Coordinated Optimization of Generation and Compensation to Enhance Short-Term Voltage Security of Power Systems Using Accelerated Multi-Objective Reinforcement Learning

ZHUOMING DENG¹, ZHILIN LU², ZHIFEI GUO¹, WENFENG YAO¹, WENMENG ZHAO¹, BAORONG ZHOU¹, AND CHAO HONG¹

¹Electric Power Research Institute, China Southern Power Grid, Guangzhou 510663, China
²School of Electric Power Engineering, South China University of Technology, Guangzhou 510640, China

Corresponding author: Zhuoming Deng (zhuomingupup@163.com)

This work was supported by the Comparative Study on the Characteristics and Applications of Synchronous Condensers and STATCOMs in AC-DC Hybrid Power Grids through the Science and Technology Project of China Southern Power Grid under Grant ZBKJXM20180183.

ABSTRACT High proportions of asynchronous motors in demand-side have pressured heavily on short-term voltage security of receiving-end power systems. To enhance short-term voltage security, this paper coordinates the optimal outputs of generation and compensation in a multi-objective dynamic optimization model. With equipment dynamics, network load flows, lower and upper limitations, and security constraints considered, this model simultaneously minimizes two objectives: the expense of control decision and the voltage deviation. The Radau collocation method is employed to handle dynamics, by transforming all differential algebraic equations into algebraic ones. Most importantly, Pareto solutions are obtained through an accelerated multi-objective reinforcement learning (AMORL) method by filtering the dominated solutions. The entire feasible region is partitioned into small independent regions, to eliminate the scope for Pareto solutions. Besides, the AMORL method redefines the state functions and introduces creative state sensitivities, which accelerate the switch from learning to applying, once the agent accumulates sufficient knowledge. Furthermore, Pareto solutions are diversified via introducing some potential solutions. Lastly, the Fuzzy decision-making methodology picks up the tradeoff solution. Case studies are implemented on a practical 748-node power grid, which validate the acceleration and efficiency of the AMORL method. The AMORL method is overall superior to conventional reinforcement learning (RL) method with more optimal non-dominated objective values, much shorter CPU time, and better convergence to accurate values. Moreover, compared with another three state-of-the-art RL methods, the AMORL method takes almost the same CPU time of several seconds, but is slightly superior to the state-of-the-art methods in terms of optimal objective values. Additionally, the calculated values of the AMORL method fit the best with the accurate values during each iteration, resulting in a good convergence.

INDEX TERMS Accelerated multi-objective reinforcement learning, dynamic optimization, Pareto solutions, short-term voltage security.

NOMENCLATURE

A. PARAMETERS

$t_0$ Initial moment of the short-term process.
$t_c$ Moment of tripping short-circuit contingency.
$t_f$ Final moment of the short-term process.
$V_{\text{ref}}$ Node reference voltage.
$N_B$ Amount of network nodes.
$N_C$ Amount of decision variables.

$u_{i0}$ Initial value of decision variable $i$ without coordinated optimization.
$a_i$ Unit cost of decision variable $i$.
$\delta_{\text{max}}$ Maximum allowable difference of the rotor angle to the inertia center.
$M_g$ Inertia of generator $g$.

B. VARIABLES

$V_i(t)$ Voltage of node $i$.
$V_l(t)$ Voltage of key load node $l$.
I. INTRODUCTION

With growing proportions of cooling loads equipped in power grids, such as air conditioners in the summer, large amounts of reactive power are demanded, especially in some receiving-end power grids. The standardized models of asynchronous motors can be employed for cooling loads universally, and they are the main reactive power loads. Once if severe short-circuit contingencies occur, load voltages drop sharply, resulting in deceleration of the rotors of these asynchronous motors. The slower the rotors spin, the more reactive power should be required, and the lower voltages drop. Especially when the rotors stall, the asynchronous motors are equated with another short-circuit contingency [1], threatening the short-term voltage security. Hence, a vicious circle exists among rotors, reactive power, and voltages. Such receiving-end power grids (e.g., Guangdong power grid in China), have suffered from severe and outstanding short-term voltage security problems. In general, there are two frequently-used and effective means preventing short-term voltage security problems: (1) installing appropriate number and capacity of statcoms in advance [2], and (2) urgent load shedding after short-circuit contingencies [3]. Although effectively, installing STATCOMs and load shedding respectively suffer from expensive installing cost and load loss. On the other hand, utilizing high side var-voltage control of generators [4] and switching capacitor banks [5] sufficiently take advantage of the existing reactive power supports of excitation systems and capacitor banks, respectively. Consequently, the short-term voltage security can be enhanced by these two much more economic means for practical power grids. High side var-voltage control and switch-able capacitor banks respectively belong to the source-side and the demand-side, and provide control and compensation for generators and loads. Hence, coordinated optimization of generation and compensation is significant to make sufficient use of these two means and to enhance short-term voltage security.

To coordinate the optimization of generation and compensation at different sides, the reference voltages of automatic voltage regulators (AVRs) and the numbers of capacitor units in operation are both regarded as decision variables in this paper. Their optimal coordinated settings are obtained by means of solving a coordinated multi-objective dynamic optimization (MODO) model, simultaneously considering the index of short-term voltage security and the cost of control decisions. How to obtain the optimal solution of multi-objective optimization received significant concerns in the past. Nowadays, artificial intelligence (AI) has become a research hotspot from all walks of life, especially its increasingly wide utilization in the optimization of power grids. Based on its searching methodologies for optimal solutions, the AI can accordingly be summarized and classified as many branches or methods, including generic algorithm [6], non-dominated sorting genetic algorithm II [7], artificial neural network [8], particle swarm optimization [9], and reinforcement learning (RL) method [10]. All of these AI methods are extended into multi-objective optimization formulations by comparing the dominating relationships among feasible solutions and obtaining Pareto solutions. AI is increasingly generalized and extended to the coordinated optimization problems of universal power grids. One of the most popular topics of AI should be the machine learning, whose typical representation is the multi-objective reinforcement learning (MORL) method for MODO models [11]. Hence, this paper mainly focuses on and discusses the MORL method in the follows. As we know, conventional MORL methods require a totally throughout search (i.e., trial and learning by means of comparing the dominating relationships) of the complete feasible region. This helps to obtain sufficient knowledge and guides the searching directions of the agent. However, since the feasible region in practice can be extremely huge, a totally throughout search requires rather long computation time. In deed, the probability of Pareto solutions in some feasible regions may be quite small, which should be ignored for simplification. This can be achieved by two means: (1) directly eliminating the feasible regions with small probability of Pareto solutions; and (2) accelerating the switch from trial and learning to deterministic searching directions based on the existing knowledge. Accordingly, significant accelerations can be put forward for the conventional MORL method to eliminate the searching areas of feasible region and to switch the agent to from trial and learning to deterministic searching directions, once the agent accumulates sufficient knowledge [12], [13].

This paper proposes a coordinated MODO model for coordinated optimization of generation at the source-side and compensation at the demand-side, to enhance short-term voltage security. This MODO model is converted into a nonlinear algebraic optimization by the Radau collocation method. Its Pareto solutions are further obtained by a novel accelerated multi-objective reinforcement learning (AMORL) method. The main contributions (i.e., the accelerations) are clarified as the following three aspects. (1) The complete feasible region is broken into many independent small ones by means of dichotomy. These small regions are further filtered and only those with potential non-dominated solutions are remained for searching. This reduction of feasible region significantly eliminates the computational burdens of the trial and learning. (2) The state function and state sensitivity are recommended and redefined for the AMORL method. They provide significant evaluation of the convergence of trial and learning. According to several settled threshold values, the agent easily judges whether it has accumulated sufficient knowledge. (3) Once sufficient knowledge has been achieved, the agent intelligently switches from trial and learning to a deterministic searching direction for the remaining search of Pareto
solutions according to existing knowledge. This makes full use of the knowledge obtained by trial and learning, while avoiding unnecessary searching of the huge feasible region and accelerating convergence.

Moreover, it is noteworthy that our work is not restricted to the enhancement of short-term voltage security. It can be applied to other fields such as optimal scheduling, and is able to obtain optimal solutions in scalable infrastructures. The MODO modeling framework can be generalized to common multi-objective optimization/planning over multiple time periods. Besides, the Radau collocation method for numerical integration can carry over to general dynamic optimization problems. Most importantly, the proposed three contributions are insensitive to the model and only care about the state space to achieve acceleration. Benefiting from the scalability and generalization of machine learning, the AMORL method always obtains relatively optimal solutions for scalable multi-objective optimization problems. Detailed analysis about the scalability and generalization of the AMORL method will be provided in Section III-C.

The remaining of this paper is organized as follows. In Section II, the coordinated MODO framework to enhance short-term voltage security is introduced, fully utilizing the var supports of generators and capacitor banks. In Section III, the solution methodologies for the MODO model are presented, including continuous/discrete variable coordination, Radau collocation method, the most significant AMORL method, and the Fuzzy decision-making method. In Section IV, case studies are conducted on a 748-node power grid in China to confirm the efficiency of the AMORL method. Conclusions are available in Section V.

II. COORDINATED MODO FRAMEWORK TO ENHANCE SHORT-TERM VOLTAGE SECURITY

This paper proposes a coordinated MODO framework to enhance short-term voltage security by re-adjusting the reference voltages of AVR and the numbers of capacitor units in operation. The coordinated control decision is achieved by obtaining and implementing the Pareto solutions of the above MODO model. Both the index of short-term voltage security (i.e., voltage deviation) \( J_1 (\cdot) \) and the cost of control decision \( J_2 (\cdot) \) are considered as multiple objective functions:

\[
J_1 (\cdot) = \sum_{i=1}^{N_R} \int_{t_0}^{t_f} (V_i(t) - V_{ref})^2 dt, \quad (1a)
\]

\[
J_2 (\cdot) = \sum_{i=1}^{N_C} \int_{t_0}^{t_f} a_i \cdot (u_i(t) - u_{i0})^2 dt, \quad (1b)
\]

where the decision variable \( u_i(t) \) includes the continuous reference voltages of AVR and the discrete numbers of capacitor units in operation.

According to practical criterion of power grid security in China, short-term voltages are regarded secure, if the following two conditions are both satisfied: (a) voltages of key load nodes are restored shortly after tripping short-circuit contingencies, usually no less than 0.75 p.u. in the duration of 1 s; (b) network synchronisms are guaranteed, usually by means of keeping the deviations of rotor angles between the generators and the inertia center no larger than a threshold. Conditions (a) and (b) are respectively represented as (2a) and (2b) as follows:

\[
V_i(t) \geq 0.75, \quad t \in [t_c + 1, t_f], \quad (2a)
\]

\[
|\delta(t) - \delta_{COI}| \leq \delta_{max}, \quad t \in [t_0, t_f], \quad (2b)
\]

where \( \delta_{max} \) is set to 120° in this paper; and \( \delta_{COI} \) denotes the rotor angle of inertia center, calculated as

\[
\delta_{COI}(t) = \sum_{g=1}^{N_G} M_g \delta_g(t)/ \sum_{g=1}^{N_G} M_g. \quad (3)
\]

The coordinated optimal generation and compensation framework to enhance short-term voltage security in this paper is represented as the following compact MODO formulation (4a)–(4e). This formulation simultaneously minimizes the short-term voltage deviation and the cost of control decision, with constraints of element dynamics, network constraints, upper and lower limits, and security constraints:

\[
\begin{align*}
\min_{x(t), y(t), u(t)} & \left\{ J_1 (x(t), y(t), u(t)) , J_2 (x(t), y(t), u(t)) \right\}, \\
\text{s.t.} & \quad \dot{x}(t) = f (x(t), y(t), u(t)), \quad (4a) \\
& \quad 0 = g (x(t), y(t), u(t)), \quad (4b) \\
& \quad 0 \leq h (x(t), y(t), u(t)), \quad (4c) \\
& \quad t \in [t_0, t_f]. \quad (4e)
\end{align*}
\]

where \( x(t) \) denotes the vector of differential variables, including transient electric potentials, rotor angles, rotational speeds, \( y(t) \) denotes the vector of algebraic variables, including node voltages, phase angles, active powers, \( u(t) \) denotes the vector of continuous/discrete decision variables, including the reference voltages of AVR and the numbers of capacitor units in operation; (4a) denotes the above two objective functions; (4b) denotes differential equations, representing the dynamics of asynchronous motor loads, synchronous generators, and their excitation systems; (4c) denotes network constraints, representing power flow equations; (4d) denotes two kinds of inequality constraints: security constraints for short-term voltages (2a) and short-term rotor angles (2b), and upper and lower limits for all types of variables; (4e) limits the time boundary of simulation and optimization.

III. SOLUTION METHODOLOGY

The short-term voltage security problem can be regarded as a Markov decision process, whose strategy can be settled by RL algorithms. However, conventional RL algorithms suffer from state space explosion because of the curse of dimensionality. To overcome the problem, many state-of-the-art RL algorithms have put forward significant
promotions. The deep Q-network (DQN) [14], [15] adopted neural network incorporating target network and experience replay to efficiently approximate the Q-function. Based on the least squares policy iteration methodology, a sequential learning algorithm [16] was employed by the batch reinforcement learning (BRL) method to narrow the state-action space. Additionally, to complement one another with their advantages, the DQN method handled continuous states and low-dimension discrete action spaces, while the deep deterministic policy gradient (DDPG) method dealt with high-dimensional action spaces [17].

In this paper, the AMORL method is based on the Q-learning algorithm and proposes several significant promotions. To overcome the curse of dimensionality, we redefine the state function and introduce the state sensitivity, which provide significant evaluation for the learning process and efficiently switch the searching directions. Additionally, the AMORL method provides a complete accelerated searching strategy for Pareto solutions, including feasible regions reduction and non-dominated solutions search.

A. VARIABLE COORDINATION FOR MIXED-INTEGER OPTIMIZATION

In this paper, the coordinated optimization framework contains both continuous-valued decision variables (i.e., the reference voltages of AVRs), and discrete-valued decision variables (i.e., the numbers of switch-able capacitor units in operation). Actually, to coordinate the optimization of generation and compensation and to obtain a unified control decision without difficult mixed-integer optimization, these two types of decision variables should be discretized to discrete values, or be relaxed to continuous values together.

The AMORL method belongs to AI optimization field, rather than conventional mathematical numerical optimization field. These kinds of AI optimization methods do not need information about continuous derivatives. Instead, they depend on agents to search out the feasible regions by means of trial and error, which rely on discrete-valued variables [11]. Consequently, all of the continuous-valued decision variables should be discretized and transformed into discrete-valued ones for variable coordination as the following. As we know, to prevent system oscillation, the power outputs of generators should ramp up and down stably and slowly. The deviations between the original AVR reference voltages and generators should ramp up and down stably and slowly. The deviations between the original AVR reference voltages and their coordinated control values are recommended to remain within a continuous range of \([-0.1, 0.1]\). This range is discretized with a value resolution of \(\Delta h = 0.005\). As illustrated in Fig. 1, Set A contains 41 integer points, including \(0, \pm 1, \pm 2, \ldots, K, \pm 3, \ldots, 20\), which denote the corresponding gear positions of AVR reference voltages. During each iteration, the AVR reference voltages are updated by determining their deviations \(\Delta V_{\text{ref}}\), which are calculated as the products of the corresponding discrete points of Set A and the discretizing resolution \(\Delta h\). Furthermore, the updated AVR reference voltages are represented as the sum of their original values and their deviations [12], [13].

B. RADAU COLLOCATION METHOD

To enhance short-term voltage security, the coordinated MODO model has already taken system dynamics into account, which are represented as differential equation (4b). Numerical solutions are convenient for differential equations, but they must be incorporated into nonlinear algebraic equations in advance. This is achieved by numerical integration; for instance, the Radau collocation method declared in the following.

\[
\begin{align*}
\mathbf{x}(t) &= \mathbf{x}(t_{p-1}) + T_p \sum_{r=1}^{N_q} \Omega_r \left( \frac{t - t_{p-1}}{T_p} \right) f \left( \mathbf{x}(t_{p,r}), \mathbf{y}(t_{p,r}), \mathbf{u}(t_{p,r}) \right), \quad (5a) \\
\mathbf{y}(t) &= \sum_{r=1}^{N_q} L_r \left( \frac{t - t_{p-1}}{T_p} \right) \mathbf{y}(t_{p,r}), \quad (5b)
\end{align*}
\]

where \(T_p\) is the corresponding time duration of time period \(p\); \(t_{p,r}\) is the normalized location of collocation point \(r\) within time period \(p\); and \(L_r(\cdot)\) and \(\Omega_r(\cdot)\) respectively denote the Lagrange polynomial and its corresponding integration [2], [18].

Generally for short-term voltage security problems, \(\mathbf{u}(t)\) is assumed to operate twice during a short-term duration.
Both generators and capacitor banks should supply strong var supports for their first operation to retain voltages, soon after short-circuit contingencies are tripped. Later, overvoltage catches attentions when voltages recover. Additionally, and capacitor banks should decrease their var outputs for their second operation. Hence, \( u(t) \) is assumed to switch/operate twice, where the first time for strong var supports, and the second time for overvoltage preventions. Additionally, as illustrated in Fig. 2, \( u(t) \) should remain constant during each operation.

The numerical integration is conducted using the Radau collocation method described before. The model (4a)–(4e) with dynamics, is incorporated into the following non-linear programming model, which contains only algebraic equations:

\[
\min_{x(t_p, q), y(t_p, q), u(t_p, q)} \{ J_1(\cdot), J_2(\cdot) \}, \tag{6a}
\]
\[
\text{s.t.} \quad 0 = \mathbf{G}(x(t_p, q), y(t_p, q), u(t_p, q)), \tag{6b}
\]
\[
0 \leq \mathbf{H}(x(t_p, q), y(t_p, q), u(t_p, q)), \tag{6c}
\]
\[
t_p, q \in [0, T_f], \tag{6d}
\]

where \( \mathbf{G}(\cdot) \) and \( \mathbf{H}(\cdot) \) are respectively the sets of equality and inequality algebraic equations.

### C. ACCELERATED MULTI-OBJECTIVE REINFORCEMENT LEARNING

Usually, the conventional RL method achieves optimization and obtains optimal solutions by means of trial and learning. Assume that there is an agent locating within the feasible regions. The agent searches throughout the feasible regions and compares all of the feasible solutions. Once if the agent discovers a solution better than before, it gains a reward. With trial and learning for a long time, the agent is familiar with the way to the largest reward and obtains the optimal solution.

For multi-objective optimization, the reward depends on non-dominated solutions. The agent should compare each potential solution \( u(t_p, q) \) against all other potential ones \( u'(t_p, q) \) were regarded as a non-dominated solution (Pareto solution) and a reward should be paid, if and only if no other potential solutions \( u'(t_p, q) \) are better (i.e., smaller) in terms of both objective functions:

\[
R(u) = \begin{cases} 
1, & \text{non-dominated solution} \\
0, & \text{otherwise} 
\end{cases} \tag{7}
\]

However, a totally throughout search of feasible regions is requested utilizing conventional MORL methods, to obtain all of the non-dominated solutions. This throughout search may be impossible for large-scale power grids with many decision variables. Hence, this paper proposes an AMORL method, which bases on state sensitivities for the significant evaluation, to intelligently accelerate the switch from trial and learning to a deterministic searching direction according to existing knowledge. To a certain extent, the long-term trial and learning are prevented, as long as the agent has gained sufficient knowledge and switches to deterministic searching directions.

Before introducing state sensitivities, the state functions should be redefined to gain information about trial and learning as much as possible. Assume that the coordinated optimization contains \( n \) decision variables with \( m \) statuses each. The state functions have already been defined in conventional MORL methods, but they are only defined for each individual decision variable and summarize the rewards gained. Consequently, \( n \) decision variables need \( n \) state functions. However, this kind of state functions cannot obtain complete reward information about each status. Hence, the AMORL method extends the state function to all of the \( m \) statuses. In other words, \( n \) decision variables totally need \( nm \) state functions:

\[
V_{i,j} = V_{i,j} + \gamma R(u), \tag{8}
\]

where \( V_{i,j} \) denotes the state function of the \( j \)th \( (j = 1, 2, \ldots, m) \) status of decision variable \( i (i = 1, 2, \ldots, n) \); and \( \gamma \) denotes the learning step size. The re-definition (8) allows the agent to record the accurate source (i.e., the exact status) of rewards and facilitates the trial and learning.

The state sensitivity \( S_i \) provides significant evaluation of the convergence of trial and learning and concludes whether sufficient knowledge has been achieved. The state sensitivity \( S_i \) is based on the above state function which summarizes the rewards gained. However, to ensure positive state sensitivity, the state function must be transformed to its natural index form in advance:

\[
\hat{V}_{i,j} = e^{V_{i,j}}, \tag{9}
\]

where \( \hat{V}_{i,j} \) denotes the natural index form of the state function.

The more rewards gained at the current status, the larger probability of this status to be selected. Consequently, the probability of the agent choosing the current status is evaluated as the proportion of \( \hat{V}_{i,j} \) in their sum:

\[
P_{i,j} = \hat{V}_{i,j} / \sum_{j=1}^{m} \hat{V}_{i,j}, \tag{10}
\]

where \( P_{i,j} \) is the probability of the agent choosing the \( j \)th status of decision variable \( i \).

Last but not least, the state sensitivity \( S_i \) can be defined for each decision variable as

\[
S_i = \sum_{j=1}^{m} P_{i,j} \ln P_{i,j} + \ln m, \tag{11}
\]

where \( S_i \) is the state sensitivity for decision variable \( i \).

According to the definition of the state sensitivity (9)–(11), the values of state sensitivity depend on the disparity among state functions of decision variables. With increasing disparity among state functions, the state sensitivity becomes larger and larger. Hence, the state sensitivity \( S_i \) provides significant evaluation about the disparity.

34774 VOLUME 8, 2020
among state functions. Moreover, it would be easier to distinguish between good and bad of the current status. Once \( S_i \) becomes larger than a settled threshold \( S_c \), the disparity among state functions is regarded large enough, and the convergence of trial and learning has been achieved. The agent has already accumulated sufficient knowledge, which significantly guides the remaining search for Pareto solutions. Consequently, the agent should break from trial and learning, and switch to deterministic searching directions. This switching significantly mitigates the computational burdens by providing a deterministic searching direction.

Accordingly, the AMORL method further provides an accelerated searching strategy, to effectively reduce the computational burdens by means of eliminating the feasible regions and intelligently switching from trial and learning to deterministic searching directions. Fig. 3 illustrates the flowchart of the accelerated searching strategy for Pareto solutions of the AMORL method. This strategy is broken into two accelerated steps on the whole: feasible regions reduction and non-dominated solutions search.

1) STEP 1: FEASIBLE REGION REDUCTION
Based on the variable coordination in Section III-A, all of the decision variables are discretized into two types of values: positive and negative. Consequently, the amount of their combinations (\( n \) decision variables each with two types of potential values) would be \( 2^n \). The AMORL method accordingly divides the complete feasible region into \( 2^n \) small independent regions. Afterwards, the positive and negative values of each decision variable are represented by the half of upper and lower limits (i.e., \( \frac{1}{2}u_{i_{\text{max}}}(t,p,q) \) for positive and \( \frac{1}{2}u_{i_{\text{min}}}(t,p,q) \) for negative), respectively. It means, \( 2^n \) combinations are generated for \( 2^n \) small independent regions. These \( 2^n \) combinations are compared to each other with their representations, in terms of their objective values. Only the non-dominated regions containing potential non-dominated optimal solutions are retained, by means of filtering the above small feasible regions with non-dominated comparison. The dominated regions are filtered out, which reduces the feasible regions and accelerates the search for non-dominated solutions.

2) STEP 2: NON-DOMINATED SOLUTION SEARCH WITH ACCELERATION
Within each non-dominated region, the agent further searches for the non-dominated solutions. Once the agent obtains a non-dominated solution, it gains a reward. The AMORL method introduces and updates state functions and state sensitivities every time after a non-dominated comparison. Mainly two types of searching directions are utilized: stochastic and deterministic. The state sensitivities evaluate whether sufficient knowledge has been accumulated and decide to utilize which searching directions. The stochastic searching direction is designed for those decision variables with state sensitivities no greater than a threshold. The stochastic searching direction randomly determines the values of decision variables during each non-dominated comparison and accumulates knowledge about the optimization model. This stochastic searching direction is actually aimless and a waste of time. Once the state sensitivity of the corresponding decision variable gets greater than a threshold, the agent intelligently accelerates and switches from stochastic searching direction to the deterministic direction, where the current decision variable is kept unchanged in the remaining search. This intelligent and accelerated switch significantly mitigates the computational burdens by providing deterministic searching direction. The threshold \( S_c \) is settled as follows. As mentioned in Section III-C, the increasing disparity of probabilities among different statuses results in a growing state sensitivity. Hence, this paper settles the threshold \( S_c \) according to a case with relatively large probability disparity. In this case, the chosen probability is mainly concentrated on a certain status, so the agent can easily select the certain status based on the probability. Assume that \( P_{\text{max}} \) denotes the probability of the above concentrated status. Because all of

![Figure 3. Accelerated searching strategy of AMORL method.](image-url)
the other status probabilities are much smaller than \( P_{\text{max}} \), they are regarded as the same value of \( P_c = (1 - P_{\text{max}}) / (m - 1) \) for simplification. When the status probabilities vary dramatically (for example, the ratio of \( P_{\text{max}} \) and \( P_c \) becomes larger than a certain threshold ratio, 9 adopted by this paper), sufficient knowledge is regarded to be achieved. With this threshold ratio of \( P_{\text{max}} \) and \( P_c \) being determined, all of the probability thresholds can be calculated, because the sum of \( m \) probability thresholds \( (P_{\text{max}}, P_c, P_c, \ldots, P_c) \) must be 1. Practically for AVR reference voltages and switch-able capacitor banks, their numbers of status \( m \) are respectively 41 and 7. Substituting the above probability thresholds into Eq. (11), the threshold \( S_c \) is obtained as 0.2253 and 0.5562 for the two kinds of decision variable.

Furthermore, several potential solutions are introduced for non-dominated comparison to diversify the Pareto solutions. These potential solutions are randomly selected and generated between the current solution and the mean value as provided in (12)

\[
u'_i(t_p, q) = u_i(t_p, q) + \text{rand}(-1, 1) \cdot (u_i(t_p, q) - u_{\text{iz}}),\]

where \( u_i(t_p, q) \) and \( u'_i(t_p, q) \) are values of decision variable \( i \) within, respectively, the current non-dominated solution and the generated potential solution; and \( u_{\text{iz}} \) denotes the mean value of decision variable \( i \).

All solutions mentioned above must pass through a comparison about their domination relationship, to filter out the dominated solutions and to retain the non-dominated ones as Pareto solutions. Both state functions and state sensitivities must be updated to accumulate increasing knowledge, once the agent discovers a non-dominated solution. The flowchart of the accelerated searching strategy using the AMORL method is illustrated in Fig. 3.

All of the reward function, state function, state sensitivity, and accelerated searching strategy are defined from a general perspective, according to the state space, regardless of the nonlinear programming model (6a)–(6d). Hence, the AMORL method can be applied to other general fields. The three contributions: feasible region reduction, function re-definition, switch of searching direction are specially designed to promote classical MORL methods, with better convergence. Because of the high efficiency and good generalization, the AMORL method can obtain optimal solutions even in scalable infrastructures.

**D. TRADEOFF SOLUTION DETERMINED BY FUZZY DECISION-MAKING**

The Fuzzy decision-making [19] method determines the final tradeoff solution for enhanced short-term voltage security, after obtaining a series of non-dominated solutions. Firstly, the objective values of non-dominated solutions are normalized as

\[
\hat{J}^k_i = \frac{J^k_i - J^\text{min}_i}{J^\text{max}_i - J^\text{min}_i},
\]

where \( J^k_i \), \( J^\text{min}_i \), and \( J^\text{max}_i \) respectively denote the values of objective \( i \) with respect to non-dominated solution \( k \), its minimum, and its maximum; and \( \hat{J}^k_i \) is the objective value of non-dominated solution \( k \) after being normalized.

Secondly, the straight distance between the non-dominated solution and the origin point is calculated as

\[
D^k = \left( \left( J^1_k \right)^2 + \left( J^2_k \right)^2 \right)^{\frac{1}{2}}.
\]

Thirdly, select the non-dominated solution, which is the closest to the origin point, as the tradeoff solution. The selected solution is conducted on power grids to enhance short-term voltage security.

**IV. TESTS AND RESULTS**

The proposed AMORL method was implemented on China 748-node provincial power grid for case study. It consisted of 748 nodes, 140 synchronous generators, and 1155 transmission lines. The main 500-kV network topology was illustrated in Fig. 4, where the two test 220-kV regions of Cases 1 and 2 were respectively marked in olive and pink colors. Loads were presented using the combination of 45% constant impedances and 55% induction motors. This power grid contained three types of dynamic elements: four-order dynamic generators, four-order dynamic excitation systems, and three-order dynamic induction motors [20]. The AMORL method was implemented on the platform of Matlab R2018b. In terms of hardware, the computer was configured with 16 GB RAM and 3.2 GHz Core i5 processor.

**FIGURE 4. 500-kV network topology of the 748-node power grid.**

**A. VOLATGES WITHOUT COORDINATED OPTIMIZATION**

Fig. 5 further illustrated two simplified diagrams of the 220-kV regions of the 748-node power grid. The AMORL method was implemented on two short-term voltage security cases: Cases 1 and 2. Both regions suffered from heavy...
loads, so their short-term voltage security problems were outstanding. The most severe contingencies, three-phase short circuits, were similarly employed for the two case studies. The short circuits were respectively located on nodes RB and YF in Cases 1 and 2, and occurred at the moment of $t = 1$ s. In both cases, the short circuits lasted for a duration of 0.1 s. After that, the short circuits were cleaned at the moment of $t = 1.1$ s by tripping the double lines between nodes RB and GN in Case 1, and tripping the double lines between nodes YF and ZJ in Case 2.

The system performances of Cases 1 and 2 without any coordinated optimization were respectively illustrated in Figs. 6 and 7. After short circuits happened, the node voltages dropped to quite a low level sharply. The voltages could not recover to a normal level, and the induction motors rapidly stalled, even if tripping the contingencies. Consequently, the 748-node power grid was insecure in terms of short-term voltages without coordinated optimization.

B. PARETO OPTIMAL SOLUTIONS

The widely used normal boundary intersection (NBI) method [21] was regarded as a numerical benchmark to verify the correctness and acceleration of the AMORL method. The NBI method employed the Baron solver on the GAMS platform [22] to directly handle the mixed-integer nonlinear programming. Figs. 8 and 9 illustrated the Pareto solutions for Cases 1 and 2, respectively, using the two methods: AMORL and NBI. In both figures, the Pareto solutions of these two methods were quite similar, especially their distributions/locations in the space. Consequently, the correctness of the AMORL method was verified. Furthermore, the Pareto solutions of the AMORL method in fact were more evenly distributed and were superior to those of the NBI method. Much more Pareto solutions were easily obtained by means of comparing their domination relationship utilizing the AMORL method, rather than numerical optimization conducted by the NBI method. With much more evenly distributed Pareto solutions, the AMORL outperformed the NBI method in the middle of the Pareto frontier. Note that all of the Pareto solutions have passed through a comparison about their domination relationship and have been verified to be non-dominated solutions.

Moreover, to further quantify the Pareto frontiers of the two methods, the Index C was introduced. Assume that $S_1$ and $S_2$ respectively denote the sets of the Pareto solutions of the AMORL and NBI methods. The Index C is defined as

$$C(S_1, S_2) = \frac{N(S_1)}{N(S_2)}$$

where $N(S_2)$ is the number of non-dominated solutions in $S_2$; and $N(S_1, S_2)$ is the number of non-dominated solutions in $S_2$, which are actually dominated by $S_1$, and vice versa.

The detailed values of Index C between the NBI and AMORL methods were listed in Table 1. The values of
C(\(S_1, S_2\)) in the two cases were always greater than those of \(C(S_2, S_1)\). Based on the definition of Index C in (15), more dominated solutions existed in the Pareto frontier of NBI method with a greater value of \(C(S_1, S_2)\). Although the AMORL and NBI methods obtained quite similar Pareto solutions in Figs. 8 and 9, lots of the solutions of the NBI method were actually dominated by those of AMORL, because \(C(S_1, S_2)\) is much greater than \(C(S_2, S_1)\). Consequently, with a small value of \(C(S_2, S_1)\), the solutions of AMORL were closer to those really existing solutions, and were more likely to be the truly non-dominated solutions.

**TABLE 1.** Index C of Pareto solutions of NBI and AMORL methods (%).

| Case | \(C(S_1, S_2)\) | \(C(S_2, S_1)\) |
|------|----------------|----------------|
| 1    | 48.73          | 11.52          |
| 2    | 44.36          | 29.61          |

**C. SYSTEM PERFORMANCES UNDER COORDINATED TRADEOFF SOLUTIONS**

After obtaining the complete Pareto frontiers in Figs. 8 and 9, their tradeoff solutions were further obtained with the Fuzzy decision-making method, which were listed in Tables 2 and 3. These tradeoff solutions were implemented for the two cases, and the corresponding node voltages and induction motor slips were illustrated in Figs. 10 and 11. After contingencies were tripped, all of the node voltages recovered to a reasonable and normal level of near 1.0 p.u.. The slips of induction motors rapidly dropped to nearly zero and remained safe. Consequently, the 748-node power grid retained short-term voltage security, which verified the effectiveness of coordinated optimization with the AMORL method.

Moreover, Figs. 12 and 13 illustrated the standard deviations of generator reactive outputs in the two cases, before and after optimization. After implementing the coordinated optimization methods, the standard deviations of generator reactive outputs were significantly reduced, which further enhanced the short-term voltage security of the power grid.
optimization of generation and compensation, the standard deviations evidently decreased. Particularly after the second switch/operation, the standard deviations rapidly decreased and got much smaller than those before optimization. The smaller standard deviations were, the more even generator reactive outputs were, which were of benefit to the operation of power grid. It demonstrated that the coordinated optimization of generation and compensation optimized the steady operations of generators and capacitor banks (i.e., the operating status of the power grid), while enhancing short-term voltage security.

D. COMPARISONS WITH OTHER RL METHODS
To verify the efficiency, the AMORL method was compared with two types of RL methods. Type 1 was the classical RL, which was represented by the conventional multi-objective reinforcement learning (CMORL) [11] without any proposed accelerations (e.g., feasible region reduction to eliminate calculation burden or state sensitivities to determine searching directions). Type 2 was the state-of-the-art RL, which was represented by three methods from the latest literatures: DQN [14], BRL [16], and DQN + DDPG [17]. The comparisons were conducted around three aspects: their Pareto objective values, their CPU time, and their root mean squared errors (RMSEs) of the state functions. Note that the DQN, BRL, and DQN + DDPG methods in the literatures were all conducted for single objective optimization. Similar to this paper, they must be extended into multi-objective optimization formulations. This was achieved by returning different rewards when comparing the dominating relationships among the objective vectors of the feasible solutions. The non-dominated solutions were finally remained in a list as the Pareto solutions.

1) TRADEOFF PARETO OBJECTIVE VALUES
Figs. 14 and 15 illustrated the tradeoff Pareto objective values of the above five methods in Cases 1 and 2, respectively. In both cases, the tradeoff Pareto objective values obtained by the five methods were quite close and distributed in a certain range, which verified the correctness of solutions obtained by the AMORL method. Furthermore, compared with the largest objective values obtained by the CMORL method, the objective values of the AMORL method were the smallest among the five methods. The other three state-of-the-art methods obtained objective values between those of the CMORL and AMORL methods. The smallest objective values of the AMORL method declared that our proposed promotions for RL were effective and helped to find the optimal solutions. The AMORL method was slightly superior to the state-of-the-art methods (DQN, BRL, and DQN + DDPG) in terms of optimal objective values.
Z. Deng et al.: Coordinated Optimization of Generation and Compensation to Enhance Short-Term Voltage Security

FIGURE 16. CPU time of five RL methods.

2) CPU TIME
Fig. 16 has illustrated the CPU time of the coordinated optimization utilizing the five methods. Among all of the five methods, the CMORL method underwent the longest CPU time of minutes, while the DQN, BRL, DQN + DDPG, and AMORL methods enjoyed much rapider optimizations within almost equal CPU time of several seconds. The CMORL method did not include feasible region reduction and auto-switching searching directions, so its agent had to explore the optimal solutions in explosive exponential state space. Hence, the trial and learning were actually inefficient and suffered from longer CPU time than the other four methods. Consequently, the AMORL and three state-of-the-art methods were much more efficient and overcame the curse of dimensionality in large-scale coordinated optimization problems. Further compared with the three state-of-the-art methods, the AMORL method also achieved highly efficient performance, since their CPU time was almost the same and it was sufficient for practical operation.

3) ROOT MEAN SQUARED ERRORS OF STATE FUNCTIONS
To eliminate the influence of hardware and better evaluate the convergence of the above five methods, the RMSEs of state functions were calculated. The calculation and approximation of state function (value function) can be referred to Refs. [23] and [24], which are usually model-free and are available for general RL methods. Note that in order to plot the RMSEs of different decision variables in the same figure, we have normalized the values of decision variables within the range from 0 to 1. In this section, the RMSE represents the root mean squared error between calculated values during finite iterations and accurate values after convergence. Small RMSE indicates that the calculated values fit well with the accurate values. Figs. 17 and 18 illustrate the RMSEs of state function after 600 iterations in Cases 1 and 2, respectively. Compared with the state-of-the-art RL methods, the AMORL method enjoyed the smallest RMSEs in general. The order of RMSEs from the largest to the smallest was CMORL, DQN, BRL, DQN + DDPG, and AMORL. Although the CPU time taken by DQN, BRL, DQN + DDPG, and AMORL was almost the same, within several seconds, the AMORL method approached closer to the accurate values than the other three methods in each iteration, and fewer iterations were required. The state-of-the-art methods, including DQN, BRL, and DQN + DDPG, mainly concerned on the approximation of Q-values and the reduction of state spaces, but they rarely involved the switch of searching directions. However, the AMORL method has involved all of the above three aspects, and resulted in a better convergence during each iteration.

V. CONCLUSION
A novel AMORL method was proposed to handle the coordinated optimization of generation and compensation to mitigate short-term voltage security. Its accelerations mainly lay on the feasible region reduction to eliminate calculation burden and the state sensitivities to determine searching directions. The feasible region was eliminated significantly by breaking the full region into many independent small ones and filtering out the dominated ones. The agent intelligently
switched from trial and learning to deterministic searching direction once the state sensitivities were greater than thresholds, which made full use of existing knowledge and avoided excessive learning.

Through case studies on a 748-node power grid, the coordinated optimization was demonstrated to effectively enhance short-term voltage security and optimize the steady operating status of the power grid simultaneously. Moreover, the Pareto solutions of the AMORL method outperformed those of the numerical NBI method, in terms of both their even distributions and domination relationships. Additionally, the AMORL method was compared with another four RL methods to verify its acceleration. Compared to the CMORL method, the AMORL method enjoyed more optimal non-dominated objective values, much shorter CPU time, and better convergence to accurate values. Compared with three state-of-the-art RL methods including DQN, BRL, and DQN + DDPG, the AMORL method not only concerned on the approximation of Q-values and the reduction of state spaces, but also involved the intelligent switch of searching directions. The above promotions of AMORL resulted in CPU time nearly as short as the state-of-the-art methods, and slightly smaller objective values. With the smallest RMSEs among all of the RL methods involved, the AMORL method approached the nearest and most efficiently to the convergence values.

REFERENCES

[1] Y. Dong, X. Xie, W. Shi, B. Zhou, and Q. Jiang, “Demand-response-based distributed preventive control to improve short-term voltage stability,” IEEE Trans. Smart Grid, vol. 9, no. 5, pp. 4785–4795, Sep. 2018.

[2] Z. Deng, M. Liu, Y. Qian, S. Lin, and M. Xie, “Multi-objective mixed-integer dynamic optimization method applied to optimal allocation of dynamic VAR sources of power systems,” IEEE Trans. Power Syst., vol. 33, no. 2, pp. 1683–1697, Mar. 2018.

[3] S. Nourollah and G. B. Gharabpetian, “Coordinated load shedding strategy to restore voltage and frequency of microgrid to secure region,” IEEE Trans. Smart Grid, vol. 10, no. 4, pp. 35182–35193, 2018.

[4] Y. Dong, X. Xie, B. Zhou, W. Shi, and Q. Jiang, “An integrated high side-var-voltage control strategy to improve short-term voltage stability of receiving-end power systems,” IEEE Trans. Power Syst., vol. 31, no. 3, pp. 2105–2115, May 2016.

[5] K. Wang, M. Ye, W. Xiong, F. Wang, and J. Hou, “Coordinated control of STATCOM and mechanically switched capacitors to improve short-term voltage stability,” in Proc. IEEE Int. Conf. Power Syst. Technol. (POWERCON), Wollongong, NSW, Australia, Sep. 2016, pp. 1–5.

[6] B. Liu, J.-H. Zhai, and H.-B. Liu, “Experimental comparison of two feature selection methods based on generic algorithm,” in Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC), Ningbo, China, Jul. 2017, pp. 241–245.

[7] H. Farzin, M. Foutahi-Firuzabad, and M. Moeini-Aghtaie, “A stochastic multi-objective framework for optimal scheduling of energy storage systems in microgrids,” IEEE Trans. Smart Grid, vol. 8, no. 1, pp. 117–127, Jan. 2017.

[8] Y. Sun, S. Li, B. Lin, X. Fu, M. Ramezani, and I. Jathiwa, “Artificial neural network for control and grid integration of residential solar photovoltaic systems,” IEEE Trans. Sustain. Energy, vol. 8, no. 4, pp. 1484–1495, Oct. 2017.

[9] W. Yu, Z. Zeng, B. Peng, S. Yan, Y. Huang, H. Jiang, X. Li, and T. Fan, “Multi-objective optimum design of high-speed backplane connector using particle swarm optimization,” IEEE Access, vol. 6, pp. 35182–35193, 2018.

[10] T. Chen and W. Su, “Indirect customer-to-customer energy trading with reinforcement learning,” IEEE Trans. Smart Grid, vol. 10, no. 4, pp. 4338–4348, Jul. 2019.

[11] 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.

[12] T. Li and M. Liu, “Reduced reinforcement learning method applied to multi-objective coordinated secondary voltage control,” (in Chinese), Proc. CSEE, vol. 33, no. 31, pp. 130–139, Nov. 2013.

[13] Y. Zhao, “Research on several issues related to reinforcement learning,” M.S. dissertation, School Automat., Nanjing Univ. Sci. Technol., Nanjing, Jiangsu, China, 2009.

[14] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, “Adaptive power system emergency control using deep reinforcement learning,” IEEE Trans. Smart Grid, to be published, doi: 10.1109/TSG.2019.2933191.

[15] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, “Two-timescale voltage control in distribution grids using deep reinforcement learning,” IEEE Trans. Smart Grid, to be published, doi: 10.1109/TSG.2019.2951769.

[16] H. Xu, A. D. Domínguez-García, and P. W. Sauer, “Optimal tap setting of voltage regulation transformers using batch reinforcement learning,” IEEE Trans. Power Syst., to be published, doi: 10.1109/TPWRS.2019.2948132.

[17] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, “Deep-reinforcement-learning-based autonomous voltage control for power grid operations,” IEEE Trans. Power Syst., vol. 35, no. 1, pp. 814–817, Jan. 2020.

[18] L. T. Biegler, A. M. Cervantes, and A. Wächter, “Advances in simultaneous dynamic Var sources of power systems,” Energy, vol. 115, pp. 1022–1041, Nov. 2016.

[19] P. Kundur, Power System Stability and Control. New York, NY, USA: McGraw-Hill, 2009.

[20] C. Roman and W. Rosehart, “Evenly distributed Pareto points in multi-objective optimal power flow,” IEEE Trans. Power Syst., vol. 21, no. 2, pp. 1011–1012, May 2006.

[21] The GAMS Solver Manuals, Washington, DC, USA. GAMS Development Corporation. Accessed: Nov. 2019. [Online]. Available: https://www.gams.com/latest/docs/S_MAIN.html

[22] M. Geist and O. Pietquin, “Algorithmic survey of parametric value function approximation,” IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 6, pp. 845–867, Jun. 2013.

[23] G.-F. Wang, Z. Fang, B. Li, and P. Li, “Integrating symmetry of environment in reinforcement learning,” in Proc. 14th Int. Conf. Control, Autom., Robot. Vis. (ICARCV), Phuket, Thailand, Nov. 2016, pp. 1–6.

ZHOUMING DENG received the B.S. and Ph.D. degrees in electrical engineering from the South China University of Technology, Guangzhou, China, in 2014 and 2019, respectively. He is currently with the Electric Power Research Institute, China Southern Power Grid, Guangzhou. His research interests include voltage stability analysis, planning, and optimal scheduling of power systems.

ZHILIN LU received the B.S. degree in electrical engineering from the South China University of Technology, Guangzhou, China, in 2018, where he is currently pursuing the M.S. degree with the School of Electric Power Engineering. His research interests include energy management and operation control of power systems.
ZHIFEI GUO received the M.S. degree from Hunan University, Changsha, China, in 2012. He is currently with the Electric Power Research Institute, China Southern Power Grid, Guangzhou, China. His research interests include operation control, voltage stability analysis, and flexible DC control in power systems.

WENFENG YAO received the M.S. degree from Hunan University, Changsha, China, in 2006. He is currently with the Electric Power Research Institute, China Southern Power Grid, Guangzhou, China. His research interests include operation control, voltage stability analysis, and flexible DC control in power systems.

WENMENG ZHAO received the Ph.D. degree in electrical engineering from the South China University of Technology, Guangzhou, China, in 2016. He is currently with the Electric Power Research Institute, China Southern Power Grid, Guangzhou. His research interests include electricity market and economic dispatch of power systems.

BAORONG ZHOU received the B.S. degree from Wuhan University, Wuhan, China, in 1996, and the Ph.D. degree from Tianjin University, Tianjin, China, in 2004. He is currently with the Electric Power Research Institute, China Southern Power Grid, Guangzhou, China. He is also a senior engineer at professor grade. His research interests include power system planning, power system analysis and control, and renewable energy resources.

CHAO HONG received the B.S. and M.S. degrees from the Wuhan University of Hydraulic and Electric Engineering, Wuhan, China, in 1987 and 1990, respectively, and the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2000. He is currently the Director of the Power System Research Department, Electric Power Research Institute, China Southern Power Grid, Guangzhou, China. His research interests include AC–DC hybrid power system analysis and DC transmission system dynamic performance analysis.