光学学报

基于深度强化学习的自由电子激光优化研究

吴嘉程[1,2]，蔡萌[3]，陆宇杰[1,3]，黄楠顺[4*]，冯超[2,3]，赵振堂[1,2,3]

[1]上海科技大学物质科学与技术学院，上海 201210；

[2]中国科学院上海高等研究院，上海 201210；

[3]中国科学院上海应用物理研究所，上海 201800；

[4]张江实验室，上海 201210

摘要 束流轨道优化是短波长自由电子激光调试放大过程的关键环节。在实际实验中，需要花费大量的时间来调整参数，以校正轨道。为简化该多参数调优过程，研究了基于深度强化学习的自动优化技术，在仿真环境中使用SAC、TD3和DDPG算法调整多个校正磁铁，以优化自由电子激光的输出功率。为模拟实际实验中非理想的轨道状态，在第一节波荡器入口处设置一磁铁以偏转束流轨道。随后利用深度强化学习算法自动调节后续7个磁铁以校正轨道。结果表明，通过引入偏差将输出功率降低一个数量级后，基于最大熵原理的SAC算法将功率恢复到初始值的98.7%，优于TD3与DDPG算法。此外，SAC算法表现出更强的鲁棒性，有望后续应用在我国X射线自由电子激光装置中实现自动调束。

关键词 激光光学；自由电子激光；轨道校正；输出功率；深度强化学习；多参数优化

中图分类号 O436 **文献标志码** A **DOI**: 10.3788/AOS230893

1 引言

自由电子激光可以产生超短脉冲长度、超高峰值亮度、波长连续可调的X射线，为材料、生物、化学和原子物理等学科的应用开辟了新的篇章[1-3]。近年来，世界各地有许多自由电子激光装置在建设或者运行中，以软X射线自由电子激光装置(SXFEL)为例，在其日常运行中，为保证激光质量，必须精准控制加速器状态，这就需要对磁铁参数进行高维、高频、闭环的控制。此外，由于科学实验对光脉冲的波长、脉宽、亮度等特性提出了复杂的要求，自由电子激光装置的控制和优化任务变得更为繁复，这项任务通常由经验丰富的调试人员完成，耗时耗力。因此，需要先进的在线优化算法来改进调试过程。

深度强化学习是深度学习(DL)和强化学习(RL)的结合。深度学习具有强大的表征能力，可以凭借神经网络拟合强化学习的各个组成部分，包括策略、动作价值函数、状态价值函数等。强化学习受到动物学习中试错法的启发，将智能体与环境交互得到的奖励值作为反馈信号对智能体进行训练[4]。深度强化学习将深度学习和强化学习结合，以解决连续控制、组合优化等问题。

策略梯度是强化学习中控制策略的经典方法。Silver等[5]提出了确定性策略梯度(DPG)算法，在动作选择时采用确定性方法，每一状态下的动作通过最优动作策略μ直接获得确定的值。为进一步提升算法的适用性，Lillicrap等[6]将深度Q网络(DQN)与DPG算法结合，提出了深度确定性策略梯度(DDPG)算法，通过神经网络代替线性函数进行值函数预测[7]，提高了动作网络与评价网络的稳定性与收敛性，实现了端到端的策略学习，但DDPG存在动作价值高估、超参数敏感等问题，很难在实际应用中取得效果。为解决这两大问题，Fujimoto等[8]提出了双延迟深度确定性策略梯度(TD3)算法，在DDPG的基础上，提出了双重网络、目标策略平滑正则化、延迟更新3种关键技术，缓解了动作价值高估对策略更新的影响，提高了训练过程稳定性。为进一步提高算法在复杂任务中的训练效果，Haarnoja等[9]提出一种基于最大熵的执行器-评价器(SAC)算法，将代表策略随机性的熵值引入到策略更新中，实现最大化奖励的同时最大化熵值，在完成任务的基础上尽可能使动作随机。与其他算法相比，SAC算法稳定性强、空间探索度高，采样效率明显优于DDPG算法。实验结果表明，SAC算法在复杂任务上的完成效果优于DDPG、TD3等算法[10]。

强化学习已经在控制拥有数百个变量的大型数据中心表现出卓越的能力[11]。近年来，人工智能在加速器光源领域获得了广泛的应用，强化学习算法在加速器控制、调优、诊断和建模方面发挥着重要作用。结合

收稿日期：2023-04-28；**修回日期**：2023-05-22；**录用日期**：2023-05-31；**网络首发日期**：2023-06-28

通信作者：*huangns@zjlab.ac.cn

神经网络的深度强化学习方法已在直线加速器相干光源(LCLS)中用于优化波荡器taper,在自种子模式下使输出功率提高了1倍[12]。Edelen等[13]使用神经网络定义策略和自由电子激光装置(FEL)模型,并让策略和学习模型之间进行互动,FEL在3~6 MeV范围内切换束流能量时,学习好的策略后仅通过一步迭代便可输出正确的Twiss参数,但整个研究只通过模拟进行。Bruchon等[14]将强化学习算法应用到FERMI的EOS系统,在仅知道所检测的激光束强度下,Q-learning算法可以学习正确的状态-动作关联来控制激光校正过程。Ramirez等[15]介绍了在德国的BESSY Ⅱ光源中应用机器学习来分析和优化工具集的情况,使用深度强化学习调整机器参数(助推器电流、注入效率、轨道校正等)以优化不同情况。O'Shea等[16]在FERMI上使用强化学习方法进行自由电子激光和太赫兹源的优化和稳定,并简化了一种策略梯度算法,在几百步内优化所需信号。为了解决无模型强化学习方法样本效率低的问题,Kain等[17]开发了基于归一化优势函数的无模型强化学习方法,该方法具有多达16个自由度,并在不同设施上成功测试。Hirlaender等[18]成功地将两种不同的算法——Q-learning算法和基于模型的算法应用于FEL优化。基于模型的强化学习算法具有更强的表征能力和更高的样本效率,有望成为未来热门的优化方法。

强化学习利用智能体与环境互动所获得的正负奖励来更新参数,不需要输入环境的固有性质,不依赖数据集,理论上可以随时用于在线装置中任何参数的优化,具有成为先进自动控制算法的巨大潜力。本文利用TD3算法、SAC算法优化带轨道偏差的FEL放大过程。结果表明,当算法执行到400步时,轨道已被校正到理想位置,激光功率得到大幅提升。

2 自由电子激光模拟

我国自由电子激光研究与工程建设正蓬勃发展。2016年开始建设的上海软X射线自由电子激光试验装置(SXFEL-TF)是我国首台X射线自由电子激光装置,该装置于2020年11月通过国家验收,之后开始进行用户装置升级改造。改造后的上海软X射线自由电子激光用户装置(SXFEL-UF)将采用SASE和外种子型FEL的运行模式,可以获得高相干度、高亮度、短脉冲的水窗波段[19-20](2.34~4.4 nm)自由电子激光。

使用GENESIS 1.3软件[21]模拟自由电子激光的SASE模式,模拟设置的参数如表1所示,均为上海软X射线自由电子激光用户装置的典型参数。

表1 模拟所使用的主要参数

Table 1 Main parameters of the simulation

Parameter	Value
Beam average energy /GeV	1.5
Peak current /A	800
Energy spread /%	0.014
FEL wavelength /nm	3.72
Average beam radius /μm	50
Undulator length /m	3
Period length /cm	2.35

在自由电子激光的放大饱和过程中,需要电子束和辐射光在波荡器内重合以进行持续的相互作用。在实际工程中,电子束的轨道受到各种因素的影响,处于非理想状态,其横向位置与波荡器中心存在偏差,使得FEL峰值功率降低。为模拟非理想的轨道状态,在束流出口处设置单个二极磁铁以偏转电子束,在此设置下,电子束中心位置在波荡器中的最大偏差将达到200 μm,大于FEL光斑尺寸。模拟结果显示,FEL输出辐射功率下降约一个数量级。束流轨道优化部分在偏转磁铁后211.5 cm处开始设置,为了让每台波荡器中电子束与光斑重合,在每节波荡器后总共设置了7个二级磁铁来优化束流轨道。在本次模拟中,采用DDPG、TD3、SAC 3种算法来控制校正磁铁电流值,以实现SASE模式的优化。整个波荡器系统的结构如图1所示。

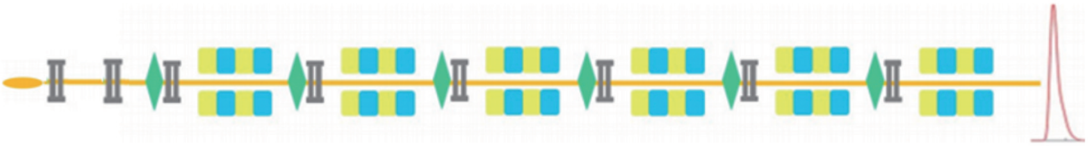

图1 波荡器系统的布局

Fig. 1 Layout of the undulator system

在优化任务中,将7个校正磁铁在水平和垂直方向上的电流值设置为智能体的动作,将束流轨道位置,即7个校正磁铁后电子束沿波荡器的x、y方向位置坐标设置为环境的状态,将光斑的强度与圆度确定为激光质量的评价指标。在模拟过程中,使用Python修改GENESIS 1.3软件的输入文件(电子束流和辐射光计算信息)与磁结构文件(波荡器、聚焦四极铁和校正二极铁排布)以完成动作执行过程,并通过读取分析GENESIS 1.3软件的输出功率和辐射光横向分布文件得到状态与奖励参数。

对于优化过程中的每一步,智能体首先执行动作,调控14个变量以校正轨道,此时环境发生变化并依据

激光质量的评价指标返给智能体一个奖励，智能体以最大化累计奖励为目标来优化动作。为了使FEL功率最大化，将奖励函数设置为$\left(I/I_0-1\right)/22$，以确保奖励值在(−1，1)区间，其中I为每一步的实时功率，I_0为轨道校正前的初始功率。

3　结果与讨论

在FEL仿真环境中，使用具有表2所列参数的SAC、TD3和DDPG算法在不同的随机数种子下对束流轨道进行优化。3种算法的策略网络学习率与评价网络学习率均为0.0003，策略网络与评价网络各有两层，分别有256个节点与512个节点，每次训练约950步，在不同随机数种子下共训练10次。图2给出了所提算法的训练结果，可以看到：随着SAC、TD3算法的学习过程累积，奖励函数大约在400步后收敛，FEL功率最终饱和；DDPG算法无法在有限的步骤内达到稳定条件。其原因在于TD3算法在DDPG算法的基础上缓解了动作价值高估对策略更新的影响，提高了训练过程稳定性，而SAC算法在最大化期望奖励的同时最大化熵值，增强了策略的随机性以及鲁棒性。随机性体现在信息熵的最大化会使输出的各个动作都趋于平均，有效防止策略过早收敛到局部最优值；鲁棒性体现在环境出现噪声时，算法有多个动作可供选择，使得收敛结果趋于稳定。由此可以看到，SAC算法收敛后的功率均值高于TD3算法，其置信区间相较于TD3算法也更小，表现出更好的稳定性。

表2　SAC、TD3和DDPG算法网络参数的设置

Table 2　Network parameter settings of SAC, DDPG, and TD3 algorithms

Parameter	Value
Actor learning rate	0.0003
Critic learning rate	0.0003
Neural network size	256×512
Batch size	64
Optimizer	Adam
Discount factor	0.99
Alpha learning rate	0.0003
Police noise	0.2

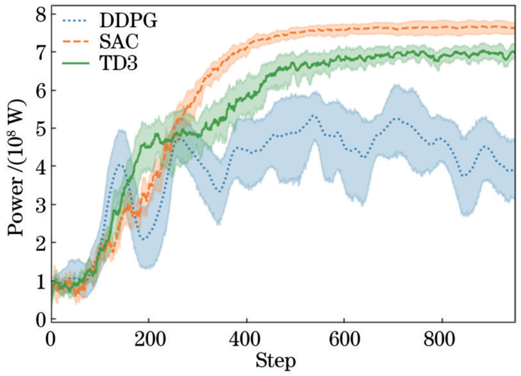

图2　SAC、TD3和DDPG算法在FEL优化中的功率曲线比较

Fig. 2　Comparison of power curves among SAC, TD3, and DDPG algorithms for FEL tuning tasks

3种算法在调优任务中的增益曲线与初始曲线如图3(a)所示，在优化任务中，采用在束流出口处设置偏转磁铁的方法来模拟具有非理想轨道的实际FEL设施。在设置偏转磁铁前，束流经过一段211.5 cm长的漂移段后进入SASE过程，所能达到的激光功率约为0.78 GW，这一数值在设置单个偏转磁铁后下降为0.08 GW；随后利用3种算法调整后续7个校正磁铁的电流值，以恢复自由电子激光功率。从图3(a)可以看到，SAC算法将输出功率由0.08 GW近似优化到0.77 GW，略高于TD3算法的0.71 GW，明显高于DDPG算法。

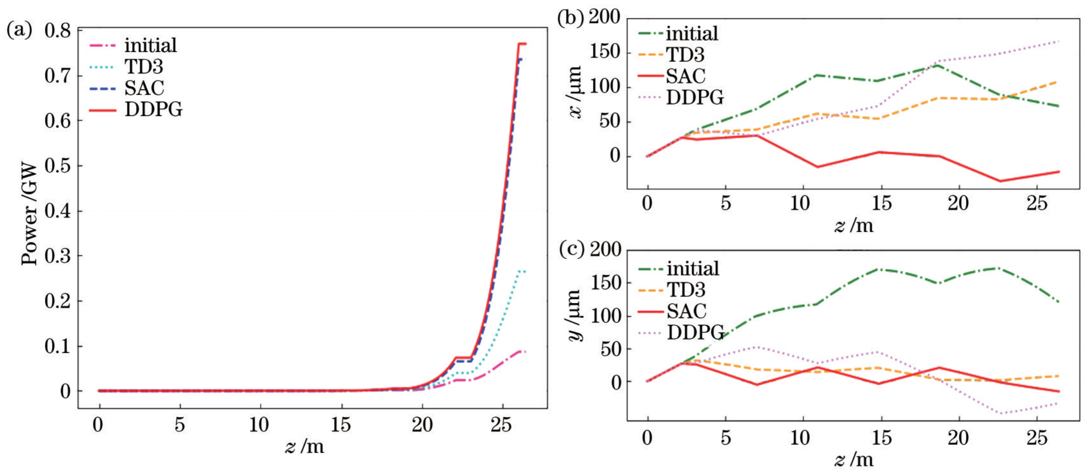

图3　SAC、TD3和DDPG算法在FEL优化中增益曲线与校正轨道比较。(a) 增益曲线；(b)水平方向和(c)垂直方向的校正轨道

Fig. 3　Comparison of gain curves and optimized orbits among SAC, TD3, and DDPG algorithms in the FEL optimization. (a) Gain curves; optimized orbits along (b) x and (c) y directions

3种算法所优化的轨道与初始轨道如图3(b)所示，由于在系统入口处施加了偏转磁铁并设置了漂移段，束流在波荡器结构的前2.115 m处于偏转状态，未校正轨道在此后继续发散，激光功率下降了一个数量级。SAC、TD3与DDPG算法均对轨道进行调整，图3(b)显示，SAC算法所优化的轨道在水平与垂直两个方向上更接近波荡器中心即理想轨道，这也就解释了SAC算法优化后的输出功率要高于TD3与DDPG算法。为了更直观地体现轨道优化结果，在图4中对比了波荡器出口处的初始光斑与3种算法优化后的光斑。可以看到：初始光斑在x、y方向上各有偏移且强度较弱；TD3算法在DDPG算法的基础上进行了改进，相较于DDPG算法，其光斑强度增大，但其在x方向上仍有偏移；SAC算法优化的光斑完全处于波荡器中心位置，且强度最大。在优化任务中，设置奖励为光斑强度与圆度的函数，算法以最大化奖励为目标，明显改善了光斑的亮度与横向模式。

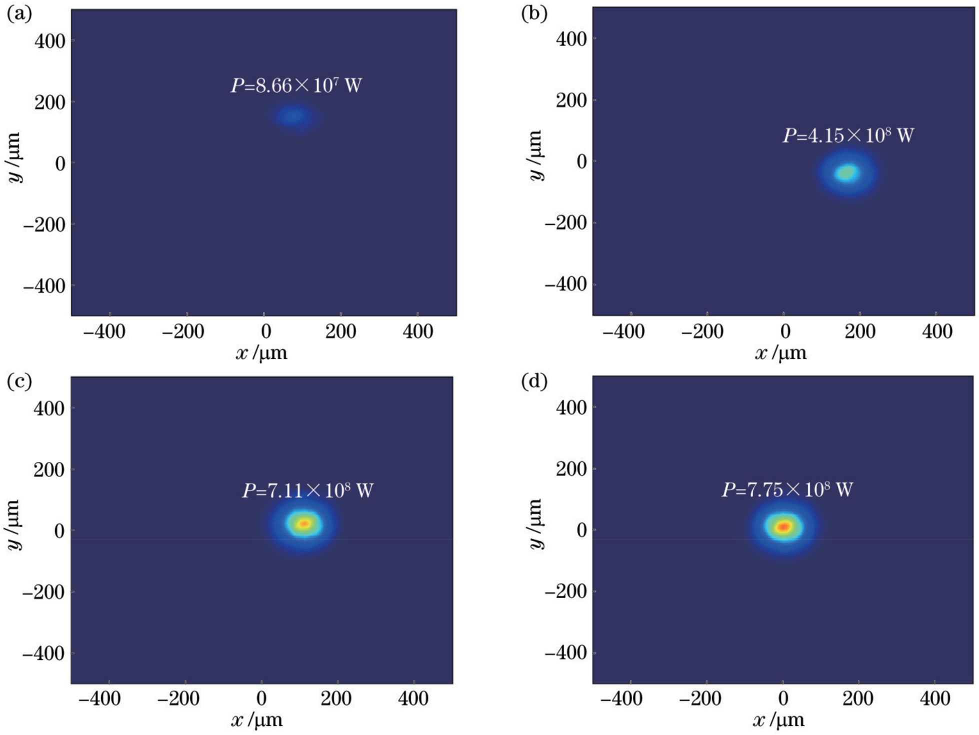

图4 DDPG、TD3和SAC算法优化后光斑与初始光斑对比。(a) 初始光斑；(b)DDPG优化后光斑；(c)TD3优化后光斑；(d)SAC优化后光斑

Fig. 4 Comparison between the optimized spots of DDPG, TD3, and SAC algorithms and the initial spot. (a) Initial spot; (b) DDPG optimized spot; (c) TD3 optimized spot; (d) SAC optimized spot

4 结论

在仿真实验中利用深度强化学习方法同时调整多个校正磁铁来优化波荡器内的束流轨道，SAC与TD3算法通过分析状态、权衡奖励、优化动作，高效地优化了束流轨道，改善了光斑品质。深度强化学习方法不需要标定数据集训练，而是从历史经验中学习规律，相较于启发式算法有更高的效率且不易陷入局部最优，通过修改网络结构就可以控制更多的参数来完成更加复杂的任务，有较大的应用潜力。模拟结果表明：TD3算法将激光功率优化到0.71 GW，解决了DDPG算法高估动作价值而造成偏差的问题，显著提升了DDPG算法的学习效率与性能；SAC算法将激光功率优化到0.77 GW，得到了最好的光斑形态，基于最大熵原理的算法体现出更好的训练效果与稳定性，其在不同随机数种子下可以近似得到相同的结果，有望后续应用在SXFEL中以实现自动调束。

参考文献

[1] Chapman H N. X-ray free-electron lasers for the structure and dynamics of macromolecules[J]. Annual Review of Biochemistry, 2019, 88: 35-58.

[2] Coffee R N, Cryan J P, Duris J, et al. Development of ultrafast capabilities for X-ray free-electron lasers at the linac coherent light source[J]. Philosophical Transactions A, 2019, 377(2145): 20180386.

[3] Marangos J P. The measurement of ultrafast electronic and structural dynamics with X-rays[J]. Philosophical Transactions A, 2019, 377(2145): 20170481.

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

[5] Sliver D, Lever G, Heess N, et al. Deterministic policy

gradient algorithms[C]//The 31st International Conference on Machine Learning, June 21-26, 2014, Beijing, China. Cambridge: JMLR, 2014: 387-395.

[6] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[EB/OL]. (2015-09-09)[2023-02-05]. https://arxiv.org/abs/1509.02971.

[7] 杨思明，单征，丁煜，等. 深度强化学习研究综述[J]. 计算机工程，2021，47(12): 19-29.
Yang S M, Shan Z, Ding Y, et al. Survey of research on deep reinforcement learning[J]. Computer Engineering, 2021, 47(12): 19-29.

[8] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods[EB/OL]. (2018-02-26)[2023-02-03]. https://arxiv.org/abs/1802.09477.

[9] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[EB/OL]. (2018-01-04) [2023-02-05]. https://arxiv.org/abs/1801.01290.

[10] Haarnoja T, Zhou A, Hartikainen K, et al. Soft actor-critic algorithms and applications[EB/OL]. (2018-12-13)[2023-02-03]. https://arxiv.org/abs/1812.05905.

[11] Evans R, Gao J. DeepMind AI reduces energy used for cooling google data centers by 40%[EB/OL]. (2016-07-20) [2023-02-05]. https://blog.google/outreach-initiatives/environment/deepmind-ai-reduces-energy-used-for/.

[12] Wu J, Huang X, Raubenheimer T, et al. Recent on-line taper optimization on LCLS[EB/OL]. [2023-05-03]. https://accelconf.web.cern.ch/fel2017/papers/tub04.pdf.

[13] Edelen A L, Milton S V, Biedron S G, et al. Using a neural network control policy for rapid switching between beam parameters in an fel [R]. Los Alamos: Los Alamos National Lab, 2017.

[14] Bruchon N, Fenu G. Free-electron laser optimization with reinforcement learning[C]//17th International Conference on Accelerator and Large Experimental Physics Control Systems (ICALEPCS' 19), October 5-11, 2019, New York, USA. Geneva: JACOW Publishing, 2020: 1122-1126.

[15] Ramirez L V, Mertens T, Mueller R, et al. Adding machine learning to the analysis and optimization toolsets at the light source BESSY II[C]//Proceedings of the 17th International Conference on Accelerator and Large Experimental Physics Control Systems, ICALEPCS2019, October 5-11, 2019, New York, USA. Geneva: JACOW Publishing, 2019: 5-11.

[16] O' Shea F, Bruchon N, Gaio G. Policy gradient methods for free-electron laser and terahertz source optimization and stabilization at the FERMI free-electron laser at Elettra[J]. Physical Review Accelerators and Beams, 2020, 23(12): 122802.

[17] Kain V, Hirlander S, Goddard B, et al. Sample-efficient reinforcement learning for CERN accelerator control[J]. Physical Review Accelerators and Beams, 2020, 23(12): 124801.

[18] Hirlaender S, Bruchon N. Model-free and Bayesian ensembling model-based deep reinforcement learning for particle accelerator control demonstrated on the FERMI FEL[EB/OL]. (2020-12-17)[2023-02-04]. https://arxiv.org/abs/2012.09737.

[19] Feng C, Deng H X. Review of fully coherent free-electron lasers [J]. Nuclear Science and Techniques, 2018, 29(11): 160.

[20] 周开尚. 超高亮度X射线自由电子激光物理研究[D]. 上海：中国科学院上海应用物理研究所，2018.
Zhou K S. Physical study of ultra-high brightness X-ray free electron laser[D]. Shanghai: Shanghai Institute of Applied Physics, Chinese Academy of Sciences, 2018.

[21] Reiche S. GENESIS 1.3: a fully 3D time-dependent FEL simulation code[J]. Nuclear Instruments and Methods in Physics Research Section A, 1999, 429(1/2/3): 243-248.

Reinforcement Learning for Free Electron Laser Online Optimization

Wu Jiacheng[1,2], Cai Meng[3], Lu Yujie[1,3], Huang Nanshun[4*], Feng Chao[2,3], Zhao Zhentang[1,2,3]

[1]*School of Physical Science and Technology, ShanghaiTech University, Shanghai* 201210, *China*;
[2]*Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai* 201210, *China*;
[3]*Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai* 201800, *China*;
[4]*Zhangjiang Laboratory, Shanghai* 201210, *China*

Abstract

Objective The X-ray free-electron lasers (FELs) have undergone a significant transformation in the fields of biology, chemistry, and material science. The capacity to produce femtosecond and nanoscale pulses with gigawatt peak power and tunable wavelengths down to less than 0.1 nm has stimulated the construction and operation of numerous FEL user facilities worldwide. Shanghai soft X-ray free-electron laser (SXFEL) is the first X-ray FEL user facility in China. Its daily operation requires precise control of the accelerator state to ensure laser quality and stability. This necessitates high-dimensional, high-frequency, and closed-loop control of beam parameters. Furthermore, the intricate demands of scientific experiments on FEL characteristics such as wavelength, bandwidth, and brightness make the control and optimization task of FEL devices even more challenging. This activity is usually carried out by proficient commissioning personnel and requires a significant investment of time. Therefore, the utilization of automated online optimization algorithms is crucial in enhancing the commissioning procedure.

Methods A deep reinforcement learning method combined with a neural network is employed in this study. Reinforcement learning uses positive and negative rewards obtained from the interaction between agents and the environment to update parameters. It does not require input from the inherent nature of the environment and is not dependent on data sets. In theory, this methodology has the potential to be implemented in various scenarios to optimize any given parameter in online devices. We employ SAC, TD3, and DDPG algorithms to adjust multiple correction magnets and optimize the output power of the free electron laser in a simulation environment. To simulate non-ideal orbit conditions, the beam trajectory is deflected by a magnet at the entrance of the first undulator. In the optimization task, we set the current values of seven correction magnets in both horizontal and vertical directions as the agent's action. The position coordinates of the electron beam along the x and y directions of the undulator line after passing through the seven correction magnets are set as the environment's state. The intensity and roundness of the spot are used as evaluation criteria for laser quality. During the simulation, Python is used to modify the input file and magnetic structure file of Genesis 1.3 to execute the action. The status and reward are obtained by reading and analyzing the power output and radiation field of Genesis 1.3. For each step in the optimization process, the agent first performs an action and adjusts 14 magnet parameters to correct the orbit. At this time, the environment changes and returns a reward to the agent according to evaluation criteria for laser quality. The agent optimizes its action to maximize cumulative reward.

Results and Discussions In the FEL simulation environment, we use SAC, TD3, and DDPG algorithms with parameters listed in Table 2 to optimize the beam orbit under different random number seeds. Figure 2 shows the training results of the proposed algorithm. As the learning process of SAC and TD3 algorithms progresses, the reward function converges, and the FEL power eventually reaches saturation. SAC and TD3 algorithms maximize FEL intensity at about 400 steps, with the convergence results of the SAC algorithm being better than those of the TD3 algorithm. This is because the TD3 algorithm, built on the DDPG algorithm, mitigates the impact of overestimation of action value on strategy updating and enhances the stability of the training process. The SAC algorithm maximizes the entropy while maximizing the expected reward, enhances the randomness of the strategy, and prevents the strategy from prematurely converging to the local optimal value. Furthermore, after convergence, the power mean of the SAC algorithm is noticeably more stable compared to that of the TD3 algorithm. Its confidence interval is also smaller, indicating better stability. The gain curve and initial curve of the three algorithms in the tuning task are shown in Fig. 3(a). The SAC algorithm approximately optimizes the output power from 0.08 GW to 0.77 GW, slightly higher than that of TD3 algorithm and significantly higher than that of DDPG algorithm. The optimized orbits and initial orbits of the three algorithms are shown in Fig. 3(b). Due to the deflection magnet applied at the entrance of the system and the drift section set, the beam is deflected and divergent in the first 2.115 m of the undulator structure, with the uncorrected orbits maintaining this state. The SAC, TD3, and DDPG algorithms all make adjustments to the orbits. Figure 3(b) shows that the orbits optimized by the SAC algorithm are closer to the center of the undulator, namely the ideal orbits, in both horizontal and vertical directions, which can also explain that the output power optimized by SAC is higher than that of TD3 and DDPG. To more directly reflect the results of orbit optimization, we compare the initial light spot at the outlet of the undulator with the optimized light spots of three algorithms (Fig. 4). The initial light spot is offset in both x and y directions and has weak intensity. However, the light spot optimized by SAC is completely centered in the undulator with the highest intensity, while it remains offset in the x direction for the other two algorithms.

Conclusions We employ deep reinforcement learning techniques to simultaneously control multiple correction magnets to optimize the beam orbit within the undulator. The deep reinforcement learning approach acquires rules from past experiences, avoiding the need for training with a calibration dataset. In contrast to heuristic algorithms, this approach exhibits superior efficiency and less proneness to local optima. In this study, the SAC and TD3 algorithms have been shown to effectively optimize beam orbit and improve spot quality through the analysis of system state, reward balancing, and action optimization. Results of the simulation indicate that the TD3 algorithm effectively optimizes the laser power to 0.71 GW, thereby resolving the issue of bias that arises from overestimating the action value of DDPG. Furthermore, the SAC algorithm has been utilized to optimize laser power to a value of 0.77 GW, demonstrating a marked improvement in the learning efficiency and performance of DDPG. The SAC optimization is based on the maximum entropy principle and is indicative of improved training effectiveness and stability. Thus, the SAC algorithm exhibits strong robustness and holds the potential to be utilized for the automated light optimization of SXFEL.

Key words laser optics; free electron laser; orbit correction; output power; deep reinforcement learning; multiparameter optimization