第 49 卷 第 6 期：2329-2339 高电压技术 Vol.49, No.6: 2329-2339
2023 年 6 月 30 日 High Voltage Engineering June 30, 2023
DOI: 10.13336/j.1003-6520.hve.20230511

基于鲁棒强化学习的配网潮流优化方法

李晓旭[1]，田　猛[1]，朱紫阳[1]，董政呈[2]，龚　立[1]，王先培[1]
（1. 武汉大学电子信息学院，武汉 430072；2. 武汉理工大学自动化学院，武汉 430072）

摘　要：传统深度强化学习在优化配网潮流时易受传感器观测误差等干扰，鲁棒性较差。对此，提出一种基于鲁棒强化学习的配网潮流优化方法。首先以最小化配网网损为目标，电压、潮流越限为安全约束，建立包含分布式发电、储能及负荷单元的配网潮流优化模型。然后将干扰建模为攻击智能体，对配网潮流优化主智能体的观测状态施加扰动，构建双智能体零和博弈鲁棒强化学习模型。最后提出一种双智能体–拉格朗日乘子–信任区域策略优化算法，配网潮流优化主智能体与攻击智能体同步训练、异步学习，相互对抗博弈。仿真结果表明，通过该方法训练的配网潮流优化智能体，能在不同类型的干扰下做出安全决策，提高了配网潮流优化的鲁棒性和安全性。
关键词：配网潮流优化；鲁棒强化学习；零和博弈；状态扰动；安全决策

Distribution Network Power Flow Optimization Method Based on Robust Reinforcement Learning

LI Xiaoxu[1], TIAN Meng[1], ZHU Ziyang[1], DONG Zhengcheng[2], GONG Li[1], WANG Xianpei[1]
(1. Electronic Information School, Wuhan University, Wuhan 430072, China;
2. School of Automation, Wuhan University of Technology, Wuhan 430072, China)

Abstract: Traditional deep reinforcement learning is prone to interference such as sensor observation errors when optimizing distribution network power flow, and its robustness is poor. Therefore, a distribution network power flow optimization method based on robust reinforcement learning is proposed. Firstly, a distribution network power flow optimization model including distributed generation, energy storage, and load cells is established with the goal of minimizing distribution network power losses and the security constraints of voltage and power flow exceeding limits. Then, the interference is modeled as an attack agent, which perturbs the observed state of the main agent for distribution network power flow optimization, thus a robust reinforcement learning model for a dual agent zero sum game is constructed. Finally, a dual agent Lagrange multiplier trust region strategy optimization algorithm is proposed, in which the main agent and the attacking agent of distribution network power flow optimization can train synchronously, learn asynchronously, and play games against each other. The simulation results show that the distribution network power flow optimization agent trained by this method can make security decisions under different types of interference, improving the robustness and security of distribution network power flow optimization.
Key words: distribution network power flow optimization; robust reinforcement learning; zero-sum game; state disturbance; security decisions

0　引言

配网潮流优化是配网分析、规划与调度运行的重要基础，对保证电力系统可靠、安全、高效运行具有重要意义[1]。传统配网潮流优化方法包括梯度下降算法、牛顿法以及启发式算法等[2]。然而，分布式发电单元的接入增加了配网调度运行的复杂程度。传统优化方法在面对复杂配网环境时存在计算效率低下、无法保证收敛精度的缺陷，难以应对愈发复杂的现代电力系统[3]。

近年来，大量研究利用深度强化学习(deep reinforcement learning, DRL)理论的非线性数据拟合能力和自主学习能力，有效提升了复杂配网环境下潮流优化计算的计算效率和精度[4]。文献[5]针对复杂配网潮流计算不收敛的问题，提出基于 DRL 的潮流计算收敛自动调整方法，并利用知识经验提高算法效率。文献[6]提出了基于 DRL 的分布式电源

基金资助项目：国家自然科学基金(52177109)；湖北省重点研发计划(2020BAB109)。
Project supported by National Natural Science Foundation of China (52177109), Key R&D Program of Hubei Province (2020BAB109).

就地电压控制框架，使得调度智能体能够实时制定分布式电源运行策略，自适应地应对电压波动问题。文献[7]针对接入分布式可再生能源的配电网络中线路过载、母线电压违规等安全问题，提出了一种基于 DRL 的在线网络重构方案。文献[8]利用 DRL 框架建立了以最小化系统购能成本和弃风弃光成本为目标的综合能源系统经济调度模型。文献[9]提出一种多智能体迁移强化学习算法，旨在实现配网最优碳−能复合流模型的快速、高质量求解。综上，DRL 理论与算法在配网潮流计算、网络安全、规划调度等领域表现出优秀的计算效率和控制性能。

然而，随着 DRL 技术在电力系统中的深入应用，其面临的鲁棒性问题日益凸显[10]。目前，相关研究针对这一问题仅对配网环境的不确定性进行了简化和假设。文献[11]基于历史数据将随机的分布式电源出力视为满足一定分布，以减小对出力的预测误差。文献[12]忽略负荷需求和光伏发电的间歇与时变性，将用户电热负荷需求和预测的光伏发电出力进行简化。文献[13]对光储充电站储能系统全寿命周期优化运行进行研究，简化了储能系统的可用容量衰减过程。文献[14]基于历史数据对输电线路负荷的不确定性进行建模，简化了 DRL 优化调度场景的不确定因素。总的说来，现有研究均假设配网调度中心掌握准确无误的观测状态信息。

实际上，为了保障电网安全，DRL 方法先在模拟电网环境中进行训练，然后再部署到实际电网中。模拟电网环境和现实电网环境之间存在一定差距。采用模拟环境进行训练存在如下干扰：(1)配网中各传感器观测误差导致 DRL 训练数据出现偏差[15]。例如，空心线圈电流传感器角差易出现单向偏离问题，电容分压电压传感器角差最大幅值波动达到 7.6%，电阻分压电压传感器易受杂散电容影响，幅值超出 5%时角差变化大[16]。这些误差直接影响 DRL 模型输出的准确性和稳定性，进而影响配网潮流优化效果[17]。(2)新能源发电单元的出力预测以及负荷预测存在偏差[18]。2021 年我国省级电网风电、光伏日前电力预测准确率分别约为 88%、87%，风光资源趋势判断(中长期预测)准确率约为 78%；美国德克萨斯州电力可靠性委员会日前风电预测误差中位数为 4.4%，平均值为 5.6%，最大值为 32.3%[19]。这些误差加剧了配网的复杂性和不确定性。(3)极端天气或硬件故障给配网正常运行带来严峻挑战[20]。此时，配网运行的可靠性和稳定性会受到很大的威胁。因此，现有基于 DRL 的配网调度、潮流优化方法难以直接推广到实际配网中[21-22]。DRL 在配网潮流优化领域的鲁棒性问题亟待解决。

针对上述问题，本文提出一种基于鲁棒强化学习的配网潮流优化方法。首先，建立一个配网潮流优化模型，该模型以配网网损为目标，电压、潮流越限为安全约束，并包含分布式发电、储能及负荷单元。然后构建一个对配网潮流优化主智能体施加观测状态扰动的攻击智能体。该攻击智能体的训练目标是干扰和破坏配网潮流优化主智能体的学习和训练。最后，根据本文提出的双智能体−拉格朗日乘子−信任区域策略优化(two agents-Lagrange multiplier-trust region policy optimization，TA-Lag-TRPO)算法，配网潮流优化主智能体与攻击智能体形成一个 Maximin 策略的双智能体零和博弈鲁棒强化学习模型。仿真结果表明，所提出的算法和模型框架能够提高配网潮流优化的鲁棒性。

1 配网潮流优化模型

三相不平衡在配网潮流优化模型中普遍存在，其增加了配电变压器的电能损耗，对配网线损造成较大影响[23]。治理三相不平衡的主要研究方法是负载平衡，但这类方法对设备通讯与主机算力提出了较高要求[24]。为减小算力要求，使用平衡三相系统来供电。平衡三相系统可用等效的单相表示法进行分析，只考虑其中一个相位。配网模型如图 1 所示，其中配网模型包括：

1）发电单元 U_{G}，这些单元在理想状态下只向配网中注入电能，具体包括传统发电单元 U_{TG}、分布式发电单元 U_{DG}。为便于分析，分布式发电单元具体指风力发电单元 U_{WDG} 和光伏发电单元 U_{SDG}。

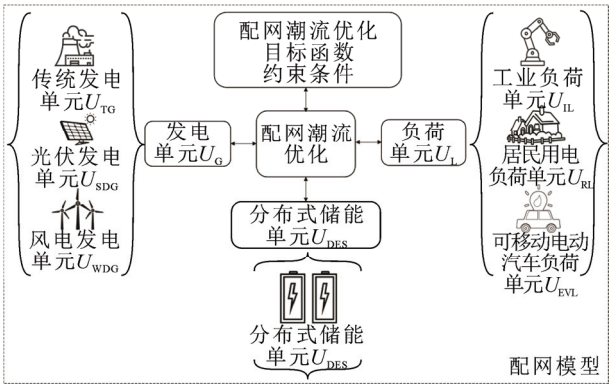


图 1 配网模型示意图

Fig.1 Schematic diagram of distribution network model

2）分布式储电单元 U_{DES}，储能单元既可以从配网中提取电能并存储，也可以将存储中的电能注入到配网中。

3）负荷单元 U_{L}，具体包括：工业用电负荷单元 U_{IL}、居民用电负荷单元 U_{RL}、可移动电动汽车负荷单元 U_{EVL}。不同类型的负荷具有不同的特性，工业负荷的比重在用电构成中居于首位，一般比较恒定；居民用电负荷具有极强的时间性波动；而电动汽车负荷最大的特点是可移动。

配网潮流优化模型目标函数为

$$\min\left(\sum_{t=0}^{T}\left(\Delta E_{\mathrm{cur}}^{t}+\Delta E_{\mathrm{pj}}^{t}+\Delta E_{\mathrm{des}}^{t}\right)\right) \tag{1}$$

配网潮流优化模型应满足的约束条件[25]如下：

$$P_{i,t}=\sum_{u_{\mathrm{g}}\in U_{\mathrm{G}}}P_{i,t}^{u_{\mathrm{g}}}+\sum_{u_{\mathrm{des}}\in U_{\mathrm{DES}}}P_{i,t}^{u_{\mathrm{des}}}+\sum_{u_{\mathrm{l}}\in U_{\mathrm{L}}}P_{i,t}^{u_{\mathrm{l}}} \tag{2}$$

$$P_{ij}=\left(V_i^2-V_iV_j\cos\theta_{ij}\right)g_{ij}-V_iV_jb_{ij}\sin\theta_{ij} \tag{3}$$

$$Q_{ij}=-V_iV_jg_{ij}\sin\theta_{ij}-\left(V_i^2-V_iV_j\cos\theta_{ij}\right)b_{ij} \tag{4}$$

$$S_{ij}^{\min}\le S_{ij}\le S_{ij}^{\max} \tag{5}$$

$$V_i^{\min}\le V_i\le V_i^{\max} \tag{6}$$

$$\left(P_i^{\min}\right)_{u_{\mathrm{g}}}\le\left(P_{i,t}^{u_{\mathrm{g}}}\right)_{u_{\mathrm{g}}\in U_{\mathrm{G}}}\le\left(P_{i,t}^{\max}\right)_{u_{\mathrm{g}}} \tag{7}$$

$$\left(Q_i^{\min}\right)_{u_{\mathrm{g}}}\le\left(Q_{i,t}^{u_{\mathrm{g}}}\right)_{u_{\mathrm{g}}\in U_{\mathrm{G}}}\le\left(Q_i^{\max}\right)_{u_{\mathrm{g}}} \tag{8}$$

$$\left(P_{i,t}^{\max}\right)_{u_{\mathrm{dg}}\in U_{\mathrm{DG}}}\in\left[\left(P_i^{\min}\right)_{u_{\mathrm{dg}}},\left(\hat{P}_{i,t}^{\max}\right)_{u_{\mathrm{dg}}}\right] \tag{9}$$

$$0\le\left(\sqrt{\left(P_{i,t}^{u_{\mathrm{dg}}}\right)^2+\left(Q_{i,t}^{u_{\mathrm{dg}}}\right)^2}\right)_{u_{\mathrm{dg}}\in U_{\mathrm{DG}}}\le\left(S_i^{\max}\right)_{u_{\mathrm{dg}}} \tag{10}$$

$$\frac{1}{\Delta t\eta}\left(E_{i,t}^{\mathrm{SoC}}-E_i^{\max}\right)\le\left(P_{i,t+1}^{u_{\mathrm{des}}}\right)_{u_{\mathrm{des}}\in U_{\mathrm{DES}}}\le\frac{\eta}{\Delta t}\left(E_{i,t}^{\mathrm{SoC}}-E_i^{\min}\right) \tag{11}$$

$$E_{i,t}^{\mathrm{SoC}}\in\left[E_i^{\min},E_i^{\max}\right] \tag{12}$$

$$0\le\left(\sqrt{\left(P_{i,t}^{u_{\mathrm{des}}}\right)^2+\left(Q_{i,t}^{u_{\mathrm{des}}}\right)^2}\right)_{u_{\mathrm{des}}\in U_{\mathrm{DES}}}\le\left(S_i^{\max}\right)_{u_{\mathrm{des}}} \tag{13}$$

$$\left(P_{i,t}^{\min}\right)_{u_{\mathrm{l}}}\le\hat{P}_{i,t}^{u_{\mathrm{l}}}\le 0 \tag{14}$$

式(1)表示配网潮流优化模型的目标函数即配网网损，利用配网能量守恒原理由 3 部分叠加求得：$\Delta E_{\mathrm{cur}}^{t}$ 代表 U_{DG} 的有功削减；$\Delta E_{\mathrm{pj}}^{t}$ 代表 U_{G}、U_{DES}、U_{L} 的有功注入(消耗)和；$\Delta E_{\mathrm{des}}^{t}$ 表示流入 U_{DES} 的净电能；T 表示时段总数量；$\Delta(\cdot)$表示求某物理量在 t 到 t+1 时段的变化量。上述 3 项的计算式分别是：

$$\Delta E_{\mathrm{cur}}^{t}=\Delta\left(\sum_{u_{\mathrm{dg}}\in U_{\mathrm{DG}}}\left(P_i^{\max}-P_{i,t+1}^{u_{\mathrm{dg}}}\right)_{u_{\mathrm{dg}}\in U_{\mathrm{DG}}}\right) \tag{15}$$

$$\Delta E_{\mathrm{tel}}^{t}=\Delta\left(\sum_{u_{\mathrm{g}}\in U_{\mathrm{G}}}P_{i,t+1}^{u_{\mathrm{g}}}+\sum_{u_{\mathrm{des}}\in U_{\mathrm{DES}}}P_{i,t+1}^{u_{\mathrm{des}}}+\sum_{u_{\mathrm{l}}\in U_{\mathrm{L}}}P_{i,t+1}^{u_{\mathrm{l}}}\right) \tag{16}$$

$$\Delta E_{\mathrm{des}}^{t}=-\Delta\left(\sum_{u_{\mathrm{des}}\in U_{\mathrm{DES}}}P_{i,t+1}^{u_{\mathrm{des}}}\right) \tag{17}$$

式(2)—式(14)依次表示配网潮流优化模型应满足的约束项。式(2)表示功率平衡约束。式(3)、(4)表示配网支路有功、无功潮流约束方程。式(5)表示视在潮流安全约束，若支路包含配电变压器，则式(5)包含有配电变压器的容量约束。式(6)表示电压安全约束。式(7)、(8)表示 U_{G} 有功、无功出力限值约束。式(9)表示 U_{DG} 的最大有功出力约束。式(10)表示 U_{DG} 的容量约束。式(11)表示 U_{DES} 注入的有功功率与储能充放电功率、荷电状态的关系约束。式(12)表示 U_{DES} 的荷电状态限值约束。式(13)表示 U_{DES} 的容量约束。式(14)表示 U_{L} 注入的有功限值约束。其中，$P_{i,t}$ 表示 t 时段节点 i 处各单元注入的有功功率；u_{g}、u_{dg}、u_{des}、u_{l} 依次表示发电、分布式发电、分布式储能、负荷单元的元素；$P_{i,t}^{u_{\mathrm{g}}}$、$P_{i,t}^{u_{\mathrm{dg}}}$、$P_{i,t}^{u_{\mathrm{des}}}$、$P_{i,t}^{u_{\mathrm{l}}}$ 分别表示 t 时段节点 i 处 U_{G}、U_{DG}、U_{DES}、U_{L} 注入(消耗)的有功功率；$Q_{i,t}^{u_{\mathrm{g}}}$、$Q_{i,t}^{u_{\mathrm{dg}}}$、$Q_{i,t}^{u_{\mathrm{des}}}$、$Q_{i,t}^{u_{\mathrm{l}}}$ 分别表示 t 时段节点 i 处 U_{G}、U_{DG}、U_{DES}、U_{L} 注入(消耗)的无功功率；P_{ij}、Q_{ij}、S_{ij} 分别表示支路(i,j)的有功、无功、视在潮流；V_i 为节点 i 处的节点电压幅值；θ_{ij}、g_{ij} 及 b_{ij} 分别为支路(i,j)的相角、电导和电纳；$S_{ij}^{\max}$ 和 $S_{ij}^{\min}$ 分别为支路(i,j)的最大、最小视在潮流，若支路(i,j)上存在配电变压器，则 $S_{ij}^{\max}$ 和 $S_{ij}^{\min}$ 包含有配电变压器的容量极限；$V_i^{\max}$ 和 $V_i^{\min}$ 分别为节点 i 的最大、最小电压幅值；$\left(P_{i,t}^{\max}\right)_{u_{\mathrm{g}}}$ 和 $\left(P_i^{\min}\right)_{u_{\mathrm{g}}}$ 分别为节点 i 处 U_{G} 注入的最大、最小有功限制；$\left(Q_i^{\max}\right)_{u_{\mathrm{g}}}$ 和 $\left(Q_i^{\min}\right)_{u_{\mathrm{g}}}$ 分别为节点 i 处 U_{G} 注入的最大、最小无功限制；$\left(\hat{P}_{i,t}^{\max}\right)_{u_{\mathrm{dg}}}$ 表示 t 时段节点 i 处 U_{DG} 最大有功出力预测值，该值随自然天气变化具有一定间歇性和波动性；$\left(S_i^{\max}\right)_{u_{\mathrm{dg}}}$ 为 U_{DG} 的容量极限；$E_{i,t}^{\mathrm{SoC}}$ 表示 t 时段节点 i 处 U_{DES} 的荷电状态(state of charge, SoC)；η为充放电效率系数，$\eta\in[0,1]$；Δt 表示 t 时段变化量；$E_i^{\max}$ 和 $E_i^{\min}$ 分别为节点 i 处的 U_{DES} 的最大、最小荷电状态限制，$P_{i,t}^{u_{\mathrm{des}}}$ 受限于到当前的荷电状态 $E_{i,t}^{\mathrm{SoC}}$；$\left(S_i^{\max}\right)_{u_{\mathrm{des}}}$ 为 U_{DES} 的容量极限；$\hat{P}_{i,t}^{u_{\mathrm{l}}}$ 表示 t 时段节点 i 处 U_{L} 消耗的有功预测值，该值随时间变化，存在一定程度的周期性波动。

2 基于鲁棒强化学习的配网潮流优化模型

2.1 配网潮流优化马尔可夫零和博弈模型

为模拟现实电网环境中传感器观测误差、分布式发电单元出力预测误差等对配网正常运行的干扰，提出了如图 2 所示的配网潮流优化马尔可夫零和博弈模型框架。配网潮流优化马尔可夫零和博弈模型框架主要实现配网潮流优化主智能体 N_1 与攻击智能体 N_2 之间的双智能体零和博弈。其中，将配网仿真模型与现实配网环境之间的差距设置为攻击智能体 N_2[26]。其训练目标是对配网主智能体 N_1 的马尔可夫观测状态施加外界扰动，企图干扰配网潮流优化过程[27]。两智能体具有相反的奖励函数和相同的约束条件，构成零和博弈。

配网潮流优化马尔可夫零和博弈模型框架可以表示为双智能体零和博弈约束型马尔可夫决策模型。该模型可以由 8 元组$(S, A_1, A_2, P, R, C, \gamma, \rho_0)$表示。

1）状态集合 S。配网潮流优化主智能体 N_1 与攻击智能体 N_2 的马尔可夫状态 $s_t \in S$，具有相同形式，均定义为

$$s_t = \left\{ \begin{array}{c} \left[\left(P_{i,t}^{u_{\mathrm{g}}} \right)_{u_{\mathrm{g}} \in U_{\mathrm{G}}}, \left(Q_{i,t}^{u_{\mathrm{g}}} \right)_{u_{\mathrm{g}} \in U_{\mathrm{G}}} \right], \\ \left[\left(P_{i,t}^{u_{\mathrm{des}}} \right)_{u_{\mathrm{des}} \in U_{\mathrm{DES}}}, \left(Q_{i,t}^{u_{\mathrm{des}}} \right)_{u_{\mathrm{des}} \in U_{\mathrm{DES}}} \right], \\ \left[\left(P_{i,t}^{u_{\mathrm{l}}} \right)_{u_{\mathrm{l}} \in U_{\mathrm{L}}}, \left(Q_{i,t}^{u_{\mathrm{l}}} \right)_{u_{\mathrm{l}} \in U_{\mathrm{L}}} \right], \\ \left[\left(P_{i,t}^{\max} \right)_{u_{\mathrm{dg}} \in U_{\mathrm{DG}}}, \left(Q_{i,t}^{\max} \right)_{u_{\mathrm{dg}} \in U_{\mathrm{DG}}} \right], \\ \left(E_{i,t}^{\mathrm{SoC}} \right)_{u_{\mathrm{des}} \in U_{\mathrm{DES}}} \end{array} \right\} \tag{18}$$

2）动作集合 A_1。配网潮流优化主智能体 N_1 的马尔可夫动作集合 $a_{1,t} \in A_1$，定义为

$$a_{1,t} = \left\{ \begin{array}{c} \left(P_{u_{\mathrm{dg}},t}^{\mathrm{cur}} \right)_{u_{\mathrm{dg}} \in U_{\mathrm{DG}}}, \left(Q_{u_{\mathrm{dg}},t}^{\mathrm{cur}} \right)_{u_{\mathrm{dg}} \in U_{\mathrm{DG}}}, \\ \left(P_{u_{\mathrm{des}},t}^{\mathrm{inj}} \right)_{u_{\mathrm{des}} \in U_{\mathrm{DES}}}, \left(Q_{u_{\mathrm{des}},t}^{\mathrm{inj}} \right)_{u_{\mathrm{des}} \in U_{\mathrm{DES}}} \end{array} \right\} \tag{19}$$

式中：$P_{u_{\mathrm{dg}},t}^{\mathrm{cur}}$、$Q_{u_{\mathrm{dg}},t}^{\mathrm{cur}}$ 分别表示 t 时段 U_{DG} 的有、无功削减值；$P_{u_{\mathrm{des}},t}^{\mathrm{inj}}$、$Q_{u_{\mathrm{des}},t}^{\mathrm{inj}}$ 分别表示 t 时段 U_{DES} 的有、无功注入(或提取)值。

3）动作集合 A_2。攻击智能体 N_2 的动作是对主智能体 N_1 的马尔可夫观测状态施加扰动，因此其动作集合 A_2 与 s_t 的形式、维度相同。现实配网环境中，扰动动作的物理含义是：传感器观测误差、分布式发电单元出力预测误差、负荷单元负荷预测误差以及极端条件对配网正常运行的干扰。

4）转移概率密度函数 P。即根据当前时刻马尔可夫状态 s_t 和两智能体的马尔可夫动作 A_1、A_2 得到下一状态 s_{t+1} 的概率转换函数，具体形式为 $S \times A_1 \times A_2 \times S \rightarrow [0,1]$。

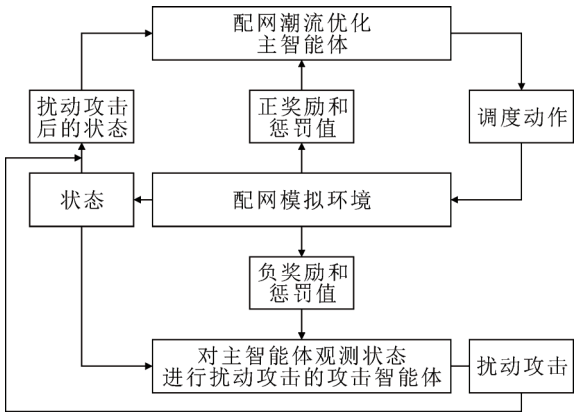


图 2 配网潮流优化马尔可夫零和博弈模型框架图

Fig.2 Framework diagram of Markov zero sum game model for distribution network power flow optimization

5）奖励函数 R。对于配网潮流优化马尔可夫零和博弈模型，两智能体的奖励函数应当互为相反数。$|r_t| \in R$，定义为

$$\left| r_t \right| = \left| \Delta E_{\mathrm{cur}}^t + \Delta E_{\mathrm{pj}}^t + \Delta E_{\mathrm{des}}^t \right| \tag{20}$$

奖励函数由配网运行过程中的能量损耗构成，式(20)中的 3 项已在式(15)—式(17)中定义。

6）惩罚项集合 C。在配网潮流优化过程中任何违反相关安全约束的操作都有可能导致某些单元的停机或永久性损坏，造成巨大经济损失。因此，应确保所应用的控制决策是安全的，以保障驱动配网的运行状态不会违反关键的物理约束。配网潮流优化模型有 2 个最典型的安全问题，利用惩罚集合 C 量化这些违反安全调度约束的惩罚反馈。

(1) 对每个节点 i 的电压值范围的安全限制。节点电压安全约束对于保障整个配网各个单元的稳定运行尤其关键。式(21)设置节点电压越限惩罚函数项 $c_1 \in C$。

$$c_1 = \sum_i \left(\max(0, |V_i| - V_i^{\max}) + \max(0, V_i^{\min} - |V_i|) \right) \tag{21}$$

则 c_1 为衡量节点电压幅值越限的惩罚值。

(2) 配网中传输线路存在视在潮流安全约束。设置此安全约束以确保线路不会过载而引起故障。式(22)设置视在潮流越限惩罚函数项 $c_2 \in C$。

$$c_2 = \sum_{i,j} \left(\max(0, \left| S_{ij} \right| - S_{ij}^{\max}) \right) \tag{22}$$

则 c_2 为衡量支路潮流越限的惩罚值。

7）折扣因子$\gamma\in(0,1]$，γ接近 1 的程度表示未来收益的重要程度；反之γ接近 0 的程度表示当前收益的重要程度。

8）$\rho_0\in\Delta(S)$是初始状态所属的分布函数，抽样该分布函数得到初始状态 s_0。

2.2　鲁棒强化学习目标和约束

鲁棒强化学习让智能体在面对已知或未知扰动攻击时，仍有良好的性能表现。通过设置对抗博弈，每个智能体都需要不断适应对手的策略和动作，单个智能体将具有更好的适应性和鲁棒性。

本文所提方法旨在通过配网潮流优化主智能体 N_1 和攻击智能体 N_2 的零和博弈训练，学习到主智能体 N_1 的最优调度策略 μ。最优的调度策略 μ 能够在鲁棒性测试过程中获得更高的期望奖励值以及更低的期望惩罚值。这等价于在外界干扰下，配网潮流优化运行过程中具有更低的能量损耗以及更低的安全越限惩罚。配网潮流优化主智能体 N_1 通过 DRL 算法和前述双智能体零、博弈约束型马尔可夫决策过程模型迭代更新策略 μ_θ 的网络参数θ，使式(23)的策略 μ_θ 下的奖励函数期望值 R_{μ_θ} 最大化，同时使得式(24)的策略 μ_θ 下的惩罚函数期望值 C_{μ_θ} 低于约束阈值 K。

$$R_{\mu_\theta}=\max_{\mu_\theta}\left(E_{\substack{a_{1,t}\sim\mu_\theta(s)\\ a_{2,t}\sim v_{\theta'}(s)}}\left(\sum_{t=0}^{T}\gamma^t r\left(s_t+a_{2,t},a_{1,t}\right)\right)\right) \tag{23}$$

$$\text{s.t.}\quad C_{\mu_\theta}=E_{\substack{a_{1,t}\sim\mu_\theta(s)\\ a_{2,t}\sim v_{\theta'}(s)}}\left(\sum_{t=0}^{T}\gamma^t c\left(s_t+a_{2,t},a_{1,t}\right)\right)\le K \tag{24}$$

式中：$a_{1,t}\sim\mu_\theta(s)$ 表示输入为状态集合 s 时利用策略 μ_θ 输出动作 $a_{1,t}$；$a_{2,t}\sim v_{\theta'}(s)$ 表示输入为状态集合 s 时利用策略 $v_{\theta'}$ 输出动作 $a_{2,t}$；$E_{\substack{a_{1,t}\sim\mu_\theta(s)\\ a_{2,t}\sim v_{\theta'}(s)}}$ 表示动作 $a_{1,t}$、$a_{2,t}$ 时求期望值。攻击智能体 N_2 的目标是学习到状态扰动策略 $v_{\theta'}$ 的最优网络参数 θ'，对主智能体 N_1 的观测状态进行扰动。与配网潮流优化主智能体 N_1 相反，攻击智能体 N_2 使式(23)的奖励函数期望值最小化，同时满足式(24)的惩罚函数期望值低于约束阈值 K。本质上，两智能体的目标、约束函数共同构成了鲁棒强化学习目标约束形式：

$$\max_{\mu_\theta}\min_{v_{\theta'}}R_{\mu_\theta,v_{\theta'}} \tag{25}$$

$$\text{s.t.}\quad C_{\mu_\theta,v_{\theta'}}\le K \tag{26}$$

式中 $R_{\mu_\theta,v_{\theta'}}$ 和 $C_{\mu_\theta,v_{\theta'}}$ 分别表示策略 μ_θ 和 $v_{\theta'}$ 共同作用下的奖励函数、惩罚函数期望值。式(25)、(26)得到了带约束的目标优化问题。因此，将拉格朗日乘子法[28]应用于上述目标约束优化问题，使其转换为无约束优化问题。配网潮流优化主智能体 N_1 得到新的无约束目标形式为

$$J_{\mu_\theta}=\max_{a_{1,t}\sim\mu_\theta}\min_{\beta}\left(R_{\mu_\theta}+\beta\left(C_{\mu_\theta}-K\right)\right) \tag{27}$$

攻击智能体 N_2 得到

$$J_{v_{\theta'}}=\min_{a_{2,t}\sim v_{\theta'}}\max_{\lambda}\left(R_{v_{\theta'}}+\lambda\left(C_{v_{\theta'}}-K\right)\right) \tag{28}$$

式中 $R_{v_{\theta'}}$ 和 $C_{v_{\theta'}}$ 分别表示扰动策略 $v_{\theta'}$ 作用下的奖励函数、惩罚函数期望值。式中拉格朗日乘子β、λ是一个可迭代更新的参数，用于权衡智能体的决策策略神经网络参数在奖励和惩罚之间的取舍更新。

根据上述分析，式(25)和(26)可以转化为如下鲁棒强化学习无约束目标形式[29]。

$$\max_{\mu_\theta}\min_{v_{\theta'}}J_{\mu_\theta,v_{\theta'}} \tag{29}$$

3　基于 TA-Lag-TRPO 的配网潮流优化方法

信任区域策略优化(trust region policy optimization, TRPO)算法是强化学习中的一种策略梯度算法，该算法在网络参数迭代更新时需要划定一块库尔贝克–莱布勒(Kullback-Leibler，KL)信任区域，将决策参数的更新限制 KL 区域内[30]。

然而，TRPO 算法本身不考虑环境的安全约束，直接将其应用于配网潮流优化模型会出现电压越限等安全问题。为保证决策的安全性，Lag-TRPO 算法将拉格朗日乘子法应用于式(24)的惩罚约束项，使得约束条件函数与原函数联系到一起，则有约束目标求解问题转化为无约束问题[28]。该算法的目标函数形式为：

$$\theta_{\text{new}}=\arg\max_{\theta_{\text{new}}}E_{a\sim\mu_\theta(s)}\left(\frac{\mu_{\theta_{\text{new}}}(a|s)}{\mu_\theta(a|s)}A_r^{\mu_\theta}\right) \tag{30}$$

$$\text{s.t. }E_{\theta_{\text{new}}}\left(D_{\text{KL}}\left(\mu_\theta(a|s),\mu_{\theta_{\text{new}}}(a|s)\right)\right)\le\delta \tag{31}$$

$$C_{\mu_\theta}\le K \tag{32}$$

式中：$E_{a\sim\mu_\theta(s)}$ 表示输入为状态集合 s 利用策略 μ_θ 输出动作 a 时求期望值；$A_r^{\mu_\theta}$ 为奖励函数的时序差分残差，将其定义为优势函数；$\mu_\theta(a|s)$ 表示输入状态为 s 时，利用潮流优化策略 μ_θ 输出动作 a 的概率；$\mu_{\theta_{\text{new}}}(a|s)$ 表示输入状态为 s 时，利用潮流优化策略 $\mu_{\theta_{\text{new}}}$ 输出动作 a 的概率；式(31)使用 KL 散度来衡量新旧人工神经网络参数θ_{new}和θ之间的更新距离，

δ 为 KL 散度距离的阈值；$E_{\theta\text{new}}$ 表示人工神经网络参数为θ_{new}时求期望值；D_{KL} 表示求 KL 距离。式(32)为前述带惩罚函数期望值的约束项。

首先利用泰勒展开公式对式(30)、(31)进行近似处理[30]：

$$\theta_{\text{new}} = \arg\max_{\theta_{\text{new}}} \boldsymbol{g}^{\text{T}}(\theta_{\text{new}} - \theta) \tag{33}$$

$$\text{s.t.} \quad \frac{1}{2}(\theta_{\text{new}} - \theta)^{\text{T}} \boldsymbol{H}(\theta_{\text{new}} - \theta) \le \delta \tag{34}$$

式中：$\boldsymbol{g}$ 表示求解目标项的梯度，$\boldsymbol{g} = \nabla_{\theta_{\text{new}}} \underset{a\sim\mu_\theta(s)}{E}\left(\frac{\mu_{\theta_{\text{new}}}(a|s)}{\mu_\theta(a|s)} A_r^{\mu_\theta}\right)$；$\boldsymbol{H}$ 表示新旧网络参数 θ_{new} 和 θ 之间平均 KL 距离的黑塞矩阵(Hessian matrix)。

再利用拉格朗日乘子法将带惩罚函数期望值约束项的式(32)变换到求解目标项上成为无约束形式。通过人工神经网络梯度下降法使得拉格朗日乘子β自适应收敛到合适的值，从而实现有约束到无约束优化问题的等价转换[28]。Lag-TRPO 算法的具体过程见附录 A 算法 A1。

针对式(29)的鲁棒强化学习无约束目标形式，结合 TRPO 算法以及拉格朗日乘子法，提出了一种 TA-Lag-TRPO 算法来实现配网潮流优化主智能体 N_1 和攻击智能体 N_2 的博弈训练，求解双智能体零和博弈鲁棒强化学习模型，具体过程见附录 A 算法 A2。

算法 A2 中实现的是配网潮流优化主智能体 N_1 和攻击智能体 N_2 之间的零和博弈训练过程。为保障训练效果，两个智能体在每一轮迭代训练中均交替优化更新各自网络参数[31]。首先，固定攻击智能体 N_2 的扰动策略 v，配网潮流优化智能体 N_1 学习决策策略 μ 以及如何防御攻击智能体 N_2 的状态扰动攻击，n 表示迭代轮数。在 T 步决策中，Lag-TRPO 算法的 Actor 决策网络根据决策参数 θ_n、扰动参数 θ'_{n-1} 依次做出决策 $a_{1,t}$，$a_{2,t}$；Critic 策略优化器网络对训练序列日志数据 $\{s_t, a_{1,t}, a_{2,t}, r_t, c_t\}$ 进行采样，评价所做决策并更新优化决策参数 θ_n。T 步决策后，固定住主智能体 N_1 的调度决策策略 μ，攻击智能体 N_2 学习如何攻击和干扰智能体 N_1 的决策调度，与上述过程相似。

TA-Lag-TRPO 算法的优势是：构建同步训练、异步学习的智能体策略更新框架，两智能体共享交互得到的训练数据进行训练。同时利用拉格朗日乘子法转换非凸优化问题，并考虑信任区域来限制更新步长。

4 算例分析

在文献[25]提出的配网测试系统上，本节对提出的双智能体零和博弈鲁棒强化学习模型和 TA-Lag-TRPO 算法进行了仿真验证。相关参数如下：Actor 和 Critic 网络均包含 2 个隐藏层，每层均为 64 个神经元；训练共进行 600 轮次的迭代，其中每一轮次的最大决策步数为 5000，单次 Episode 仿真的最大决策步数为 1000；折扣因子γ=0.99，学习率α=0.5。

该仿真实验基于 Python 3.8 和 Pytorch 1.13.0 实现算法编写。使用操作系统为 Ubuntu 18.04.6 的计算机进行实验，硬件配置为 Intel Xeon(R) E5-2630 v4 @2.2GHz×20、GTX 1080TI×2、RAM 32GB。

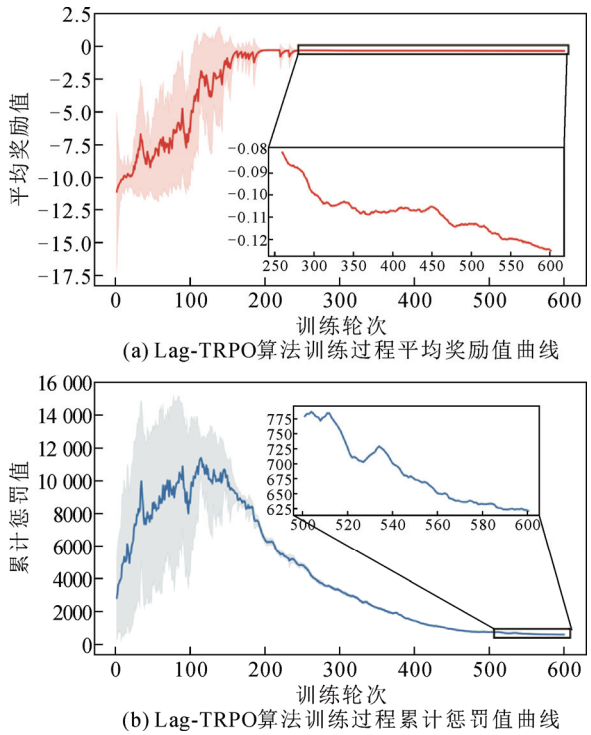
4.1 TA-Lag-TRPO 算法训练效果分析

为展示提出的双智能体零和博弈鲁棒强化学习模型和 TA-Lag-TRPO 算法的鲁棒性增强效果，首先利用算法 A1 训练配网潮流优化主智能体 N_1(不设置攻击智能体 N_2)。图 3 展示了该仿真实验的训练效果。

图 3(a)中的曲线表明，随着训练轮次的增加，配网潮流优化主智能体 N_1 输出的决策动作所获得的平均奖励值逐渐上升并迅速收敛。图 3(b)中的曲线表明，违反安全越限约束的累积惩罚值在训练前期呈现震荡上升的趋势，在训练末期逐渐下降至相对微小值。综合 2 条曲线的表现可知，训练末期因为拉格朗日乘子在奖励与惩罚值之间的取舍更新，平均奖励值随着累积惩罚值的收敛下降而略微降低。根据上述实验现象，认为配网潮流优化主智能体 N_1 利用 Lag-TRPO 算法学习到了较好的决策策略 μ。在没有外界干扰的情况下能够使得配网潮流优化过程取得较低的网损值，同时减少安全越限情况。

然后，设置攻击智能体 N_2，利用算法 A2 使其与配网潮流优化主智能体 N_1 进行双智能体零和博弈训练。两智能体的训练超参数保持一致，图 4 展示了双智能体零和博弈仿真实验的训练对比效果。

综合分析图 4(a)、(b)的效果对比曲线可知：(1)训练前期，在攻击智能体 N_2 的状态干扰下，利用 TA-Lag-TRPO 算法训练的潮流优化主智能体 N_1 决策表现劣于未设置攻击智能体 N_2 的 Lag-TRPO 算法；(2)训练后期，TA-Lag-TRPO 算法表现迅速提升，平均奖励值和累积惩罚值收敛情况均优于

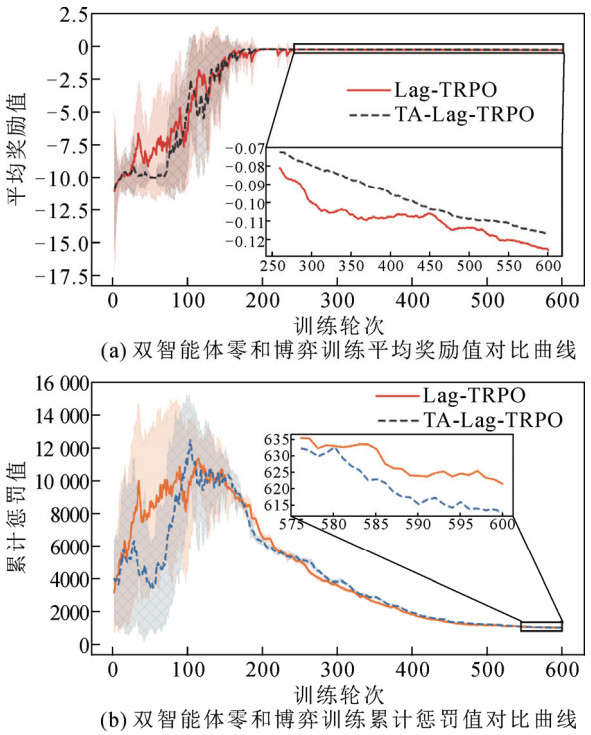


(a) Lag-TRPO算法训练过程平均奖励值曲线

(b) Lag-TRPO算法训练过程累计惩罚值曲线

图 3　Lag-TRPO 算法训练主智能体 N_1 效果图

Fig.3　Effect diagram of Lag-TRPO algorithm training main agent N_1



(a) 双智能体零和博弈训练平均奖励值对比曲线

(b) 双智能体零和博弈训练累计惩罚值对比曲线

图 4　双智能体零和博弈训练效果对比图

Fig.4　Comparison chart of training effects of Dual agent zero sum game

Lag-TRPO 算法。实现双智能体零和博弈训练的 TA-Lag-TRPO 算法使得配网潮流优化主智能体 N_1 学习到了更好的决策 μ。

4.2　惩罚约束限制阈值参数 K 影响分析

违反安全越限约束的惩罚项 C 是分析评价潮流优化智能体 N_1 鲁棒性和安全性的重要指标。因此，首先分析式(32)提出的配网潮流优化惩罚项约束阈值 K。附录 A 图 A1 展示了设置不同 K 值对配网潮流优化主智能体 N_1 的训练效果影响。

附录 A 图 A1 表明，在采取不同惩罚项约束阈值 K 训练时，对最终的平均奖励值 R 与累积惩罚值 C 产生了不同的影响。附录 A 图 A1 (a)显示过大的阈值 K=2000 导致智能体 N_1 获取到了较低的奖励值；附录 A 图 A1 (b)显示过小的阈值 K=500 反而使得训练过程中的惩罚项 C 收敛不到较低的水平。因此，结合上述实验现象，应选择适中大小的惩罚项约束阈值 K。根据附录 A 图 A1 中适中阈值 K 的训练表现以及智能体 N_1 在训练结束时的惩罚项 C 实际收敛结果，本文所有仿真训练实验均选择适中的惩罚约束限制阈值 K=1000。

4.3　鲁棒性测试及效果分析

表 1 展示了无观测状态扰动下，利用配网潮流优化主智能体 N_1 的最佳决策 μ^*进行 10 个 Episode 测试的 TA-Lag-TRPO 与 Lag-TRPO 算法效果对比。

表 1　TA-Lag-TRPO 与 Lag-TRPO 算法的性能比较

Table 1　Performance comparison between TA-Lag-TRPO and Lag-TRPO algorithms

算法	累积奖励值	累积惩罚值
Lag-TRPO	–46.77±4.23	617.54±10.23
TA-Lag-TRPO	**–45.22±1.76**	**553.29±9.36**

对累积奖励值而言，TA-Lag-TRPO 训练下的配网潮流优化主智能体 N_1 表现略优于 Lag-TRPO 算法 3.3%；从累积惩罚值来分析，主智能体 N_1 潮流优化过程中违反安全约束带来的累积惩罚值优于 Lag-TRPO 算法 10.4%。在无观测状态扰动的情况下，双智能体零和博弈鲁棒强化学习模型能够加强配网潮流优化主智能体 N_1 决策的安全性。

本文提出的 TA-Lag-TRPO 算法和双智能体零和博弈鲁棒强化学习模型旨在提高配网潮流优化过程中对现实配网中干扰的鲁棒性。因此，本节重点展示利用 TA-Lag-TRPO 算法训练的配网潮流优化主智能体 N_1 在不同程度、不同类型干扰下的鲁棒性提升效果。

4.3.1 不同程度状态扰动鲁棒性测试效果分析

本节模拟配网各传感器观测误差以及分布式发电单元出力、负荷需求预测不准确对配网正常工作运行的干扰。首先，在鲁棒性测试中对配网潮流优化主智能体 N_1 的马尔可夫观测状态设置多组不同程度的高斯噪声扰动。高斯噪声扰动的均值设置为 0，观测状态扰动程度对应高斯噪声扰动的标准差σ。

图 5 展示了 TA-Lag-TRPO、Lag-TRPO、原始对偶优化(primal-dual optimization, PDO)[32]和受限策略优化(constraint policy optimization, CPO)[33]4 种算法在不同程度状态扰动下鲁棒性测试对比效果。

由图 5 可知，对观测状态施加不同程度的扰动时(扰动程度依次为 0、0.3、0.5、0.8、1)，使用 TA-Lag-TRPO 算法和双智能体零和博弈鲁棒强化学习模型框架进行训练的配网潮流优化主智能体 N_1 与其他算法相比，其累计奖励值和累积惩罚值 2 个评价指标均优于各算法。该鲁棒性测试实验表明，面对配网各传感器观测误差以及分布式发电单元出力、负荷需求预测不准确带来的干扰时，经历双智能体零和博弈训练的潮流优化主智能体 N_1 做出了更优的决策。配网潮流优化过程中对观测状态扰动的鲁棒性和安全性得到提升。

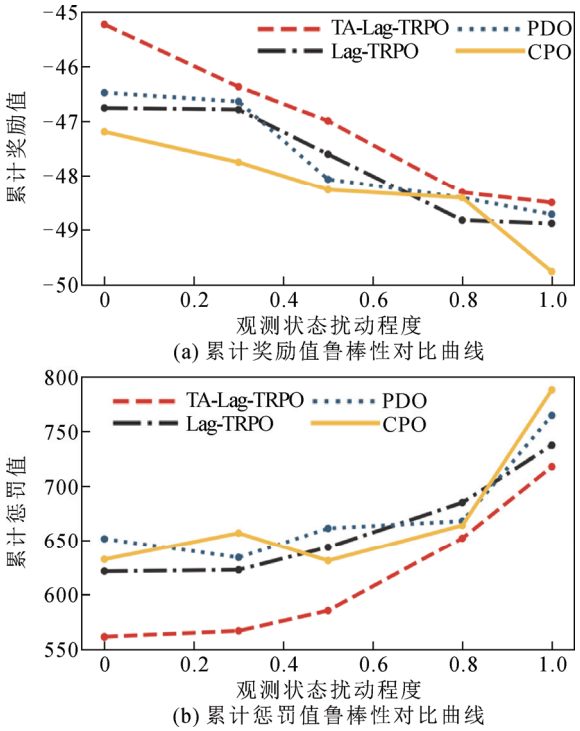
4.3.2 配网极端条件鲁棒性测试效果分析

针对极端条件的干扰，本节模拟实验了极端天气和储能单元故障条件下对配网的扰动攻击。

极端天气条件选取夜间台风场景，该条件下配网 U_{WDG} 持续高发，U_{SDG} 持续停发。对应观测状态中 U_{WDG} 最大有功出力持续处于较高数值，U_{SDG} 最大有功出力的值持续为 0。表 2 展示了极端自然天气下配网潮流优化主智能体 N_1 鲁棒性测试效果。

表 2 中的数据表明，夜间台风极端天气条件对配网的正常安全运行造成极大威胁。TA-Lag-TRPO 算法与 Lag-TRPO、PDO 以及 CPO 算法相比，潮流优化主智能体 N_1 取得了更低的安全越限累积惩罚值。在该极端天气条件下，使用 TA-Lag-TRPO 算法和双智能体零和博弈鲁棒强化学习模型框架进行学习和训练的潮流优化主智能体 N_1 具有更强的鲁棒性和安全性。

储能单元故障时，对应观测状态中 U_{DES} 的荷电状态 $E_{i,t}^{\mathrm{SoC}}$ 的值持续为 0。此时配网中 U_{DES} 的充、放电功能受到影响。表 3 展示了储能单元故障条件下配网潮流优化主智能体 N_1 鲁棒性测试效果。



(a) 累计奖励值鲁棒性对比曲线

(b) 累计惩罚值鲁棒性对比曲线

图 5 不同程度观测状态扰动下的鲁棒性测试对比图

Fig.5 Comparison chart of robustness tests under different degree of observed state disturbance

表 2 极端天气下各算法鲁棒性测试效果对比

Table 2 Comparison of robustness test results of various algorithms in extreme weather

算法	累积惩罚值
Lag-TRPO	3254.91±17.89
TA-Lag-TRPO	**1201.89±13.65**
CPO	2029.55±46.66
PDO	1300.26±11.29

表 3 储能单元故障时各算法鲁棒性测试效果对比

Table 3 Comparison of robustness test results of various algorithms when energy storage unit fails

算法	累积惩罚值
Lag-TRPO	951.81±19.81
TA-Lag-TRPO	**604.44±8.69**
CPO	2324.55±126.64
PDO	655.03±22.77

表 3 中的数据表明，配网储能单元故障时，与 Lag-TRPO、PDO 以及 CPO 算法相比，使用 TA-Lag-TRPO 算法和双智能体零和博弈鲁棒强化学习模型框架进行训练的配网潮流优化主智能体 N_1 具有更低的安全越限累积惩罚值。

综合分析以上 2 组模拟实验数据可知：相较

PDO 及 CPO 算法，TA-Lag-TRPO 算法和双智能体零和博弈鲁棒强化学习模型框架在对抗性训练下适应了未知的外部攻击；获得了更广泛的决策经验；自适应地考虑安全约束；兼顾了策略的稳步提升。潮流优化主智能体 N_1 在极端条件下的鲁棒性和安全性得到了提升。

5　结论

针对现实配网中干扰导致的 DRL 鲁棒性问题，本文提出了一种基于鲁棒强化学习的配网潮流优化方法。在文献[24]提出的配网测试系统中进行了仿真验证。通过对比实验和鲁棒性分析，可以得出以下结论：

1）提出的 TA-Lag-TRPO 算法和双智能体零和博弈鲁棒强化学习模型框架能够在不同程度的高斯噪声观测状态扰动下保持安全的决策动作，与其他算法相比能取得更低的网损值和惩罚值。

2）提出的算法和模型框架在极端自然天气或配网硬件故障的条件下仍能做出更安全的决策动作，与其他算法相比能取得更低的安全越限惩罚值。

3）提出的算法和模型框架可以有效提高配网潮流优化过程中的鲁棒性和安全性，保证配网安全运行可靠性、提高配网的经济投资效益。

本文重点关注配网中传感器观测误差、分布式发电出力预测误差和负荷预测误差等带来的鲁棒性问题，但未考虑配网潮流的三相不平衡性。下一步工作的重点是在现有成果上研究计及配网三相不平衡的潮流优化鲁棒性增强问题，充分考虑三相不平衡性带来的配网电压、电流不均衡影响，进一步提升所提方法和模型框架的应用参考价值。

附录见本刊网络版(http://hve.epri.sgcc.com.cn)。

参考文献　References

[1] 董旭柱，华祝虎，尚　磊，等. 新型配电系统形态特征与技术展望[J]. 高电压技术，2021，47(9)：3021-3035.
DONG Xuzhu, HUA Zhuhu, SHANG Lei, et al. Morphological characteristics and technology prospect of new distribution system[J]. High Voltage Engineering, 2021, 47(9): 3021-3035.

[2] 王甜婧，汤　涌，王　兵，等. 传统方法与人工智能：潮流控制优化算法的现状、挑战与未来方向[J]. 中国电机工程学报，2023，43(5)：1799-1817.
WANG Tianjing, TANG Yong, WANG Bing, et al. Traditional methods versus artificial intelligence: optimization algorithms for power flow control in state of the art, challenge and future directions[J]. Proceedings of the CSEE, 2023, 43(5): 1799-1817.

[3] PARK M G, LEE M C. A new technique to escape local minimum in artificial potential field based path planning[J]. KSME International Journal, 2003, 17(12): 1876-1885.

[4] 唐文虎，牛哲文，赵柏宁，等. 数据驱动的人工智能技术在电力设备状态分析中的研究与应用[J]. 高电压技术，2020，46(9)：2985-2999.
TANG Wenhu, NIU Zhewen, ZHAO Boning, et al. Research and application of data-driven artificial intelligence technology for condition analysis of power equipment[J]. High Voltage Engineering, 2020, 46(9): 2985-2999.

[5] 王甜婧，汤　涌，郭　强，等. 基于知识经验和深度强化学习的大电网潮流计算收敛自动调整方法[J]. 中国电机工程学报，2020，40(8)：2396-2405.
WANG Tianjing, TANG Yong, GUO Qiang, et al. Automatic adjustment method of power flow calculation convergence for large-scale power grid based on knowledge experience and deep reinforcement learning[J]. Proceedings of the CSEE, 2020, 40(8): 2396-2405.

[6] 习　伟，李　鹏，李　鹏，等. 基于深度强化学习的分布式电源就地自适应电压控制方法[J]. 电力系统自动化，2022，46(22)：25-31.
XI Wei, LI Peng, LI Peng, et al. Adaptive local voltage control method for distributed generator based on deep reinforcement learning[J]. Automation of Electric Power Systems, 2022, 46(22): 25-31.

[7] OH S H, YOON Y T, KIM S W. Online reconfiguration scheme of self-sufficient distribution network based on a reinforcement learning approach[J]. Applied Energy, 2020, 280: 115900.

[8] 朱振山，陈哲盛，盛明鼎. 基于柔性行动器−评判器的园区综合能源系统运行优化[J]. 高电压技术，2022，48(12)：4949-4958.
ZHU Zhenshan, CHEN Zhesheng, SHENG Mingding. Operation optimization of park-level integrated energy system based on soft actor-critic[J]. High Voltage Engineering, 2022, 48(12): 4949-4958.

[9] 陈艺璇，张孝顺，郭乐欣，等. 基于多智能体迁移强化学习算法的电力系统最优碳−能复合流求解[J]. 高电压技术，2019，45(3)：863-872.
CHEN Yixuan, ZHANG Xiaoshun, GUO Lexin, et al. Optimal carbon-energy combined flow in power system based on multi-agent transfer reinforcement learning[J]. High Voltage Engineering, 2019, 45(3): 863-872.

[10] 乔　骥，赵紫璇，王晓辉，等. 面向电力系统智能分析的机器学习可解释性方法研究(二)：电网稳定分析的物理内嵌式机器学习[J/OL]. 中国电机工程学报，2023：1-13[2023-05-25]. https://doi.org/10.13334/j. 0258-8013. pcsee. 221721.
QIAO Ji, ZHAO Zixuan, WANG Xiaohui, et al. Research on machine learning interpretability methods for intelligent analysis of power systems (II): physical embedded machine learning for power system stability analysis[J/OL]. Chinese Journal of Electrical Engineering, 2023: 1-13[2023-05-25]. https: //doi. org/10.13334/j. 0258-8013. pcsee. 221721.

[11] 彭刘阳，孙元章，徐　箭，等. 基于深度强化学习的自适应不确定性经济调度[J]. 电力系统自动化，2020，44(9)：33-42.
PENG Liuyang, SUN Yuanzhang, XU Jian, et al. Self-adaptive uncertainty economic dispatch based on deep reinforcement learning[J]. Automation of Electric Power Systems, 2020, 44(9): 33-42.

[12] 杨　挺，赵黎媛，刘亚闯，等. 基于深度强化学习的综合能源系统动态经济调度[J]. 电力系统自动化，2021，45(5)：39-47.
YANG Ting, ZHAO Liyuan, LIU Yachuang, et al. Dynamic economic dispatch for integrated energy system based on deep reinforcement learning[J]. Automation of Electric Power Systems, 2021, 45(5): 39-47.

[13] 陈亭轩，徐潇源，严　正，等. 基于深度强化学习的光储充电站储能系统优化运行[J]. 电力自动化设备，2021，41(10)：90-98.
CHEN Tingxuan, XU Xiaoyuan, YAN Zheng, et al. Optimal operation

based on deep reinforcement learning for energy storage system in photovoltaic-storage charging station[J]. Electric Power Automation Equipment, 2021, 41(10): 90-98.

[14] 邱 高，刘友波，许立雄，等. 基于深度确定性策略梯度的电网断面极限传输能力动态趋优控制[J]. 中国电机工程学报，2021，41(15)：5128-5138.
QIU Gao, LIU Youbo, XU Lixiong, et al. A deep deterministic policy gradient based-dynamic optimizing control for power system total transfer capability[J]. Proceedings of the CSEE, 2021, 41(15): 5128-5138.

[15] 韩世军，杜 帅，李真娣. 配电网不停电状态下电能分项计量误差测试方法[J]. 电网与清洁能源，2022，38(8)：24-29，39.
HAN Shijun, DU Shuai, LI Zhendi. Test method for power metering error in distribution network without power failur[J]. Power System and Clean Energy, 2022, 38(8): 24-29, 39.

[16] 韩筛根，陈 驰，王自桢. 配网用电压电流传感器误差在线测试装置设计及应用[J]. 高压电器，2020，56(12)：235-241.
HAN Saigen, CHEN Chi, WANG Zizhen. Design and application of error online testing device for voltage and current sensors in distribution network[J]. High Voltage Apparatus, 2020, 56(12): 235-241.

[17] ZHANG H, CHEN H G, XIAO C W, et al. Robust deep reinforcement learning against adversarial perturbations on state observations[J]. Advances in Neural Information Processing Systems, 2020, 33: 21024-21037.

[18] 黎静华，骆怡辰，杨舒惠，等. 可再生能源电力不确定性预测方法综述[J]. 高电压技术，2021，47(4)：1144-1155.
LI Jinghua, LUO Yichen, YANG Shuhui, et al. Review of uncertainty forecasting methods for renewable energy power[J]. High Voltage Engineering, 2021, 47(4): 1144-1155.

[19] 邓韦斯，孟子超，王皓怀，等. 新能源功率预测特性分析及精度提升措施[J]. 南方电网技术，2023，17(2)：11-23.
DENG Weisi, MENG Zichao, WANG Haohuai, et al. Renewable energy power prediction characteristics analyses and accuracy improvement measures[J]. Southern Power System Technology, 2023, 17(2): 11-23.

[20] 何 俊，于 华，邓长虹，等. 极端天气下基于态势感知的重点区域电网负荷供电保障策略[J]. 高电压技术，2022，48(4)：1277-1285.
HE Jun, YU Hua, DENG Changhong, et al. Power supply guarantee strategy for key regional power grid load based on situation awareness in extreme weather[J]. High Voltage Engineering, 2022, 48(4): 1277-1285.

[21] 毕聪博，唐聿劼，罗永红，等. 电力系统优化控制中强化学习方法应用及挑战[J/OL]. 中国电机工程学报，2023：1-24[2023-05-25]. https://doi. org/10.13334/j. 0258-8013. pcsee. 223433.
BI Congbo, TANG Yujie, LUO Yonghong, et al. Application and challenges of reinforcement learning methods in optimal control of power systems[J/OL]. Chinese Journal of Electrical Engineering, 2023: 1-24 [2023-05-25]. https://doi.org/10.13334/j.0258-8013.pcsee. 223433.

[22] XU M D, LIU Z X, HUANG P D, et al. Trustworthy reinforcement learning against intrinsic vulnerabilities: robustness, safety, and generalizability[J]. arXiv preprint arXiv, 2022: 2209.08025.

[23] 谢荣斌，杜 帆，程 湘，等. 三相不平衡及谐波对三相四线低压配电网线损的影响[J]. 电力系统保护与控制，2020，48(21)：22-30.
XIE Rongbin, DU Fan, CHENG Xiang, et al. Influence of three-phase imbalance and harmonic on line loss of three-phase four-wire low-voltage distribution network[J]. Power System Protection and Control, 2020, 48(21): 22-30.

[24] 郑在涛，叶保璇，钟康骅，等. 基于预测滚动的有源低压配电网三相不平衡治理方法[J]. 电网与清洁能源，2023，39(1)：29-36.
ZHENG Zaitao, YE Baoxuan, ZHONG Kanghua, et al. A predictive rolling-based three-phase unbalance management method of low-voltage active distribution network[J]. Power System and Clean Energy, 2023, 39(1): 29-36.

[25] HENRY R, ERNST D. Gym-ANM: reinforcement learning environments for active network management tasks in electricity distribution systems[J]. Energy and AI, 2021, 5: 100092.

[26] PINTO L, DAVIDSON J, SUKTHANKAR R, et al. Robust adversarial reinforcement learning[C]// Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR. org, 2017: 2817-2826.

[27] ZHANG H, CHEN H G, BONING D S, et al. Robust reinforcement learning on state observations with learned optimal adversary[J]. arXiv preprint arXiv, 2021: 2101.08452.

[28] FUJIMOTO S, CONTI E, GHAVAMZADEH M, et al. Benchmarking batch deep reinforcement learning algorithms[J]. arXiv preprint arXiv, 2019: 1910.01708.

[29] LI S H, WU Y, CUI X Y, et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient[C] // Proceedings of the 33rd AAAI conference on artificial intelligence. Honolulu, USA: AAAI, 2019: 4213-4220.

[30] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR. org, 2015: 1889-1897.

[31] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning[J]. arXiv preprint arXiv, 2013: 1312.5602.

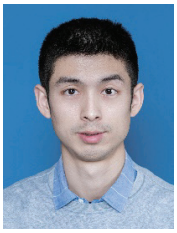
[32] LIANG Q K, QUE F, MODIANO E. Accelerated primal-dual policy optimization for safe reinforcement learning[J]. arXiv preprint arXiv, 2018: 1802.06480.

[33] ACHIAM J, HELD D, TAMAR A, et al. Constrained policy optimization[C] // Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 22-31.



李晓旭
1999—，男，硕士生
主要从事电力系统鲁棒性增强策略方面的研究
E-mail: xylonlee@whu.edu.cn

LI Xiaoxu



田 猛(通信作者)
1989—，男，博士，副研究员
主要从事电网信息物理系统安全方面的研究
E-mail: mengtian@whu.edu.cn

TIAN Meng
Ph.D.
Associate professor
Corresponding author

收稿日期 2023-03-31 修回日期 2023-05-26 编辑 程子丰

附录 A

算法 A1　Lag-TRPO 算法伪代码

Alg.A1　Lag-TRPO algorithm pseudocode

算法 A1　Lag-TRPO 算法

输入：初始决策策略 μ 参数θ_0，拉格朗日乘子β。

for n=1,2,…N do

固定参数θ执行 T 步决策，保存执行序列日志$\{s_1,a_1,r_1,c_2,\ldots,s_T,a_T,r_T,c_T\}$，估计奖励优势函数 $\hat{A}_r^{\mu_\theta}(s,a)$。

for m=1,2,…M do

计算拉格朗日乘子β更新其人工神经网络的损失函数 $\beta\{C_{\mu_\theta}-K\}$。

梯度下降法更新拉格朗日乘子β。

初始化 θ_{new} 并根据保存的决策序列日志数据计算实际奖励优势函数 $A_r^{\mu_\theta}$、计算目标策略梯度 $g=\nabla_{\theta_{new}}\underset{a\sim\mu_\theta(s)}{E}\left[\frac{\mu_{\theta_{new}}(a|s)}{\mu_\theta(a|s)}A_r^{\mu_\theta}(s,a)\right]$。

使用共轭梯度法求解 $x=H^{-1}g$、线性搜索法按照 $\theta_{new}=\theta+\sqrt{\frac{2\delta}{x^THx}}x$ 更新决策参数 θ_{new}，同时确保新旧参数之间满足 KL 距离限制。

end for

end for

返回：新的决策策略参数 θ_{new}

算法 A2　TA-Lag-TRPO 算法伪代码

Alg.A2　TA-Lag-TRPO algorithm pseudocode

算法 A2　TA-Lag-TRPO

输入：基于配网的双智能体马尔可夫零和博弈模型，配网潮流优化主智能体 N_1 初始决策策略参数 μ 参数 θ_0，攻击智能体 N_2 初始扰动策略 ν 参数 θ_0'。

for n=1,2,…N do

$\theta_n \leftarrow \theta_{n-1}$

for t=1,2,…T do

固定决策参数 θ_n、扰动参数 θ_{n-1}' 执行决策，攻击智能体 N_2 输出 $a_{2,t}\sim v_{\theta'}(s_t)$ 对 s_t 进行扰动攻击，主智能体 N_1 输出 $a_{1,t}\sim\mu_\theta(s_t+a_{2,t})$，保存执行序列日志$\{s_t,a_{1,t},a_{2,t},r_t,c_t\}$

利用决策日志$\{s_t,a_{1,t},-r_t,c_t\}$、决策参数 θ_n、算法 A1 Lag-TRPO 学习训练，更新决策策略 μ 参数 θ_n。

end for

$\theta_n' \leftarrow \theta_{n-1}'$

for t=1,2,…T do

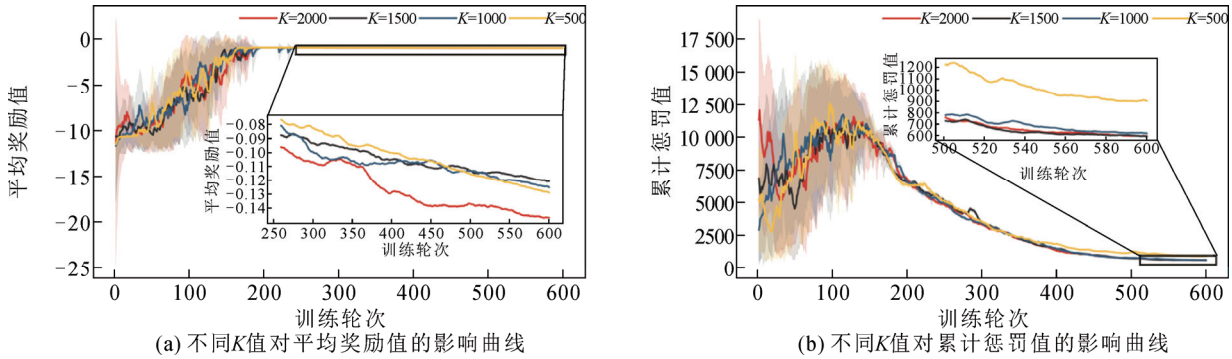
固定决策参数 θ_n、扰动参数 θ_n' 执行决策，攻击智能体 N_2 输出 $a_{2,t}\sim v_{\theta'}(s_t)$ 对 s_t 进行扰动攻击，主智能体 N_1 输出 $a_{1,t}\sim\mu_\theta(s_t+a_{2,t})$，保存执行序列日志$\{s_t,a_{1,t},a_{2,t},r_t,c_t\}$

利用决策日志$\{s_t,a_{2,t},r_t,c_t\}$、扰动参数 θ_n'、算法 A1 Lag-TRPO 学习训练，更新扰动策略 ν 参数 θ_n'。

end for

end for

返回：配网潮流优化主智能体 N_1 决策策略 μ 参数 θ_N，攻击智能体 N_2 扰动策略 ν 参数 θ_N'。



(a) 不同K值对平均奖励值的影响曲线　(b) 不同K值对累计惩罚值的影响曲线

图 A1　不同 K 值对配网潮流优化主智能体的训练效果影响

Fig.A1　Influence diagram of different K values on training effect of distribution system power flow optimization main agent