Adaptive Offline Data Replay in Offline-to-Online Reinforcement Learning

Extended Abstract

Xu Liu
Shanghai Jiao Tong University
Shanghai, China
liu_skywalker@sjtu.edu.cn

Tong Yu
Adobe Research
San Jose, CA, USA
tyu@adobe.com

Shuai Li
Shanghai Jiao Tong University
Shanghai, China
shuaili8@sjtu.edu.cn

ABSTRACT

Offline-to-online reinforcement learning combines the advantages of offline data utilization with online exploration to enhance sample efficiency and performance. A primary challenge lies in managing the distribution shift between offline and online data, which significantly impacts training effectiveness. Existing methods often employ fixed mixing ratios for data replay, but these require task-specific tuning and may fail to generalize across different environments. To address this, we introduce a metric that evaluates policy quality relative to offline and online data, and propose a bandit-based strategy to adjust the mixing ratio adaptively, optimizing policy quality during training. Experiments across diverse environments demonstrate that our approach outperforms static methods, offering robust adaptability and minimizing manual tuning.

KEYWORDS

Reinforcement Learning; Data Replay

ACM Reference Format:
Xu Liu, Tong Yu, and Shuai Li. 2025. Adaptive Offline Data Replay in Offline-to-Online Reinforcement Learning: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement learning (RL) has achieved remarkable success in domains such as strategic gameplay, robotic control, and autonomous navigation [1, 8, 9]. However, its application in real-world scenarios is hindered by sample inefficiency, requiring extensive and often costly interactions with the environment [5, 6]. This limitation is particularly critical in sensitive applications like healthcare and autonomous driving [13, 14].

The offline-to-online RL paradigm addresses these challenges by combining offline pretraining with online fine-tuning, leveraging static datasets for robust initialization and dynamic interactions for policy adaptation [4, 12]. While promising, this transition is complicated by distribution shifts between offline and online data, which affect the agent's ability to integrate learned behaviors effectively [3, 7]. Common solutions include replaying offline data during online training using fixed mixing ratios or more balanced schemes, but these methods often require extensive task-specific tuning [9, 11].

Our empirical studies reveal that no single mixing ratio universally optimizes performance across tasks and datasets. Instead, the optimal data replay strategy depends on task-specific factors such as offline data quality and policy performance. For instance, higher reliance on offline data benefits high-quality datasets, while prioritizing online exploration can improve performance when the agent's policy surpasses the offline dataset.

To address these challenges, we propose the **R**einforcement **L**earning with **O**ptimized **A**daptive **D**ata-mixing (**ROAD**) framework. ROAD dynamically adjusts the mixing ratio during training using a metric that quantifies the agent's policy quality relative to both offline and online data. This metric is derived from the agent's state-action value function, allowing for adaptive replay through a bandit-based mechanism that balances exploration and exploitation. ROAD minimizes empirical tuning, enhances learning efficiency, and is theoretically robust with cumulative suboptimality guarantees.

2 ROAD: ADAPTIVE OFFLINE DATA REPLAY FOR OFFLINE-TO-ONLINE RL

We introduce **R**einforcement learning with **O**ptimized **A**daptive **D**ata-mixing (**ROAD**), a framework for adaptively replaying offline data in offline-to-online reinforcement learning (RL). ROAD dynamically adjusts the mixing ratio of offline and online data based on a metric quantifying the agent's policy quality relative to these two sources, enhancing learning efficiency and minimizing manual tuning. ROAD is algorithm-agnostic and integrates seamlessly with value-based RL methods. ROAD quantifies the relative policy quality using the following metric:

$$\begin{aligned} R_q =& \underbrace{\mathbb{E}_{s,a\sim\mathcal{D}_{\text{offline}}}[Q_\phi(s,a)] - \mathbb{E}_{s,a\sim\mathcal{D}_{\text{offline}}}[Q_\phi(s,a)]}_{\text{Offline Policy Quality}} \\ &- \underbrace{\mathbb{E}_{s,a\sim\mathcal{D}_{\text{online}}}[Q_\phi(s,a)] - \mathbb{E}_{s,a\sim\mathcal{D}_{\text{online}}}[Q_\phi(s,a)]}_{\text{Online Policy Quality}} \end{aligned} \tag{1}$$

where R_q measures the policy's effectiveness relative to offline and online data. ROAD dynamically adjusts the offline-to-online data mixing ratio m by optimizing R_q using a bandit-based mechanism with Upper Confidence Bound (UCB) exploration. This ensures an adaptive balance between leveraging high-quality offline data and exploring dynamic online environments.

ROAD maintains a value estimate $\bar{R}_{q,m}$ for each ratio m, calculated as the average observed R_q values, and selects the next ratio

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Table 1: Performance of ROAD and the baseline methods for offline data replay under Antmaze tasks, Locomotion tasks and Kitchen tasks. The best performance for fixed mixing ratios is underlined, and the best-performed score is bolded.

	Fixed Mixing Ratios									
Tasks	0.0	0.1	0.2	0.3	0.4	0.5	Uniform	Decreasing	BR	ROAD
antmaze-large-diverse	56.98±3.33	60.32±4.08	55.98±2.26	56.01±2.29	50.66±3.79	50.35±0.86	60.83±4.76	52.15±0.25	46.83±6.74	**63.13±2.24**
antmaze-large-play	58.16±2.63	46.51±1.63	53.16±5.70	53.66±0.61	46.82±2.02	55.12±3.63	46.32±9.76	52.17±1.28	38.34±1.24	**59.75±2.97**
antmaze-medium-diverse	80.00±4.55	78.68±4.48	80.50±5.00	81.66±2.01	82.67±1.85	83.29±3.71	82.30±2.01	81.82±2.52	81.84±0.75	**83.67±1.39**
antmaze-medium-play	82.50±3.11	79.17±2.40	77.32±1.92	80.16±1.66	82.00±2.95	78.62±3.53	82.67±0.26	77.16±3.02	80.99±0.99	**83.49±3.62**
antmaze-umaze-diverse	63.49±13.79	69.66±15.78	7.83±2.24	38.50±33.15	46.51±21.76	16.13±8.51	25.65±10.24	48.82±23.49	**80.66±3.99**	72.12±23.29
antmaze-umaze	92.16±0.94	93.32±0.96	93.16±0.85	93.34±1.32	94.33±0.47	94.37±0.24	91.32±1.76	92.99±0.50	94.33±0.75	**95.83±0.30**
halfcheetah-random	41.48±2.21	43.50±6.28	48.25±3.30	39.62±3.10	43.71±6.07	44.39±4.37	45.91±0.06	42.72±0.22	47.43±1.12	**49.37±3.57**
halfcheetah-medium-replay	50.70±3.38	54.54±1.81	52.55±1.72	51.55±1.90	50.11±1.14	47.41±0.34	50.07±0.05	53.55±3.13	49.28±2.29	**55.82±1.07**
halfcheetah-medium	69.61±1.58	73.53±1.74	72.23±1.72	69.47±4.25	69.39±2.61	67.87±1.08	67.56±4.61	70.46±0.30	72.49±1.51	**74.57±1.95**
halfcheetah-medium-expert	63.75±5.65	92.75±1.40	92.45±1.48	94.52±1.13	93.07±1.16	91.54±5.56	92.83±1.18	93.88±1.70	93.34±2.33	**95.06±0.55**
halfcheetah-expert	78.28±3.58	94.17±1.65	94.64±0.66	95.65±0.36	95.94±0.47	96.61±0.45	93.83±1.76	93.15±0.32	94.91±0.77	**96.86±0.39**
walker2d-random	6.57±1.50	8.72±2.36	8.13±0.38	10.00±1.52	9.20±1.08	10.65±1.66	8.58±0.31	9.56±1.00	7.58±0.86	**12.43±0.69**
walker2d-medium-replay	46.55±24.38	34.56±7.61	34.94±3.47	63.26±13.31	43.33±9.41	40.98±20.32	57.70±15.01	39.16±2.18	70.26±7.95	**78.15±15.74**
walker2d-medium	50.28±7.17	59.07±14.22	56.72±9.78	44.59±9.07	55.74±16.19	31.52±8.42	39.78±1.67	**64.30±10.13**	40.11±19.46	60.17±8.51
walker2d-medium-expert	79.40±11.42	76.15±13.66	68.21±6.43	73.54±11.20	62.06±28.48	60.48±4.65	59.12±22.35	82.93±7.30	81.88±6.36	**83.60±13.06**
walker2d-expert	86.02±12.41	**89.75±10.80**	41.59±23.56	56.49±18.25	73.62±9.23	58.19±9.67	52.01±8.22	74.28±6.07	79.64±2.37	88.06±2.09
hopper-random	17.43±10.05	18.54±6.79	19.71±2.87	12.82±3.29	13.51±5.57	15.51±5.41	13.77±3.61	19.80±1.37	21.48±2.84	**31.77±4.44**
hopper-medium-replay	54.46±29.79	80.14±25.89	67.24±13.88	71.00±25.77	67.35±9.05	77.74±29.92	97.93±13.56	77.12±9.42	88.27±24.20	**99.14±9.97**
hopper-medium	87.25±19.06	81.84±29.83	61.15±18.09	77.46±18.12	87.85±7.86	73.30±22.99	86.17±2.53	85.37±13.13	88.56±20.03	**99.23±1.05**
hopper-medium-expert	46.63±4.13	49.86±8.83	47.01±12.77	75.97±29.32	57.68±12.16	73.46±22.75	66.02±23.94	61.76±4.93	66.21±4.65	**94.08±16.78**
hopper-expert	71.28±24.84	72.74±12.96	62.33±14.68	57.80±3.67	52.69±10.61	54.07±1.54	50.81±5.70	58.80±4.32	39.62±5.91	**85.10±4.27**
kitchen-partial	4.16±13.27	38.17±11.61	39.55±4.71	23.76±9.43	18.74±14.94	41.87±18.71	22.04±0.00	32.93±28.74	29.99±29.37	**42.65±5.13**
kitchen-mixed	0.41±0.59	44.34±1.58	45.44±8.12	40.02±7.12	39.14±15.11	45.00±9.17	55.83±11.90	29.58±13.81	47.56±0.64	**56.62±6.18**
kitchen-complete	10.04±1.56	51.65±10.22	52.07±6.64	47.08±12.38	19.59±24.76	41.86±11.94	53.75±13.69	21.65±16.24	41.64±27.49	**66.25±17.63**
Average	54.07	62.15	57.28	58.66	56.49	56.26	58.45	59.00	61.80	**71.95**

m_t as:

$$m_t = \arg\max_{m'} \left\{ \bar{R}_{q,m'} + \sqrt{\frac{c \log(t \wedge \tau)}{N_t(\tau, m')}} \right\}, \tag{2}$$

where $N_t(\tau, m')$ tracks the selection count for m' within a sliding window τ, and c controls exploration. ROAD adaptively samples batches with $m \times 100\%$ offline and $(1 - m) \times 100\%$ online data for gradient updates, dynamically optimizing the mixing ratio for changing task demands. ROAD is compatible with standard offline-to-online RL pipelines, where the agent is pretrained using offline data $\mathcal{D}_{\text{offline}}$ and fine-tuned through interactions with the environment, generating online data $\mathcal{D}_{\text{online}}$. ROAD evaluates R_q, updates UCB estimates, and adjusts m at the end of each episode, ensuring effective replay patterns.

3 EXPERIMENTS

We evaluate ROAD's effectiveness in offline-to-online RL through diverse benchmarks, including AntMaze, MuJoCo Locomotion, and FrankaKitchen tasks. In AntMaze [2], an ant robot navigates mazes to reach predefined goals, with rewards assigned as +1 for success and 0 otherwise. The MuJoCo Locomotion tasks [10], such as HalfCheetah, Walker2D, and Hopper, simulate dynamic robotic movements under varying offline data qualities, ranging from random to expert. FrankaKitchen tasks [9] involve controlling a robotic arm to complete sub-tasks in a kitchen environment, rewarding agents for successful task execution. To benchmark ROAD, we compare it with fixed-ratio strategies, which statically mix offline and online data, and decreasing-ratio methods, which linearly reduce the offline ratio from 50% to 10%. Additionally, we test uniform sampling of ratios from $\mathcal{M}$ and a balanced replay approach that samples based on density ratio estimates [7]. These baselines represent common strategies for integrating offline data during online training, but often require manual tuning to adapt to specific tasks and datasets.

ROAD consistently outperforms baselines across a variety of tasks and environments. Table 1 highlights the variability of fixed-ratio strategies, which perform well in some settings but fail in others. Decreasing and uniform sampling improve robustness but cannot consistently match the best fixed ratio. Balanced replay occasionally surpasses fixed strategies but lacks generalizability. In contrast, ROAD dynamically adjusts the mixing ratio, achieving robust and adaptive performance across tasks and data qualities.

4 CONCLUSIONS AND FUTURE WORK

We presented ROAD, an adaptive data replay strategy for offline-to-online RL that dynamically adjusts mixing ratios based on policy quality relative to offline and online data. ROAD leverages a bandit-based mechanism to optimize data utilization, reducing manual tuning and enhancing adaptability across diverse tasks, datasets, and algorithms. Experimental results demonstrate ROAD's robustness and effectiveness in improving learning outcomes through dynamic replay patterns.

Despite its strengths, ROAD has limitations. The selection of mixing ratios is restricted to a predefined set $\mathcal{M}$, which, while effective, limits flexibility. Future work could explore more dynamic selection strategies or advanced bandit algorithms to enhance adaptability. Additionally, ROAD's reliance on the quality of offline data highlights a potential risk: poor or misleading data could negatively impact performance. Addressing this issue through data quality assessment or robust policy evaluation techniques could further improve ROAD's reliability in real-world applications.

REFERENCES

[1] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34, 6 (2017), 26–38.

[2] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219* (2020).

[3] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. 2018. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[4] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2021. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169* (2021).

[5] Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. 2023. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning* 112, 7 (2023), 2433–2467.

[6] Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. 2022. Exploration in deep reinforcement learning: A survey. *Information Fusion* 85 (2022), 1–22.

[7] Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. 2022. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*. PMLR, 1702–1712.

[8] Nina Mazyavkina, Sergey Sviridov, Sergei Ivanov, and Evgeny Burnaev. 2021. Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research* 134 (2021), 105400.

[9] Mitsuhiko Nakamoto, Yuexiang Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. 2023. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *arXiv preprint arXiv:2303.05479* (2023).

[10] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 5026–5033.

[11] Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin Riedmiller. 2017. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817* (2017).

[12] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. 2021. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems* 34 (2021), 27395–27407.

[13] Yang Yu. 2018. Towards Sample Efficient Reinforcement Learning.. In *IJCAI*. 5739–5743.

[14] Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. 2021. Sample efficient reinforcement learning with REINFORCE. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10887–10895.