subsection compares the performance of the proposed algorithm and another algorithm performing united training only (denoted as algorithm 2). The performance comparison of different algorithms is shown in Fig. 9.

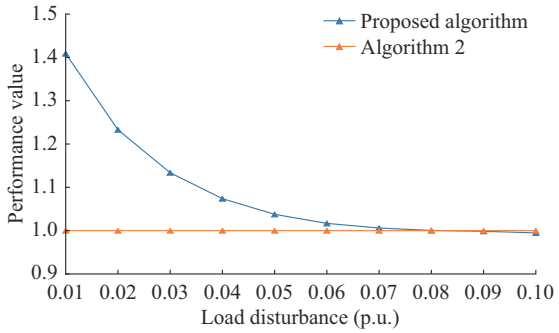


Fig. 9. Performance comparison of different algorithms.

It can be observed from Fig. 9 that the proposed algorithm with the executor training stage outperforms algorithm 2 in most cases. Intuitively, the executor training stage helps the executor acquire a cluster of meaningful skills. In comparison, performing united training from the beginning may cause insufficient or meaningless exploration and lead to poor training effect.

### E. Sensitivity Analysis

The objective of the optimal control problem is formulated as the weighted sum of different terms in (3) to balance the control cost and frequency deviations. Different values of weight coefficients $q_1$, $q_2$, and $q_3$ in (3) result in different trade-offs. This subsection takes the coefficient $q_1$ as an example to show the impact of weight coefficients on the optimization results of the proposed method. The value of $q_1$ is set to be 0.4, 0.1, and 0.025, respectively. The dynamics of frequencies $\omega$ and FFR signals $u$ after step load disturbances with size $l=0.1$ p.u. and $l=0.05$ p.u. are plotted in Fig. 10.

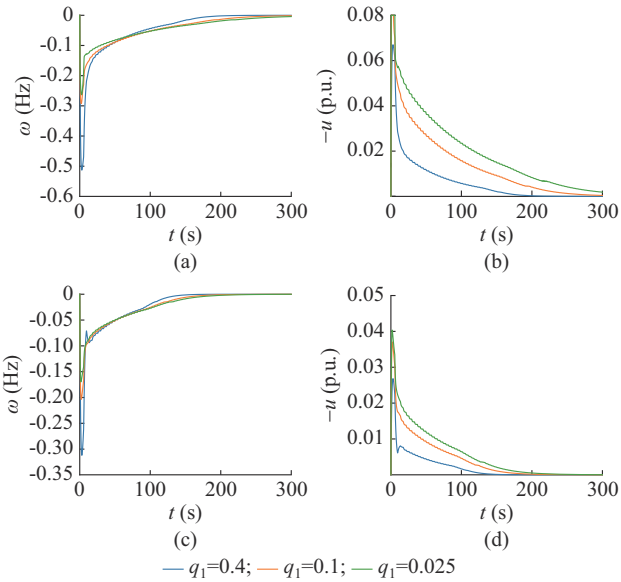


Fig. 10. Dynamics of frequencies and FFR signals after step load disturbances with size $l=0.1$ p.u. and $l=0.05$ p.u.. (a) Dynamics of frequencies with $l=0.1$ p.u.. (b) Dynamics of FFR signals with $l=0.1$ p.u.. (c) Dynamics of frequencies with $l=0.05$ p.u.. (d) Dynamics of FFR signals with $l=0.05$ p.u..

A larger $q_1$ value denotes a higher cost of flexible resource-based FFR service. As shown in Fig. 10, the proposed method optimized with a higher $q_1$ value tends to utilize less frequency regulation resources at the cost of larger frequency deviations. Consequently, the transmission system operators should fine-tune the weight coefficients according to the actual regulation cost of flexible resources and requirements for frequency quality based on numerical simulations before practical implementations.

### F. Method Applicability in Other System Types

Although the SFR model depicted in Fig. 1 incorporates only two types of frequency regulation resources, the proposed method is applicable to larger load frequency control systems with diverse resource types. To validate such applicability, we modify the SFR model in Fig. 1 and conduct simulations under the same settings as introduced in Section VI-A. This modified SFR model incorporates an additional

type of frequency regulation resource, namely an aggregated non-reheat generator, into the original system model by substituting the synchronous generator block in Fig. 1 with Fig. 11. $T_{g,nr}$ and $T_{ch,nr}$ are the time constants of the equivalent governor and turbine, respectively, for the aggregated non-reheat generator. The proportion of reheat and non-reheat generators can be adjusted by modifying the values of $K_r$ and $K_{nr}$, respectively. In this case study, we set $K_r = K_{nr} = 0.5$.

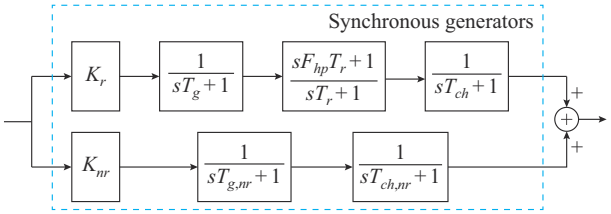


Fig. 11. Block diagram of reheat and non-reheat generators.

We also compare the proposed method with the two benchmark methods as detailed in Section VI-C. Method 1 maintains its typical droop value of 1%. Method 2 and the proposed method undergo training using PPO and the proposed algorithm, respectively, under the modified SFR model. The control objective value $J$ under various load disturbances are presented in Table III. Additionally, the relative performance of three different methods, calculated using (11), is illustrated in Fig. 12. Based on the simulation results, the proposed method shows significant superiority over the benchmarks as in Section VI-C, validating its adaptability to different types of power systems.

TABLE III
CONTROL PERFORMANCE IN MODIFIED SFR MODEL UNDER VARIOUS LOAD DISTURBANCES

| $l$ (p.u.) | $J$ | | |
|---|---|---|---|
| | Method 1 | Method 2 | Proposed |
| 0.01 | –0.28 | –0.23 | –0.11 |
| 0.02 | –0.57 | –0.49 | –0.33 |
| 0.03 | –0.89 | –0.79 | –0.64 |
| 0.04 | –1.25 | –1.15 | –1.03 |
| 0.05 | –1.66 | –1.56 | –1.47 |
| 0.06 | –2.11 | –2.02 | –1.97 |
| 0.07 | –2.61 | –2.53 | –2.51 |
| 0.08 | –3.15 | –3.10 | –3.09 |
| 0.09 | –3.73 | –3.82 | –3.72 |
| 0.10 | –4.41 | –4.69 | –4.43 |

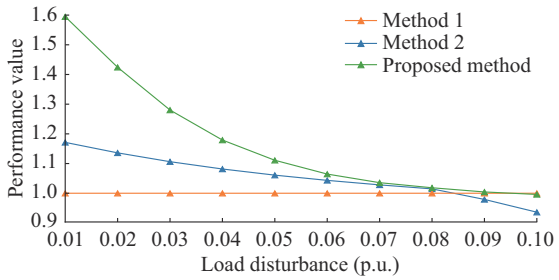


Fig. 12. Performance comparisons of different methods in modified SFR model.

## VII. CONCLUSION

This paper investigates the flexible resource-based FFR optimization problem considering the guarantee of system frequency stability. A new meta-RL approach is proposed to realize dynamic nonlinear *P-f* droop-based FFR with rapid adaptability to different operating conditions.

We first formulate a frequency stability-constrained meta-RL problem, then reformulate it into a more tractable HNN-based form with the well-designed network constraint and trigger condition. A two-stage algorithm is proposed to enhance the optimality in solving the HNN-based meta-RL problem. Simulation results validate that the proposed method can adapt rapidly to different operating conditions with the system frequency stability guaranteed. Compared with benchmarks including static linear control and static nonlinear control methods, the proposed method achieves better trade-offs between frequency quality and regulation cost. Future research directions include the coordinated FFR optimization of multiple inter-connected control areas and the differentiated utilization of heterogeneous flexible resources in FFR.

## REFERENCES

[1] R. W. Kenyon, M. Bossart, M. Marković *et al.*, "Stability and control of power systems with high penetrations of inverter-based resources: an accessible review of current knowledge and open questions," *Solar Energy*, vol. 210, pp. 149-168, Nov. 2020.
[2] J. Boyle, T. Littler, S. M. Muyeen *et al.*, "An alternative frequency-droop scheme for wind turbines that provide primary frequency regulation via rotor speed control," *International Journal of Electrical Power & Energy Systems*, vol. 133, p. 107219, Dec. 2021.
[3] F. Sattar, S. Ghosh, Y. J. Isbeih *et al.*, "A predictive tool for power system operators to ensure frequency stability for power grids with renewable energy integration," *Applied Energy*, vol. 353, p. 122226, Jan. 2024.
[4] M. H. Marzebali, M. Mazidi, and M. Mohiti, "An adaptive droop-based control strategy for fuel cell-battery hybrid energy storage system to support primary frequency in stand-alone microgrids," *Journal of Energy Storage*, vol. 27, p. 101127, Feb. 2020.
[5] M. Mousavizade, F. Bai, R. Garmabdari *et al.*, "Adaptive control of V2Gs in islanded microgrids incorporating EV owner expectations," *Applied Energy*, vol. 341, p. 121118, Jul. 2023.
[6] C. Christiansen and N. Hillmann. (2017, May). Feasibility of fast frequency response obligations of new generators. [Online]. Available: https://www. aemc. gov. au/sites/default/files/content/661d5402-3ce5-4775-bb8a-9965f6d93a94/AECOM-Report-Feasibility-of-FFR-Obligations-of-New-Generators.pdf
[7] L. Meng, J. Zafar, S. K. Khadem *et al.*, "Fast frequency response from energy storage systems – a review of grid standards, projects and technical issues," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1566-1581, Mar. 2020.
[8] National Grid Group. (2016, Mar.). Enhanced frequency response: frequently asked questions. [Online]. Available: https://www.nationalgrid.com/sites/default/files/documents/Enhanced% 20Frequency% 20Response%20FAQs%20v5.0_.pdf
[9] P. Du, N. V. Mago, W. Li *et al.*, "New ancillary service market for ERCOT," *IEEE Access*, vol. 8, pp. 178391-178401, Sept. 2020.
[10] Y. Yuan, Y. Zhang, J. Wang *et al.*, "Enhanced frequency-constrained unit commitment considering variable-droop frequency control from converter-based generator," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1094-1110, Mar. 2023.
[11] M. F. M. Arani and Y. A. R I. Mohamed, "Cooperative control of wind power generator and electric vehicles for microgrid primary frequency regulation," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5677-5686, Nov. 2018.
[12] W. Cui, Y. Jiang, and B. Zhang, "Reinforcement learning for optimal primary frequency control: a Lyapunov approach," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1676-1688, Mar. 2023.
[13] C. Zhao, U. Topcu, N. Li *et al.*, "Design and stability of load-side pri-

mary frequency control in power systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1177-1189, May 2014.
[14] Y. Liu, Y. Song, Z. Wang *et al*., "Optimal emergency frequency control based on coordinated droop in multi-infeed hybrid AC-DC system," *IEEE Transactions on Power Systems*, vol. 36, no. 4, pp. 3305-3316, Jul. 2021.
[15] Z. Ding, K. Yuan, J. Qi *et al*., "Robust and cost-efficient coordinated primary frequency control of wind power and demand response based on their complementary regulation characteristics," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4436-4448, Nov. 2022.
[16] E. Ekomwenrenren, J. W. Simpson-Porco, E. Farantatos *et al*. (2022, Aug.). Data-driven fast frequency control using inverter-based resources. [Online]. Available: https://arxiv.org/abs/2208.01761
[17] E. Ekomwenrenren, Z. Tang, J. W. Simpson-Porco *et al*., "Hierarchical coordinated fast frequency control using inverter-based resources," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 4992-5005, Nov. 2021.
[18] R. Chakraborty, A. Chakrabortty, E. Farantatos *et al*., "Hierarchical frequency control in multi-area power systems with prioritized utilization of inverter based resources," in *Proceedings of 2020 IEEE PES General Meeting*, Montreal, Canada, Aug. 2020, pp. 1-5.
[19] Q. Yang, L. Yan, X. Chen *et al*., "A distributed dynamic inertia-droop control strategy based on multi-agent deep reinforcement learning for multiple paralleled VSGs," *IEEE Transactions on Power Systems*, vol. 38, no. 6, pp. 5598-5612, Nov. 2023.
[20] Z. Yan, Y. Xu, Y. Wang *et al*., "Deep reinforcement learning-based optimal data-driven control of battery energy storage for power system frequency support," *IET Generation*, *Transmission & Distribution*, vol. 14, no. 25, pp. 6071-6078, Dec. 2020.
[21] J. Beck, R. Vuorio, E. Z. Liu *et al*. (2023, Jan.). A survey of meta-reinforcement learning. [Online]. Available: https://arxiv. org/abs/2301. 08028
[22] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *International Conference on Machine Learning*, Sydney, Australia, Aug. 2017, pp. 1126-1135.
[23] Y. Duan, J. Schulman, X. Chen *et al*. (2016, Nov.). $RL^2$: fast reinforcement learning via slow reinforcement learning. [Online]. Available: https://arxiv.org/abs/1611.02779
[24] J. Li, T. Zhou, K. He *et al*., "Distributed quantum multiagent deep meta reinforcement learning for area autonomy energy management of a multiarea microgrid," *Applied Energy*, vol. 343, p. 121181, Aug. 2023.
[25] R. Huang, Y. Chen, T. Yin *et al*., "Learning and fast adaptation for grid emergency control via deep meta reinforcement learning," *IEEE Transactions on Power Systems*, vol. 37, no. 6, pp. 4168-4178, Nov. 2022.
[26] Q. Shi, F. Li, and H. Cui, "Analytical method to aggregate multi-machine SFR model with applications in power system dynamic studies," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6355-6367, Nov. 2018.
[27] D. L. Poole and A. K. Mackworth, *Artificial Intelligence*. Cambridge, UK: Cambridge University Press, 2010.
[28] V. Mnih, K. Kavukcuoglu, D. Silver *et al*., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, Feb. 2015.
[29] T. P. Lillicrap, J. J. Hunt, A. Pritzel *et al*. (2015, Sept.). Continuous control with deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1509.02971
[30] J. Schulman, F. Wolski, P. Dhariwal *et al*. (2017, Jul.). Proximal policy optimization algorithms. [Online]. Available: https://arxiv. org/abs/1707.06347
[31] A. Wehenkel and G. Louppe, "Unconstrained monotonic neural networks," in *Proceedings of 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, Jun. 2019, pp. 1545-1555.
[32] K. Cho, B. van Merriënboer, D. Bahdanau *et al*. (2014, Sept.). On the properties of neural machine translation: encoder-decoder approaches. [Online]. Available: https://arxiv.org/abs/1409.1259
[33] K. Frans, J. Ho, and X. Chen. (2017, Oct.). Meta learning shared hierarchies. [Online]. Available: https://arxiv.org/abs/1710.09767

**Yuxin Ma** received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2021, where she is currently working toward the Ph.D degree in electrical engineering. Her current research interests include optimal operation and control of energy storage system and power system.

**Zechun Hu** received the B.S. and Ph.D. degrees in electrical engineering from Xi'an Jiao Tong University, Xi'an, China, in 2000 and 2006, respectively. He was with Shanghai Jiao Tong University, Shanghai, China and University of Bath, Bath, UK as a Research Officer from 2009 to 2010. He joined the Department of Electrical Engineering, Tsinghua University, Beijing, China, in 2010, where he is currently an Associate Professor. His major research interests include optimal planning and operation of power system, electric vehicle, and energy storage system.

**Yonghua Song** received the B.E. degree from the Chengdu University of Science and Technology, Chengdu, China, in 1984, and the Ph.D. degree from the China Electric Power Research Institute, Beijing, China, in 1989, all in electrical engineering. From 1989 to 1991, he was a Post-doctoral Fellow with Tsinghua University, Beijing, China. He then held various positions with Bristol University, Bristol, UK, Bath University, Bath, UK, and John Moores University, Liverpool, UK, from 1991 to 1996. In 1997, he was a Professor of power systems with Brunel University, London, UK, where he has been a Pro-Vice Chancellor for Graduate Studies since 2004. In 2007, he took up a Pro-Vice Chancellorship and Professorship of electrical engineering with the University of Liverpool, Liverpool, UK. In 2009, he joined Tsinghua University as a Professor of electrical engineering and an Assistant President and the Deputy Director of the Laboratory of Low-carbon Energy. From 2012 to 2017, he was the Executive Vice President of Zhejiang University, Hangzhou, China, as well as the Founding Dean of the International Campus and a Professor of electrical engineering and higher education. Since 2018, he has been a Rector of the University of Macau, Macau, China. He was elected as the Vice President of Chinese Society for Electrical Engineering (CSEE) and appointed as the Chairman of the International Affairs Committee of the CSEE in 2009. In 2004, he was elected as a Fellow of the Royal Academy of Engineering, UK. He was a recipient of the D.Sc. Award by Brunel University, in 2002, for his original achievements in power system research. His current research interests include smart grid, electricity economics, and operation and control of power system.