ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOPT013

TWIN DELAYED DEEP DETERMINISTIC POLICY GRADIENT FOR FREE-ELECTRON LASER ONLINE OPTIMIZATION

M. Cai[1,2], Z. H. Zhu[1,2], K. Q. Zhang [3*], C. Feng[1,2,3†], L. J. Tu[1,2], D. Gu [3], Z. T. Zhao[1,2,3]
[1]Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China

Abstract

X-ray free-electron lasers (FEL) have contributed to many frontier applications of nanoscale science which benefit from its extraordinary properties. During FEL commissioning, the beam status optimization especially orbits correction is particularly significant for FEL amplification. For example, the deviation between beam orbit and the magnetic center of undulator can affect the interaction between the electron beam and the FEL pulse. Usually, FEL commissioning requires a lot of efforts for multi-dimensional parameters optimization in a time-varying system. Therefore, advanced algorithms are needed to facilitate the commissioning procedure. In this paper, we propose an online method to optimize the FEL power and transverse coherence by using a twin delayed deep deterministic policy gradient (TD3) algorithm. The algorithm exhibits more stable learning convergence and improves learning performance because the overestimation bias of policy gradient methods is suppressed.

INTRODUCTION

X-ray free-electron lasers (FEL) open new chapters to various frontiers of scientific applications in biology, chemistry, and material science for its abilities to generate femtosecond and nanoscale pulses with gigawatt peak power and tunable wavelength down to less than 0.1 nm [1–3]. In recent years, several FEL facilities worldwide are constructed and operated successfully, which indicates a new era of X-ray science. To ensure the stable operation of FEL facilities, a robust and collimated beam orbit is generally required to achieve precise overlaps between the electron beam and radiation. The deviation between the beam orbit and the center of magnetic elements can induce a significant decrease in the peak power and transverse coherence of FEL radiation [4].

For a traditional FEL commissioning, beam orbit alignment can be roughly achieved by adjusting the current of correctors. However, the precise beam orbit alignment is usually difficult to be implemented and the beam orbit can change with the variation of beam status. Moreover, the effect of beam orbit alignment will also rely on the beam orbit stability. Therefore, manual beam orbit alignment and optimization require a lot of efforts in the accelerators which is a time-varying dynamics system of multi-dimensional parameters [5, 6]. In recent years, deep reinforcement learning (DRL) methods have been adopted in the commissioning and optimization of FEL facilities since they can work at different operating points and do not require labeled datasets compared to supervised learning methods [7, 8].

This work proposes an online optimization method in FEL based on a model-free off-policy actor-critic algorithm tailored to Markov decision processes. The modified DRL method can achieve good convergence and reduce overestimation by improving policy gradient approaches. Due to practical restrictions such as radiation safety and the potential for damage to the hardware of facility from erroneous online state, optimization with previous FEL simulation is usually the preferred solution. In this paper, we assess the algorithm in a simulated FEL environment firstly, and the simulation results of two methods are compared.

TWIN DELAYED DEEP DETERMINISTIC POLICY GRADIENT

As a subfield of machine learning, reinforcement learning (RL) has advantages in solving control tasks that conform to Markov decision processes [9]. In tasks with sufficient nonlinearity, complexity and time-varying systems like the FEL tuning process, RL is a more appropriate consideration than traditional control methods.

Reinforcement learning aims to find the optimal policy π_φ, with parameters φ, by maximizing the discounted sum of rewards $R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i)$, where γ is a discount factor. The action-value function describes the expected return after taking an action in state s_t following policy π:

$$Q^\pi(s_t, a_t) = E_{s_i \sim E, a_i \sim \pi}[R_t | s_t, a_t]$$

The function can be estimated recursively through Bellman equation:

$$Q^{\pi_\varphi}(s_t, a_t) = E_{s_{t+1} \sim E}[r(s_t, a_t) + \gamma Q^{\pi_\varphi}(s_{t+1}, \pi_\varphi(s_{t+1}))]$$

A neural network function approximator parameterized by θ^Q can be trained by minimizing the loss:

$$L(\theta^Q) = E_{s_t \sim E, a_t \sim \pi_{\varphi_i}}[(Q(s_t, a_t | \theta^Q) - y_t)^2]$$

where $y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, \pi_\varphi(s_{t+1}) | \theta^Q)$.

Deep Deterministic Policy Gradient (DDPG) is a model-free Q-Learning method in continuous action space [10] by combining Deep Q-network [11] and Deterministic Policy Gradient [12]. However, the performance of the Q-learning algorithm is known to be influenced by the systematic overestimation of values because of the output prediction noise [13].

* zhangkaiqing@zjlab.org.cn
† fengchao@zjlab.org.cn

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOPT013

If overestimation bias accumulates, policy updates will be negatively affected during the FEL online optimization with relatively large orbit jitter. As a potential solution, Twin Delayed Deep Deterministic Policy Gradient (TD3) [14] solves the problem by using several techniques on the DDPG. Figure 1 describes the structure of TD3.

The first technique is Clipped Double-Q Learning. In TD3, additional bias can be reduced by learning two Q-functions and selecting the smaller Q value:

$$y^{target} = r + \gamma \min_{i=1,2} Q_{\theta_{i'}}(s', \pi_{\varphi_1}(s'))$$

Unlike the actor-critic network, the target update is presented by two critic networks. The value target in Clipped Double Q-learning contributes no additional overestimation as compared to the standard Q-learning target.

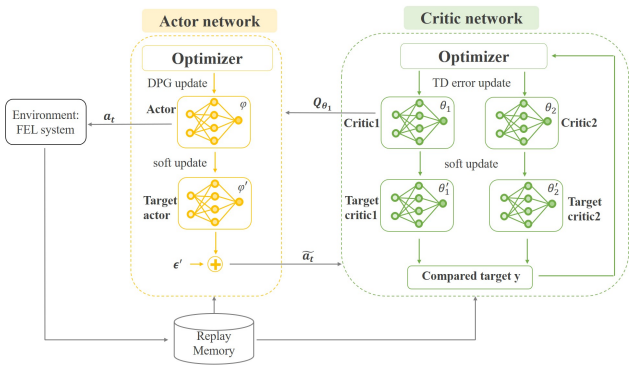

Figure 1: Structure of Twin Delayed Deep Deterministic Policy Gradient.

Less frequent policy updates in TD3 allow the value network to become more reliable and error-free, resulting in a lower variance value estimate and a better policy network. The parameters are updated according to $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$, which maintains a small temporal difference error and slows down the updating process.

Adding noise to the target policy and averaging over mini-batches can reduce the impact of inaccuracies induced by function approximation error:

$$y = r + \gamma Q_{\theta'}(s', \pi_{\varphi'}(s') + \epsilon)$$

$$\epsilon \sim clip(\mathcal{N}(0, \sigma), -c, c)$$

Moreover, a regularization approach can be used in the training stage to smooth the target policy and solve the overfitting issue of value estimate in the deterministic policy.

FEL SIMULATION

Simulations with GENESIS 1.3 [15] are carried out using the typical parameters of the Shanghai Soft X-Ray Free Electron Laser User Facility (SXFEL-UF), as given in Table 1, to test RL methods at different setpoints. In this simulation, the methods of TD3 and DDPG are used to correct the orbit between the undulators and achieve the SASE FEL optimization. According to the simulated result with an ideal orbit, the radiation can get saturation with 6 undulators as shown in Fig. 2, thus 6 correctors are used to correct the orbit and ensure a better FEL amplification. Before optimization, we have added horizontal and vertical offsets of about 0.12-0.15 mm at the very beginning of the undulator because the entrance parameters of the actual facility are not always ideal. With these offsets, the laser power dropped from 1.16 GW to around 0.05-0.18 GW. Misalignments between undulators are not considered during the training.

Table 1: Main parameters of the simulation

Parameter	Value
Beam average energy	1.5 GeV
Peak current	800 A
Energy spread	0.014%
Average beam radius (RMS)	50 μm
Undulator length	126 × 2.35 cm
FEL wavelength	3.72 nm

Figure 2: Layout of the undulator system. Six undulator segments are used for SASE amplification to saturate at the typical wavelength.

In the optimization process, actions performed by the agent are the magnetic strengths of six correctors in the horizontal and vertical directions. The electron beam trajectory is a group of position coordinates of the electron beam along the undulator line, and it is also the state of the environment. To maximize the FEL power, we set the reward function as $I/I_0 - 1$, where I is the current power of each step and I_0 is the initial power before optimization with orbit offsets.

RESULTS AND DISCUSSION

In the FEL simulation environment, we utilize TD3 and DDPG with the parameters listed in Table 2 to conduct beam tuning trails with different random seeds. The learning rate of actor network is 0.0001 and critic network is 0.0003. Actor and critic networks have two layers, with 256 and 512 nodes respectively. Each trail runs for about 2000 steps. The learning results with TD3 and DDPG agents are shown in Fig. 3. According to Fig. 3, the reward function converges as the learning process of TD3 agent accumulates, and the FEL power gets saturation finally. Whereas, the DDPG agent cannot be able to get the stabilization condition in limited episodes.

Figure 4 shows the initial FEL gain curve and optimized FEL gain curves with these two methods in a tuning task. In our optimization, a special case with an initial offset of 150 μm at both horizontal and vertical positions is adopted to simulate the actual FEL facility with a nonideal entrance orbit. For this case, there will be an offset between the optimal

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOPT013

Table 2: Network parameter settings of DDPG and TD3 algorithm

Parameter	Value
Actor learning rate	0.0001
Critic learning rate	0.0003
Deep neural network size	256×512
Batch size	128
Optimizer	Adam
Discount factor	0.99

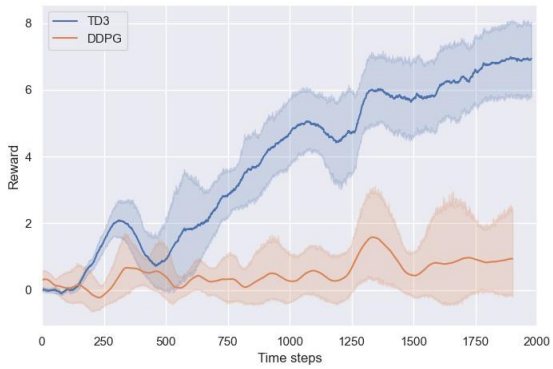

Figure 3: Learning curves comparison between TD3 and DDPG agents for the FEL tuning tasks. The shaded area represents the range of confidence intervals when aggregating 10 training curves with randomly initialized network parameters by an estimator.

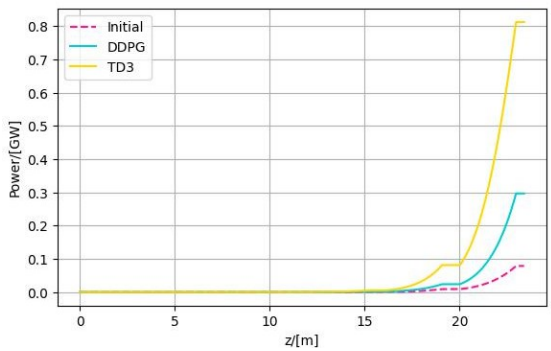

Figure 4: The initial FEL gain curve and optimized FEL gain curves with DDPG and TD3 optimization.

orbit and the magnetic center of undulators as well. According to Fig. 4, TD3 algorithm optimizes the output power from the initial 0.078 GW to 0.813 GW approximately, which is significantly higher than the DDPG algorithm. In Fig. 5, the true value is estimated using the average discounted sum of rewards over randomly initialized 10 steps following the current policy. Compared with TD3, there exists an apparent overestimation bias that occurs on the DDPG learning procedure. Rapidly rising values but limited laser power gains during DDPG optimization indicate inaccurate action evaluation and problematic strategy updates.

Figure 6 shows the initial beam orbit and optimized beam orbit with DDPG and TD3 method. For an actual FEL amplification, a relatively straight orbit can ensure better interaction between the electron beam and radiation. Therefore, we try to find a straight and optimized orbit to obtain a better FEL amplification in our simulation, as shown in Fig. 6. According to Fig. 4, we can also find that the final peak power with TD3 optimization is larger than that with no optimization and DDPG optimization.

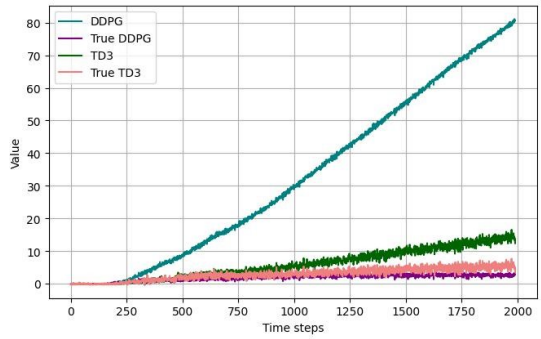

Figure 5: Estimated value and true value of DDPG and TD3 in FEL optimization.

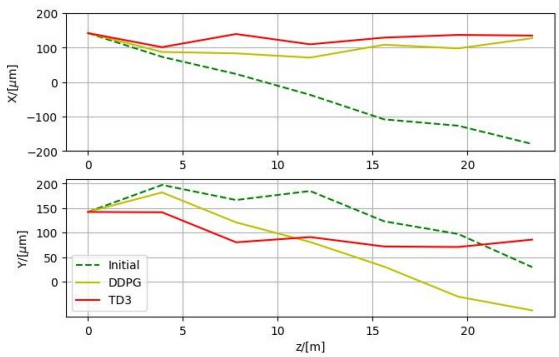

Figure 6: The initial beam orbit and optimized beam orbit with DDPG and TD3 methods.

CONCLUSION

We have tried to optimize the electron beam trajectory in undulators by using policy gradient methods to simultaneously control multiple corrector magnets. According to the preliminary simulation results, TD3 can considerably enhance the learning speed and performance of DDPG because it addresses the overestimation bias issue. The proposed method can perform tasks that involve management of more magnetic parameters and more complex requirements by simply modifying network structure since no prior physical knowledge and datasets are required. This method is being tested in the SXFEL facility, and it may be applied to the FEL optimization for the whole undulators and varying wavelengths in the future.

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-TUPOPT013

REFERENCES

[1] H. N. Chapman, "X-ray free-electron lasers for the structure and dynamics of macromolecules," *Annu. Rev. Biochem.*, vol. 88, pp. 35–58, 2019. doi:10.1146/annurev-biochem-013118-110744

[2] R. N. Coffee, J. P. Cryan, J. Duris, W. Helml, S. Li, and A. Marinelli, "Development of ultrafast capabilities for x-ray free-electron lasers at the linac coherent light source," *Philos. Trans. R. Soc. London, Ser. A*, vol. 377, no. 2145, p. 20180386, 2019. doi:10.1098/rsta.2018.0386

[3] J. Marangos, "The measurement of ultrafast electronic and structural dynamics with x-rays," *Philos. Trans. R. Soc. London, Ser. A*, vol. 377, no. 2145, 2019. doi:10.1098/rsta.2017.0481

[4] L. Zeng, C. Feng, D. Gu, J. Li, and Z. Zhao, "The beam-based alignment for soft x-ray free-electron lasers via genetic algorithm," *Nucl. Instrum. Methods Phys. Res., Sect. A*, vol. 905, pp. 104–111, 2018. doi:10.1016/j.nima.2018.07.033

[5] J. Duris, D. Kennedy, A. Hanuka, J. Shtalenkova, A. Edelen, P. Baxevanis, A. Egger, T. Cope, M. McIntire, S. Ermon, *et al.*, "Bayesian optimization of a free-electron laser," *Phys. Rev. Lett.*, vol. 124, no. 12, p. 124801, 2020. doi:10.1103/PhysRevLett.124.124801

[6] A. Scheinker, D. Bohler, S. Tomin, R. Kammering, I. Zagorodnov, H. Schlarb, M. Scholz, B. Beutner, and W. Decking, "Model-independent tuning for maximizing free electron laser pulse energy," *Phys. Rev. Accel. Beams*, vol. 22, no. 8, p. 082802, 2019. doi:10.1103/PhysRevAccelBeams.22.082802

[7] F. O'Shea, N. Bruchon, and G. Gaio, "Policy gradient methods for free-electron laser and terahertz source optimization and stabilization at the FERMI free-electron laser at Elettra," *Phys. Rev. Accel. Beams*, vol. 23, no. 12, p. 122802, 2020. doi:10.1103/PhysRevAccelBeams.23.122802

[8] N. Bruchon, G. Fenu, G. Gaio, M. Lonza, F. H. O'Shea, F. A. Pellegrino, and E. Salvato, "Basic reinforcement learning techniques to control the intensity of a seeded free-electron laser," *Electronics*, vol. 9, no. 5, p. 781, 2020. doi:10.3390/electronics9050781

[9] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 2014.

[10] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015. doi:10.48550/arXiv.1509.02971

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013. doi:10.48550/arXiv.1312.5602

[12] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*, pp. 387–395, PMLR, 2014.

[13] S. Thrun and A. Schwartz, "Issues in using function approximation for reinforcement learning," in *Proc. 1993 Connectionist Models Summer School Hillsdale, NJ. Lawrence Erlbaum*, vol. 6, 1993.

[14] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*, pp. 1587–1596, PMLR, 2018.

[15] S. Reiche, "Genesis 1.3: a fully 3d time-dependent fel simulation code," *Nucl. Instrum. Methods Phys. Res., Sect. A*, vol. 429, no. 1-3, pp. 243–248, 1999. doi:10.1016/S0168-9002(99)00114-X