第XX卷 第XX期 XXXX年XX月XX日
DOI：10.7500/AEPS20240919003
电力系统自动化
Automation of Electric Power Systems
Vol. XX No. XX XX XX, XXXX

基于融合经验安全强化学习的配电网电压控制

冯昌森[1]，汤飞霞[1]，王国烽[1]，文福拴[2]，张有兵[1]
（1. 浙江工业大学信息工程学院，浙江省杭州市 310023；2. 浙江大学电气工程学院，浙江省杭州市 310007）

摘要： 随着分布式可再生能源在配电网中的渗透率逐渐提高，分布式并网逆变器参与电压-无功控制对提升电力系统运行的安全性和经济性具有重要意义。然而，在基于强化学习的电压-无功控制模型中，安全运行约束难以建模，且无法确保控制策略满足运行约束。针对上述问题，文中提出一种基于安全强化学习的配电网电压控制策略。首先，将带约束的电压控制问题建模为约束马尔可夫决策过程。然后，采用原始-对偶方法学习最优策略，确保控制策略满足系统运行约束。随后，引入增强经验融合方法来改进强化学习经验利用方式，从而提高算法样本效率。最后，通过配电系统算例验证了所提方法的有效性。

关键词： 电压-无功控制；安全强化学习；约束马尔可夫决策过程；增强经验融合

0 引言

随着配电网中以可再生能源为主的分布式电源（distributed generator，DG）渗透率逐渐提高，可再生能源的随机性、波动性、间歇性给配电网运行带来挑战[1-2]。同时，基于逆变器的并网特点也赋予其参与无功功率调节的能力。此外，通过电压-无功控制（volt-VAR control，VVC）也可提高配电系统新能源消纳、电能质量和供电可靠性[3-4]。

VVC一般被建模为混合整数规划或混合整数非线性规划问题[5-7]。该类方法基于配电网拓扑和电气设备物理模型，精确的模型和参数在实际电网中可能难以获得。为了消除对网络拓扑和参数信息的依赖，有研究提出了基于强化学习（reinforcement learning，RL）算法的无模型方法。RL算法一般可分为基于价值的算法和基于策略的算法。基于价值的算法包括Q学习算法、深度Q网络（deep Q network，DQN）算法[8]等。基于策略的算法包括策略梯度（policy gradient，PG）算法[9]、信任域策略优化（trust region policy optimization，TRPO）算法[10]、近端策略优化（proximal policy optimization，PPO）算法[11]等。行动-评价（actor-critic，AC）框架是最常见的结合策略和价值的算法，例如，深度确定性策略梯度（deep deterministic policy gradient，DDPG）算法[12]、双延迟深度确定性策略梯度（twin delayed deep deterministic policy gradient，TD3）算法[13]和软行动-评价（soft actor-critic，SAC）算法[14]。此外，AC框架也可应用到TRPO和PPO等算法。基于价值的算法需要估计动作价值，在VVC问题中，通常用于长时间尺度下离散调节设备的控制，例如，有载调压变压器、分组投切电容器等[8]。基于策略的算法可输出连续动作，常用于短时间尺度下连续调节设备的控制。DG逆变器能够连续调节无功功率，但DG的出力特点要求控制策略具有较高的适应性和实时性。因此，为了改进策略学习过程，通常采用AC框架来解决无功功率优化等拥有连续、高维状态和动作空间的问题[12-14]。

RL算法允许智能体自由探索环境，采取能够提升奖励的动作，故智能体学习的策略不一定能保证系统安全性。RL算法可以处理奖励驱动的问题，无法直接对约束条件进行建模。因此，在RL算法中，部分研究提出安全强化学习（safe reinforcement learning，SRL）算法的解决方案，通过建模和处理配电网运行优化问题的不等式约束实现安全、可靠运行。文献[15]采用基于惩罚函数法的RL模型，将约束定义为奖励的惩罚项，对多区域VVC问题进行求解。文献[16]提出一种基于安全探索机制的方法，即用一个较小的正实数代表安全边界，当最大或最小电压接近安全边界时，停止迭代更新动作，防止电压违规。文献[13]采用基于物理屏蔽机制来确保

收稿日期：2024-07-10；修回日期：2024-11-19。
上网日期：2025-02-08。
国家自然科学基金资助项目（52107129，U22B20116），浙江省自然科学基金资助项目（LQ22E070007）。

动作安全，即当储能系统荷电状态趋于危险状态时，屏蔽机制将会改变电池的充放电功率，以保证储能系统安全运行。文献[17]将VVC问题表述为约束马尔可夫决策过程(constrained Markov decision process，CMDP)，在每次迭代过程中，采用约束策略优化(constrained policy optimization，CPO)算法将信任域策略梯度投影在约束构成的可行域上，确保策略安全。上述文献中，惩罚函数法最为直观，但是惩罚系数对结果影响较大，并且该方法通常用于软约束，即偶尔违反约束是可以容忍的。安全探索是集成到RL算法框架内的保障机制，会影响算法探索效率和收敛速度。物理屏蔽机制限制了智能体动作可行域，在复杂环境中的优化性能会大幅下降。CPO算法实现过程较为复杂，尤其是在高维策略空间中的计算开销较大。因此，有必要开发一种可以在保证策略安全的前提下，提高经验利用效率的解决方案。

根据经验的利用方式，RL算法一般可分为在轨策略(on-policy)算法和离轨策略(off-policy)算法。离轨策略算法可以从过去的经验中学习，样本效率更高。该类算法在解决复杂问题时较为常用(例如，DDPG算法[12])。然而，离轨策略算法需要对不同策略产生的经验进行随机采样，学习过程波动性较大[18]。为使该类算法在利用大量样本数据时更为稳定、高效地学习，可通过改进经验利用方式增强有效经验对于策略优化的引导。例如，部分研究通过优先级经验采样在学习过程中更快地修正错误预测，加速算法收敛[19-20]。考虑到在轨策略算法可不断学习当前最新样本，及时响应环境的变化，故可在离轨策略经验中融合在轨策略样本，以提升学习过程的稳定性[21]。

在上述背景下，本文提出一种基于SRL算法的配电网VVC策略，在短时间尺度利用多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient，MADDPG)算法训练具有连续调节能力的DG逆变器。首先，将配电网优化运行问题建模为CMDP，解决了约束条件难以建模的问题。然后，采用基于原始-对偶的RL算法寻找最优策略，从而确保控制策略满足系统安全运行约束。随后，本文提出一种增强经验融合方法，将离轨策略样本与最新在轨策略经验有机结合，增强有效经验对策略学习的引导效果进而提高算法样本效率。最后，采用算例对所提模型和算法的有效性进行了验证。

1 基于CMDP的电压控制问题

1.1 CMDP

CMDP在马尔可夫决策过程(Markov decision process，MDP)的基础上增加了代价函数集合C，可表示为一个六元组$<S,\Pi,A,R,C,\gamma>$。其中，S为状态集合；Π为策略集合；A为动作集合；R为奖励函数；C为代价函数集合，表示智能体违反相应约束条件而产生的代价；γ为折扣因子，反映了对未来奖励和代价的重视程度。

定义$C=\{C_m|m=1,2,\cdots,M\}$为约束对应的代价函数集合。其中，C_m为第m个约束对应的代价；M为约束的总数。策略π下的累积折扣代价函数可被定义为$C_m(\pi)=E\left[\sum_{t=0}^{T}\gamma^t C_{m,t}(s_t,a_t,s_{t+1})\right]$，其中，$E[\cdot]$为期望函数；$C_{m,t}$为时段$t$第$m$个约束对应的代价；$T$为CMDP中的总时段数；$s_t$和$a_t$分别为时段$t$的状态和动作。CMDP的目标是在满足累积折扣代价小于上限的条件下，最大化奖励收益，如式(1)所示。

$$\begin{cases}\max\limits_{\pi\in\Pi} R(\pi)\\ \text{s.t.}\quad C_m(\pi)\leqslant\overline{C}_m\end{cases}\tag{1}$$

式中：$\overline{C}_m$为第m个约束对应的累积折扣代价的上限。

为求解CMDP问题，可采用拉格朗日迭代方法求解，其拉格朗日函数为：

$$L(\pi,\lambda)=R(\pi)-\sum_{m=1}^{M}\lambda_m(C_m(\pi)-\overline{C}_m)\tag{2}$$

式中：$L(\cdot)$为拉格朗日函数；$\lambda=\{\lambda_m|m=1,2,\cdots,M\}$为拉格朗日乘子集合；$\lambda_m$为第$m$个约束对应的拉格朗日乘子，且$\lambda_m\geqslant 0$。

根据强对偶定理，式(1)可转变为无约束的对偶问题，最优策略π^*和最优策略对应的拉格朗日乘子λ^*可根据式(3)求得。

$$(\pi^*,\lambda^*)=\arg\min_{\lambda\geqslant 0}\max_{\pi\in\Pi}L(\pi,\lambda)\tag{3}$$

1.2 基于多智能体的VVC模型

1)多智能体学习框架

多智能体强化学习(multi-agent reinforcement learning，MARL)的“集中训练、分布式执行”架构具有降低计算与通信负担的优点。本文采用MARL框架进行算法训练。MARL框架如附录A图A1所示。首先，将电网拓扑划分为N个区域，并将每个区域建模为一个智能体。在集中训练阶段，每个智

能体根据所有智能体共享的观测状态和动作进行价值评估。然后，选择最优动作，实现全局优化。在分布式执行阶段，每个智能体仅根据自己的策略和状态独立决策，不依赖于其他智能体的信息，可以降低在线执行时的通信和计算负担。

2）状态集合

状态集合S由各个智能体的状态S_n组成，即$S=\{S_n|n=1,2,\cdots,N\}$，其中，N为智能体的总数。由于节点电压直接受到各节点注入功率的影响，状态空间必须包含各节点的功率信息。智能体状态空间包括区域内的负荷功率集合、DG有功功率集合和节点电压集合。因此，智能体n在时段t的状态集合$S_{n,t}$为：

$$S_{n,t}=\{P_{n,t}^{\mathrm{L}},Q_{n,t}^{\mathrm{L}},P_{n,t}^{\mathrm{DG}},V_{n,t}\} \tag{4}$$

式中：$P_{n,t}^{\mathrm{L}}$和$Q_{n,t}^{\mathrm{L}}$分别为时段t智能体n中负荷有功、无功功率的集合；$P_{n,t}^{\mathrm{DG}}$为时段t智能体n中光伏的有功功率；$V_{n,t}$为时段t智能体n中电压的集合。

3）动作集合

动作集合A由各个智能体的动作空间A_n组成，即$A=\{A_n|n=1,2,\cdots,N\}$。将DG逆变器的无功功率作为连续动作变量，智能体n在时段t的动作$A_{n,t}$由智能体n内所有逆变器的动作集构成，如式（5）所示。

$$A_{n,t}=\{Q_{n,t}^{\mathrm{DG}},i\in\Omega_n^{\mathrm{DG}}\} \tag{5}$$

式中：$Q_{i,t}^{\mathrm{DG}}$为时段t节点i处逆变器的无功功率；Ω_n^{DG}为智能体n内所有DG节点的集合。

逆变器的无功功率上、下限分别如式（6）和式（7）所示。

$$\overline{Q}_{i,t}^{\mathrm{DG}}=\sqrt{(S_{i,t}^{\mathrm{DG}})^2-(P_{i,t}^{\mathrm{DG}})^2} \tag{6}$$

$$\underline{Q}_{i,t}^{\mathrm{DG}}=-\overline{Q}_{i,t}^{\mathrm{DG}} \tag{7}$$

式中：$\overline{Q}_{i,t}^{\mathrm{DG}}$和$\underline{Q}_{i,t}^{\mathrm{DG}}$分别为时段$t$节点$i$处逆变器的最大、最小无功功率；$S_{i,t}^{\mathrm{DG}}$为时段$t$节点$i$处逆变器的额定容量；$P_{i,t}^{\mathrm{DG}}$为时段$t$节点$i$处逆变器的有功功率。

动作约束为：

$$\underline{Q}_{i,t}^{\mathrm{DG}}\leqslant Q_{i,t}^{\mathrm{DG}}\leqslant\overline{Q}_{i,t}^{\mathrm{DG}} \tag{8}$$

4）奖励函数

奖励函数R由各个智能体的奖励函数R_n组成，即$R=\{R_n|n=1,2,\cdots,N\}$。将智能体n在时段t的奖励函数定义为该区域网损的负值，如式（9）所示。

$$R_{n,t}=-P_{n,t}^{\mathrm{loss}}=-\sum_{ij\in\Omega_n^{ij}}I_{ij,t}^2r_{ij} \tag{9}$$

式中：$P_{n,t}^{\mathrm{loss}}$为智能体n在时段t的网损；Ω_n^{ij}为智能体n中的支路ij的集合；$I_{ij,t}$为支路ij在时段t的电流；r_{ij}为支路ij的电阻。

因此，智能体n的累积折扣奖励J_n^{R}为：

$$J_n^{\mathrm{R}}=E\left[\sum_{t=0}^{T}\gamma^tR_{n,t}\right] \tag{10}$$

5）代价函数

C由各个智能体的代价C_n组成，即$C=\{C_n|n=1,2,\cdots,N\}$。代价函数体现CMDP问题对约束条件的满足程度，本文将节点电压的上下限作为约束条件，当区域内节点电压越限时，就会产生相应代价。电压违规代价值的计算方法为：

$$C_{i,t}^{\mathrm{V}}=\max(0,V_{i,t}-\overline{V}_i)+\max(0,\underline{V}_i-V_{i,t}) \tag{11}$$

式中：$C_{i,t}^{\mathrm{V}}$为时段t节点i的电压违规代价；$\max(\cdot)$为取大值函数；$V_{i,t}$为时段t节点i的电压；$\overline{V}_i$和$\underline{V}_i$分别为节点i处电压幅值的上、下限。

将智能体n在时段t的代价函数$C_{n,t}$定义为区域各节点电压违规代价值总和，如式（12）所示。

$$C_{n,t}=\sum_{i\in\Omega_n}C_{i,t}^{\mathrm{V}} \tag{12}$$

式中：Ω_n为智能体n内所有节点的集合。

因此，智能体n的累积折扣代价J_n^{C}为：

$$J_n^{\mathrm{C}}=E\left[\sum_{t=0}^{T}\gamma^tC_{n,t}\right] \tag{13}$$

为计算上述奖励和代价函数，采用潮流方程模拟电网运行。基于CMDP，配电网电压无功优化控制的目标是在满足累积代价值不超过上限的条件下，最小化网损，智能体n的目标函数为：

$$\begin{cases}\max\limits_{\pi_n\in\Pi_n}J_n^{\mathrm{R}}\\ \text{s.t.}\quad J_n^{\mathrm{C}}(\pi_n)\leqslant\overline{J}_n^{\mathrm{C}}(\pi_n)\end{cases} \tag{14}$$

式中：π_n为智能体n的策略；Π_n为智能体n的策略集合；$\overline{J}_n^{\mathrm{C}}$为智能体$n$的代价上限。

在基于惩罚函数的SRL算法中，约束条件通常作为负奖励加入目标函数，如式（15）所示。

$$\max_{\pi_n\in\Pi_n}E\left[\sum_{t=0}^{T}\gamma^t(R_{n,t}+\eta C_{n,t})\right] \tag{15}$$

式中：η为违反约束的惩罚系数，$\eta\leqslant0$。

2　多智能体SRL算法

2.1　基于原始-对偶的RL算法

多智能体算法的任务是每个智能体通过合作学习最优策略π_n^*。拉格朗日函数为：

$$L(\pi_n,\lambda_n)=J_n^{\mathrm{R}}+\lambda_n(\bar{J}_n^{\mathrm{C}}(\pi_n)-J_n^{\mathrm{C}}(\pi_n)) \quad (16)$$

式中：λ_n为智能体n的约束对应的拉格朗日乘子，$\lambda_n\geqslant 0$。

因此，原问题的对偶问题如式(17)所示。

$$\min_{\lambda_n\geqslant 0}\max_{\pi_n\in\Pi_n}[J_n^{\mathrm{R}}+\lambda_n(\bar{J}_n^{\mathrm{C}}(\pi_n)-J_n^{\mathrm{C}}(\pi_n))] \quad (17)$$

采用原始-对偶的RL算法对CMDP问题进行迭代求解。以第x次迭代为例，更新过程如下：令$\lambda_n=\lambda_n^{(x)}$，实现策略优化，即求解式(17)得到$\pi_n^{(x)}$，其中，上标$x$表示该物理量的第$x$次迭代；令$\pi_n=\pi_n^{(x)}$，实现对偶变量更新，更新完成后再进行下一次迭代。

本文基于MADDPG构建原始-对偶的RL算法，算法结构如图1所示。图1中：$Q_{n,\mathrm{R}}$和$Q_{n,\mathrm{C}}$分别为奖励评价网络和代价评价网络对状态-动作对的值估计，a_n为智能体n策略输出的动作。MADDPG算法中每个智能体的深度神经网络包括：行动网络、奖励评价网络、代价评价网络，以及这3个网络对应的目标网络。行动网络通过输出策略$\mu_n(s_n|\theta_n)$来近似智能体n的策略$\pi_n(s_n)$，其中，θ_n为行动网络的参数。每个智能体与环境交互过程中，其评价网络根据所有智能体的状态和动作估计预期回报和预期代价。此处，评价网络包括奖励评价网络和代价评价网络，其参数分别为$\theta_{n,\mathrm{R}}^{\mathrm{Q}}$和$\theta_{n,\mathrm{C}}^{\mathrm{Q}}$。

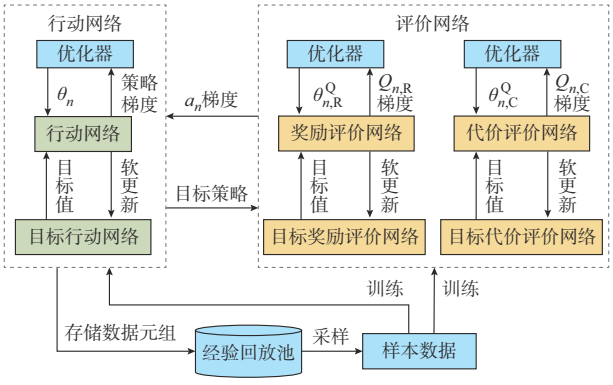


图1　原始-对偶的RL算法的网络结构图

Fig. 1　Network structure diagram of primal-dual RL algorithm

智能体与环境互动得到的样本被存入经验回放池D中。在每次训练中，从经验回放池随机抽样K条转移元组$\{(s_{n,k},a_{n,k},r_{n,k},c_{n,k},s'_{n,k})\}$进行学习，其中，$s'_{n,k}$为状态$s_{n,k}$经动作$a_{n,k}$后的转移状态；$r_{n,k}$为即时奖励；$c_{n,k}$为即时代价。在智能体学习过程中，需要通过样本数据计算当前估计与目标估计，用于更新评价网络参数。令$y_{n,k}$和$z_{n,k}$分别为策略产生的目标奖励值和目标代价值，其计算公式分别如式(18)和式(19)所示。

$$y_{n,k}=r_{n,k}+\gamma Q_{n,\mathrm{R}}(s'_{n,k},\mu_n(s'_{n,k}|\theta'_n)|\theta_{n,\mathrm{R}}^{Q'}) \quad (18)$$

$$z_{n,k}=c_{n,k}+\gamma Q_{n,\mathrm{C}}(s'_{n,k},\mu_n(s'_{n,k}|\theta'_n)|\theta_{n,\mathrm{C}}^{Q'}) \quad (19)$$

式中：θ'_n为目标行动网络的参数；$\theta_{n,\mathrm{R}}^{Q'}$和$\theta_{n,\mathrm{C}}^{Q'}$分别为目标奖励评价网络和目标代价评价网络的参数。

损失函数为目标估计与当前估计差值的平方，如式(20)和式(21)所示。

$$\delta_{n,\mathrm{R}}=\frac{1}{K}\sum_{k=1}^{K}(y_{n,k}-Q_{n,\mathrm{R}}(s_{n,k},a_{n,k}|\theta_{n,\mathrm{R}}^{\mathrm{Q}}))^2 \quad (20)$$

$$\delta_{n,\mathrm{C}}=\frac{1}{K}\sum_{k=1}^{K}(z_{n,k}-Q_{n,\mathrm{C}}(s_{n,k},a_{n,k}|\theta_{n,\mathrm{C}}^{\mathrm{Q}}))^2 \quad (21)$$

式中：$\delta_{n,\mathrm{R}}$和$\delta_{n,\mathrm{C}}$分别为奖励损失和代价损失。

为最小化$\delta_{n,\mathrm{R}}$和$\delta_{n,\mathrm{C}}$，需要计算其关于相应评价网络参数$\theta_{n,\mathrm{R}}^{\mathrm{Q}}$和$\theta_{n,\mathrm{C}}^{\mathrm{Q}}$的梯度，分别如式(22)和式(23)所示。深度神经网络通过反向传播过程将损失函数的梯度信息向神经网络的各层逐步传播。评价网络优化器通过式(24)更新参数$\theta_{n,\mathrm{R}}^{\mathrm{Q}}$和$\theta_{n,\mathrm{C}}^{\mathrm{Q}}$，从而提高价值估计的准确性。

$$\nabla\delta_{n,\mathrm{R}}=\frac{2}{K}\sum_{k=1}^{K}(y_{n,k}-Q_{n,\mathrm{R}}(s_{n,k},a_{n,k}|\theta_{n,\mathrm{R}}^{\mathrm{Q}}))\cdot \nabla Q_{n,\mathrm{R}}(s_{n,k},a_{n,k}|\theta_{n,\mathrm{R}}^{\mathrm{Q}}) \quad (22)$$

$$\nabla\delta_{n,\mathrm{C}}=\frac{2}{K}\sum_{k}^{K}(z_{n,k}-Q_{n,\mathrm{C}}(s_{n,k},a_{n,k}|\theta_{n,\mathrm{C}}^{\mathrm{Q}}))\cdot \nabla Q_{n,\mathrm{C}}(s_{n,k},a_{n,k}|\theta_{n,\mathrm{C}}^{\mathrm{Q}}) \quad (23)$$

$$\begin{cases}\theta_{n,\mathrm{R}}^{\mathrm{Q}}\leftarrow\theta_{n,\mathrm{R}}^{\mathrm{Q}}-\alpha_n^{\mathrm{Q}}\nabla\delta_{n,\mathrm{R}}\\ \theta_{n,\mathrm{C}}^{\mathrm{Q}}\leftarrow\theta_{n,\mathrm{C}}^{\mathrm{Q}}-\alpha_n^{\mathrm{Q}}\nabla\delta_{n,\mathrm{C}}\end{cases} \quad (24)$$

式中：∇为求梯度符号；α_n^{Q}为评价网络参数的更新步长。

一般情况下，更新步长等超参数通过手动调参和经验调参的方法确定，本文算例中设置的具体值见附录A表A1。

在行动网络更新过程中，一般通过梯度上升法更新网络参数。因此，需要根据链式法则分别求解拉格朗日函数中累积折扣奖励和累积折扣代价关于行动网络参数的梯度，其计算过程如式(25)和式(26)所示。

$$\nabla J_n^{\mathrm{R}}=\frac{\partial J_n^{\mathrm{R}}}{\partial a_n}\frac{\partial a_n}{\partial\mu_n(s_n|\theta_n)}\frac{\partial\mu_n(s_n|\theta_n)}{\partial\theta_n} \quad (25)$$

$$\nabla J_n^{\mathrm{C}}=\frac{\partial J_n^{\mathrm{C}}}{\partial a_n}\frac{\partial a_n}{\partial\mu_n(s_n|\theta_n)}\frac{\partial\mu_n(s_n|\theta_n)}{\partial\theta_n} \quad (26)$$

为计算上述梯度数值，评价网络利用从经验回放池中抽取的样本来估计奖励和代价值，并以此计

算策略梯度，更新行动网络的参数θ_n。拉格朗日函数关于行动网络参数的梯度为：

$$\nabla L(\theta_n^{\mu},\lambda_n)=\frac{1}{K}\sum_{k=1}^{K}\nabla\left[Q_{n,\mathrm{R}}(s_n,a_n|\theta_{n,\mathrm{R}}^{Q})-\lambda_n Q_{n,\mathrm{C}}(s_n,a_n|\theta_{n,\mathrm{C}}^{Q})\right] \tag{27}$$

由于MADDPG算法采用确定性策略，对于给定的状态，行动网络会输出一个确定的动作，而不是动作的概率分布。因此，根据式(27)，智能体n的策略梯度为：

$$\nabla L(\theta_n^{\mu},\lambda_n)=\frac{1}{K}\sum_{k=1}^{K}\nabla\left[\nabla\mu_n(s_n|\theta_n^{\mu})\cdot\nabla(Q_{n,\mathrm{R}}(s_n,a_n)-\lambda_n Q_{n,\mathrm{C}}(s_n,a_n))\right] \tag{28}$$

行动网络参数根据式(29)进行更新。

$$\theta_n\leftarrow\theta_n+\alpha_n^{\mu}\nabla L(\theta_n,\lambda_n) \tag{29}$$

式中：α_n^{μ}为行动网络参数的更新步长。

拉格朗日函数关于λ_n的梯度为：

$$\nabla L(\theta_n,\lambda_n)=\frac{1}{K}\sum_{k=1}^{K}\left[Q_{n,\mathrm{C}}(s_{n,k},\mu_n(s_{n,k}|\theta_n)|\theta_{n,\mathrm{C}}^{Q})-\bar{J}_{n,k}^{\mathrm{C}}(\pi_n)\right] \tag{30}$$

在学习过程中，拉格朗日乘子λ_n的更新方法为：

$$\lambda_n\leftarrow\max(\lambda_n+\alpha_n^{\lambda}\nabla L(\theta_n,\lambda_n),0) \tag{31}$$

式中：α_n^{λ}为拉格朗日乘子的更新步长。

当训练到一定步数时，智能体n的目标网络参数需要进行软更新，以使目标评价网络的估计值更为准确。软更新通过逐渐调整目标网络的参数，减少训练过程中的震荡。更新过程如式(32)所示。

$$\begin{cases}\theta_{n,\mathrm{R}}^{Q'}\leftarrow\tau\theta_{n,\mathrm{R}}^{Q}+(1-\tau)\theta_{n,\mathrm{R}}^{Q'}\\ \theta_{n,\mathrm{C}}^{Q'}\leftarrow\tau\theta_{n,\mathrm{C}}^{Q}+(1-\tau)\theta_{n,\mathrm{C}}^{Q'}\\ \theta_n^{\mu'}\leftarrow\tau\theta_n^{\mu}+(1-\tau)\theta_n'\end{cases} \tag{32}$$

式中：τ为软更新系数。

2.2　增强经验融合方法

上述MADDPG算法中的原始策略更新与对偶变量更新均利用了经验回放池中的离轨策略数据样本。传统的离轨策略算法从经验回放池中随机采样一批经验样本以训练网络。本文提出增强经验融合方法改进经验利用过程，主要采用了延迟存储、优先级选择和经验融合技术，实现过程如附录A图A2所示。

在经验收集阶段，推迟向经验回放池中添加最新经验，即将从环境中获得的最新经验保存到临时存放池D_{temp}，间隔一定的回合数ξ再将D_{temp}的经验存入经验回放池中。这主要是为了减少短期内策略波动对学习过程的影响。如果$\xi=1$，即在训练时将策略经验较快地添加到经验回放池中，可能会导致过拟合问题，降低模型在测试集上的泛化能力。如果ξ非常大，即在较长的时间内不向经验回放池中添加新的经验数据，神经网络学习过程会很缓慢。

在参数更新阶段，对所有收集到的转移元组数据添加优先级权重参数，优先级权重参数根据奖励与代价的差值来确定。从经验回放池中随机抽样两批数据，根据余弦相似性计算两批数据优先级权重的相似性，如果两批数据的优先级权重相似度不高，选择K个优先级权重较大的转移元组作为训练用的批数据。如果两批数据的优先级权重相似度较高，则使用随机抽样的批数据进行训练。这主要是为了使智能体关注那些对提升性能更有帮助的经验，并维持样本的多样性。最后，随机抽取批数据中的某条转移元组，用最新的在轨策略元组替换，形成融合样本数据。这主要是为了在充分利用历史经验的基础上，通过最新的在轨策略经验捕捉环境的即时变化。增强经验融合方法的SRL算法流程如附录A所示。

3　算例分析

3.1　配电网系统和参数

在IEEE 33节点测试系统中，对增强经验融合方法的可行性与有效性进行验证。根据无功平衡度指标，将该测试系统划为4个区域智能体(如附录A图A3所示)，并在每个区域内接入光伏机组。该测试系统的基准电压为12.66 kV，节点电压标幺值的安全范围区间为[0.95，1.05]p.u.。基于Python搭建配电网仿真环境，根据智能体的动作进行潮流计算，智能体依据潮流结果计算实时奖励和代价。优化周期为24 h，智能体进行决策的时间尺度为5 min。本文多智能体框架可以满足实时电压的控制要求，在执行阶段，智能体实时获取电网的当前状态信息，仅根据自身状态进行决策，实现分布式实时调度。

在训练时，设置光伏渗透率为60%，即光伏最大出力为负荷最大功率的60%。训练数据来源为中国某地区电网1年的负荷和光伏数据[22]，选取30个典型日，对数据进行适当处理，并添加噪声。随机挑选20天数据作为训练集，其余10天数据作为测试集，相关参数设置如附录A表A1所示。

3.2　算法训练过程对比

为验证本文方法的有效性，设计以下5种算法求解测试系统电压控制问题以对比分析：

1)优化算法,基于决策时间尺度将电压控制问题建模为二阶锥优化模型,并采用Gurobi求解器求解。

2)DDPG算法,使用基于MDP的惩罚函数方法,如式(15)所示。

3)原始-对偶DDPG算法,使用基于CMDP的原始-对偶方法,其目标函数如式(14)所示。

4)原始-对偶PPO算法,使用基于CMDP的原始-对偶方法,其目标函数等相关公式在附录B中说明。需指出PPO算法属于在轨策略算法。

5)本文方法,采用增强经验融合方法的原始-对偶DDPG算法。

上述算法的奖励和约束代价在训练过程中的变化分别如图2和图3所示。

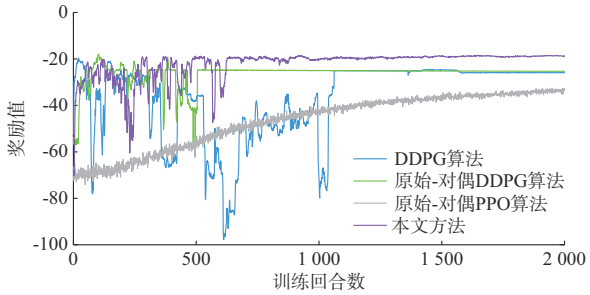


图2　训练过程中使用不同算法的奖励代价对比
Fig. 2　Comparison of reward cost of training process with different algorithms

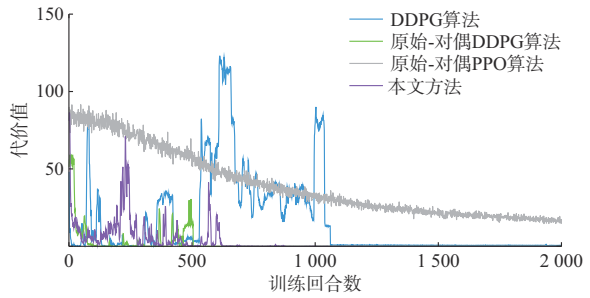


图3　训练过程中使用不同算法的约束代价对比
Fig. 3　Comparison of constraint cost of training process with different algorithms

由图2和图3可知,本文方法在网络收敛后所得累积奖励值最高,说明本文方法拥有最好的性能和效率。原始-对偶DDPG算法和本文方法的累积代价均趋于0,智能体动作策略被限制在安全的区间内。这说明基于CMDP的原始-对偶的RL算法可以更好地平衡奖励和约束,保证策略满足安全约束。另外,原始-对偶DDPG算法在收敛后所得累积奖励代价高于原始-对偶PPO算法,且收敛速度明显更快。这说明在轨策略的PPO算法通过当前策略与环境交互产生的样本更新策略,训练过程稳定,但是样本效率低。离轨策略算法可以从更多的数据样本中学习,因而算法性能更好、效率更高。本文方法在离轨策略数据中融入在轨策略样本,由于实时样本在一定程度上反映了当前环境状态和动作的最新信息,且延迟存储、优先级权重等技术也进一步提高了样本使用效率,因而具有最佳性能。

3.3　测试结果对比

选取5个典型日负荷和光伏数据作为测试数据集,对训练好的模型进行测试。测试指标包括:平均网损、平均电压偏差、电压合格率。平均网损是一天内各时段所有支路功率损耗的平均值。平均电压偏差为一天内所有时段各节点电压与基准电压偏差的平均值。电压合格率是一天内所有时段中电压处于安全约束范围时段的比例。上述5种算法的测试结果如表1所示。选取电压波动相对较大的节点进行分析,这里选择馈线末端节点16分析其在一天内的电压变化,如图4所示。

表1　不同算法性能测试结果
Table 1　Performance test results of different algorithms

名称	模型框架	平均网损/MW	平均电压偏差/p.u.	电压合格率/%
优化算法		0.072	0.023 6	100.00
DDPG算法	MDP	0.089	0.015 9	98.26
原始-对偶DDPG算法	CMDP	0.089	0.011 9	100.00
原始-对偶PPO算法	CMDP	0.102	0.020 2	96.53
本文方法	CMDP	0.074	0.008 9	100.00

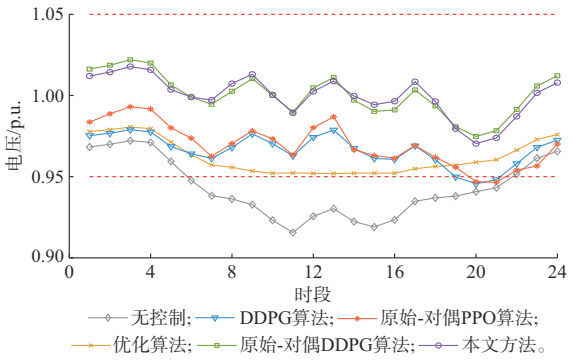


图4　不同算法所训练模型在节点16的电压曲线
Fig. 4　Voltage curves of models trained by different algorithms for node 16

对测试系统33个节点的电压在时段18时的分布进行分析,因为该时段的净负荷较大,较为容易出现电压越限的情况,如图5所示。通过所有光伏逆

变器的总无功功率反映逆变器的调节能力，如附录A图A4所示。

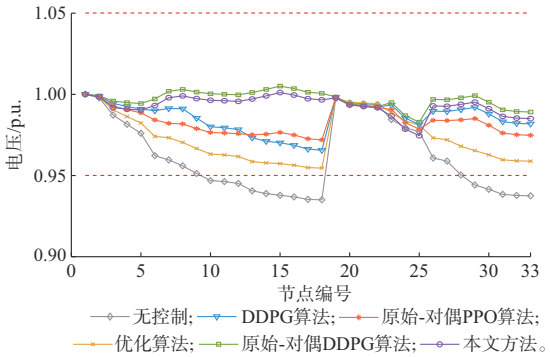


图5　在时段18时测试系统33个节点的电压分布
Fig. 5　Voltage distribution of 33-bus test system at time period 18

由表1可知，优化算法的网损最小，因为其目标函数为网损最小化，可求出全局最优解。相比于惩罚函数DDPG算法，采用CMDP的DDPG算法可以保证电压满足安全约束，且本文方法模型测试的平均网损接近于优化方法所得结果。由图4可知，节点16在基于CMDP的DDPG算法模型下一天内的电压均不越限，而在本文方法模型下相应的电压波动较小。由图5可知，33个节点在基于CMDP的DDPG算法模型下的电压均不越限，而在本文方法模型下相应的电压波动较小。以上算例测试结果说明，本文方法能够在实现较小网损的同时，更好地平衡多项指标性能，使得节点电压波动更小，可以保障配电网安全与经济运行。

本文模型在固定拓扑演算，泛化能力侧重于适应变化多样的负荷和光伏的场景。为了进一步验证算法的泛化能力，本文从该地区电网不同年份的负荷和光伏数据中选取10个典型日，作为第2组测试集，测试结果如表2所示。由附录A表A2可知，本文方法能够保证电压满足安全约束，在新数据上的测试表现最好，说明本文方法具有较好的泛化能力。

3.4　鲁棒性测试

为验证所提方法对光伏渗透率的鲁棒性，设置3种不同的光伏渗透率场景对所提算法进行压力测试，结果如附录A表A3所示。采用本文方法模型时，测试节点16在不同光伏渗透率下的全天电压变化曲线如图A5所示。

由表3可知，在不同光伏渗透率下，本文方法的电压合格率较高，平均网损接近最优值。由附录A图A5可知，本文方法所训练的模型在不同光伏渗透率下的节点电压变化比较平稳。这说明本文方法能较好地适应配电网的环境变化，在光伏渗透率改变时仍能实时生成较优的电压控制策略，具有较强的鲁棒性和较强泛化能力。

4　结语

随着分布式可再生能源渗透率不断提高，基于DG逆变器的VVC对配电网安全经济运行具有重要意义。本文将带约束的电压控制问题表述为CMDP，并通过原始-对偶的RL算法为智能体训练了安全的调度策略，确保系统安全可靠运行。然后，构建了“集中训练、分布式运行”的多智能体框架，可实现全局优化，并降低通信和计算负担，具有良好的实时性能。此外，采用一种增强经验融合方法，改进经验利用过程，有效提升了算法样本效率和性能。最后，通过算例对比验证算法性能，结果表明本文所提的多智能体SRL控制策略可使配电网在达到较低功率损耗的同时，更好地满足节点电压约束。

本文所提出的SRL算法可以集成到拓扑变化的模型架构中，在未来工作中，将考虑拓扑变化的差异性环境，进一步提高模型在配电网应用中的泛化能力。

附录见本刊网络版（http://www.aeps-info.com/aeps/ch/index.aspx），扫英文摘要后二维码可以阅读网络全文。

参 考 文 献

[1] 卓振宇，张宁，康重庆，等. 面向双碳目标的电力系统规划方案量化归因分析方法[J]. 电力系统自动化，2023，47(2)：1-14.
ZHUO Zhenyu，ZHANG Ning，KANG Chongqing，et al. Quantitative attribution analysis method of power system planning scheme for carbon emission peak and carbon neutrality goals[J]. Automation of Electric Power Systems，2023，47(2)：1-14.

[2] 路帅超，孙英云，赵鹏飞，等. 新型电力系统多阶段输-储协同分布鲁棒规划[J]. 电力系统自动化，2024，48(15)：15-24.
LU Shuaichao，SUN Yingyun，ZHAO Pengfei，et al. Multi-stage transmission-storage cooperative distributionally robust planning for new power system [J]. Automation of Electric Power Systems，2024，48(15)：15-24.

[3] Measuring，and verifying volt-VAR control and optimization on distribution systems：IEEE 1885-2022[S]. IEEE，2022.

[4] SINGH S，BABU PAMSHETTI V，THAKUR A K，et al. Multistage multiobjective volt/var control for smart grid-enabled CVR with solar PV penetration[J]. IEEE Systems Journal，15(2)：2767-2778.

[5] NIU T，LIU H T，GUO Q L，et al. Wind farm side optimal power flow based on distflow and SOCP：model and case study

[C]// 2014 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), December 7-10, 2014, Hong Kong, China.

[6] ZHU H, LIU H J. Fast local voltage control under limited reactive power: optimality and stability analysis [J]. IEEE Transactions on Power Systems, 31(5): 3794-3803.

[7] XU T, WU W C. Accelerated ADMM-based fully distributed inverter-based volt/var control strategy for active distribution networks [J]. IEEE Transactions on Industrial Informatics, 16(12): 7532-7543.

[8] 倪爽,崔承刚,杨宁,等.基于深度强化学习的配电网多时间尺度在线无功优化[J].电力系统自动化,2021,45(10):77-85.
NI Shuang, CUI Chenggang, YANG Ning, et al. Multi-time-scale online optimization for reactive power of distribution network based on deep reinforcement learning [J]. Automation of Electric Power Systems, 2021, 45(10): 77-85.

[9] 冯斌,胡轶婕,黄刚,等.基于深度强化学习的新型电力系统调度优化方法综述[J].电力系统自动化,2023,47(17):187-199.
FENG Bin, HU Yijie, HUANG Gang, et al. Review on optimization methods for new power system dispatch based on deep reinforcement learning [J]. Automation of Electric Power Systems, 2023, 47(17): 187-199.

[10] ZENG S Q, HUANG C Y, WANG F, et al. Trust region policy optimization-based secondary frequency regulation for isolated microgrid with voltage constraint [C]// 12th International Conference on Renewable Power Generation (RPG 2023), October 14-15, 2023, Shanghai, China.

[11] 张沛,朱驻军,谢桦.基于深度强化学习近端策略优化的电网无功优化方法[J].电网技术,2023,47(2):562-570.
ZHANG Pei, ZHU Zhujun, XIE Hua. Reactive power optimization based on proximal policy optimization of deep reinforcement learning [J]. Power System Technology, 2023, 47(2): 562-570.

[12] 胡丹尔,彭勇刚,韦巍,等.多时间尺度的配电网深度强化学习无功优化策略[J].中国电机工程学报,2022,42(14):5034-5044.
HU Daner, PENG Yonggang, WEI Wei, et al. Multi-timescale deep reinforcement learning for reactive power optimization of distribution network [J]. Proceedings of the CSEE, 2022, 42(14): 5034-5044.

[13] CHEN P C, LIU S C, WANG X Z, et al. Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems [J]. IEEE Transactions on Smart Grid, 14(4): 2656-2667.

[14] 刘林鹏,朱建全,陈嘉俊,等.基于柔性策略-评价网络的微电网源储协同优化调度策略[J].电力自动化设备,2022,42(1):79-85.
LIU Linpeng, ZHU Jianquan, CHEN Jiajun, et al. Cooperative optimal scheduling strategy of source and storage in microgrid based on soft actor-critic [J]. Electric Power Automation Equipment, 2022, 42(1): 79-85.

[15] LIU H Y, ZHANG C, CHAI Q M, et al. Robust regional coordination of inverter-based volt-VAR control via multi-agent deep reinforcement learning [J]. IEEE Transactions on Smart Grid, 12(6): 5420-5433.

[16] NGUYEN H T, CHOI D H. Three-stage inverter-based peak shaving and volt-VAR control in active distribution networks using online safe deep reinforcement learning [J]. IEEE Transactions on Smart Grid, 13(4): 3266-3277.

[17] LI H P, HE H B. Learning to operate distribution networks with safe deep reinforcement learning [J]. IEEE Transactions on Smart Grid, 13(3): 1860-1872.

[18] SESHAGIRI S, PREMA K V. An empirical study of on-policy and off-policy actor-critic algorithms in the context of exploration-exploitation dilemma [C]// 2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), September 21-23, 2023, Hyderabad, India.

[19] 陈池瑶,苗世洪,姚福星,等.基于多智能体算法的多微电网-配电网分层协同调度策略[J].电力系统自动化,2023,47(10):57-65.
CHEN Chiyao, MIAO Shihong, YAO Fuxing, et al. Hierarchical cooperative dispatching strategy of multi-microgrid and distribution networks based on multi-agent algorithm [J]. Automation of Electric Power Systems, 2023, 47(10): 57-65.

[20] 顾雪平,刘彤,李少岩,等.基于改进双延迟深度确定性策略梯度算法的电网有功安全校正控制[J].电工技术学报,2023,38(8):2162-2177.
GU Xueping, LIU Tong, LI Shaoyan, et al. Active power correction control of power grid based on improved twin delayed deep deterministic policy gradient algorithm [J]. Transactions of China Electrotechnical Society, 2023, 38(8): 2162-2177.

[21] BANERJEE C, CHEN Z, NOMAN N. Improved soft actor-critic: mixing prioritized off-policy samples with on-policy experiences [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 3121-3129.

[22] 郭鸿业,郑可迪,唐庆虎,等.数据驱动的电力市场研究:挑战与展望[J].电力系统自动化,2023,47(1):200-215.
GUO Hongye, ZHENG Kedi, TANG Qinghu, et al. Data-driven research on electricity markets: challenges and prospects [J]. Automation of Electric Power Systems, 2023, 47(1): 200-215.

冯昌森(1990—),男,博士,副教授,主要研究方向:电力系统优化与控制、电力市场、机器学习等。E-mail:fcs@zjut.edu.cn

汤飞霞(1999—),女,硕士,主要研究方向:电力系统优化与控制、机器学习等。E-mail:fxtang2024@163.com

文福拴(1965—),男,教授,博士生导师,主要研究方向:电力系统故障诊断与系统恢复、电力市场与电力经济、智能电网与电动汽车等。E-mail:fushuan.wen@gmail.com

张有兵(1971—),男,通信作者,教授,博士生导师,主要研究方向:智能电网、分布式发电及新能源优化控制、电动汽车入网、电力系统通信、电能质量监控。E-mail:youbingzhang@zjut.edu.cn

(编辑　杨松迎)

Volt-VAR Control for Distribution Network Based on Safe Reinforcement Learning with Mixed Experiences

FENG Changsen[1], *TANG Feixia*[1], *WANG Guofeng*[1], *WEN Fushuan*[2], *ZHANG Youbing*[1]

(1. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China;
2. School of Electrical Engineering, Zhejiang University, Hangzhou 310007, China)

Abstract: With the growing integration of distributed renewable energy sources into the distribution network, the participation of distributed grid-connected inverters in volt-VAR control (VVC) is of great significance to improve the safety and economy of power system operation. However, the safety operation constraints are difficult to model in the VVC control model based on reinforcement learning, and it is not possible to ensure that the control strategy satisfies the operation constraints. To address the above problems, this paper proposes a voltage control strategy for distribution networks based on the safe reinforcement learning. Firstly, the voltage control problem with constraints is modeled as a constrained Markov decision process. Then, a primal-dual approach is used to learn the optimization policy to ensure that the control policy satisfies the system operation constraints. Furthermore, an enhanced experience fusion method is introduced to improve the utilization of reinforcement learning experience, so as to improve the sample efficiency of the algorithm. Finally, the effectiveness of the proposed method is verified by the distribution system example.

This work is supported by National Natural Science Foundation of China (No. 52107129, No. U22B20116) and Zhejiang Provincial Natural Science Foundation of China (No. LQ22E070007).

Key words: volt-VAR control; safe reinforcement learning; constrained Markov decision process; enhanced experience fusion

附录 A

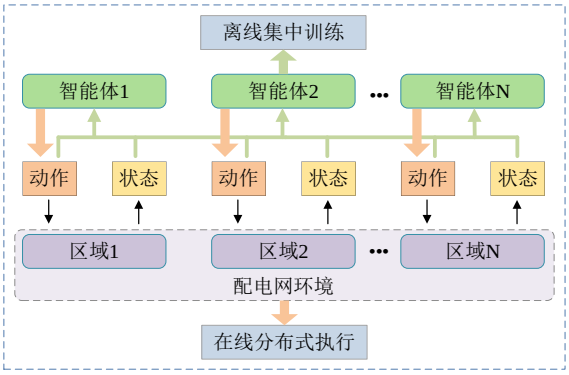


图 A1　MARL 框架
Fig. A1　Framework of MARL

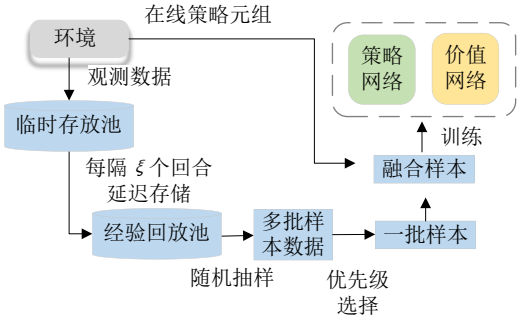


图 A2　增强经验融合过程
Fig. A2　Enhanced experience fusion process

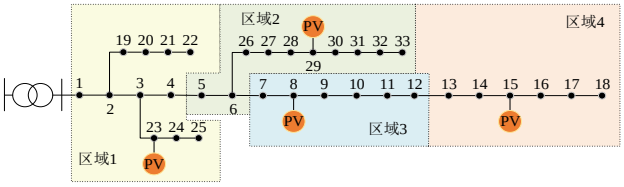


图 A3　IEEE 33 节点配电系统结构
Fig. A3　IEEE 33 Node power distribution system

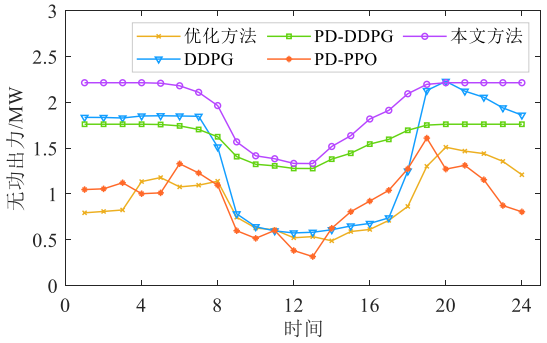
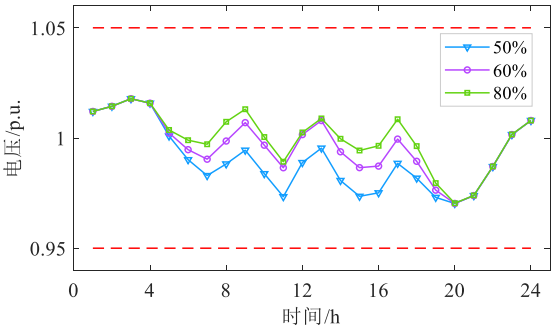


图 A4　光伏逆变器的无功出力总和
Fig. A4　Total reactive power output of photovoltaic inverters



图A5　不同光伏渗透率下节点16的24h电压曲线
Fig. A5　24h voltage profile of node 16 under different photovoltaic penetration rates

表A1　RL算法参数
Table A1　Parameters of RL algorithm

参数	数值	参数	数值
折扣系数γ	0.99	批数据规模K	32
隐藏层激活函数	ReLU	行动网络学习率α_n^{μ}	1e-5
隐藏层神经元数	64	评价网络学习率α_n^{Q}	1e-3
延迟存储间隔ξ	3	惩罚系数η	-10
经验回放池规模	6000	软更新系数τ	0.01
对偶变量学习率α_n^{λ}	1e-3	代价值上限$\bar{J}_n^{C}$	1e-3

表A2　第二组测试集性能测试结果
Table A2　Performance test results of the second test set

训练模型	模型框架	平均网损(MW)	平均电压偏差	电压合格率
优化方法	/	0.075	0.0257	100%
DDPG	MDP	0.108	0.0231	90.63%
PD-DDPG	CMDP	0.079	0.0163	100%
PD-PPO	CMDP	0.104	0.0245	94.44%
本文方法	CMDP	0.079	0.0111	100%

表A3　不同算法所训练模型在不同光伏渗透率下测试结果
Table A3　Test results of models trained by different algorithms under different photovoltaic penetration rates

算法	渗透率	电压合格率	平均网损(MW)
优化方法	50%	100%	0.075
	60%	100%	0.072
	80%	100%	0.065
DDPG	50%	95.83%	0.105
	60%	98.26%	0.089
	80%	99.31%	0.078
PD-DDPG	50%	100%	0.103
	60%	100%	0.089
	80%	100%	0.078
PD-PPO	50%	93.40%	0.136
	60%	96.53%	0.102
	80%	97.92%	0.083
本文方法	50%	100%	0.079
	60%	100%	0.074
	80%	100%	0.068

附录 B

PPO的前身是TRPO，使用KL散度来控制策略更新的幅度，在满足KL散度约束的前提下，通过策略梯度方法更新策略参数，可避免更新幅度过大导致策略发生剧烈变化。TRPO的目标函数为式(B1)。

$$\begin{cases} \max \mathrm{E}\left[\dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta^{old}}(a_t|s_t)}A_t\right] \\ \text{s.t.} \quad \mathrm{E}\left[\mathrm{KL}\left[\pi_{\theta^{old}}(\cdot|s_t),\pi_\theta(\cdot|s_t)\right]\right]\leqslant\Delta \end{cases} \tag{B1}$$

式中：π为策略；A_t为优势函数，用于判断在某个状态下采取哪些动作对策略的改进更有价值；θ为策略参数；θ^{old}为旧的策略参数；Δ为KL散度的阈值。

本文采用PPO-clip算法，用clip方法代替KL散度约束惩罚项。PPO-clip算法的目标为式(B2)。

$$\max \ \mathrm{E}\left[\min\left(r_t(\theta)A_t,\mathrm{clip}\left(r_t(\theta),1-\epsilon,1+\epsilon\right)A_t\right)\right] \tag{B2}$$

式中：clip(·)为剪切函数，能够限制新旧策略之间的差异；ϵ为剪切函数使用的剪切比例，是一个介于0和1之间的数值；$r_t(\theta)$为新旧策略的比率，如式(B3)所示。

$$r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta^{old}}(a_t|s_t)} \tag{B3}$$

因此，式(B2)可以写为式(B4)。

$$J^{\theta^{\mathrm{old}}}_{\mathrm{PPO-clip}}(\theta)\approx\sum_{(s_t,a_t)}\min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta^{old}}(a_t|s_t)}A^{\theta^{\mathrm{old}}}(s_t,a_t),\mathrm{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta^{old}}(a_t|s_t)},1-\varepsilon,1+\varepsilon\right)A^{\theta^{\mathrm{old}}}(s_t,a_t)\right) \tag{B4}$$

本文算例中采用原始-对偶的PPO算法，基于CMDP，在PPO-clip算法中添加安全约束，智能体n目标函数的拉格朗日函数为式(B5)。

$$\begin{aligned} \mathrm{L}(\theta_n,\lambda_n)=&\mathrm{E}\left[\min\left(r_t(\theta_n)A_{n,t},\mathrm{clip}\left(r_t(\theta_n),1-\epsilon,1+\epsilon\right)A_{n,t}\right)\right]+\\ &\lambda_n\left[\bar{J}^{\mathrm{C}}_n-\mathrm{E}\left[\min\left(r_t(\theta_n)A^{\mathrm{C}}_{n,t},\mathrm{clip}\left(r_t(\theta_n),1-\epsilon,1+\epsilon\right)A^{\mathrm{C}}_{n,t}\right)\right]\right] \end{aligned} \tag{B5}$$

式中：$A_{n,t}$为奖励优势函数；$A^{\mathrm{C}}_{n,t}$为代价优势函数。