基于多智能体深度强化学习的配电网无功优化策略

邓清唐[1], 胡丹尔[2], 蔡田田[1], 李肖博[1], 徐贤民[2], 彭勇刚[2]

(1.南方电网数字电网研究院有限公司, 广东 广州 510663;
2.浙江大学电气工程学院, 浙江 杭州 310027)

摘要: 配电网中光伏、风机设备出力随机波动以及负荷波动带来的电压波动、网损增加等问题,给配电网在线无功优化带来了挑战。本文采用一种无模型的深度确定性策略梯度(MADDPG)算法多智能体强化学习框架, 采用集中训练、分散执行的方式解决无功优化问题。MADDPG算法将每一个智能体当作一个行动者(Actor),在离线训练过程中每个Actor可以借助一个评论家(Critic)进行训练。所提策略用深度神经网络拟合可投切电容器、有载调压变压器分接头以及分布式电源逆变器的动作函数,在和配电网环境交互过程中完成深度神经网络的训练。利用该强化学习算法在线实时决策无功调节设备的调度方案,此方法不需要通过精确的潮流建模,也不依赖于日前的数据预测,适用于通信能力较弱的部分观测配电网。最后,通过算例来验证MADDPG算法的有效性。

关键词: 多智能体; 深度强化学习; 无功优化; 数据驱动; 低感知度配电网

DOI: 10.12067/ATEEE2103019 **文章编号**: 1003-3076(2022)02-0010-11 **中图分类号**: TM761

1 引言

随着大量的可再生分布式电源接入配电网中,风力设备和光伏设备出力的随机波动、负荷的不确定性波动会引发配电网运行电压波动大、电压越线、网损提高等问题,影响电能质量[1]。

配电网无功优化的目标就是在充分满足电网安全运营约束下,有效地保证各个节点电压的稳定,减少电压波动和减轻电网的网损。配电网的无功优化往往包含了多个不同的变量、一个约束,通常被认为是非线性计划中的一个问题。在对于传统配电网络中的有功和优化研究中,常用的计算方法包括非线性规划[2]、二次规划[3]、牛顿法[4]等;此外,用于非线性复杂空间中采取随机或近似随机方式寻找最优求解的算法,比如遗传算法[5,6]、模拟退火算法[7,8]、粒子群算法[9,10]等也广泛应用于对无功优化的求解中。这些方法存在计算速度慢、易陷入局部最优、依赖于模型与预测数据等问题[11-14]。随着配电网规模的增加以及无功可控设备装置数量的增多,使得传统方法求解无功优化问题的复杂度大大提高,不再适用于在线控制的无功优化求解。

近年来,人工智能、数据驱动相关技术的推进,使得基于人工智能的优化方法在配电网无功优化中得到了广泛应用。文献[15,16]针对配电网低感知、无模型等特点,以降低网损和成本为目标,提出了一种行动者批评家的深度强化学习算法,实现在线连续无功优化。文献[17]考虑了分布式光伏电源接入和配电网电压波动问题,建立了深度高速公路神经网络拟合的注入功率与节点电压之间的关系。但上述方法无功调节方法单一,未考虑具有一定无功调节能力的设备协同优化,以节约成本,提升电力系统的可靠性、安全性。

考虑接入光伏和风机的实际配电网系统,本文提出一种基于行动者-评论家(actor-critic)的多智能体深度强化学习(Multi-Agent Deep Reinforcement Learning, MADRL)方法用来解决配电网无功优化和电压波动问题[18-21]。结合离散投切电容器(Switching Capacitor, SC)、有载调压变压器(On-Line Tap Changer, OLTC)、分布式电源(Distributed Generation, DG)作为多个智能体进行协调控制和优

收稿日期: 2021-07-19

基金项目: 国家重点研发计划资助(2020YFB0906000, 2020YFB0906002)

作者简介: 邓清唐(1987-), 男, 广西籍, 工程师, 硕士,研究方向为数字电网配用电安全运行技术;
胡丹尔(1996-), 女, 浙江籍, 工程师, 博士研究生,研究方向为智能配电网运行与优化。

化。无功调节设备的动作变量与配电网环境进行交互,通过使用数学中的离散时间序列把交互过程描述成一个称为马尔可夫决策过程(Markov Decision Process, MDP),智能体最终能够实现对于外部环境的最优响应,从而获得最大的回报值[13]。用神经网络方法来分析和拟合每一个智能体的战略函数和动作价值函数,并用深度确定性战略梯度(Multi-Agent Deep Deterministic Policy Gradient, MADDPG)算法对模型进行训练。训练过程不依赖于预测数据结果和精确的潮流建模,多个智能体之间完成协调优化,实现在线的无功优化。在改进的 IEEE-33 配电网系统上进行仿真模拟,验证了所提深度强化学习算法的可行性。

2 多智能体强化学习

强化学习方式的本质在于互动性学习,即可以让一个智能体(agent)与外部环境(environment)之间进行交互。智能体根据自己所感知的环境状态(state)选择响应的动作(action),以对环境做出响应,然后通过观测动作所导致的结果,并依据该动作的结果进行调整,最后通过智能体的动作选择策略对环境做出最优反应,使其获得最大的奖励值(reward)。多智能体强化学习是指多个具有自我控制能力、能够相互作用的智能体,在同一个环境中通过传感器感知状态,执行操作,其算法框架如图 1 所示。

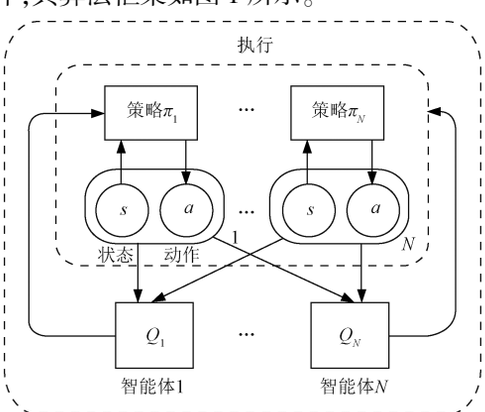


图 1 多智能体强化学习算法框架

Fig.1 Multi-agent reinforcement learning algorithm framework

单个函数智能体在强化学习中一般需要通过 MDP 来对其进行描述,而多个智能体在强化学习中则需要用马尔可夫博弈方法来对其进行强化描述。马尔可夫博弈的方法也被称为随机函数博弈(stochastic game)。一个典型的多智能体马尔可夫博弈用元组来表示,其表达形式为 $\langle N, s, a_1, a_2, \cdots, a_N, T, \gamma, r_1, \cdots, r_N \rangle$,其中 N 表示智能体的个数,s 表示系统状态,$a_1, a_2, \cdots, a_N$ 为智能体的动作集合。T 表示状态转移函数,$T = s \times a_1 \times \cdots \times a_N \times s$,$T \in [0,1]$,根据当前系统状态和联合动作,给出下一个状态的概率。$r_i(s^t, a_1, \cdots, a_N, s^{t+1})$ 表示在 t 时刻下,智能体 i 在状态 s 时,执行联合动作 $a_1, \cdots, a_N$ 所得到的奖励。具体的奖励函数需要根据环境和学习目标设计。γ 是折扣因子,保证越后面的奖励,对奖励函数的影响越小,包含对未来奖励的不确定性。每个智能体都有一个核心目标是为了要求找到一个可以做到最大化的折扣回报。用期望的形式表示为:

$$\max_{\pi_i} \mathop{\mathbb{E}}_{\substack{a_i^t \sim \pi_i \\ s^{t+1} \sim p(s^{t+1} \mid s^t, a_i^t)}} \left(\sum_{t=0}^{T_1} \gamma^t r_i^t \right) \tag{1}$$

式中,π_i 为智能体的策略;γ 为奖励折扣因子;r_i^t 为智能体 i 在 t 时刻下的奖励;a_i^t 为智能体 i 在 t 时刻下的动作;下一时刻状态 s^{t+1} 可以从状态分布 p 中采样;T_1 为决策的周期。

MADDPG 算法可以用于解决多个智能体交互的问题,是一种基于行动者-评论家框架下的 MADRL,需要同时学习一个策略和一个值函数。行动者网络能够教会智能体如何选择动作,而评论家网络是用来评价智能体采取行动可能带来的回报。行动者网络的参数通常是基于评论家网络给出的回报通过策略梯度法完成的。

对于一个具有 N 个智能体的任务,MADDPG 会包含 N 个策略函数和 N 个评价函数。$\pi = \{\pi_1, \cdots, \pi_N\}$ 表示智能体采用 N 个随机策略,其对应的参数为 $\theta = \{\theta_1, \cdots, \theta_N\}$。第 i 个智能体在某个观察下的动作可以表示为 $\pi_i(a_i \mid s_i)$。每一个动作价值函数 Q_i 都是单独学习的,任何一个智能体的奖励都是可以单独进行设计的。MADDPG 的主要目的是已知每个智能体所执行的动作,如果策略进行改变,环境也可以是稳定的。

3 配电网无功优化强化学习建模

3.1 方案描述

结合 SC、OLTC、DG 作为多个智能体进行潮流计算,通过调节无功调节设备观测系统的运行状态。考虑到 SC 和 OLTC 是离散调节,而 DG 的无功是连续调节。本文设计 DG 的逆变器运行在母线上,其视在功率容量为 $S_{\mathrm{DG,bus}}$。DG 逆变器装置在母线上提供或吸收的无功功率可表示为:

$$\underline{Q}_{\mathrm{DG,bus}} \leqslant Q_{\mathrm{DG,bus}} \leqslant \bar{Q}_{\mathrm{DG,bus}} \tag{2}$$

式中，$\bar{Q}_{\mathrm{DG,bus}}$为运行在母线上的最大无功功率值，$\underline{Q}_{\mathrm{DG,bus}} = -\bar{Q}_{\mathrm{DG,bus}}$；$\bar{Q}_{\mathrm{DG,bus}} = \sqrt{(S_{\mathrm{DG,bus}})^2-(P_{\mathrm{DG,bus}})^2}$；$P_{\mathrm{DG,bus}}$为有功功率值。定义控制变量$\alpha_{\mathrm{DG}} \in [-1,1]$，并且$Q_{\mathrm{DG,bus}} = \alpha_{\mathrm{DG}}\bar{Q}_{\mathrm{DG,bus}}$。$Q_{\mathrm{DG,bus}}$的可调节范围相对较小，因为在 DG 运行期间会优先选择更高的功率因数，例如 0.95。

3.2 无功优化模型设计

配电网无功优化的目标是要在最小化有功网损的同时保证电压能在正常范围内运行（0.95pu ~ 1.05pu）。对于无功优化的目标函数 *obj* 定义为

$$obj = \min \sum_{i=1}^{N_{\mathrm{D}}} P_{\mathrm{loss}i} \tag{3}$$

式中，N_{D}为日内指令周期的个数；$P_{\mathrm{loss}i}$为配电网的有功网损。

约束条件包括节点电压、无功功率和动作量变化的上、下限约束以及潮流方程的约束，如下所示：

$$\begin{cases} U_{\min} \leqslant U_{\mathrm{d}} \leqslant U_{\max} \\ Q_{\min} \leqslant Q_{\mathrm{d}} \leqslant Q_{\max} \\ W_{\min} \leqslant W_{\mathrm{d}} \leqslant W_{\max} \\ G_i(T_{\mathrm{d}}) = 0 \quad i = 1,2,\cdots,N_{\mathrm{D}} \end{cases} \tag{4}$$

在低感知度配电网中，式(4)中的潮流方程无法精确计算。T_{d}为控制变量的上下限约束。该配电网中只有部分节点可以实时测量，式(4)只适用于可以测量到的节点；而在部分可观测配电网中，精确潮流模型是无法求解的，因此，网损（式(3)）需要通过部分可测节点的数据进行理论推算得出[22]。

3.3 马尔可夫决策过程

所有满足马尔可夫属性的强化学习过程称为 MDP。在学习过程中，SC、OLTC、DG 被定义为智能体（agent）。agent 执行操作与配电网环境交互。在对 agent 进行训练过程中会根据配电系统中的状态调整策略函数，针对给定的运行条件采取控制措施，以实现无功优化。随着可控装置数量的增多，动作空间的维数呈爆炸式增长。由于联合行动空间的维数极高，单智能体强化学习模型很难有效地提供策略[23]。

3.3.1 状态和动作

配电网无功调节设备动作可以表示为$[a_1,a_2,\cdots,a_N]^{\mathrm{T}}$。$a_i$为 SC、OLTC、DG 的动作集，$a_i \in \mathscr{A}_i$，$\mathscr{A}_i$为第$i$个动作的搜索空间，$i \in \{1,2,\cdots,N\}$。在学习过程中，智能体之间的有效通信，可以通过它们当前状态 s 和最新动作 a 的共享观察来选择最优动作。在每一个训练幕（episode）中，agent 会根据当前状态向环境提供新的控制动作。例如，对于智能体 j 采取的新动作为 $a(j) = [a_1,a_2,\cdots,a_{j-1},a'_j,a_j,\cdots,a_N]^{\mathrm{T}}$。在一个幕内作用于配电网的新动作集定义为 $a' = [a'_1,a'_2,\cdots,a'_N]^{\mathrm{T}}$。

对于多智能体给定的动作集 a，环境提供配电系统中所有母线上的电压，作为 MADDPG 的状态，可以表示为：

$$s = \{\boldsymbol{U}_i,\boldsymbol{W}_i,\boldsymbol{C}_i\} \tag{5}$$

式中，$\boldsymbol{U}_i$为第 i 个决策阶段的配电网的节点电压矩阵，维度为 $n \times m$，n 为可量测的节点个数，m 为调度周期的测量次数；$\boldsymbol{W}_i$为第 i 个调度周期内各个调节设备的投切档位，为了方便训练，本文用 one hot 的编码方法[16]，如图 2 所示；$\boldsymbol{C}_i$为 i 个调度周期内各个调节设备已经完成的动作，也用 one hot 编码格式。举例说明 one hot 编码方法，假设配电网系统的动作决策周期时间为 15 min，无功设备的采样时间在 15 min，OLTC 的可调比在 0.9 pu ~ 1.1 pu 之间，共有 9 个档位；SC1 和 SC2 分别有 3、4 档可调。$\boldsymbol{W}_i$可以用一个 9 + 3 + 4 = 16 位的 one hot 编码来表示。1 代表接入该档位，0 代表不接入该档位。$\boldsymbol{W}_i$的组成由图 2 所示。

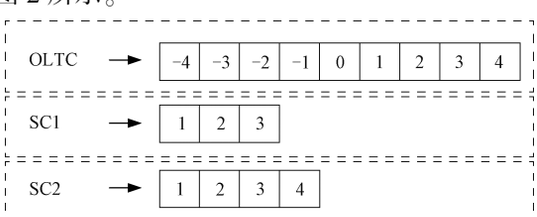


图 2 $\boldsymbol{W}_i$ 的 one hot 编码组成示意图

Fig.2 Diagram of one hot coding of $\boldsymbol{W}_i$

假设并联电容器最大投切次数是 5，OLTC 的累计变化档位上限为 8。在一个决策周期内，调节设备 OLTC 选择分接头为 5 号档位，SC1 和 SC2 档位数分别为 3 档、2 档，当前决策周期内，OLTC、SC1、SC2 的累积档位变化分别为 6、4、3，则 $\boldsymbol{W}_i$ 和 $\boldsymbol{C}_i$ 的 one hot 编码为：

$$\boldsymbol{W}_i = [0000100000010100] \tag{6}$$

$$\boldsymbol{C}_i = [000001000001000100] \tag{7}$$

动作空间 $\mathscr{A}_i$ 定义为下一个指令周期的无功调节设备的档位状态，也采用 one hot 编码表示，即：

$$a_i = \boldsymbol{W}_{i+1} \tag{8}$$

将配电网的潮流信息矩阵 $\boldsymbol{U}_i$ 与投切档位矩阵 $\boldsymbol{W}_i$ 以及累计投切档位矩阵 $\boldsymbol{C}_i$ 拼接后得到状态矩阵的维度为 33+16+18=67。

3.3.2 奖励函数

根据 3.2 节建立的无功模型,为了求解优化模型,将电压约束引入到目标函数中,构造深度强化学习的奖励函数。对每个智能体的奖励函数进行准确量化设计,来保证强化学习算法高效运行。无功优化的目标必须要保证所有电压幅值在正常工作范围内,否则该无功优化问题就没有可行解。因此,在计算功率损耗降低之前,所提出的奖励函数首先检查各智能体动作是否导致电压违规,智能体需尽可能避免产生导致电压越限的操作。

SC 智能体节点电压需满足约束条件,并且在当前幕内没有超过允许动作次数情况下,将即时回报 r_i 设定成网损和动作成本之和的相反数。如果违反了约束条件(4),则会对智能体的奖励施加重大的惩罚因子,在当前调度时刻获得的即时奖励定义为:

$$r_{i,1}=-P_{\text{loss}i}-\lambda_{\text{c}}\sum_{j=0}^{i}\left|B_{\text{bus},j}-B_{\text{bus},j-1}\right|-\eta_1\sigma(\left|B_{\text{bus},j}-B_{\text{bus},j-1}\right|\leqslant N_{T,\max}) \tag{9}$$

式中,$\sigma(\cdot)$为判断函数,无功补偿设备的调节累计档位变化次数若超过约束调节,则函数值为 1,未超过时,则值为 0;λ_{c} 为动作调节代价;$B_{\text{bus},j}$ 为第 j 次决策时 SC 的投切挡位;$N_{T,\max}$ 为优化周期内的累计档位变化上限;η_1 为惩罚因子。

OLTC 智能体在当前调度时刻获得的即时奖励可以定义为:

$$r_{i,2}=-P_{\text{loss}i}-\lambda_{\text{o}}\sum_{j=0}^{i}\left|N_{L,j}-N_{L,j-1}\right|-\eta_2\sigma\left(N_{L,j}-N_{L,j-1}\leqslant N_{L,\max}\right) \tag{10}$$

式中,λ_{o} 为 OLTC 智能体档位动作代价;$N_{L,j}$ 为第 j 次决策时 OLTC 的投切挡位;L 为 OLTC 所在的节点序号;$N_{L,\max}$ 为优化周期内 OLTC 累计档位变化上限;η_2 为惩罚因子。

DG 智能体在当前调度时刻获得的即时奖励可以定义为:

$$r_{i,3}=-P_{\text{loss}i}-\lambda_{\text{d}}\sum_{k=1}^{N_{\text{e}}}\left|\frac{U_{k,j}-U_{k,\text{base}}}{U_{\max}-U_{\min}}\right|-\eta_3\sigma(U_{\min}\leqslant U_{k,j}\leqslant U_{\max}) \tag{11}$$

式中,λ_{d} 为 DG 智能体档位调整设定值;$U_{k,\text{base}}$ 为电压基准值;$U_{k,j}$ 为 DG 智能体所连母线的电压量测值;$U_{\max}$ 和 $U_{\min}$ 分别为电压上、下限;N_{e} 为可观测的节点总数;η_3 为惩罚因子。

3.4 配电网多智能体深度强化学习算法

3.4.1 深度神经网络算法分析

目前常用的多智能体强化学习算法主要有两种:基于值函数的多智能体深度强化学习(Deep Reinforcement Learning,DRL)和基于策略的多智能体 DRL。基于值函数的方法是通过得到一个值函数,根据该值函数可以生成相应的策略。基于策略的方法则直接在策略空间利用梯度上升找出最优的策略[24]。

本文中采用的 MADDPG 算法构建了两个神经网络:行动者网络(actor network)和评论家网络(critic network)。行动者网络是将状态-行为值函数和策略梯度法相结合。θ_1 和 θ_2 分别为行动者网络和评论家网络的参数。通过调节神经网络的参数 θ 来确定某状态下的最佳行动。评论家网络通过计算时间差分的误差(temporal difference error)来评估行动者网络的行为。每个行动者网络和评论家网络中同时构建了两个结构相同、参数不同的神经网络,即估值网络(evaluation network)和目标网络(target network)。估值网络的参数可以根据训练不断调整更新,目标网络则不参与训练过程,其参数会根据估值网络的参数进行迭代更新。因此本文设计了行动者-评论家的深度神经网络结构如图 3 所示。图 3 中所示的深度神经网络结构主要由 3 个部分组成:①卷积神经网络用来提取关键特征;②行动者网络可以拟合状态到动作的映射;③评论家网络可以拟合状态价值函数。根据 3.3.1 节分析,配电网无功优化的马尔可夫决策的状态矩阵记作 $[\boldsymbol{U}_i,\boldsymbol{W}_i,\boldsymbol{C}_i]$,作为模型的输入。

行动者网络是一个具有三个全连接层的神经网络。输入维度为状态矩阵的维度,两个隐含层各有 128 个神经元;输出维度为矩阵 $\boldsymbol{W}_i$ 的维度,该层的激活函数为 ReLU。评论家网络与行动者网络输入相同,且同样是用三层全连接网络,与行动者网络不同的是评论家网络用来拟合状态价值函数 $V(s)$,输出的维度是 1,代表对每一个状态-动作的值估计。

3.4.2 深度神经网络优化

在深度神经网络训练过程中,第 i 个智能体的策略梯度可以表示为:

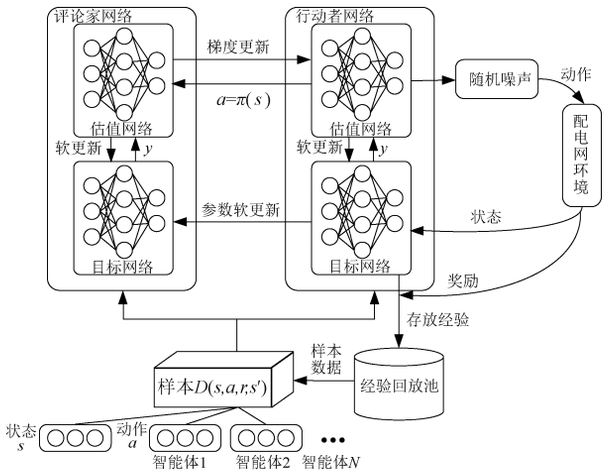


图 3 MADDPG 算法神经网络结构

Fig.3 Neural network structure of MADDPG algorithm

$$\nabla_{\theta_i} J(\theta_i) = \mathop{\mathbb{E}}_{\substack{s\sim\rho,\\ a_i\sim\pi_i}} [\nabla_{\theta_i} \log \pi_i(a_i|s_i) Q_i(s,a_1,\cdots,a_N)] \tag{12}$$

式中，s 为整个环境的状态；s_i 为智能体 i 的观测值；$Q_i(s,a_1,\cdots,a_N)$ 为第 i 个中心化的评价函数，其输入是每个智能体采取的动作 a_i 以及环境状态 s，其输出为第 i 个智能体的动作价值函数 Q 值。

由于根据概率选择动作会存在收敛效率低的困难，MADDPG 方法将随机策略替换为确定性策略，用 μ_i 表示第 i 个智能体的策略，θ_i 表示价值网络的参数，则智能体的梯度可以描述为：

$$\begin{aligned}\nabla_{\theta_i} J(\mu_i) &= \mathop{\mathbb{E}}_{\substack{x,\\ a_i\sim\mu_i}} [\nabla_{\theta_i} \log \mu_i(a_i|s_i) Q_i(s,a)] \\ &= \mathop{\mathbb{E}}_{\substack{x,\\ a_i\sim\mu_i}} \left[\frac{1}{\mu_i(a_i|s_i)} \nabla_{\theta_i} \mu_i(a_i|s_i) Q_i(s,a)\right] \\ &= \mathop{\mathbb{E}}_{\substack{s,\\ a_i\sim D}} [\nabla_{\theta_i} \mu_i(a_i|s_i) \nabla_{a_i} Q_i(s,a) |_{a_i=\mu_i(s_i)}]\end{aligned} \tag{13}$$

式中，$D(s,a,r,s')$ 为经验回放缓存区（experience replay buffer），记录了所有智能体的经验，$a=\{a_1,\cdots,a_N\}$，$r=\{r_1,\cdots,r_N\}$。

决策主体与环境交互过程中，评论家网络的损失函数定义为：

$$\mathscr{L}(\theta_i) = \mathop{\mathbb{E}}_{x,a,r,x'} [(Q_i(s,a) - y)^2] \tag{14}$$

其中

$$y = r_i + \gamma Q_i[s_{t+1}, a_{t+1}] |_{a_{t+1}=\mu'(s_i)} \tag{15}$$

式中，$\mu'=\{\mu_{\theta'_1},\cdots,\mu_{\theta'_N}\}$ 定义为带延迟参数的策略集，$\mu_{\theta'_1},\cdots,\mu_{\theta'_N}$ 为智能体 $1\sim N$ 的网络参数；s_{t+1} 和 a_{t+1} 分别为 $t+1$ 时刻的状态和动作。

4 算例分析

4.1 算例系统介绍

算例基于 IEEE-33 节点配电网系统进行改动，基准电压调整为 12.66 kV。算例系统中包含了 3 个 DG，每个容量为 750 kW，采用下垂控制用以调节电压，如图 4 所示。将整个系统的观测率设定为 60%。图 4 中能观测到的节点标为实心，不能观测到的节点标为空心。在系统中节点 21 号、32 号加上并联电容器。每个 SC 每日动作次数上限是 5 次。SC1 有 6 个档位，每档为 0.4 MVar，总容量为 2 MVar，SC2 有 5 个档位，每档为 0.3 MVar，总容量为 1.2 MVar。OLTC 的可调节比例为 0.95 pu ~ 1.05 pu，设定 10 个调节档位。考虑所有设备的动作变量会产生一个高位的联合决策空间，例如，在 33 节点的总线系统中，假设 SC1 和 SC2 各有 5、4 个档位（各 6、5 种投切状态，其中包含全退状态）指定为智能体 1 的动作，记作 $dim(\xi_1)=6\times5=30$。OLTC 有 11 个离散档位指定为智能体 2 的动作，记作 $dim(\xi_2)=11$。为了训练收敛效果更好，通过节点 15 的 DG 来参与调节无功功率，设置 DG 有 20 个离散档位为智能体 3 的动作，记作 $dim(\xi_3)=20$。则所有动作通过矩阵拼接所产生的动作组合共有 $30\times11\times20=6600$ 个。

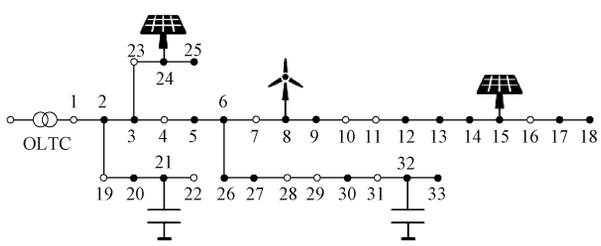


图 4 配电网算例系统和分布式电源无功优化

Fig.4 Distribution network system and reactive power optimization of distributed generation

选取浙江某地区半年的分布式光伏、风机以及负荷出力的实际数据进行深度强化学习算力分析。随机挑选两个典型日（冬季典型日和夏季典型日）共 30 天作为强化学习的测试集，其余时间作为训练集，取决策间隔时长为 15 min。

4.2 深度学习训练过程

深度强化学习训练过程需要在多智能体决策和环境的交互过程下实现。本文设计搭建了 Matlab 仿真环境来代替深度强化学习实际配电网的环境，通过计算潮流的方式来代替训练过程中的无功优化

环境。搭建 Python 中的 Pytorch 框架完成多智能体深度强化学习算法,设置网络中各个参数值,训练过程可以通过软件之间的接口来完成,模型的神经网络结构训练参数详见表 1。

表 1 神经网络结构以及参数设定

Tab.1 Neural network structure and hyperparameters

	评论家网络	行动者网络
网络层数	3	3
学习率	0.01	0.001
折扣因子	0.99	0.99
隐藏层激活函数	ReLU 函数	ReLU 函数
输出层激活函数	双曲正切函数	双曲正切函数
每层神经元个数	128	128
训练的最大幕数	200	200
每幕训练次数	200	200
经验回放池规模	10 000	10 000

在本文例子中,例如在图 1 所示多智能体的强化机器学习训练算法中,训练以及学习也都应该是统一进行。在训练开始阶段,对各个智能体的动作进行初始化准备,仿真环境会根据行动者评论家网络给出的动作指令进行潮流计算,并且完成该动作 a。根据各个智能体的奖励函数得到即时的奖励 r_i。

各个行动者收集数据 (s,a,r,s',a'),并存入经验回放池中,当缓存池的数量超过了预热的阈值,就会开始进行学习。每个行动者分别更新策略 π,与深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)算法相似,只需要当前 s,$a=\mu(s)$[25],$\mu(s)$ 为智能体在环境 s 下的策略。每个评论家分别更新动作价值参数,每个评论家都可以看到所有的行动者收集到的数据,更新参数的时候会考虑所有的行动者自己生成的数据,即优化后的结果是每一个评论家对于全局的贡献最大。在这个过程中反复地训练,最后神经网络会达到收敛的效果,其过程用流程图描述如图 5 所示。

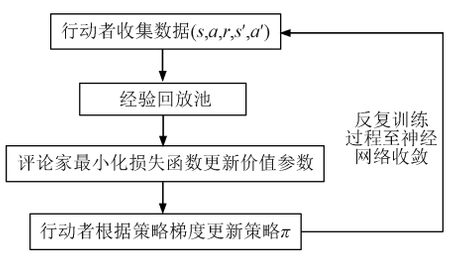


图 5 行动者-评论家的神经网络训练过程

Fig.5 Training process of actor and critc neural network

第 3.3.2 节中的式(9) ~式(11)中,设备动作成本的系数 λ_c、λ_o 和 λ_d 分别为 5、6 和 7。η_1、η_2 和 η_3 惩罚因子均为 1 000。一天 24 h 96 个时段作为一幕,即该天结束后本回合结束。在多智能体与环境交互过程中,累计的回报值不断地变大,在训练结束后智能体的动作选择也趋于稳定。训练结果显示如图 6 所示,其中,L_{oss} 为损失函数,为单个训练样本与真实值之间的误差。网络的回报值在经过2 000 幕左右逐渐收敛,达到了比较理想的控制效果,也验证了 MADDPG 算法在文中对于配电网环境下的无功功率控制和实际应用的可行性和有效性。

4.3 仿真结果分析

4.3.1 算法有效性分析

为了验证 MADDPG 算法的有效性,本文采用以下两种算法进行对比:①为了证明 MADDPG 方法比传统的无功功率优化算法粒子群(Particle Swarm Optimization,PSO)更具有效性,利用文献[26]中的算法来调节典型日下 24 h 内多个无功调节设备,选择离散的动作方案;② 将 MADDPG 与其他强化学习算法进行对比,选择文献[27]提出的基于值的深度 Q 网络(Deep Q Network, DQN)无功优化方法进行对比,观察网损和节点电压的情况。

利用所提出的不同优化方法的典型日(夏季日和冬季日)的 12:00 这一典型时刻进行电压分布情况和电压偏差的对比,结果如图 7 和表 2 所示。

表 2 典型日电压和网损对比

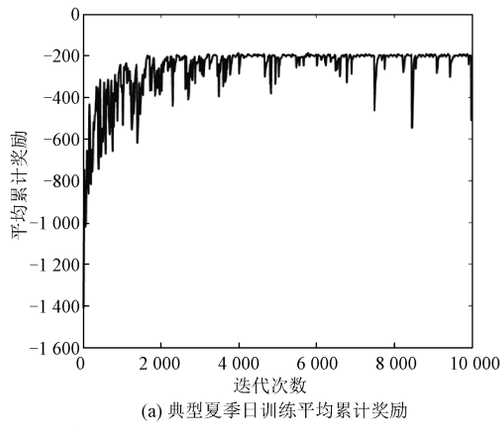
Tab.2 Comparison of typical daily voltage and network loss

方法	平均网损/kW		电压偏差(pu)	
	夏季日	冬季日	夏季日	冬季日
PSO	128.9	139.7	7.313	6.871
DQN	121.1	122.2	5.874	5.593
MADDPG	112.1	115.7	3.327	3.102

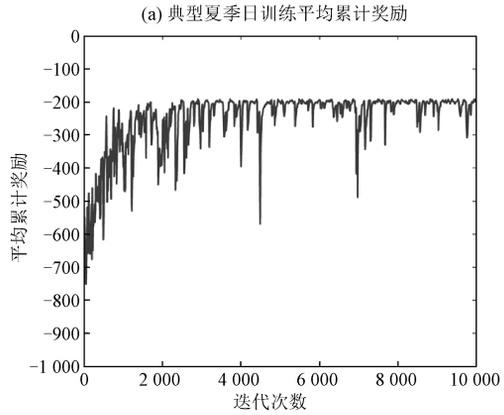
由图 7 和表 2 可知,本文采用的 MADDPG 算法优化以后得到的电压偏差最小,即保证了电压运行的稳定性,最小化电压波动。本文采用深度强化学习算法,多智能体可以最大化奖励,有效提高设备动作的合理性。

利用不用方法对两个典型日下网络的损耗结果进行对比分析,如图 8 和表 2 所示。

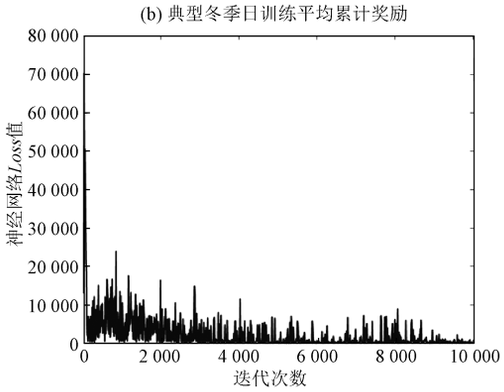
从图 8 和表 2 中可知,采用本文的 MADDPG 算法得到的典型日下的网损更低。将 MADDPG 算法与其他两种方法的结果进行对比,在夏季日的平均网损分别降低了 14.98% 和 8.03%;在冬季日的平



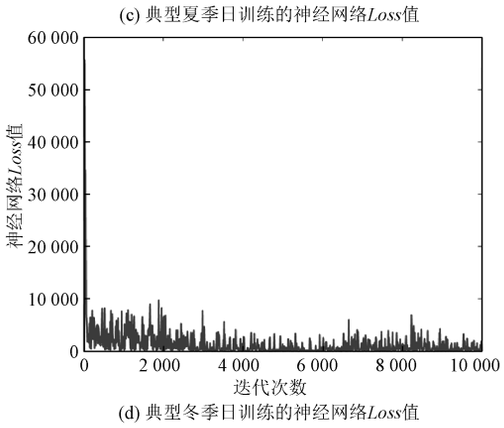
(a) 典型夏季日训练平均累计奖励



(b) 典型冬季日训练平均累计奖励



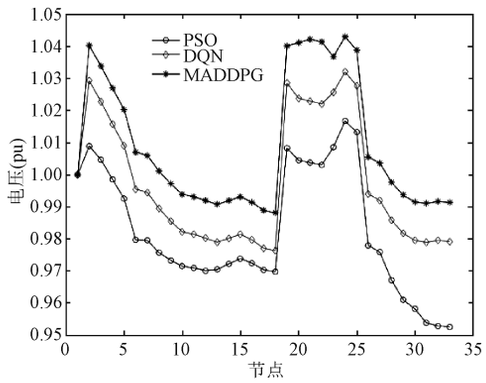
(c) 典型夏季日训练的神经网络*Loss*值



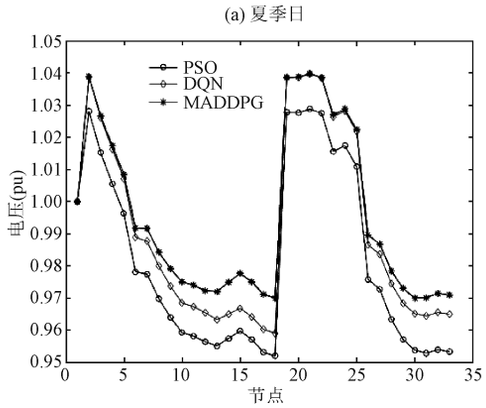
(d) 典型冬季日训练的神经网络*Loss*值

图 6 平均累计奖励和神经网络 *Loss* 值训练结果

Fig.6 Average cumulative reward and neural network *Loss* value training results



(a) 夏季日



(b) 冬季日

图 7 典型日下电压分布情况

Fig.7 Voltage distribution under typical day

均网损分别降低了 20.74%和 5.62%。所以可以验证本文 MADDPG 算法在 2 种典型日下,都可以更大效率减小系统网损,即说明了本文提出算法的有效性和优越性。

不同算法的 2 个典型日下,OLTC、SC 调节设备的日累计档位变化数如图 9 所示。采用本文的方法优化以后的设备累计档位变化比其他两种方法更小,说明本文方法优化设备的动作成本更小,具备更好的经济性。各优化决策算法下,日内的 DG 调节设备日均补偿无功的出力见表 3。

表 3 调压装置参数设定

Tab.3 Parameters of voltage regulation devices

设备	夏季日日均 无功补偿/kVar	冬季日日均 无功补偿/kVar
DG1	54.70	57.89
DG2	28.74	89.65
DG3	97.22	85.34

4.3.2 复杂配电网算法有效性对比分析

为了验证该方法的可扩展性和适用性,还对改进的 IEEE-123 节点配电网进行了仿真。调压

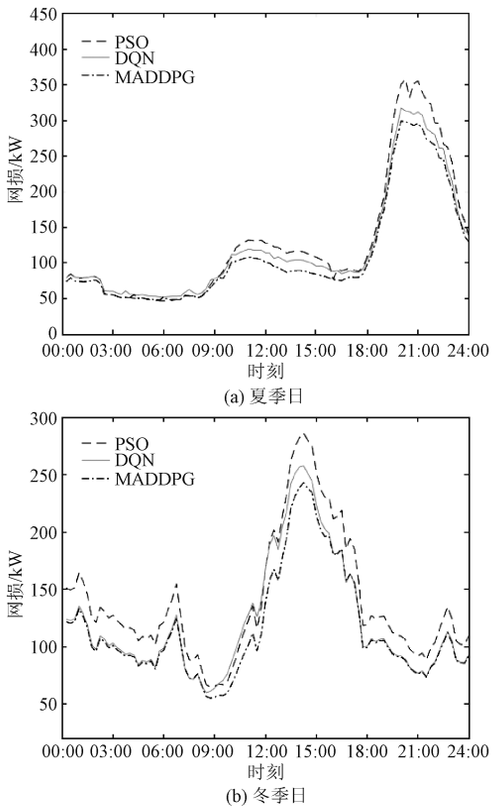


图 8 典型日下 IEEE-33 网损情况

Fig.8 Typical daily network loss of IEEE-33

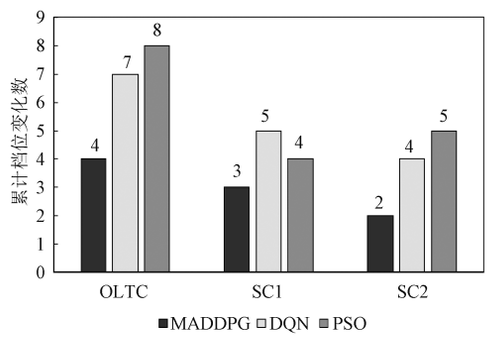


图 9 不同算法离散调节设备的日累计档位变化数

Fig.9 Daily change number of discrete adjustment devices

装置的节点位置、有载调压变压器 OLTC、离散投切电容器(Capacitor Bank,CB)和分布式电源 DG 的详细参数见表 4。采用本文不同算法对两个典型日 IEEE-123 节点配电网的网络损耗结果进行了对比分析,如图 10 所示。可以看出本文提出的 MADDPG 算法在两种典型日下,在更复杂的配电网拓扑 IEEE-123 中也可以更大效率地减小系统网损,再次证明了本文提出的 MADDPG 算法的有效性。

表 4 调压装置参数设定

Tab.4 Parameters of voltage regulation devices

	参数	运行限制	节点位置
OLTC	± 10 × 0.01	5	1
CB1~CB4	5×100 kVar	6	16,43,58,103
DG1~DG6	750 kW	–	28,48,67,89,95,112

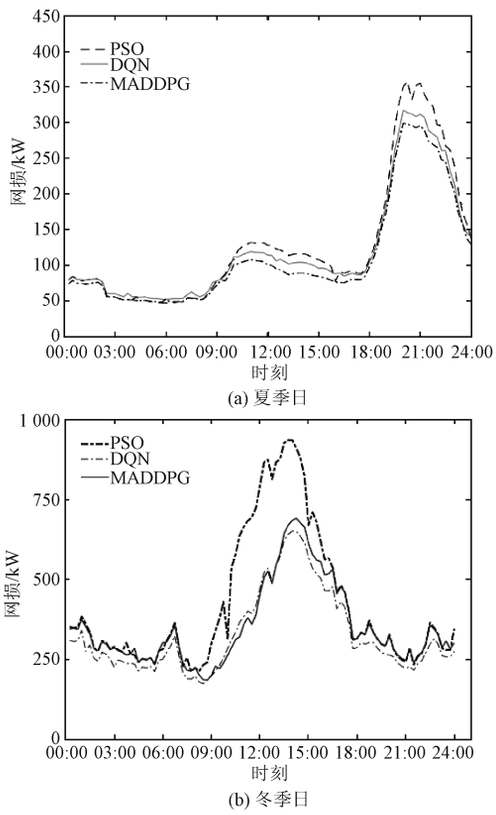


图 10 典型日下 IEEE-123 网损情况

Fig.10 Typical daily network loss of IEEE-123

4.3.3 计算性能分析

本文进行的仿真测试硬件平台包括:Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz ; 32GB RAM; GPU: NVIDIA GTX 2080 Ti ; 软件平台包括:ubuntu18.04(Linux) ; Python 3.7.6; Pytorch 1.2.0。

在线测试阶段,33 节点的配电网系统中,本文所用 MADDPG 算法的测试用例平均执行时间为 29.5 ms,能够满足电力系统实时性的要求。因此,本文提出的 MADDPG 算法在处理动作空间为高维空间时具有一定的合理性。

5 结论

本文设计了一种基于多智能体深度强化学习 MADDPG 算法的配电网无功优化算法。采用集中训练、分散执行的方式,通过环境与多个智能体之间

的交互,自适应地选择调度动作指令,实现节点电压调节和降低网损。主要结论如下。

(1)设计了多个智能体,根据不同的奖励函数来训练调节无功补偿装置的动作,通过多个智能体的协调合作,达到更好的优化效果。

(2)MADDPG 训练是需要通过与电网交互来实现的。该算法可以满足实际电网中无法进行精确潮流建模的情况。完成训练后的网络不需要预测日前的分布式电源、负荷等数据,就可以进行在线优化决策。

(3)与传统优化方法 PSO 以及强化学习 DQN 方法相比,本文使用的 MADDPG 算法可以使得网损更小、电压平抑效果更佳,对提升配网安全可靠性有更显著的效果。

参考文献(References):

[1] 陈虹(Chen Hong). 含分布式电源的配电网无功补偿优化研究(Reactive power compensation optimization of distribution network including distributed power supply)[J]. 信息技术(Information Technology), 2020, 44(11): 132-136.

[2] 周玮, 胡姝博, 孙辉, 等(Zhou Wei, Hu Shubo, Sun Hui, et al.). 考虑大规模风电并网的电力系统区间非线性经济调度研究(Interval nonlinear economic dispatch in large scale wind power integrated system)[J]. 中国电机工程学报(Proceedings of the CSEE), 2017, 37(2): 557-564.

[3] 许多红, 郭靖琪, 丁筱筠, 等(Xu Duohong, Guo Jingqi, Ding Xiaojun, et al.). 基于协同进化遗传算法的配电网风光储联合经济调度(Economic dispatch of distribution network with wind-solar-battery system based on co-evolutionary genetic algorithm)[J]. 电工电能新技术(Advanced Technology of Electrical Engineering and Energy), 2020, 39(6): 51-57.

[4] Zhang Chen, Jian Jin, Yang Linfeng, et al. Regularised primal-dual interior-point method for dynamic optimal power flow with block-angular structures [J]. IET Generation, Transmission & Distribution, 2020, 14(9): 1694-1704.

[5] 颜景斌, 夏赛, 王飞, 等(Yan Jingbin, Xia Sai, Wang Fei, et al.). 基于改进遗传算法的有源配电网故障定位分析(Analysis of fault location for active distribution network based on improved genetic algorithm)[J]. 电力系统及其自动化学报(Proceedings of the CSU-EPSA), 2019, 31(6): 107-112.

[6] Bouallaga A, Davigny A, Courtecuisse V, et al. Methodology for technical and economic assessment of electric vehicles integration in distribution grid [J]. Mathematics and Computers in Simulation, 2017, 131: 172-189.

[7] 黄俊辉, 汪惟源, 王海潜, 等(Huang Junhui, Wang Weiyuan, Wang Haiqian, et al.). 基于模拟退火遗传算法的交直流系统无功优化与电压控制研究(Study of hybrid genetic algorithm and annealing algorithm on reactive power optimization and voltage control in AC/DC transmission system)[J]. 电力系统保护与控制(Power System Protection and Control), 2016, 44(10): 37-43.

[8] Boicea V A. Distribution grid reconfiguration through simulated annealing and tabu search [A]. 2017 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE) [C]. Bucharest, Romania, 2017. 563-568.

[9] 于琳, 孙莹, 徐然, 等(Yu Lin, Sun Ying, Xu Ran, et al.). 改进粒子群优化算法及其在电网无功分区中的应用(Improved particle swarm optimization algorithm and its application in reactive power partitioning of power grid)[J]. 电力系统自动化(Automation of Electric Power System), 2017, 41(3): 89-95, 128.

[10] 周平, 孙悦, 康朋, 等(Zhou Ping, Sun Yue, Kang Peng, et al.). 光热电站的高比例可再生能源系统混合储能容量优化配置方法(Optimal hybrid energy storage capacity configuration for CSP integrated power systems with high renewable energy penetration)[J]. 电工电能新技术(Advanced Technology of Electrical Engineering and Energy), 2021, 40(3): 22-31.

[11] 任佳依, 顾伟, 王勇, 等(Ren Jiayi, Gu Wei, Wang Yong, et al.). 基于模型预测控制的主动配电网多时间尺度有功无功协调调度(Multi-time scale active and reactive power coordinated optimal dispatch in active distribution network based on model predictive control)[J]. 中国电机工程学报(Proceedings of the CSEE), 2018, 38(5): 1397-1407.

[12] 赵冬梅, 陶然, 马泰, 等(Zhao Dongmei, Tao Ran, Ma Tai, et al.). 基于多智能体深度确定策略梯度算法的有功-无功协调调度模型(Active and reactive power coordinated dispatching based on multi-agent deep deterministic policy gradient algorithm)[J]. 电工技术学报(Transactions of China Electrotechnical Society), 2021, 36(9): 1914-1925.

[13] Cao D, Hu W, Zhao J, et al. A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters [J]. IEEE Transactions on Power Systems, 2020, 35(5): 4120-4123.

[14] Wang W, Yu N, Shi J, et al. Volt-var control in power

distribution systems with deep reinforcement learning [A]. 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm) [C]. Beijing, China, 2019. 1-7.

[15] Liu H, Wu W. Two-stage deep reinforcement learning for inverter-based volt-var control in active distribution networks [J]. IEEE Transactions on Smart Grid, 2020, 12 (3): 2037-2047.

[16] 李琦, 乔颖, 张宇精 (Li Qi, Qiao Ying, Zhang Yujing). 配电网持续无功优化的深度强化学习方法 (Continuous reactive power optimization of distribution network using deep reinforcement learning) [J]. 电网技术 (Power System Technology), 2020, 44 (4): 1473-1480.

[17] 王刚军, 王承民, 李恒, 等 (Wang Gangjun, Wang Chengmin, Li Heng, et al.). 基于实测数据的配网理论网损计算方法 (Calculation method of theoretical network loss in power distribution network based on measured data) [J]. 电网技术 (Power System Technology), 2002, 26 (12): 18-20.

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

[19] 杨丰毓 (Yang Fengyu). 基于深度强化学习的电力系统无功优化策略 (Reactive power optimization strategy of power system based on deep reinforcement learning) [D]. 哈尔滨: 哈尔滨工业大学 (Harbin: Harbin Institute of Technology), 2020.

[20] 邵美阳, 吴俊勇, 石琛, 等 (Shao Meiyang, Wu Junyong, Shi Chen, et al.). 基于数据驱动和深度置信网络的配电网无功优化 (Reactive power optimization of distribution network based on data driven and deep belief network) [J]. 电网技术 (Power System Technology), 2019, 43 (6): 1874-1885.

[21] Zhang Y, Ren Z. Realtime optimal reactive power dispatch using multi-agent technique [J]. Electric Power Systems Research, 2004, 69 (3): 259-265.

[22] Buşoniu L, Babuška R, De Schutter B. Multi-agent reinforcement learning: An overview [J]. Innovations in Multi-agent Systems and Applications, 2010, 38 (2): 183-221.

[23] 龚锦霞, 刘艳敏 (Gong Jinxia, Liu Yanmin). 基于深度确定策略梯度算法的主动配电网协调优化 (Coordinated optimization of active distribution network based on deep deterministic policy gradient algorithm) [J]. 电力系统自动化 (Automation of Electric Power System), 2020, 44 (6): 113-120.

[24] 刘朝阳, 穆朝絮, 孙长银 (Liu Zhaoyang, Mu Zhaoxu, Sun Changyin). 深度强化学习算法与应用研究现状综述 (An overview on algorithms and applications of deep reinforcement learning) [J]. 智能科学与技术学报 (Chinese Journal of Intelligent Science and Technology), 2020, 2 (4): 314-326.

[25] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments [A]. Proceedings of the 31st International Conference on Neural Information Processing Systems [C]. California, USA, 2017. 6382-6393.

[26] 吕清洁, 王韶, 刘庭磊 (Lv Qingjie, Wang Shao, Liu Tinglei). 含分布式发电的配电网有功-无功综合优化 (Active/reactive power integrated optimization in distribution networks with distributed generation) [J]. 电力系统保护与控制 (Power System Protection and Control), 2012, 40 (10): 71-76, 83.

[27] Zhang Y, Wang X, Wang J, et al. Deep reinforcement learning based volt-var optimization in smart distribution systems [J]. IEEE Transactions on Smart Grid, 2020, 12 (1): 361-371.

Reactive power optimization strategy of distribution network based on multi agent deep reinforcement learning

DENG Qing-tang[1], HU Dan-er[2], CAI Tian-tian[1], LI Xiao-bo[1], XU Xian-min[2], PENG Yong-gang[2]
(1.Digital Grid Research Institute, China Southern Power Grid, Guangzhou 510663, China;
2.College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China)

Abstract: In order to overcome the problems of voltage fluctuation and network loss increase caused by random output fluctuation of photovoltaic and wind turbine equipment and load fluctuation in distribution network, it brings challenges to online reactive power optimization of distribution network. In this paper, a reactive power optimization strategy of MADDPG algorithm based on multi-agent deep reinforcement learning is designed. Compared with the traditional algorithm, this method does not need accurate power flow modeling, nor does it rely on the data prediction of day ahead load and distributed generation. The reactive power optimization problem is solved by centralized training and decentralized execution. MADDPG regards every agent as an actor. In the process of training, each actor can use a critical to train. The proposed strategy uses deep neural network to fit the action functions of switchable capacitors, voltage regulators and distributed generation inverters, and completes the training of deep neural network in the interaction process with the distribution network environment. Finally, an example is given to verify the effectiveness of MADDPG algorithm.

Key words: multi agent; deep reinforcement learning; reactive power optimization; data driven; low perception distribution network