DOI: 10.19783/j.cnki.pspc.241396

PPO 算法优化参数的微网接口变换器自抗扰控制

周雪松[1]，刘文进[1]，马幼捷[1]，陶 珑[1]，闫虎龙[2,3]，丰美丽[4]

(1. 天津市新能源电力变换传输与智能控制重点实验室(天津理工大学)，天津 300384；2. 天津瑞能电气有限公司，天津 300385；3. 天津瑞源电气有限公司，天津 300308；4. 天津安捷物联科技股份有限公司，天津 300392)

摘要：直流微电网作为新型电力系统的重要环节，因新能源的随机性和不确定性，直流微电网中负载端接口变换器的输出电压容易受到扰动影响，导致输出特性不佳。为有效消除在控制器参数保持恒定时不确定性特征对系统性能产生的不利影响，提出了一种基于近端策略优化(proximal policy optimization, PPO)算法的自抗扰控制方法。该方法利用 PPO 智能体与传统自抗扰控制系统环境进行交互，感知环境状态的变化，并依据奖励的反馈来优化控制策略。在训练过程中，智能体通过探索不同的控制动作，实现观测器参数的自适应调整，从而确保了变换器输出电压的稳定。最后，在数字仿真平台上，将 PPO-LADRC 与传统线性自抗扰控制(linear active disturbance rejection control, LADRC)、双闭环比例-积分控制方法进行了对比分析，验证了该控制策略可以显著提升系统在各种扰动下的动态性能。

关键词：直流微电网；接口变换器；深度强化学习；自抗扰控制；自适应调整

Active disturbance rejection control of microgrid interface converters using PPO algorithm for parameters optimization

ZHOU Xuesong[1], LIU Wenjin[1], MA Youjie[1], TAO Long[1], WEN Hulong[2,3], FENG Meili[4]

(1. Tianjin Key Laboratory of New Energy Power Conversion, Transmission and Intelligent Control (Tianjin University of Technology), Tianjin 300384, China; 2. Tianjin Ruineng Electric Co., Ltd., Tianjin 300385, China; 3. Tianjin Ruiyuan Electric Co., Ltd., Tianjin 300308, China; 4.Tianjin Anjie IOT Technology Co., Ltd., Tianjin 300392, China)

Abstract: As an important component of modern power systems, DC microgrids are susceptible to disturbances at the load-side interface converters due to the randomness and uncertainty of renewable energy sources, resulting in poor output characteristics. In order to effectively mitigate the adverse effects of uncertainty on system performance when the controller parameters are kept constant, this paper proposes an active disturbance rejection control method based on the proximal policy optimization (PPO) algorithm. In this method, a PPO agent interacts with the traditional active disturbance rejection control system environment to perceive changes in system states and optimizes the control strategy based on feedback from a reward. During the training process, the agent explores various control actions to adaptively tune observer parameters, thereby ensuring the stability of the converter output voltage. Finally, the proposed PPO-LADRC is compared through digital simulations with the traditional linear active disturbance rejection control (LADRC) and double-closed-loop proportional-integral control methods. The results verify that the proposed control strategy can significantly improve the dynamic performance of the system under various disturbances.

This work is supported by the Key Program of National Natural Science Foundation of China (No. U23B20142).

Key words: DC microgrid; interface converter; deep reinforcement learning; active disturbance rejection control; adaptive tuning

0 引言

在“双碳”目标的引领下，大规模分布式可再生能源正逐步占据电力供应的主导地位，迫切需要构建适应能源低碳转型、新能源占比较高的新型电力系统[1]。直流微电网成为促进可再生能源高效整合与利用的关键环节，对于保障电力系统的电能质量和安全稳定运行发挥着至关重要的作用[2]。然而，

基金项目：国家自然科学基金重点项目资助(U23B20142)

可再生能源的强不确定性等固有特性，给母线电压的稳定输出及电网向负载的平稳能量供给带来了严峻挑战，进而影响了新能源的有效消纳和利用效率[3]。在此背景下，确保 DC-DC 变换器稳定且快速的控制，对于维持直流微电网端口变换器的性能，以及系统的高效运行，具有至关重要的意义[4]。

由于 DC-DC 变换器具有非线性和时变的结构特性，所构建的模型可能会出现不准确的情况。文献[5]中比例积分(proportional-integral, PI)控制能够有效降低稳态误差。文献[6]将传统 PI 控制和模型预测控制有机结合，优化了双向 DC-DC 变换器的动态性能。文献[7]能够有效提升系统的响应速度和抗扰性。文献[8]中改进的控制算法有效抑制了变换器输出端的电压波动，并增强了系统的抗干扰能力。但文献[5-8]都需要准确的预测模型才能达到预期的控制效果。文献[9]在双向 DC-DC 变换器控制策略中引入自抗扰控制(active disturbance rejection control, ADRC)，减少预测模型的依赖性，抑制系统中电压的波动。文献[10]中使用“带宽”的思想，缩减了自抗扰控制器参数调节的数量。文献[11]对线性自抗扰控制(linear active disturbance rejection control, LADRC)结构进行改进，使电压波动的抑制效果更为显著。但文献[9-11]均需要手动调节参数，且依赖于设计人员的专业知识和经验，难以适应复杂多变的微电网系统。文献[12-14]通过智能调整 LADRC 参数，能够有效应对微电网系统中的随机性和工况不确定性。但在训练过程中容易陷入局部最优。

针对以上控制方法存在的不足，本文提出了一种近端策略优化(proximal policy optimization, PPO)算法[15]与线性自抗扰控制相结合的控制方法。LADRC 减少对复杂预测模型的依赖，同时能够有效应对扰动。PPO 通过限制策略变化和优化损失函数，促进了更全面的探索，减小算法陷入局部最优的风险。此控制方法将 PPO 算法与 LADRC 结合起来，使得 PPO 智能体与 LADRC 控制的变换器环境进行动态交互，进而学习 LADRC 的观测器参数优化策略，实现观测器参数的优化，使其能够更好地适应新能源不确定变化，并实现更好的稳压控制。

1　系统结构与传统自抗扰控制

1.1　系统结构

直流微电网系统结构如图 1 所示。由光伏、储能装置、交流电网以及负载组成。

1.2　数学模型

本文的研究对象是直流母线与负载之间的降压变换器。拓扑结构如图 2 所示。

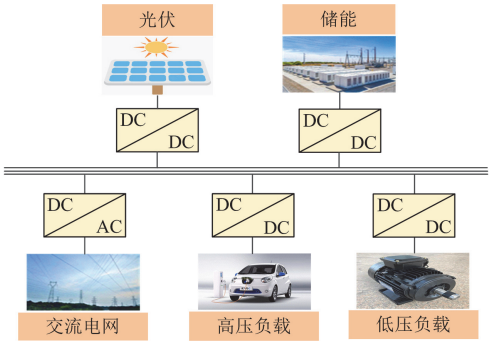


图 1　直流微电网系统结构图

Fig. 1 Structure of DC microgrid system

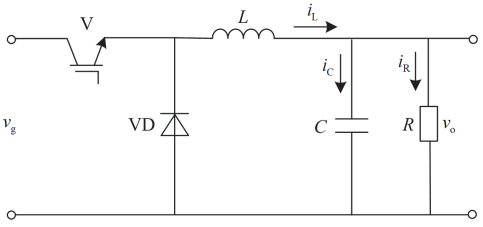


图 2　Buck 模式等效电路图

Fig. 2 Buck mode equivalent circuit diagram

图 2 中：v_g 为直流母线侧电压；v_o 为负载侧电压；L 为电感；C 为电容；R 为负载；V 为开关管；VD 为续流二极管；i_C 为电容电流；i_L 为电感电流；i_R 为电阻电流。一个周期 T_S 内，V 的开通与关断可以将电路的工作状态分为 $[0, DT_S)$ 和 $[DT_S, T_S]$ 两部分，即工作状态 1 和工作状态 2，其中 D 为占空比。

在工作状态 1，即 $[0, DT_S)$ 时段内，V 导通，VD 关断，电感电流 i_L 与电容电压 v_o 的状态方程为

$$\begin{bmatrix} \dot{i}_L \\ \dot{v}_o \end{bmatrix} = \begin{bmatrix} 0 & -\dfrac{1}{L} \\ \dfrac{1}{C} & -\dfrac{1}{RC} \end{bmatrix} \begin{bmatrix} i_L \\ v_o \end{bmatrix} + \begin{bmatrix} \dfrac{1}{L} \\ 0 \end{bmatrix} v_g \tag{1}$$

在工作状态 2，即 $[DT_S, T_S]$ 时段内，V 关断，VD 导通，电感电流 i_L 与电容电压 v_o 的状态方程为

$$\begin{bmatrix} \dot{i}_L \\ \dot{v}_o \end{bmatrix} = \begin{bmatrix} 0 & -\dfrac{1}{L} \\ \dfrac{1}{C} & -\dfrac{1}{RC} \end{bmatrix} \begin{bmatrix} i_L \\ v_o \end{bmatrix} \tag{2}$$

利用状态空间平均法求出两种工作状态下的平均状态方程[16]为

$$\begin{bmatrix} \dot{i}_L \\ \dot{v}_o \end{bmatrix} = \begin{bmatrix} 0 & -\dfrac{1}{L} \\ \dfrac{1}{C} & -\dfrac{1}{RC} \end{bmatrix} \begin{bmatrix} i_L \\ v_o \end{bmatrix} + \begin{bmatrix} \dfrac{D}{L} \\ 0 \end{bmatrix} v_g \tag{3}$$

对式(3)进行拉普拉斯变换，并利用小信号分析法[17]建立工作在 Buck 模式下的小信号模型，如式(4)—式(6)所示。

以占空比 $D(s)$ 为输入、$v_o(s)$ 为输出的传递函数为

$$G_{vd}(s)=\frac{v_o(s)}{D(s)}=\frac{v_g}{LCs^2+\frac{L}{R}s+1} \tag{4}$$

以占空比 $D(s)$ 为输入、$i_L(s)$ 为输出的传递函数为

$$G_{id}(s)=\frac{i_L(s)}{D(s)}=\frac{v_g(RCs+1)}{RLCs^2+Ls+R} \tag{5}$$

由式(4)和式(5)可推出以 $i_L(s)$ 为输入、$v_o(s)$ 为输出的传递函数 $G_{vi}(s)$ 为

$$G_{vi}(s)=\frac{v_o(s)}{i_L(s)}=\frac{R}{RCs+1} \tag{6}$$

1.3 传统二阶线性自抗扰控制策略

由式(4)可知，被控对象为二阶系统，因此选用二阶线性自抗扰控制器，控制框图如图 3 所示。

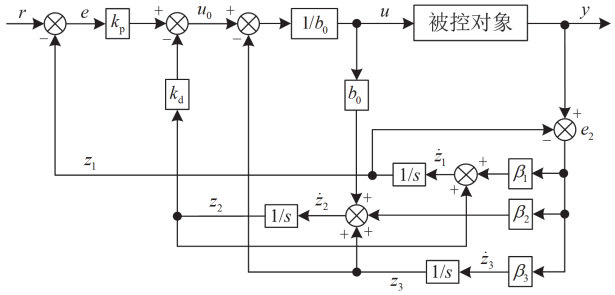


图 3 传统二阶线性自抗扰控制框图

Fig. 3 Block diagram of conventional second-order linear self-immunity control

图 3 中：r 为输出电压参考值；y 为输出电压；b_0 为控制量增益；z_1 为输出电压 y 的观测值；z_2 为输出电压 y 一阶导数的观测值；z_3 为总扰动 f 的观测值；u 为控制量；β_1、β_2、β_3 为线性扩张状态观测器增益；k_p、k_d 为控制器增益。

1.3.1 线性扩张状态观测器

二阶系统被控对象表示为

$$\ddot{y}=g(y,\dot{y},\omega,t)+bu \tag{7}$$

式中：b 为控制量增益；$g(y,\dot{y},\omega,t)$ 为含系统输出、外部扰动以及时变部分的扰动函数。在控制系统中无法精确估计参数 b 的真实值，但 b_0 可作为 b 已知的部分[18]，故将 $(b-b_0)$ 当作新的扰动项，定义 $f=g(y,\dot{y},\omega,t)+(b-b_0)u$ 为二阶系统所受到的总扰动，则式(7)等价为

$$\begin{aligned}\ddot{y}&=g(y,\dot{y},\omega,t)+(b-b_0)u+b_0u=\\&f(y,\dot{y},\omega,t)+b_0u\end{aligned} \tag{8}$$

选取状态变量：$y=x_1$，$\dot{y}=x_2$，总扰动 f 定义为二阶系统的一个扩张状态，令 $f=x_3$。假设 f 可导，定义 $h=\dot{x}_3$，则式(8)可用状态空间表示为

$$\begin{bmatrix}\dot{x}_1\\\dot{x}_2\\\dot{x}_3\end{bmatrix}=\begin{bmatrix}0&1&0\\0&0&1\\0&0&0\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}+\begin{bmatrix}0\\b_0\\0\end{bmatrix}u+\begin{bmatrix}0\\0\\1\end{bmatrix}h \tag{9}$$

由于 $\dot{f}$ 未知且可通过校正项估计出来[19]，因而式(10)中略去了 $\dot{f}$。对应的连续线性扩张状态观测器表达式为

$$\begin{cases}\dot{z}_1=z_2+\beta_1(y-z_1)\\\dot{z}_2=z_3+\beta_2(y-z_1)+b_0u\\\dot{z}_3=\beta_3(y-z_1)\end{cases} \tag{10}$$

由极点配置法可得观测器增益为

$$\beta_1=3\omega_o,\beta_2=3\omega_o^2,\beta_3=\omega_o^3 \tag{11}$$

1.3.2 线性状态误差反馈控制律

自抗扰控制的核心在于如何实时估计扰动 f 并加以消除[20]，使式(7)变为式(12)所示的线性积分器串联标准型，从而使控制变得简单。

$$\ddot{y}=u_0 \tag{12}$$

通过线性扩张状态观测器(linear extended state observer, LESO)，原对象中扩张出的扰动状态变量 x_3 被状态变量 z_3 跟踪，通过减弱 x_3 对系统的影响，可将原对象简化成式(12)的形式，由式(8)、式(12)可得控制量为

$$u=\frac{u_0-z_3}{b_0} \tag{13}$$

系统线性状态误差的反馈控制率 u_0 设计为

$$u_0=k_p(r-z_1)-k_dz_2 \tag{14}$$

式中：r 为给定值。

式(12)与式(14)结合，并进行拉普拉斯变换得到系统的闭环传递函数 $G_c(s)$ 为

$$G_c(s)=\frac{k_p}{s^2+k_ds+k_p} \tag{15}$$

采用带宽参数化的方法整定控制器增益 k_p、k_d，定义 ω_c 为控制器带宽，计算得

$$k_p=\omega_c^2,k_d=2\omega_c \tag{16}$$

2　PPO-LADRC 设计

2.1 LADRC 系统的深度强化学习训练框架

PPO-LADRC 控制策略是将 Buck 变换器的 LADRC 系统视为环境，其控制效果作为奖励评估的主要标准。PPO 智能体首先根据当前环境状态做出决策，并通过与环境交互来执行相应的动作[21]。随后，环境根据智能体输出的动作结果更新系统状态，并计算相应的奖励值。LADRC 系统的深度强化学习训练框架如图 4 所示。

强化学习的目标任务通常被建模成马尔可夫决策过程(Markov decision process, MDP)[22]，典型的 MDP 可由五元组 $\{S,A,R,P,\gamma\}$ 表示，其中 S 表示所有状态的集合，称为状态空间；A 表示智能体在环境中可以采取的动作集合，称为动作空间；R 为决策带来的回报奖励；P 为状态转移函数，即在状态 s_t 下采取动作 a_t 后，状态转移至下一状态 s_{t+1} 的概率；γ 为折扣因子，决定了智能体对未来奖励的重视程度。

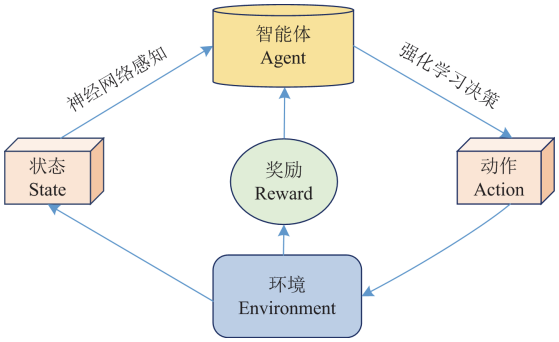


图 4 LADRC 系统的深度强化学习训练框架

Fig. 4 Deep reinforcement learning training framework for LADRC systems

2.2 深度强化学习 PPO 算法

PPO 算法旨在解决深度强化学习中的策略优化问题，通过对当前策略和参考策略之间的距离进行限制来保证算法在学习过程中的稳定性和收敛性[23]。PPO 算法本质上是基于 AC(Actor-Critic)架构，通过策略网络(Actor)和价值网络(Critic)实现的，该算法利用 Actor 网络来学习行为策略，利用 Critic 网络估计值函数。Actor 和 Critic 拟合的行为策略和值函数分别为

$$\pi_\theta = P(a_t|s_t) \tag{17}$$

$$V(s_t) = E\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k} \mid s_t\right] \tag{18}$$

式中：π_θ 为策略；θ 为 Actor 网络中可更新的参数；t 为时间步；k 为步数偏移量；s_t、a_t 和 r_t 分别为第 t 时间步下的状态、采取的动作以及得到的奖励；$P(a_t \mid s_t)$ 为当前策略在状态 s_t 下采取动作 a_t 的概率；$V(s_t)$ 为在状态 s_t 下的估计值函数；E 表示期望。

利用状态–动作的价值函数 $Q(s_t,a_t)$ 来评价策略 $\pi_\theta(a_t \mid s_t)$ 的优劣性。状态–动作的价值函数 $Q(s_t,a_t)$ 是从当前状态 s_t 执行动作 a_t 后获得的奖励值之和的期望。

$$Q(s_t,a_t) = E\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k} \mid s_t, a_t\right] \tag{19}$$

通过基于优势函数的 Actor-Critic 方法进行回报值估计，将会产生方差较小而偏差较大的问题。PPO 算法采用泛化优势估计(generalized advantage estima-tion, GAE)权衡方差和偏差的问题[24]。GAE 的表达式 $\hat{A}_t$ 为

$$\hat{A}_t = \sum_{k=0}^{\infty}(\gamma\lambda)^k\delta_{t+k} \tag{20}$$

式中：λ 和 γ 为 GAE 方法中的重要参数，可以在训练过程中平衡方差和偏差，从而更有效地估计优势函数，进而提高训练效果和性能。式(21)为时序差分误差 δ_t，可表示为

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{21}$$

Actor 网络的更新公式为

$$L(\theta) = E_t[\min(k(\theta,\theta')\hat{A}_t, \mathrm{clip}(k(\theta,\theta'),1-\varepsilon,1+\varepsilon)\hat{A}_t)] \tag{22}$$

其中

$$k(\theta,\theta') = \frac{\pi_\theta(s_t,a_t)}{\pi_{\theta'}(s_t,a_t)} \tag{23}$$

式中：$L(\theta)$ 为 Actor 网络的损失函数，为 Actor 网络更新的公式；$\pi_{\theta'}$ 为旧策略；ε 为限制新旧策略之间差异的裁剪因子；$k(\theta,\theta')$ 为新策略与旧策略之间的差异；$\mathrm{clip}(k(\theta,\theta'),1-\varepsilon,1+\varepsilon)$ 为截取函数，用于截取 $k(\theta,\theta')$ 在 $1-\varepsilon$ 和 $1+\varepsilon$ 之间值。图 5 为 PPO 截断函数 L^{Clip} 的示意图。

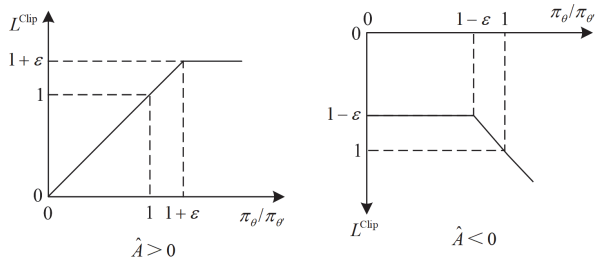


图 5 PPO 截断函数

Fig. 5 PPO truncation function

如图 5 所示，当优势函数的估计为负值时，代表当前的策略是消极的，则降低其出现的概率(以 $1-\varepsilon$ 为界限)。当优势函数的估计为正值时，表示当前策略是积极的，则增加其概率(以 $1+\varepsilon$ 为界限)。通过限制策略变化的范围，确保新策略不会偏离旧策略太远，从而稳定训练过程并提高收敛速度。

为增强 PPO 算法策略网络的探索性，在式(22)的基础上增加熵，通过熵来衡量策略的不确定性，熵在 PPO 算法中起到平衡探索和利用的作用，帮助算法更有效地优化策略，降低收敛到局部最优的风险。熵 $H(\pi(a_t \mid s_t))$ 的定义为

$$H(\pi(a_t \mid s_t)) = -\sum_{a_t} \pi_\theta(a_t \mid s_t) \log \pi_\theta(a_t \mid s_t) \tag{24}$$

Actor 网络的更新公式为

$$L'(\theta) = L(\theta) + \beta H(\pi(a_t \mid s_t)) \tag{25}$$

式中：β 为熵的系数，用于调节策略优化中熵正则项的权重。

Critic 网络的更新公式 $L(\mu)$ 为

$$L(\mu) = E_t[(r_t + \gamma V_\mu(s_{t+1}) - V_\mu(s_t))^2] \tag{26}$$

2.3 基于 PPO 算法的 LADRC 自整定流程

2.3.1 状态、动作空间和奖励函数

1) 状态空间 S 选取

状态空间主要是用来表示环境状态，以提供有效的环境状态信息，使智能体做出最优的决策。状态包括以下 3 部分：变换器的输出电压 y 与输入参考电压 r 之间的误差 e_1；变换器的输出电压 y 与观测值 z_1 之间的误差 e_2；变换器的输出电压 y。

2) 动作空间 A 选取

动作空间的选取关系到确定 LADRC 可以采取的各种可能的动作，即调节参数的范围。动作空间选取的合理性将直接影响到控制器的性能和收敛速度，因此在设计过程中需要考虑动作空间的设定，以确保算法能够有效地学习和优化控制器的参数。动作选取为 LADRC 的观测器带宽 ω_o [25]。对于线性自抗扰控制器，稳定性分析涉及到评估控制系统的闭环稳定性，确定系统的输出是否收敛到期望的状态，以及确定调节参数的范围，加快训练的收敛速度。

根据式(10)和式(11)可得

$$\begin{cases} z_1 = \dfrac{\omega_o^3 + 3\omega_o^2 s + 3\omega_o s^2}{(s+\omega_o)^3} y + \dfrac{b_0 s}{(s+\omega_o)^3} u \\ z_2 = \dfrac{\omega_o^3 s + 3\omega_o^2 s^2}{(s+\omega_o)^3} y + \dfrac{b_0 (s^2 + 3\omega_o s)}{(s+\omega_o)^3} u \\ z_3 = \dfrac{\omega_o^3 s^2}{(s+\omega_o)^3} y - \dfrac{b_0 \omega_o^3}{(s+\omega_o)^3} u \end{cases} \tag{27}$$

根据式(13)、式(14)和式(16)可得

$$u = \frac{\omega_c^2 (r - z_1) - 2\omega_c z_2 - z_3}{b_0} \tag{28}$$

式(27)代入式(28)可得到 u、r、y 之间的 s 域函数关系为

$$u = \frac{1}{b_0} G_1(s) G_2(s) \tag{29}$$

其中

$$G_1(s) = \omega_c^2 r - \frac{(3\omega_o\omega_c^2 + 6\omega_c\omega_o^2 + \omega_o^3)s^2}{(s+\omega_o)^3} y - \frac{(3\omega_o^2\omega_c^2 + 2\omega_c\omega_o^3)s + \omega_o^3\omega_c^2}{(s+\omega_o)^3} y \tag{30}$$

$$G_2(s) = \frac{(s+\omega_o)^3}{(s+\omega_o)^3 - \omega_o^3 + 2\omega_c s^2 + (\omega_c^2 + 6\omega_o\omega_c)s} \tag{31}$$

根据式(29)可得系统结构如图 6 所示。

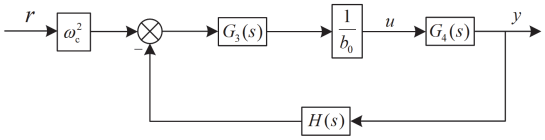


图 6 LADRC 等效控制框架

Fig. 6 LADRC equivalent control framework

$$G_3(s) = \frac{(s+\omega_o)^3}{(s+\omega_o)^3 + 2\omega_c s^2 + (\omega_c^2 + 6\omega_c\omega_o)s - \omega_o^3} \tag{32}$$

$$H(s) = (3\omega_c^2\omega_o + 6\omega_c\omega_o^2 + \omega_o^3)s^2/(s+\omega_o)^3 + [(3\omega_c^2\omega_o^2 + 2\omega_c\omega_o^3)s + \omega_c^2\omega_o^3]/(s+\omega_o)^3 \tag{33}$$

求得系统的闭环传递函数 $G_5(s)$ 为

$$G_5(s) = \frac{\omega_c^2 G_3(s) G_4(s)}{b_0 + G_3(s) G_4(s) H(s)} \tag{34}$$

式中：$G_4(s)$ 为 1.2 节中的 Buck 变换器的传递函数。

由式(34)可求得系统的特征方程为

$$a_0 s^5 + a_1 s^4 + a_2 s^3 + a_3 s^2 + a_4 s + a_5 = 0 \tag{35}$$

各项参数分别为

$$\begin{cases} a_0 = b_0 LC \\ a_1 = b_0 \left[\dfrac{L}{R} + LC(3\omega_o + 2\omega_c) \right] \\ a_2 = b_0 \left[1 + \dfrac{L}{R}(3\omega_o + 2\omega_c) + LC(3\omega_o^2 + \omega_c^2 + 6\omega_c\omega_o) \right] \\ a_3 = b_0 \left[(3\omega_o + 2\omega_c) + \dfrac{L}{R}(3\omega_o^2 + \omega_c^2 + 6\omega_c\omega_o) \right] + v_g(3\omega_o\omega_c^2 + \omega_o^3 + 6\omega_c\omega_o^2) \\ a_4 = b_0(3\omega_o^2 + \omega_c^2 + 6\omega_c\omega_o) + v_g(3\omega_c^2\omega_o^2 + 2\omega_c\omega_o^3) \\ a_5 = v_g\omega_c^2\omega_o^3 \end{cases} \tag{36}$$

根据 Lienard-Chipard 判据可知，若要系统保持稳定运行需要满足：

(1) $a_i > 0, (i = 0,1,2,3,4,5)$

(2) 偶数阶的赫尔维茨行列式：$\Delta_2 > 0$、$\Delta_4 > 0$

由于 ω_o 与 ω_c 存在 k 倍关系，$k \in (2,10)$[10]。基于上述两个条件以及多次仿真调试确定 ω_o 的范围为 $(1500,+\infty)$，缩小了参数的取值范围，加快了深度强化学习算法训练的收敛速度。

3) 奖励 R

奖励的设计会影响 PPO 智能体学习到的行为和策略。通过奖励，智能体可以从环境中获得奖励或惩罚，并据此调整其行为。合理的奖励可以加速学习过程，使 PPO 智能体更快地找到有效的策略，使得环境中的变换器输出电压 y 达到目标状态。奖励表示为

$$R = \begin{cases} -10|e_1|, |e_1| < 0.2 \\ -20|e_1|, 0.2 \leqslant |e_1| \leqslant 2 \\ -50|e_1|, 2 < |e_1| \leqslant 10 \\ -70|e_1|, |e_1| > 10 \end{cases} \tag{37}$$

2.3.2 PPO-LADRC 控制策略设计

图 7 中 PPO-LADRC 控制策略具体是将 PPO 智能体与 LADRC 结合应用于 Buck 变换器的控制。PPO 智能体与环境交互，感知变换器输出电压 y 的状态，以及获得奖励的反馈，在学习过程中朝着正确的方向优化控制策略。然后将训练好的 PPO 智能体的最优策略集成到 LADRC 控制框架中。LADRC 控制器根据智能体输出的观测器带宽生成 PWM 信号，控制变换器的开关器件，从而实现所需的输出电压 y。

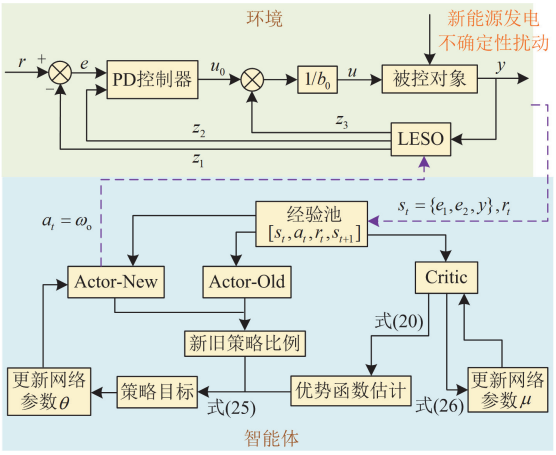


图 7 PPO-LADRC 控制策略

Fig. 7 PPO-LADRC control strategy

图 7 中，Actor-New 为新策略网络，Actor-Old 为旧策略网络。$a_t = \omega_o$ 表示动作输出为观测器带宽，$s_t = \{e_1, e_2, y\}$ 为选取的状态集合，r_t 为奖励。

在训练时，随机初始化 Actor 网络的参数 θ 与 Critic 网络的参数 μ，旧策略的参数 θ' 从新策略的参数 θ 复制过来。PPO 智能体与 LADRC 变换器模型环境展开多回合交互，在每一个回合次数内，Actor 根据当前的状态 s_t 做出相应的动作 a_t，然后得到一个奖励 r_t 并将状态转移到下一状态 s_{t+1}，上述过程进行多次交互。Actor 网络的参数 θ 通过式(25)来更新。Critic 网络的参数 μ 可以通过式(26)更新。

2.3.3 算法训练过程

在进行训练时，PPO 算法中裁剪因子 ε、折扣因子 γ 和学习率 α 等参数对算法收敛起关键作用。ε 控制每次更新的幅度。调整 γ 可以控制智能体对未来奖励的重视程度。α 用于控制模型参数更新的时间步长大小。PPO 算法参数如表 1 所示。

表 1 PPO 算法参数

Table 1 Parameters of the PPO algorithm

超参数	值
Actor 网络学习率	0.01
Critic 网络学习率	0.01
折扣因子	0.99
回合次数	700
裁剪因子	0.2
GAE 系数	0.95
熵系数	0.02

回合奖励表示在一个完整的回合(轮数)中，智能体所获得的累积奖励。平均奖励指的是智能体在多个回合中获得奖励的平均值。训练过程中，奖励的变化趋势如图 8 所示。

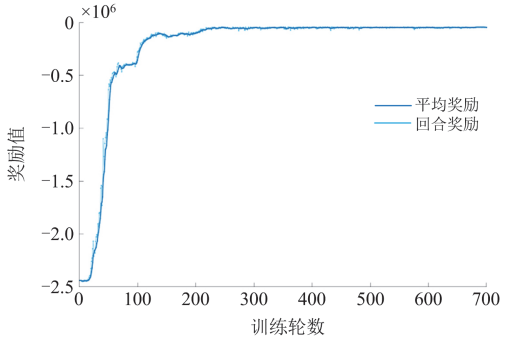


图 8 训练奖励变化曲线

Fig. 8 Training reward change curve

3 仿真验证及对比分析

为验证本文所提控制策略的正确性与有效性，

在数字仿真平台上搭建 Buck 变换器模型，并进行仿真分析。在不同的工况下，对比分析双闭环 PI、传统 LADRC 与 PPO-LADRC 这 3 种控制策略下输出电压的波形，以评估它们的性能和效果。系统参数和控制器参数分别如表 2、表 3 所示。

表 2 系统参数

Table 2 System parameters

参数类型	数值
直流母线电压/V	550
参考电压/V	440
开关频率/kHz	100
电感/μH	120
电容/μF	300
电阻/Ω	5

表 3 控制器参数

Table 3 Controller parameters

控制策略	控制器参数
双闭环 PI	电压外环：$K_{Pv}=0.9931, K_{Iv}=1136$
	电流内环：$K_{Pi}=0.0066, K_{Ii}=35$
LADRC	$K_p=7.225\times10^7, K_d=1.7\times10^4$
	$\beta_1=2.22\times10^5, \beta_2=1.6428\times10^{10}, \beta_3=4.052\,24\times10^{14}$

1) 跟踪性能对比

考虑参考电压突变对不同控制策略下输出电压的跟踪效果影响。在 0.01 s 将参考电压由 440 V 切换到 400 V。跟踪性能对比如图 9 所示，仿真结果性能指标见表 4。

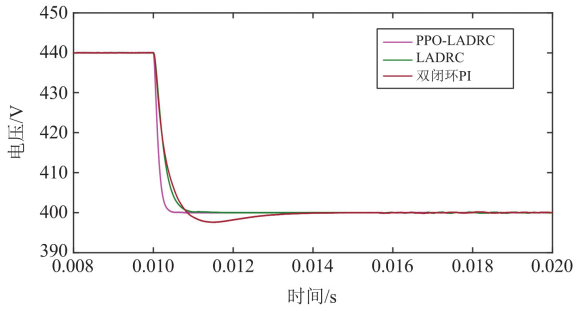


图 9 跟踪性能对比图

Fig. 9 Comparison of tracking performance

表 4 跟踪性能对比

Table 4 Tracking performance comparison

性能指标	双闭环 PI	LADRC	PPO-LADRC
调节时间/ms	3.6	1.47	0.6
超调量/%	–1.1	0	0

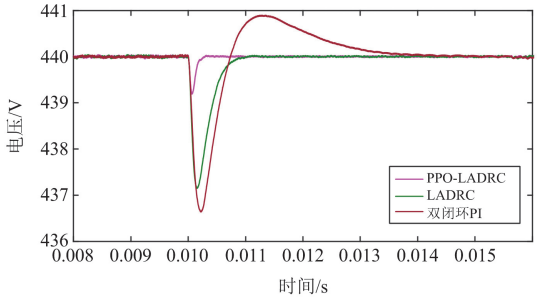
由于 PPO-LADRC 控制策略引入 PPO 算法对 LADRC 的观测器参数进行优化调整，因此，在调节过程中，相比于双闭环 PI 和传统 LADRC，调节时间最短且无超调。

2) 抗扰性分析

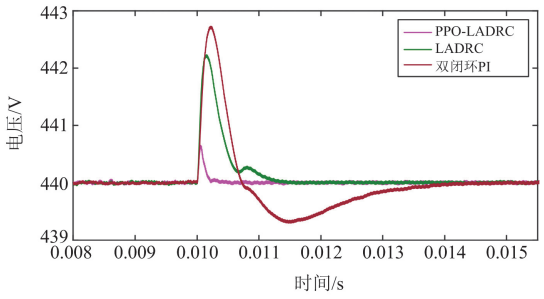
工况一：直流母线电压变化

新能源的输出功率具有较强的波动性，可能导致直流母线电压出现剧烈波动。为模拟这一工况，分别给直流母线电压施加突减 10%和突增 10%的扰动。在双闭环 PI、传统 LADRC 与 PPO-LADRC 3 种不同控制策略下，输出电压动态波形如图 10 所示。调节时间和超调量的指标对比如图 11 所示。

由图 11 可知，针对直流母线电压的变化，对比不同控制策略，在母线电压突减 10%时，PPO-LADR 相比双闭环 PI 和传统 LADRC 分别减少 0.55%、0.43%的超调量。此外，在 PPO-LADRC 控制下，调节时间也明显优于其他两种控制策略。在母线电压



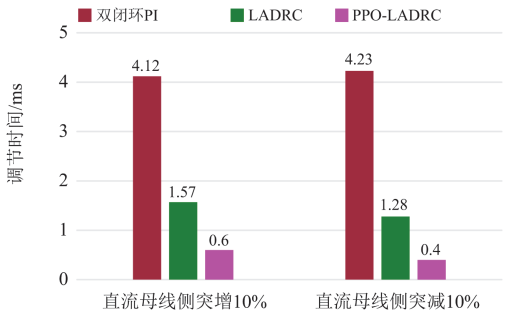
(a) 直流母线电压突减10%



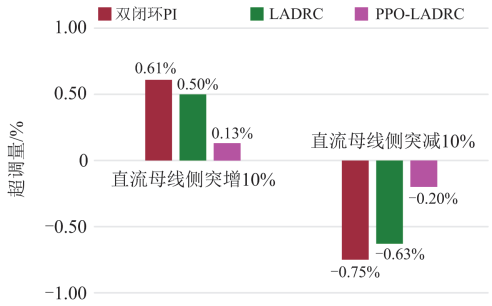
(b) 直流母线电压突增10%

图 10 直流母线侧电压扰动

Fig. 10 DC bus-side voltage disturbance



(a) 调节时间对比图



(b) 超调量对比图

图 11　直流母线侧电压扰动指标对比图

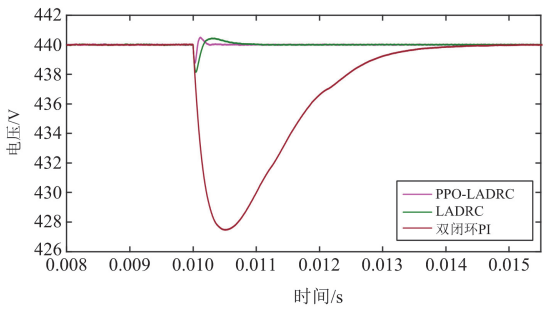
Fig. 11 Comparison of DC bus-side voltage disturbance indicators

突增 10%时，PPO-LADRC 的电压超调量分别较双闭环 PI、LADRC 减少了 0.48%、0.37%。并且在两种扰动下 PPO-LADRC 的调节时间均最短。分析表明，相较于双闭环 PI 和传统 LADRC，PPO-LADRC 有更好的扰动抑制效果。

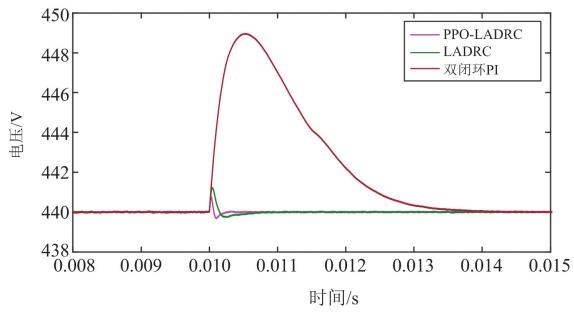
工况二：负载突变

在负载扰动的工况下，分析输出电压动态变化波形，在 0.01 s 处施加负载突减 20%和负载突增 20%的扰动，输出电压动态波形如图 12 所示。调节时间和超调量的指标对比如图 13 所示。

根据图 13 的数据显示，在负载扰动抑制方面，当负载突减 20%时，PPO-LADRC 的表现优于传统



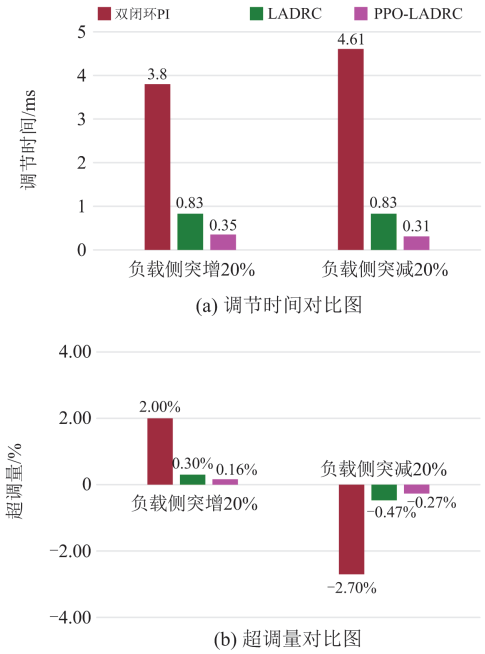
(a) 负载突减20%



(b) 负载突增20%

图 12　负载侧扰动下输出电压波形

Fig. 12 Output voltage waveform of load side disturbance



(a) 调节时间对比图

(b) 超调量对比图

图 13　负载侧扰动指标对比图

Fig. 13 Comparison of load-side disturbance metrics

LADRC 和双闭环 PI，展现出更佳的扰动抑制能力。具体而言，PPO-LADRC 相比于双闭环 PI、传统 LADRC，跌落量分别减少 2.43%、0.2%。当负载突增 20%时，相较于双闭环 PI 和传统 LADRC，PPO-LADRC 的超调量减少了 1.84%、0.14%。此外，在调节时间的性能指标上，PPO-LADRC 也优于双闭环 PI 和传统 LADRC。

4　结论

优化 DC-DC 变换器的控制方法是确保负载端接口变换器在复杂工况下能量平稳输出的核心技术手段。本文提出了一种基于深度强化学习 PPO 算法的 LADRC 观测器参数自整定控制策略，即 PPO-LADRC，具有以下特点。

1) PPO 算法通过限制策略更新步幅和促进全面探索优化了控制策略，同时结合 LADRC 的扰动估计与补偿能力，进一步提升了控制性能。

2) PPO 算法通过与 LADRC 控制的变换器环境进行交互，自动优化 LADRC 观测器的参数。

仿真结果显示，与双闭环 PI、传统 LADRC 相比，PPO 优化的 LADRC 能够在各种工况下实现更好的稳压效果。PPO 算法优化参数的微网接口变换器自抗扰控制策略不仅提升了系统的控制速度和抗扰性，还提高了其应对复杂动态环境的能力。该方法为未来微网系统的智能控制和优化提供了新的思

路和方法。

参考文献

[1] 郭庆来, 兰健, 周艳真, 等. 基于混合智能的新型电力系统运行方式分析决策架构及其关键技术[J]. 中国电力, 2023, 56(9): 1-13.
GUO Qinglai, LAN Jian, ZHOU Yanzhen, et al. Architecture and key technologies of hybrid-intelligence-based decision-making of operation modes for new type power systems[J]. Electric Power, 2023, 56(9): 1-13.

[2] 杨子龙, 宋振浩, 潘静, 等. 分布式光伏/储能系统多运行模式协调控制策略[J]. 中国电机工程学报, 2019, 39(8): 2213-2220, 4.
YANG Zilong, SONG Zhenhao, PAN Jing, et al. Multi-mode coordinated control strategy of distributed PV and energy storage system[J]. Proceedings of the CSEE, 2019, 39(8): 2213-2220, 4.

[3] 周雪松, 王博, 马幼捷, 等. 含二阶扰动补偿的交错并联变换器自抗扰控制[J]. 电机与控制学报, 2023, 27(12): 159-170.
ZHOU Xuesong, WANG Bo, MA Youjie, et al. Active disturbance rejection control of interleaving parallel converter with second-order disturbance compensation[J]. Electric Machines and Control, 2023, 27(12): 159-170.

[4] 周雪松, 王馨悦, 马幼捷, 等. 直流微网变增益专家自抗扰控制[J]. 电力系统保护与控制, 2023, 51(18): 70-80.
ZHOU Xuesong, WANG Xinyue, MA Youjie, et al. Expert system-changeable gain ADRC for a DC microgrid[J]. Power System Protection and Control, 2023, 51(18): 70-80.

[5] 吴忠, 刘朝辉. 基于电流模式的 DC/DC 升压变换器非线性 PI 控制[J]. 中国电机工程学报, 2011, 31(33): 31-36.
WU Zhong, LIU Zhaohui. Nonlinear PI control of DC/DC boost power converters based on current mode[J]. Proceedings of the CSEE, 2011, 31(33): 31-36.

[6] 杨惠, 晁凯悦, 孙向东, 等. 基于开关序列的光伏储能双向 DC-DC 变换器预测电流控制方法[J]. 电网技术, 2019, 43(1): 300-307.
YANG Hui, CHAO Kaiyue, SUN Xiangdong, et al. Predictive current control method of photovoltaic energy storage for bidirectional DC-DC converter based on switching sequence[J]. Power System Technology, 2019, 43(1): 300-307.

[7] 杨翔宇, 肖先勇, 马俊鹏, 等. 基于电感电流反馈的双向 DC-DC 变换器下垂控制[J]. 中国电机工程学报, 2020, 40(8): 2638-2647.
YANG Xiangyu, XIAO Xianyong, MA Junpeng, et al. Droop control of bi-directional DC-DC converters based on inductive current feedback[J]. Proceedings of the CSEE, 2020, 40(8): 2638-2647.

[8] 杨惠, 晁凯悦, 孙向东, 等. 基于矢量作用时间的双向 DC-DC 变换器预测电流控制方法[J]. 电工技术学报, 2020, 35(增刊 1): 70-80.
YANG Hui, CHAO Kaiyue, SUN Xiangdong, et al. Predictive current control method for bidirectional DC-DC converter based on optimal operating time of vector[J]. Transactions of China Electrotechnical Society, 2020, 35(S1): 70-80.

[9] 杨惠, 骆姗, 孙向东, 等. 光伏储能双向 DC-DC 变换器的自抗扰控制方法研究[J]. 太阳能学报, 2018, 39(5): 1342-1350.
YANG Hui, LUO Shan, SUN Xiangdong, et al. Research on ADRC method for bidirectional DC-DC converter of solar energy storage system[J]. Acta Energiae Solaris Sinica, 2018, 39(5): 1342-1350.

[10] GAO Zhiqiang. Scaling and bandwidth-parameterization based controller tuning[C] // Proceedings of the 2003 American Control Conference, June 4-6, 2003, Denver, CO, USA: 4989-4996.

[11] 马幼捷, 杨清, 周雪松, 等. 基于级联型扩张状态观测器的直流微电网低压负载接口变换器自抗扰稳压研究[J]. 电力系统保护与控制, 2024, 52(4): 121-132.
MA Youjie, YANG Qing, ZHOU Xuesong, et al. Active disturbance rejection and voltage regulation of a DC microgrid low voltage load interface converter based on a cascaded extended state observer[J]. Power System Protection and Control, 2024, 52(4): 121-132.

[12] 张世欣, 皇金锋, 杨艺. 基于平坦理论的直流微电网双向 DC-DC 变换器改进滑模自抗扰控制[J]. 电力系统保护与控制, 2023, 51(5): 107-116.
ZHANG Shixin, HUANG Jinfeng, YANG Yi. Improved sliding mode and active disturbance rejection control based on flatness theory for a bi-directional DC-DC converter in a DC microgrid[J]. Power System Protection and Control, 2023, 51(5): 107-116.

[13] 李浩东, 肖伸平, 余锦. 基于自适应 PSO 的微电网双向 DC-DC 变换器前馈自抗扰控制[J]. 电机与控制应用, 2023, 50(2): 73-81.
LI Haodong, XIAO Shenping, YU Jin. Feedforward active disturbance rejection control of microgrid bidirectional DC-DC converter based on adaptive PSO[J]. Electric Machines & Control Application, 2023, 50(2): 73-81.

[14] 周雪松, 葛建鹏, 马幼捷. 二阶 LADRC 在风电并网逆变器网侧直流母线电压控制中的运用[J]. 电测与仪表, 2024, 61(3): 182-191.
ZHOU Xuesong, GE Jianpeng, MA Youjie. Application

of second-order LADRC in grid side DC bus voltage control of wind power grid-connected inverter[J]. Electrical Measurement & Instrumentation, 2024, 61(3): 182-191.

[15] GU Yang, CHENG Yuhu, CHEN C L P, et al. Proximal policy optimization with policy feedback[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(7): 4600-4610.

[16] 栾思平, 苏适, 杨洲, 等. 适应于直流新能源/储能接入的三电平 Buck-Boost 变换器建模及控制器设计[J]. 太阳能学报, 2022, 43(4): 56-65.
LUAN Siping, SU Shi, YANG Zhou, et al. Modeling and controller design of three-level Buck-Boost converter adapted to DC new energy and energy storage access[J]. Acta Energiae Solaris Sinica, 2022, 43(4): 56-65.

[17] 赵峰, 甘延奇, 陈小强, 等. 基于频域分析的双有源桥串联谐振变换器的设计与闭环控制[J]. 高电压技术, 2022, 48(11): 4557-4567.
ZHAO Feng, GAN Yanqi, CHEN Xiaoqiang, et al. Design of dual active bridge series resonant converter based on frequency domain analysis and closed-loop control[J]. High Voltage Engineering, 2022, 48(11): 4557-4567.

[18] 韩京清. 自抗扰控制器及其应用[J]. 控制与决策, 1998, 13(1): 19-23.
HAN Jingqing. Auto-disturbances-rejection controller and it's applications[J]. Control and Decision, 1998, 13(1): 19-23.

[19] HAMMADI L, AHMED E, TAMOU N. Coordinated control by ADRC strategy for a wind farm based on SCIG considering low voltage ride-through capability[J]. Protection and Control of Modern Power Systems, 2022, 7(1): 1-18.

[20] SEYED R M, SEYED H S, HAMED M. Fault ride-through capability improvement in a DFIG-based wind turbine using modified ADRC[J]. Protection and Control of Modern Power Systems, 2022, 7(4): 1-37.

[21] 高思华, 刘宝煜, 惠康华, 等. 信息年龄约束下的无人机数据采集能耗优化路径规划算法[J]. 电子与信息学报, 2024, 46(10): 4024-4034.
GAO Sihua, LIU Baoyu, HUI Kanghua, et al. Energy-efficient UAV trajectory planning algorithm for AoI-constrained data collection[J]. Journal of Electronics & Information Technology, 2024, 46(10): 4024-4034.

[22] CHENG Yuhu, GUO Qingbang, WANG Xuesong. Proximal policy optimization with advantage reuse competition[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(8): 3915-3925.

[23] 王子昊, 王旭, 蒋传文, 等. 基于近端策略优化算法的灾后配电网韧性提升方法[J]. 电力系统自动化, 2022, 46(21): 62-70.
WANG Zihao, WANG Xu, JIANG Chuanwen, et al. Resilience improvement method for post-disaster distribution network based on proximal policy optimization algorithm[J]. Automation of Electric Power Systems, 2022, 46(21): 62-70.

[24] ZHENG Qingchun, PENG Zhi, ZHU Peihao, et al. An object recognition grasping approach using proximal policy optimization with YOLOv5[J]. IEEE Access, 2023, 11: 87330-87343.

[25] 周雪松, 张心茹, 赵洽宇, 等. 基于 DDPG 算法的微网负载端接口变换器自抗扰控制[J]. 电力系统保护与控制, 2023, 51(21): 66-75.
ZHOU Xuesong, ZHANG Xinru, ZHAO Hanyu, et al. Active disturbance rejection control of a microgrid load-side interface converter based on a DDPG algorithm[J]. Power System Protection and Control, 2023, 51(21): 66-75.

收稿日期：2024-10-21；　修回日期：2024-12-28

作者简介：

周雪松(1964—)，男，博士，教授，研究方向为新能源发电、电力系统控制；E-mail: sjteam2023@163.com

刘文进(2000—)，男，硕士研究生，研究方向为新能源发电与储能技术；E-mail: 482874253@qq.com

马幼捷(1964—)，女，通信作者，博士，教授，研究方向为微电网运行与控制。

(编辑　许　威)