https://doi.org/10.1016/j.ijepes.2015.11.057

Hierarchical Correlated Q-Learning for Multi-layer Optimal Generation Command Dispatch

T. Yu[b], X. S. Zhang[b], B. Zhou[a*] and K. W. Chan[c]

[a]College of Electrical and Information Engineering, Hunan University, Changsha 410082, China

[b]College of Electric Power, South China University of Technology, Guangzhou 510640, China

[c]Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong

***Abstract*—**This paper presents a novel hierarchical correlated Q-learning (HCEQ) algorithm to solve the dynamic optimization of generation command dispatch (GCD) in the automatic generation control (AGC). The GCD problem is to dynamically allocate the total AGC generation command from the central to each individual AGC generator. The proposed HCEQ is a novel multi-agent Q-learning algorithm based on the concept of correlated equilibrium point, and each AGC generator with an agent is to optimize its regulation participation factor and coordinate its decision with others for the overall GCD performance enhancement. In order to cope with the curse of dimensionality in the GCD problem with the increased number of AGC plants involved, a multi-layer optimum GCD framework is developed in this paper. In this hierarchical framework, the multiobjective design and a time-varying coordination factor have been formulated into the reward functions to improve the optimization efficiency and convergence of HCEQ. The application of the proposed approach has been fully verified on the China southern power grid (CSG) model to demonstrate its superior performance and dynamic optimization capability in various power system scenarios.

***Key Words*—**Hierarchical multi-agent reinforcement learning; Correlated equilibrium; Automatic generation control; Dynamic generation allocation; Control Performance Standards

1. Introduction

Automatic Generation Control (AGC) of interconnected power grids is one of the key control systems in the power dispatch centers, and its main objective is to maintain the scheduled interconnection frequency and tie-line power interchanges by regulating the generation outputs of AGC plants to accommodate the fluctuating load demands [1]. The implementation of AGC regulating commands on various AGC plants is

*Corresponding author at: College of Electrical and Information Engineering, Hunan University, 410082 Changsha, China. Tel.: +86 731 8388 9677; fax: +86 731 88664197. E-mail addresses: binzhou@hnu.edu.cn (B. Zhou).

critical to the overall control performance of AGC schemes, and this generation command dispatch (GCD) is a real-time combination optimization problem whose complexity would increase with the number of AGC committed generators being involved [2]. Further complications are the additional considerations on adjustable margin reserve and regulating cost for each AGC unit, and hence this problem cannot be solved using conventional methods. The primary objective of GCD is to dynamically tune the optimal regulation participation factors of AGC units and thus allocate the real-time central regulating command determined from load frequency control (LFC) to each dispatchable generating unit. Consequently, this paper focuses on investigating the advanced GCD methodology to solve the dynamic optimal allocation of AGC generation among various types of AGC units.

Nowadays, the control area performance of AGC in normal interconnected power system operation has been monitored and measured by area control error (ACE) and control performance standards (CPS) [3]. Over the years, extensive investigations on the AGC strategies under CPS using various mathematical and intelligent control theories, including proportional-integral (PI) control, self-tuning control, fuzzy logics and reinforcement learning (RL), and so on, have been addressed and reported in [4]-[10]. Nevertheless, the previous studies mostly focused on the optimum AGC strategies for the total regulating commands in power dispatch centers, and little attention has been paid on the GCD problem to optimally on-line distribute the total regulating command among various AGC units. So far, the existing engineering method to solve this GCD problem is called the proportional (PROP) method in which the AGC regulation participation factor for each unit is fixed and proportional to the adjustable reserve capacity of the unit [10],[11]. The PROP method has been widely adopted by most power utilities in Chinese power systems. However, this PROP method with the fixed participation factors cannot provide the satisfactory performance over a wide range of operational scenarios of power systems. For the GCD optimization problem, the authors proposed in [2] a novel hierarchical Q-learning (HQL) algorithm, which has been found to be more efficient with improved performance than the PROP method.

In recent years, a new branch of RL theory, multi-agent reinforcement learning (MARL), has been growing rapidly and applied widely in a variety of fields, including collaborative decision support systems, distributed control, robotic teams and economics [12]. Previous applications have been demonstrated that, compared with the single-agent RL methods, the overall performance of MARL can exhibit the superiority and optimality on the cooperative strategic decision making problems [13]. In general, most of the MARL algorithms concern the game theory, and the optimized payoff states in a dynamic MARL game can be solved and represented by different equilibrium points, such as Nash equilibrium [14] and Correlated equilibrium [15]. Different equilibriums express different levels of cooperation degree for the decentralized multi-agents, and a promising cooperative MARL algorithm based on the correlated equilibria point, called correlated-Q learning (CEQ), has been proposed in [15]. This paper is a follow-up research of the authors' previous investigations reported in [2], [7], [8] and [16]. The Q(λ) learning [7] and R(λ) learning [8] were applied for optimizing the total AGC regulating command, while the single-agent Q-learning was adopted in [2] for dynamic optimal GCD scheme. Besides, a distributed Q(λ) learning is proposed in [16] to solve the large-scale optimal power flow problem. Compared with the previously published works, this research further focuses on developing a novel MARL algorithm to form a significantly improved GCD scheme under CPS standards. The proposed MARL-based hierarchical correlated Q-learning (HCEQ) considers the coordination of implemented actions and information interaction among the MARL agents to optimize the joint equilibrium actions of AGC generators for the improved overall GCD performance, and it has been thoroughly tested and evaluated on the China southern power grid (CSG) model under various operational scenarios.

2. Problem Formulation

2.1. Overview of AGC Implementation

In modern AGC schemes, the generation dispatch strategies and control pulses for each interconnected control area are always determined and maintained by a central grid facility, called power dispatch center [11]. Usually, the control area is an electric power utility for an individual service area, taking provincial power grids in the CSG power system as an example. The control area's AGC scheme is implemented by two main control modules in the power dispatch center, as is shown in Fig.1. The optimal AGC controller is a closed-loop feedback control to optimize the solution of total regulating generation command $\Delta P_{C\Sigma}$ in response to the load disturbance ΔP_L. The existing AGC controllers under CPS standards are generally based on the PI control strategies as suggested in [4],[5], and most power dispatch centers in China have adopted an improved-PI based AGC controller developed by Nanjing Automation Research Institute (NARI) [10]. The AGC command $\Delta P_{C\Sigma}$ is a reference control signal and will be allocated from the central to each

AGC unit according to their regulation participation factors. On the other hand, the GCD module determines dynamically the optimized participation factors and a reference command ΔP_{Ci} will then be delivered to the ith AGC unit through Supervisory Control and Data Acquisition (SCADA) system [1].

It should be pointed out that the dynamic GCD problem in this paper is different from the economic dispatch (ED) [17] because AGC (secondary frequency control) and ED (tertiary frequency control) have different time horizons and control objectives. ED is performed to distribute the system base load amongst all dispatchable generators so that the generation costs can be minimized, and the CPS standards are not covered in ED function, while the objective of GCD function is the dynamic allocation of AGC regulating command which indicates the modification of AGC generation outputs to balance the load residuals. The implementation cycle of ED is in the range from 5 to 15 minutes, and the GCD is around from 4 to 16 seconds. In the case studies, the AGC decision cycle is set to 8 second in the CSG power system model.

2.2. GCD Objectives and Constraints

In the proposed GCD framework, multiple objectives have been considered and designed. The primary objective is to minimize the accumulated generating error between the reference AGC command ΔP_{Ci} and the actual generation variation ΔP_{Gi}. Moreover, the AGC generators with fast regulation capability, such as hydro AGC generators, should provide sufficient adjustable spinning reserve to cope with the sudden increasing load disturbances [2]. In general, hydropower is recognized as a having the ability to provide fast and efficient generation regulation for power system secondary frequency control. The raising and lowering generation rate constraint (GRC) of hydro generators are ranged from 100% to 360% p.u./min respectively, while the typical GRC of thermal generators is in the range of 3%-10% p.u./min [18],[19]. As for different types of thermal plants, the liquefied natural gas (LNG) turbine can provide faster regulation capability than oil-fired and coal-fired turbines, and hence the LNG plants could be considered as the fast-ramping generators for thermal-dominated power systems without hydropower. Lastly, the regulating cost of AGC plants should also be concerned. The three GCD objectives above can then be formulated as follows,

$$\begin{cases} F_1 = \min \sum_{k=1}^{T} \sum_{i=1}^{N} \Delta P_{ei}^2(k) \\ F_2 = \max \sum_{k=1}^{T} \left[(P_{GF}^{\max} - \sum_{i \in F} P_{Gi}(k)) \Big/ P_{GF}^{\max} \right] \times 100\% \\ F_3 = \min \sum_{k=1}^{T} \sum_{i=1}^{N} C_i[P_{Gi}(k)] \end{cases} \tag{1}$$

where T is the number of iterations in the assessment period; N is the number of AGC units; $P_{GF}^{\max}$ is the total maximum capacity of AGC units with fast regulation capability; F denotes the set of AGC units with fast regulation capability; $P_{Gi}(k)$ is the actual generation output of the ith AGC unit at the kth iteration; C_i denotes the linear cost function of the ith AGC unit, and the mechanical wear-and-tear cost caused by the maneuvering movements of AGC units has been neglected in this application; $\Delta P_{ei}(k)$ represents the power error between the actual generation output and reference command for the ith AGC unit at the kth iteration, and it can be represented as follows,

$$\Delta P_{ei}(k+1) = \Delta P_{ci}(k) - \Delta P_{Gi}(k+1) \tag{2}$$

In this paper, the linear weighted method is adopted to formulate the multiobjective GCD problem because of its simplicity of use and clarity of definition, and the method is applicable to solve the optimal correlated equilibrium (CE) solution with the efficient computational time for real-time applications. For Pareto optimization based RL in [20], each MARL agent has several Q-function matrices to represent different objective functions respectively. It is much more time-consuming for each agent to solve a family of Pareto front solutions [21], so that the real-time requirement of AGC decision cycle, 4~16 seconds, cannot be satisfied. For most of real-time control applications, the multiple objectives cannot be optimized simultaneously for Pareto optimality due to the real-time requirement. Here, the linear weighted method [7] is adopted to transform multiobjective GCD functions in (1) into an integrated objective function, and each MARL agent has only a Q-function matrix for the optimal state-action policy of multi-objective GCD scheme. Consequently, the integrated objective function of each AGC unit can be represented as follows,

$$f_i = \begin{cases} \min \sum_{k=1}^{T} \left(\Delta P_{ei}^2(k) - \mu_1 [P_{Gi}^{\max} - P_{Gi}(k)] / P_{Gi}^{\max} + \mu_2 C_i[P_{Gi}(k)] \right) & \forall i \in F \\ \min \sum_{k=1}^{T} \left(\Delta P_{ei}^2(k) + \mu_2 C_i[P_{Gi}(k)] \right) & \forall i \notin F \end{cases} \tag{3}$$

where μ_1, μ_2 are the optimum weight coefficients for GCD objectives in (3); $P_{Gi}^{\max}$ represents the maximum capacity of the ith AGC unit.

In the optimal GCD problem, four generator constraints are considered with following the problem constraints in [2], including: 1) power generation equality constraint; 2) adjustable capacity constraints of AGC units; 3) ramp rate limit constraints; and 4) generation time-delay response. The first GCD constraint requires that the sum of all reference commands of AGC units should be equal to the total AGC regulating command [22].

3. Hierarchical Correlated Q-learning for GCD Problem

3.1. Correlated Equilibrium

In a Markov game, a CE is a matrix of probability distribution over the joint space of actions from which no agent is motivated to deviate unilaterally [13]. For an action assigned from the joint action policy to every possible observation of the ith agent in state s_k, the CE action policy π can be determined by the following CE inequality constraints,

$$\begin{gathered}\sum_{\vec{a}_{-i}\in A_{-i}(s_k)} \pi(s_k,\vec{a})Q_i^k\left(s_k,(\vec{a}_{-i},a_i)\right) \geq \sum_{\vec{a}_{-i}\in A_{-i}(s_k)} \pi(s_k,\vec{a})Q_i^k\left(s_k,(\vec{a}_{-i},a_i')\right) \\ A_{-i}=\Pi_{j\neq i}A_j, \quad \vec{a}_{-i}=\Pi_{j\neq i}a_j, \quad \vec{a}=(\vec{a}_{-i},a_i), \quad a_i'\neq a_i\end{gathered} \tag{4}$$

where $\vec{a}=[a_1, \ldots, a_i, \ldots, a_n]$, a_i is the ith agent's action (the regulation participation factor of AGC unit i), and n is the number of agents in MARL; s_k is the state of MARL at the kth iteration; $A(s_k)$ is the agents' set of available joint actions in state s_k; A_i is the ith agent's set of pure actions, and A_{-i} is agents' set of joint actions except agent i; $a_i\in A_i$, $\vec{a}_{-i}\in A_{-i}$ express the ith agent's action and other agents' joint actions in the current state; a_i' is the ith agent's any other action except a_i to indicate the non-CE action; $Q_i^k(s_k,\vec{a})$ is the estimated Q-function of agent i for joint action $\vec{a}$ and state s_k at the kth iteration [23]; $\pi(s_k,\vec{a})$ is a vector of probability distribution over joint action set $A(s_k)$ to represent the optimal CE action policy of agent i in state s_k, and it can be uniquely derived from the CE point model with an equilibrium selection function [15]. Furthermore, it has been proven in [13] that there is at least a correlated equilibrium point for any Markov game.

It can be found from (4) that there may be several CE solutions with joint action policies satisfying the CE constraints. Consequently, an equilibrium selection criterion shall be designed to determine uniquely the optimum CE point from the set of CE solutions for a desirable action policy [14]. Four typical variants of equilibrium selection functions based on different equilibrium objectives, including *utilitarian*, *egalitarian*, *plutocratic* and *dictatorial*, have been designed and analyzed in [15] using comparative experiments. In this paper, further analysis on (1)-(3) show that the utilitarian function is more suitable and would be adopted for the design of optimal GCD scheme to maximize the sum of all agents' long-term objective payoffs. On the other hand, for a Markov game with *n* agents and *m* actions for each agent, the number of joint actions in MARL is m^n and the number of the CE constraints (4) is $nm(m\text{-}1)$ [13].

3.2. Correlated Q-learning

The correlated Q-learning is a newly-emerged MARL algorithm based on the CE principle to find the optimal equilibrium policies in cooperative Markov games [15]. MARL can be characterized by four basic elements: a model of the environment, a reward function, value functions and an action policy [23]. In this paper, the model of the environment can be described as a set of operating states including different ranges of AGC regulating commands as in [2], called state space *S*. The reward function is to map each perceived state-action pair of the MARL to a single value so as to express the desirability of the GCD performance. The value function (Q-function) of each state-action pair is defined to estimate the discounted sum of the future sequence of rewards starting from the current state and action policy thereafter. Finally, the action policy specifies a stimulus-response rule to select and implement a joint action from action space *A* based on value functions to maximize the expected long-term rewards in each state. Here, the joint action space *A* consists of a finite set of discrete vectors of joint AGC participation factors for generation allocation.

CEQ defines a state-value function using the linear combination of Q-functions on the basis of the CE action policy, and it expresses the CE cooperative degree of multi-agents in this state [15], as follows,

$$V_i^k(s_k) = \sum_{\vec{a}\in A(s_k)} \pi(s_k,\vec{a}) Q_i^{k-1}(s_k,\vec{a}) \tag{5}$$

where $V_i^k(s_k)$ represents the state value-function of agent *i* for state s_k at the *k*th iteration; $Q_i^{k-1}(s_k,\vec{a})$ is the estimated Q-function of agent *i* for joint action $\vec{a}$ and state s_k at the (*k*-1)th iteration. In the proposed HCEQ, the λ-return mechanism [16] is introduced to update the Q-function of each agent, as follows,

$$\delta_i^k = (1-\gamma) R_i(s_{k-1}, s_k, \vec{a}_{k-1}) + \gamma V_i^k(s_k) - Q_i^{k-1}(s_{k-1}, \vec{a}_{k-1}) \tag{6}$$

$$Q_i^k(s,\vec{a}) = Q_i^{k-1}(s,\vec{a}) + \alpha \delta_i^k e_i^k(s,\vec{a}) \tag{7}$$

where α is the learning factor, and γ is the discount factor; δ_i^k is the estimated Q-function error of the ith agent at the kth iteration; $R_i(s_{k-1}, s_k, \vec{a}_{k-1})$ is the ith agent's reward function of transition from state s_{k-1} to s_k under the selected joint action; $e_i^k(s, \vec{a})$ is the ith agent's eligibility trace for state-action pair $(s, \vec{a})$ at the kth iteration. The eligibility trace is a temporary record of the occurrence of taking actions and state trajectory [23], and it can be updated with the following policy,

$$e_i^k(s,\vec{a}) = \begin{cases} \gamma\lambda e_i^{k-1}(s,\vec{a}) + 1 & (s,\vec{a}) = (s_k, \vec{a}_k) \\ \gamma\lambda e_i^{k-1}(s,\vec{a}) & \text{otherwise} \end{cases} \tag{8}$$

where λ is the trace-decay factor. After the updation of Q-functions in each iterative step, the optimal CE solution can then be solved using the following linear programming model,

$$\begin{aligned} & f\left[\pi(s_k,\vec{a})\right] = \max \sum_{i=1}^{n} \sum_{\vec{a}\in A(s_k)} \pi(s_k,\vec{a}) Q_i^k(s_k,\vec{a}) \\ \text{s.t.} & \sum_{\vec{a}_{-i}\in A_{-i}(s_k)} \pi(s_k,\vec{a}) Q_i^k(s_k,(\vec{a}_{-i},a_i)) \geq \sum_{\vec{a}_{-i}\in A_{-i}(s_k)} \pi(s_k,\vec{a}) Q_i^k(s_k,(\vec{a}_{-i},a_i')) \\ & \sum_{\vec{a}\in A(s_k)} \pi(s_k,\vec{a}) = 1,\ 0 \leq \pi(s_k,\vec{a}) \leq 1 \end{aligned} \tag{9}$$

It can be found from (9) that the joint action policy of MARL agent can consider the other agents' decisions and Q-functions to maximize the received rewards of all MARL agents. For each iterative cycle, a list of equilibrium values can be readily obtained from (9) using linear programming solver [15], and this state-action equilibrium expresses the selection probability of joint action in a given state under the optimal CE action strategy. As a result, MARL agents will implement the joint action strategy for GCD scheme based on the probability distribution of equilibrium point, while HCEQ will recursively optimize the joint probability distribution for optimal cooperative action strategy. Rigorous proofs in [15], [23] and [25] have demonstrated that the optimal action strategy would converge to the best state-action pair with probability 1 once the action values are represented discretely and all actions are sufficiently sampled in state space.

In each iterative decision cycle, the HCEQ observes the current operating state, updates the Q-functions, solves the optimal equilibrium action policy, and then chooses and executes a joint action profile based on the optimal CE policy, as shown in Fig. 2. After the implementation of the joint action in each AGC cycle, the MARL agent will receive a reward value based on the resulting GCD performance, and the Q-functions for all the state-action pairs can then make an iterative update from the selected action and received reward while the agent's value function estimator would consider the action decisions of other cooperative agents. Therefore, the design of MARL-based GCD involves the definitions of reward function, state-action space and parameter settings to fully explore the coordinated operations among AGC generators.

3.3. MARL-based Hierarchical GCD Framework

1) Multi-layer GCD structure: The GCD problem is to optimally on-line allocate the LFC regulating commands to each individual AGC units, and this is a real-time optimal combination problem with high-dimensional complexity. Therefore, the proposed approach employs a hierarchical optimization framework to solve this real-time high dimensionality problem. In this hierarchical framework, the GCD problem can be modeled as a multi-layer hierarchy and decomposed into several multi-task MARL problems, as shown in Fig. 3. In each MARL subtask, the AGC generation command from the upper layer would be optimally assigned among various AGC units or unit groups. Firstly, the AGC committed generators are classified into different unit groups in terms of their LFC characteristics, such as coal-fired units, LNG, hydro units, and so on. The unit classification can then be further carried out based on the ramp rate, LFC time delay, adjustable capacity or unit regulating cost in the lower layers.

As illustrated in Fig. 3, the total AGC regulating commands derived from the central AGC controller can be allocated vertically from the first layer to each AGC generator in the bottom layer. Hence, the GCD problem can be transformed into several MARL subtasks, and thus the variable dimensionality of GCD can be evidently decreased through the proposed hierarchical framework. For the optimal generation allocation in each MARL subtask, the regulation participation factor of each AGC unit or unit group can be optimized, and its AGC reference command can then be determined. For example, if there are n_c AGC participation factors for coal-fired unit groups in the 2nd GCD layer as shown in Fig. 3, the AGC reference command for the ith coal-fired unit group, $\Delta P_{\mathrm{C2}\text{-}i}$, can be calculated as follows,

$$\begin{aligned}&[\Delta P_{C2-1},\Delta P_{C2-2},\cdots,\Delta P_{C2-n_c}]=\Delta P_{C2}\cdot[a_{C1},a_{C2},\cdots,a_{Cn_c}]\\&\text{s.t.}\quad\sum_{i=1}^{n_c}a_{Ci}=1,\qquad 0\le a_{Ci}\le 1\end{aligned}\tag{10}$$

where ΔP_{C2} denotes the AGC reference command for coal-fired unit groups from the 1st layer, and a_{Ci} is the optimized AGC participation factor of the ith coal-fired unit group from the MARL joint action policy.

For each MARL subtask with n agents, the control variable vector to optimize the AGC participation factors is $[a_1, a_2, \ldots, a_{n\text{-}1}]$, and the remaining participation factor can then be determined from the generation balance equality constraint. Here, AGC unit or unit group with the maximum adjustable capacity is chosen as the balancing agent, in each MARL problem. Hence, as shown in Fig. 2, the AGC participation factor of the balancing agent can be calculated as follows,

$$a_n = 1 - \sum_{i=1}^{n-1} a_i \tag{11}$$

In addition, if the sum of AGC participation factors from MARL is greater than 1, the corresponding action equilibrium value should be set to 0, as indicated in the following constraint:

$$\pi(s_k, \vec{a}) = 0 \qquad \text{if } \sum_{i=1}^{n-1} a_i > 1 \tag{12}$$

Consequently, in each iterative step, equality constraint (11)-(12) should be included in the CE point model (9) to solve the optimal joint action policy using linear programming.

2) Parameter settings: In the proposed algorithm, three parameters γ, α, and λ in (6)-(8) are critical to implement the learning control and should be set with following the generic guidelines [15],[16],[23],[24].

The discount factor, $0 < \gamma < 1$, is defined to exponentially discount the future received rewards in the Q-functions. Since later rewards in the GCD optimization process are important, a value close to 1 should be set [23]. Simulation studies indicate that a value in the range of 0.7-0.9 is recommended in this application. Here, an intermediate value of 0.8 is used.

The learning factor, $0 < \alpha < 1$, determines the amount of update in the Q-functions. A larger α tends to accelerate the convergence of algorithm but may lead to local optimum, while a smaller value can enhance the algorithm stability. In this investigation, α is set to 0.1 in the initial stage of interactive self-learning for the global exploration, and its value will decrease linearly to 0.001 after the pre-learning process for control stability of onsite application.

The trace-decay factor, $0 < \lambda < 1$, in eligibility traces is used to allocate the credit throughout sequences of state-action pairs and improve the algorithm optimization efficiency. While larger values of λ mean that more of farther backward information can be used to optimize the Q-functions, smaller ones imply that less reward will be assigned to the previous state-action pairs to estimate the Q-function errors. Our experiences show that a value in the range from 0.3 to 0.7 can work well for the dynamic performance of algorithm, and the factor is set to 0.5 in this paper.

Moreover, the learning step T_{step} of HCEQ is determined by the AGC decision cycle. In the case studies, the state-action space of GCD scheme can be specified and discretized following the space discretization in [2]. Since both state space S and action space A are finite, the values of Q-functions and eligibility traces can be stored as finite matrices and implemented in the lookup tabular forms. Following the initialization rules in [15],[23], the initial values of eligibility traces, Q-functions, and state-value functions for all MARL agents are set to zero matrices or vectors.

3) Reward function: MARL reward function determines the control objective of the GCD scheme, and has a critical influence on the algorithm performance and value function iterations. Based on the objective functions of GCD in (1)-(3), a multi-criteria reward function can be designed for the *i*th agent except the balancing agent in the MARL, as follows,

$$R_i(s_{k-1}, s_k, \vec{a}_{k-1}) = \begin{cases} \left[-\Delta P_{ei}^2(k) + \mu_1[P_{Gi}^{\max} - P_{Gi}(k)] / P_{Gi}^{\max} - \mu_2 C_i[P_{Gi}(k)]\right] + \dfrac{1}{n-1} R_b(k) & \forall i \in F \\ \left[-\Delta P_{ei}^2(k) - \mu_2 C_i[P_{Gi}(k)]\right] + \dfrac{1}{n-1} R_b(k) & \forall i \notin F \end{cases} \tag{13}$$

where $R_b(k)$ represents the received reward for the balancing agent, and it can be formulated as follows,

$$R_b(k) = \begin{cases} -\Delta P_{eb}^2(k) + \mu_1[P_{Gb}^{\max} - P_{Gb}(k)] / P_{Gb}^{\max} - \mu_2 C_b[P_{Gb}(k)] & \forall b \in F \\ -\Delta P_{eb}^2(k) - \mu_2 C_b[P_{Gb}(k)] & \forall b \notin F \end{cases} \tag{14}$$

where subscript b represents the balancing agent in a GCD subtask. In each MARL, the action as well as GCD performance of the balancing agent is determined by the joint action of other agents, as expressed in (11), and hence the reward value of the balancing agent obtained from (14) should be evenly assigned in the reward functions (13) of other agents in order to evaluate the joint action policy of HCEQ.

4) Coordination factor: With the proposed multi-layer GCD framework, the AGC generation allocation problem with various types of AGC units can be divided into several MARL optimization subproblems, and each subproblem can be solved using CEQ algorithm. Furthermore, the earlier hierarchical RL studies have demonstrated that the coordination mechanisms should be designed between adjacent layers to improve the learning efficiency and optimality of the proposed HCEQ [24]. In this paper, a time-varying coordination factor (CF) [2] is introduced and supplemented in the reward function (13) of each MARL agent for the overall coordination of the multi-layer control structure. As depicted in Fig. 3, expect for the bottom layer, the coordination factor is introduced to the MARL reward functions in other control layers. Therefore, the corresponding reward function R_i^{CF} for the ith MARL agent can be reformulated as follows,

$$\begin{cases} R_i^{CF}(s_{k-1}, s_k, \vec{a}_{k-1}) = CF_i(k) \cdot R_i(s_{k-1}, s_k, \vec{a}_{k-1}) \\ CF_i(k) = 1 \Big/ \left| \sum_{j \in L} R_j^{CF}(k) \right| \end{cases} \tag{15}$$

where $CF_i(k)$ is the coordination factor of the ith agent in the upper layers at the kth iteration; L denotes the set of MARL agents in the next lower layer under the ith unit group; R_j^{CF} represents the jth agent's rewards collected from the MARL agents in the lower layer through (13)-(15). The purpose of CF is to transmit the reward with control effects from the lower layers to the upper layer, and thus can implement a bottom-up reward flow in the proposed hierarchy. Normally, CF is a positive value less than 1, and CF would decrease with the reduction in the rewards from the lower layer. Therefore, the formulation of CF in reward function (15) can evaluate the overall control performance of the GCD strategy achieved in the top layer.

3.4. Execution Steps of the Proposed HCEQ Approach

The proposed hierarchical MARL framework provides a performance-adaptive means to implement the GCD scheme with high flexibility in specifying the equilibrium objectives, and the AGC generators would operate an optimum equilibrium state with high energy utilization under this multi-agent paradigm. To sum up, the execution steps of the HCEQ-based GCD approach for each MARL subtask can be illustrated in Table 1.

4. Simulation Studies

4.1. Simulation Environment

For the in-depth investigation of the proposed HCEQ scheme in a realistic simulation environment, the CSG power system model [26], which was previously developed by utilities for Guangdong power dispatch center projects [7],[8], is used as the benchmark system to evaluate and analyze the performance of GCD approaches. The CSG is one of the most complicated large-scale interconnected power grids over the world, the peak load of which reaches 131 GW in 2013, and the total installed capacity is approximately 174 GW [26]. Moreover, the CSG power system consists of 93 AGC generators, 1836 buses, 4519 branches, and four provincial control areas, Guangdong, Guangxi, Guizhou, and Yunnan, inter-connected by the parallel HVDC-HVAC transmission systems. All of the buses can be classified into five voltage levels, i.e. 220 kV, 230 kV, 400 kV, 500 kV, and 525 kV, respectively. In this LFC simulator, the AGC generator models for fossil-fuel-fired, LNG and hydroelectric generators are included, and each generator output is determined by the governor and the setpoint of regulating commands from AGC controller according to their regulation participation factors. Taking the Guangdong power grid as the study area, Table 2 provides the LFC parameters of AGC committed units in the studying power grid. It can be found from Table 2 that the fast regulation capability of hydro plants is obviously much higher than other AGC plants in Guangdong power grid. In the case studies, the generation capacity of hydropower in Guangdong power grid is insufficient, and thus the studied control area only consider the hydro plants as the fast regulation units in (1) for reserve requirements of fast adjustable capacity. Here, T_s represents the time delay of AGC generator in the secondary frequency control loop; UR_i and DR_i are the upper and lower ramp limits of the ith AGC unit; $\Delta P_{Gi}^{\min}$ and $\Delta P_{Gi}^{\max}$ denote the minimum and maximum adjustable capacity of the ith AGC unit, respectively. In this paper, the AGC controller, as shown in Fig. 3, adopts the NARI's improved-PI controller [10]. Moreover, all the simulations are implemented in Matlab/Simulink 7.1 by a personal computer with 3.1-GHz Core i5 Quad CPU and 4 GB of RAM, and the proposed HCEQ-based GCD scheme is built using S-function module.

As illustrated in Section 3.3, the AGC units can firstly be divided into 4 types of plant groups in the 1st layer, and then further classified into different unit groups in the 2nd layer based on their LFC response characteristics. In the bottom layer, since the AGC units have the similar LFC regulating characteristics, the PROP method can be utilized to unit groups for determining the regulation participation factor of each AGC committed unit. Therefore, the hierarchical GCD scheme can be formulated as a three-layer control structure with four MARL subproblems, and the proposed HCEQ is applied in each subproblem to optimize the AGC participation factors in real-time operation.

4.2. Study on Pre-learning Process

MARL algorithms should be scheduled to experience a series of pre-learning processes before its onsite operation, and this process is an offline preconditioning technique involving numerous exploration iterations in the state space to optimize the Q-functions and state-value function [27]. Based on the sample-average theory in [23], this pre-learning process should be carried out with a great variety of load disturbances to experience enough system scenarios for iterative policy evaluation [2] to optimize the joint equilibrium

GCD strategy. Furthermore, the termination criterion of the pre-learning process for the *i*th agent can be determined by the matrix 2-norms of Q-function differences $\|Q_i^k(s,\vec{a})-Q_i^{k-1}(s,\vec{a})\|_2 \leq \varsigma$ (ς is a given small precision factor). With the algorithm getting converged, the Q-function would be stable, and the optimal joint CE policy at various states can be gradually learned. This pre-learning phase will end once the termination criterions for all the MARL agents are satisfied.

Thereafter, all the priori knowledge obtained from the pre-learning processes would be stored and used for onsite operation in the practical AGC system, as illustrated in Fig. 4. The HCEQ-based GCD scheme, which has already benefited from the pre-learning knowledge, will continue to make steady online learning with an iterative policy evaluation during each AGC cycle, and could still improve its control behaviors by interaction with real power system.

Here, a typical sequence of square-wave load disturbances, as shown in Fig. 5b, is added in Guangdong power grid to illustrate the pre-learning process. The simulation results of the proposed multi-layer HCEQ in the convergence process have been illustrated in Fig. 5. Fig. 5a shows the regulation participation factors of two typical AGC units, oil-fired unit 1 and hydro unit 1-1, in the algorithm convergence process. It can be found from simulations that the agents in each MARL gradually converges to their deterministic GCD policy, while the AGC generation outputs and CPS compliances also tend to become stable. Furthermore, the convergence process for LNG groups and coal-fired groups of Q-function differences are given in Fig. 6. It can be found that the Q-functions tend to be stable, and the optimal CE action policy in each area can then be obtained for online optimization in real power systems.

Moreover, Table 3 provides the comparisons of average convergence time of the proposed HCEQ with other RL algorithms over 10 independent runs in the pre-learning process. It is clear to see that, the proposed approach exhibits its superiority and higher efficiency on the convergence rate than the HQL and improved HQL [2], and the time-varying CF can effectively improve the learning efficiency and optimality of GCD dispatch.

4.3. Study on Weight of Hydro Capacity Margin

For the thermal-dominated power systems, taking Guangdong power grid in the CSG as an example, the hydro power plants play an important role in the AGC performance. In general, the more the generation commands allocated to hydro units, the better the resulting AGC performance and regulating cost will be, since the hydro plants can provide fast regulating capability with less generation cost. However, the GCD scheme should maintain sufficient adjustable margin of hydro AGC capacity to cope with the potential incremental load disturbances. Consequently, Table 4 and Fig. 7 illustrate the effects of different weight μ_1 in reward function (13) on the hydro capacity margin and AGC performance.

In this case study, a series of incremental load disturbances was set in Guangdong power grid to test the dynamic behaviors of HCEQ with typical values of weight μ_1. Fig. 7 shows the plots of the total generation output of hydro plants corresponding to the prespecified load disturbances, in which ΔP_{Gh1}, ΔP_{Gh2}, ΔP_{Gh3} are the hydro generation outputs with μ_1 set to 0, 10, 50, respectively. It can be observed that a smaller weight of μ_1 will increase the generation output of hydro plants, and thus lead to less hydro capacity margin for the AGC spinning reserve. Table 4 tabulates the simulation results of AGC performance under different values of weight μ_1. Here, the weight μ_2 in (13) is set to 0, and CPS1 and ΔP_{Gh} are the average values of 10-min CPS1 metric and hydro generation output over the entire simulation period in Fig. 7. As shown in Table 4, the increased participation of hydro generation in AGC regulating commands can improve the performance metrics and reduce AGC regulating cost. In this paper, the weight μ_1 can be set to an intermediate value, 10 or 50, to maintain sufficient AGC hydro reserves, while the CPS compliances can also be ensured.

4.4. Study on Weight of Regulating Cost

The weight of regulating cost in (13), μ_2, is also critical for the GCD performance of HCEQ. In order to validate the effects of weight μ_2 on the algorithm performance, Table 5 lists the statistical simulation results with different weights of μ_2 corresponding to the step load disturbances in Fig. 7. It can be concluded that the weights μ_1 and μ_2 are equivalent to the weight parameters in linear quadratic regulator (LQR) [7], and a larger value of μ_2 would expect more fuel saving in the AGC generation costs. Thus, the weights μ_1 and μ_2 should be thoughtfully set for the trade-off and coordination among the multiple GCD objectives based on the LQR rules and system operational requirements. In the following case studies, as a compromise among the AGC cost, hydropower reserve and CPS compliances, the weight μ_1 and μ_2 are selected to 10 and 0.1 in this paper, respectively.

4.5. Statistical Experiments on CSG System

The long-term GCD performance should be thoroughly evaluated with the data statistical comparative experiments in which the CSG simulators have been implemented with the preset disturbance scenarios over a 24-hour period [7]. The adaptability and dynamic optimization of the proposed approach can be examined and analyzed under the representative stochastic load disturbances [28] and system parameter perturbations, as addressed in [2]. Furthermore, the performance of HCEQ has been benchmarked and compared with PROP method, genetic algorithm (GA), HQL and the improved HQL [2]. The resulting statistics with assessment period of 10 minutes for the studying control area on various AGC performance metrics are listed in Table 6 and 7, where $|\Delta F|$ and |ACE| are the averages of absolute values of frequency deviation and ACE over the entire simulation period; CPS1, CPS2 and CPS metrics are the daily compliance percentages. Here, the hydroelectric AGC capacity in the studying area is set to 1424 MW in July (rainy season) and it will drop to 712 MW in December (dry season). Hence, different AGC allocation strategies are required for the rainy and dry seasons in order to adapt to the load disturbances and changing hydro capacity. In this case study, the presented performance results of AGC strategies based on the RL and MARL algorithms correspond to AGC performance after the pre-learning process with sufficient training iterations for the rainy season.

It can be found from Table 6 and 7 that the dynamic optimization of GCD with the three RL methods can provide the better performance than GA and PROP with fixed AGC participation factors. On the other hand, compared with the HQL algorithm, the multi-layer coordination mechanism in HCEQ and improved HQL can also effectively enhance the optimality of GCD schemes. Also, as supported by the comparative simulation results, the MARL-based HCEQ can outperform the improved HQL in [2], and has exhibited its superior performance and dynamic optimization capability with less regulating cost. Furthermore, the above five algorithms were then implemented on the CSG power system model with a drop of hydro capacity, and the resulting statistics have been listed in Table 7. It can be seen that the AGC performance and regulating costs of all the algorithms deteriorate as the reduction in the hydro power capacity in dry season. Last but not least, in comparison with the improved HQL, the proposed HCEQ shows the fast online optimization capability to perform the best under system parameter perturbations, and the corresponding reductions on the AGC regulating costs in Table 6 and 7 are 11.17 and 8.33%, respectively.

5. Conclusion

In this paper, a novel MARL based HCEQ algorithm is proposed to solve the dynamic optimization of multi-layer GCD problem. The following are the main advantages of the proposed GCD approach.

(1) A novel hierarchical MARL algorithm based on the correlated equilibrium is proposed to optimize the regulation participation factor of each generator for the overall AGC performance enhancement, and the proposed HCEQ algorithm can adapt well to various system operation scenarios with superior adaptability and dynamic optimization capability.

(2) A multi-layer AGC generation allocation framework is also developed to overcome the curse of dimensionality in the GCD problem with the increased number of AGC plants involved. Besides, the time-varying coordination factors have been formulated among control layers to improve the convergence and optimality of dispatch solutions.

(3) The multi-criteria reward functions have been designed in the HCEQ algorithm for multiobjective equilibrium dispatch of GCD optimization problem. Simulation studies on the CSG power system model have demonstrated that, compared with the previous GCD methods, the proposed approach can effectively enhance the AGC tracking performance with less AGC regulating costs, while the reserve requirements of fast regulation capacity are ensured.

Acknowledgments

The authors gratefully acknowledge the support of National Key Basic Research Program of China (973 Program: 2013CB228205), National Natural Science Foundation of China (51177051, 51477055, 51507056), and The Hong Kong Polytechnic University under Project A-PL97.

References

[1] N. Jaleeli, L. S. VanSlyck, D. N. Ewart, L. H. Fink, and A. G. Hoffmann, "Understanding automatic generation control," *IEEE Trans. Power Syst.*, vol. 7, no. 3, pp. 1106-1122, Aug. 1992.

[2] T. Yu, Y. M. Wang, W. J. Ye, B. Zhou, and K. W. Chan, "Stochastic optimal generation command dispatch based on improved hierarchical reinforcement learning approach," *IET Gener. Transm. Distrib.*, vol. 5, no. 8, pp. 789-797, Aug. 2011.

[3] N. Jaleeli and L. S. VanSlyck, "NERC's new control performance standards," *IEEE Trans. Power Syst.*, vol. 14, no. 3, pp. 1092-1099, Aug. 1999.

[4] Ibraheem, P. Kumar and D. P. Kothari, "Recent philosophies of automatic generation control strategies in power systems," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 346-357, Feb. 2005.

[5] M. Yao, R. R. Shoults, and R. Kelm, "AGC logic based on NERC's new control performance standard and disturbance control standard," *IEEE Trans. Power Syst.*, vol. 15, no. 2, pp. 852-857, May 2000.

[6] A. Khodabakhshian and R. Hooshmand, "A new PID controller design for automatic generation control of hydro power systems," *Int. J. Elect. Power Energy Syst.*, vol. 32, no. 5, pp. 375-382, Jun. 2010.

[7] T. Yu, B. Zhou, K. W. Chan, L. Chen, and B. Yang, "Stochastic optimal relaxed automatic generation control in non-Markov environment based on multi-step Q(λ) learning," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 1272-1282, Aug. 2011.

[8] T. Yu, B. Zhou, K. W. Chan, Y. Yuan, B. Yang, and Q. H. Wu, "R(λ) imitation learning for automatic generation control of interconnected power grids," *Automatica*, vol. 48, no. 9, pp. 2130-2136, Sep. 2012.

[9] L. C. Saikia, S. Mishra, N. Sinha, and J. Nanda, "Automatic generation control of a multi area hydrothermal system using reinforced learning neural network controller," *Int. J. Elect. Power Energy Syst.*, vol. 33, no. 4, pp. 1101-1108, May 2011.

[10] Z. H. Gao, X. L. Teng, and L. Q. Tu, "Hierarchical AGC mode and CPS control strategy for interconnected power systems," *Automation of Electric Power Systems*, vol. 28, no. 1, pp. 78-81, Jan. 2004.

[11] Y. Xichang and Z. Quanren, "Practical implementation of the SCADA + AGC/EDC system of the Hunan power pool in the central China power network," *IEEE Trans. Energy Conver.*, vol. 9, no. 2, pp. 250-255, Jun. 1994.

[12] L. Busoniu, R. Babuska, and B. de Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, vol. 38, no.2, pp. 156-172, Mar. 2008.

[13] G. Chalkiadakis, E. Elkind, and M. Wooldridge, *Computational Aspects of Cooperative Game Theory*. Morgan & Claypool, 2011.

[14] J. Nash, "Non-Cooperative Games," *Ann. Math.*, vol. 54, no. 2, pp. 286-295, Sep. 1951.

[15] A. Greenwald and K. Hall, "Correlated Q-learning," *in Proc. 20th Int. Conf. Mach. Learn*. (*ICML-*03), Washington, DC, Aug. 21-24, pp. 242-249.

[16] T. Yu, J. Liu, K. W. Chan, and J. J. Wang, "Distributed multi-step Q(λ) learning for optimal power flow of large-scale power grids," *Int. J. Elect. Power Energy Syst.*, vol. 42, no. 1, pp. 614-620, Nov. 2012.

[17] B. H. Chowdhury and S. Rahman, "A review of recent advances in economic dispatch," *IEEE Trans. Power Syst.*, vol. 5, no. 4, pp. 1248-1259, Nov. 1990.

[18] K. Naidu, H. Mokhlis, A. H. A. Bakar, V. Terzija, and H. A. Illias. "Application of firefly algorithm with online wavelet filter in automatic generation control of an interconnected reheat thermal power system," *Int. J. Elect. Power Energy Syst.*, vol. 63, no. 12, pp. 401-413, Dec. 2014.

[19] R. K. Sahu, S. Panda, and G. T. C. Sekhar. "A novel hybrid PSO-PS optimized fuzzy PI controller for AGC in multi area interconnected power systems," *Int. J. Elect. Power Energy Syst.*, vol. 64, pp. 880-893, Jan. 2015.

[20] C. Liu, X. Xu, and D. Hu. "Multiobjective reinforcement learning: a comprehensive overview," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 3, pp. 385-398, Mar. 2015.

[21] B. Zhou, K. W. Chan, T. Yu, and C. Y. Chung, "Equilibrium-inspired multiple group search optimization with synergistic learning for multiobjective electric power dispatch," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 3534-3545, Nov. 2013.

[22] M. Javadi, "Dynamic models for steam and hydro turbines in power system studies," *IEEE Trans. Power App. Syst.*, vol. PAS-92, no. 6, pp. 1904-1915, Nov. 1973.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[24] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dyn. Syst.*, vol. 13, no. 4, pp. 341-379, Oct. 2003.

[25] J. N. Tsitsiklis, "Asynchronous stochastic approximation and *Q*-learning," *Mach. Learn.*, vol. 16, no. 3, pp. 185–202, Sep. 1994.

[26] The Operation Mode of China Southern Power Grid in 2013 (in Chinese), China Southern Power Grid Co. Ltd.

[27] D. Ernst, M. Glavic, and L. Wehenkel, "Power systems stability control: reinforcement learning

framework," *IEEE Trans*. *Power Syst.*, vol. 19, no. 1, pp. 427-435, Feb. 2004.

[28] Y. Manichaikul, *Industrial electric load modeling*. Cambridge, MA: MIT Press, 1978.