CoMoU: A Trust-region Model-based Method for Efficient Offline-to-online Reinforcement Learning

Extended Abstract

Dongxiang Chen
Shanghai Jiao Tong University
Shanghai, China
juicetech@163.com

Ying Wen*
Shanghai Jiao Tong University
Shanghai, China
ying.wen@sjtu.edu.cn

Rui Chen
Hohai University
Nanjing, China
2214010102@hhu.edu.cn

Kechen Li
Donghua University
Shanghai, China
2252236@mail.dhu.edu.cn

ABSTRACT

Applying dynamics model is a promising approach to further enhance the sample efficiency of offline-to-online reinforcement learning (O2O RL). The current obstacle is that distribution shift can cause drastic and erroneous updates to dynamics model, just as it has a negative impact on policy. We propose to apply trust-region constraint to the online fine-tuning of offline-trained dynamics model. We prove the avoidance of performance crashes and the unbiasedness of model updates. Experiments verify the enhanced sample efficiency and the optimal asymptotic performance.

KEYWORDS

Model-based Reinforcement Learning; Offline-to-online

ACM Reference Format:
Dongxiang Chen, Ying Wen, Rui Chen, and Kechen Li. 2026. CoMoU: A Trust-region Model-based Method for Efficient Offline-to-online Reinforcement Learning: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/JCLP2243

1 INTRODUCTION

Model-based reinforcement learning (MBRL) shows excellent sample efficiency, which benefits from dynamics model's learning transition dynamics, providing more diverse samples for policy updates [2, 3, 5, 7, 9, 10]. Yet, it is hard to apply MBRL to O2O RL [4, 6, 13] to achieve higher sample efficiency, because distribution shift will destroy the fragile dynamics model, thereby leading to performance collapse. We apply trust-region constraint [8] to model updates, achieving stable offline-to-online transfer of dynamics model. We prove the boundedness of the performance difference before and after one update, as well as the unbiasedness of constrained model updates, which ensures the avoidance of performance collapse and the optimal asymptotic performance. We design the complete method, **Co**nstrained **Mo**del **U**pdate (CoMoU). Experiments in representative benchmark show that CoMoU achieves superior sample efficiency and asymptotic performance compared to SOTA methods.

2 COMOU: CONSTRAINED MODEL UPDATE

2.1 Methodology

When we update the parametric transition dynamics $p_\phi(\cdot|s,a)$ to fit the sample transition probability $\rho^{\mathcal{D}}(\cdot|s,a)$ according to the replay buffer $\mathcal{D}$, we constrain its variation to not exceed the trust region specified by ϵ. This forms the optimization objective: $\min_\phi D_{\mathrm{KL}}(\rho^{\mathcal{D}}(\cdot|s,a)||p_\phi(\cdot|s,a)), \text{s.t.} D_{\mathrm{KL}}(p_{\phi_{\mathrm{old}}}(\cdot|s,a)||p_\phi(\cdot|s,a)) \leq \epsilon$, where $p_{\phi_{\mathrm{old}}}$ denotes the old parameters. We obtain the closed-form solution by applying the KKT conditions, and ultimately formulate the iterative update formula for the dynamics model:

$$p_{\phi_{t+1}}(s'|s,a) \leftarrow \frac{\lambda \cdot p_{\phi_t}(s'|s,a)}{\lambda+1} + \frac{\rho^{\mathcal{D}_{t+1}}(s'|s,a)}{\lambda+1}, \tag{1}$$

where λ is a hyper-parameter and t is the index variable of updates. Based on Equation (1), the complete method we propose is shown in Figure 1, which is named **co**nstrained **mo**del **u**pdate (CoMoU).

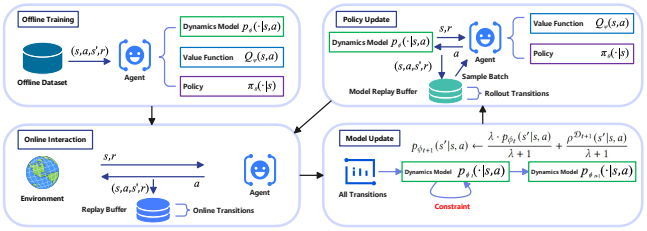


Figure 1: Complete method of CoMoU.

2.2 Theoretical Properties

We prove in Theorem 2.1 that after a model update, there is a controllable upper bound on the difference in the expected return $\eta_M(\pi^*_{\hat{M}_t})$ of the optimal policy $\pi^*_{\hat{M}_t}$ of the model $\hat{M}_t$ in the real environment M, where $\triangle\eta_{M,t} = |\eta_M(\pi^*_{\hat{M}_t}) - \eta_M(\pi^*_{\hat{M}_{t+1}})|$:

*Correspondence to Ying Wen <ying.wen@sjtu.edu.cn>.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/JCLP2243

Theorem 2.1. *If every model update is done under trust-region constraint, the difference of expected returns* $\triangle\eta_{M,t} = |\eta_M(\pi^*_{\hat{M}_t}) - \eta_M(\pi^*_{\hat{M}_{t+1}})|$ *will be bounded:*

$$\begin{aligned}\triangle\eta_{M,t} \leq & \sqrt{2\epsilon}(r_{\max} + V_{\max})H + \\ & [\mathbb{E}_{(s,a)\sim\rho^{\pi^*_{\hat{M}_t}}_{\hat{M}_t}}[\delta_{\ell_1}(p_{\hat{M}_t}(\cdot|s,a), p_M(\cdot|s,a))] + \\ & \mathbb{E}_{(s,a)\sim\rho^{\pi^*_{\hat{M}_{t+1}}}_{\hat{M}_{t+1}}}[\delta_{\ell_1}(p_{\hat{M}_{t+1}}(\cdot|s,a), p_M(\cdot|s,a))]] \cdot \frac{\gamma}{2}V_{\max},\end{aligned} \tag{2}$$

H is episodic length, $r_{\max} = \max_{s,a} r(s,a)$. $p_{\hat{M}}, p_M$ are the transition dynamics of $\hat{M}, M$. $V_{\max} = \max_{s\in\mathcal{S},\pi\in\Pi} V^\pi(s)$ (Π is policy space). $\delta_{\ell_1}(\cdot,\cdot)$ is L1-norm distance. ρ^π_M is discounted occupancy measure.

We prove in Theorem 2.2 that $p_{\phi_t}(\cdot|s,a)$ will converge to $p(\cdot|s,a)$ if the convergence process of $\rho^{\mathcal{D}_t}(s'|s,a)$ to $p(\cdot|s,a)$ satisfies that:

$$\forall(s,a,s'), \lim_{t\to\infty} \max_{t'\geq t} |\rho^{\mathcal{D}_{t'}}(s'|s,a) - p(s'|s,a)| = 0. \tag{3}$$

Theorem 2.2. *$\forall(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, the difference between the predicted probability $p_{\phi_t}(s'|s,a)$ and the real probability $p(s'|s,a)$ converges to 0 as the number of constrained model updates t increases:*

$$\forall(s,a,s'), \lim_{t\to\infty} |p_{\phi_t}(s'|s,a) - p(s'|s,a)| = 0. \tag{4}$$

In summary, Theorem 2.1 ensures that after each model update, the performance will not experience a sudden drastic decline, thereby preventing performance collapses; Theorem 2.2 ensures that trust-region constraint does not introduce bias into the estimation of real transition dynamics, thereby safeguarding the optimality of asymptotic performance.

3 EXPERIMENTS

We empirically analyze CoMoU's performance. For reproducibility, we employ four types of datasets from the openly-sourced D4RL [1]: Medium, Medium-Replay, Medium-Expert and Random. We use MOPO [10] to provide an offline-trained initialization. Our analyses focus on three MuJoCo locomotion environments: Hopper, Walker2d, and HalfCheetah. We compare to SOTA counterparts ACA [11], PEX [12], ODT [14], MOORe [5].

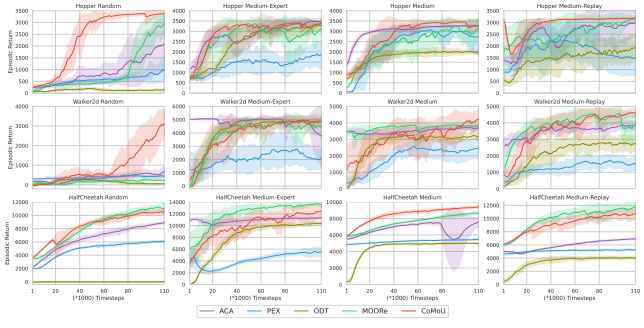


Figure 2: Performance comparison between CoMoU and SOTA methods.

The performance curves averaged over 3 random seeds are shown in Figure 2. Obviously, without sacrificing final performance, CoMoU rapidly and stably enhances its performance to high level in all the environment-datasets. We also compare to learning from scratch (MBPO) and directly finetuning (Naive Tuning) the models pretrained by MOPO through MBPO. The performance curves are shown in Figure 3.

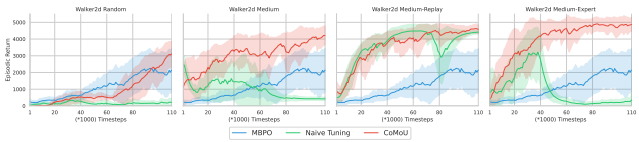
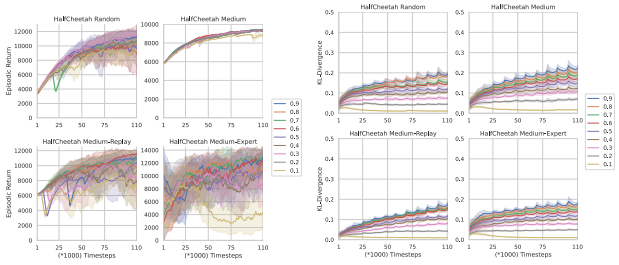


Figure 3: Necessity of CoMoU.

We find that CoMoU achieves faster performance improvements compared to MBPO, which confirms the effectiveness of CoMoU. Without trust-region constraint applied to dynamics model, performance collapses occur, which illustrates the necessity of CoMoU. We conduct an ablation analysis on hyper-parameter in HalfCheetah environment. For convenience, we directly specify $\tilde{\lambda} = \frac{1}{\lambda+1}$ instead of λ. The performance curves under different $\tilde{\lambda}$ are shown in Figure 4a. The curves of $D_{\mathrm{KL}}(p_{\phi_t}||p_{\phi_{t+1}})$ are shown in Figure 4b. We find that CoMoU achieves good performance under most $\tilde{\lambda}$. The changes of transition dynamics show obvious bounds under all $\tilde{\lambda}$.



(a) Performance under different $\tilde{\lambda}$. **(b) KL-divergence under different $\tilde{\lambda}$.**

Figure 4: Ablation analysis of $\tilde{\lambda}$.

4 CONCLUSION

We study improving offline-trained policies through online finetuning. Analyses reveal that by modulating the differences between consecutive dynamics models, we can manage the discrepancies in real-world performance of the optimal policies under these models, mitigating the negative effects of distribution shift. We develop a trust-region model update mechanism and integrate it into a complete MBRL method, CoMoU. Through rigorous examination, we confirm that this mechanism introduces no biases and the performance differences are bounded. Experiments show that CoMoU stably transfers offline-trained policies to online learning stage and achieve higher sample efficiency than SOTA methods.

ACKNOWLEDGMENTS

This work is supported by Shanghai Sailing Program (21YF1421900) and NSFC (62106141).

REFERENCES

[1] 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).

[2] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. 2019. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems* 32 (2019).

[3] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2020. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems* 33 (2020), 21810–21823.

[4] 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.

[5] Yihuan Mao, Chao Wang, Bin Wang, and Chongjie Zhang. 2022. MOORe: Model-based Offline-to-Online Reinforcement Learning. *arXiv preprint arXiv:2201.10070* (2022).

[6] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2020. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359* (2020).

[7] Rafael Rafailov, Kyle Beltran Hatch, Victor Kolev, John D Martin, Mariano Phielipp, and Chelsea Finn. 2023. MOTO: Offline to Online Fine-tuning for Model-Based Reinforcement Learning. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*.

[8] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*. PMLR, 1889–1897.

[9] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. 2021. Combo: Conservative offline model-based policy optimization. *Advances in Neural Information Processing Systems* 34 (2021).

[10] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. 2020. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems* 33 (2020), 14129–14142.

[11] Zishun Yu and Xinhua Zhang. 2023. Actor-critic alignment for offline-to-online reinforcement learning. In *International Conference on Machine Learning*. PMLR, 40452–40474.

[12] Haichao Zhang, We Xu, and Haonan Yu. 2023. Policy Expansion for Bridging Offline-to-Online Reinforcement Learning. *arXiv preprint arXiv:2302.00935* (2023).

[13] Kai Zhao, Yi Ma, Jinyi Liu, Yan Zheng, and Zhaopeng Meng. 2023. Ensemble-based Offline-to-Online Reinforcement Learning: From Pessimistic Learning to Optimistic Exploration. *arXiv preprint arXiv:2306.06871* (2023).

[14] Qinqing Zheng, Amy Zhang, and Aditya Grover. 2022. Online decision transformer. In *international conference on machine learning*. PMLR, 27042–27059.