Incorporating Low-Rank Structures into Offline Deep Reinforcement Learning

Introduction

Offline reinforcement learning (Offline RL) aims to learn optimal policies from static datasets collected by a fixed behavior policy π_β, without any environment interaction. This is critical for domains where data collection is costly or risky, such as robotics or healthcare. However, offline RL introduces challenges such as distributional shift between π and π_β, poor dataset coverage, and Q-value overestimation. These challenges are illustrated by Fujimoto et al.(2019). Result is showing in the figure 1, which compares the performance of DDPG in offline (blue) and off-policy (orange) settings.

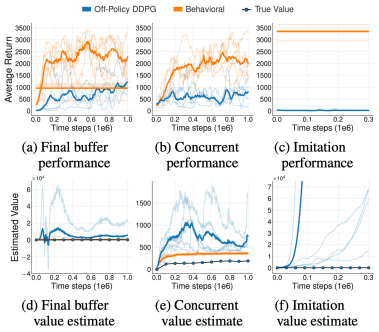


Figure 1: Offline DDPG vs. behavior policy.

Recent empirical findings in deep reinforcement learning suggest that Q-functions learned through neural approximators often evolve toward lower-rank representations during training. This phenomenon is highlighted in Figure 2, which illustrates the approximation rank of Q-matrices across time for SAC, TD3, and DDPG agents in MuJoCo continuous control tasks. As training progresses, the Q-matrix rank consistently decreases across all tasks and algorithms, suggesting that value functions gradually compress onto lower-dimensional subspaces. This supports the hypothesis that low-rank structure is not only common but may also be a consequence of optimization dynamics in deep RL.

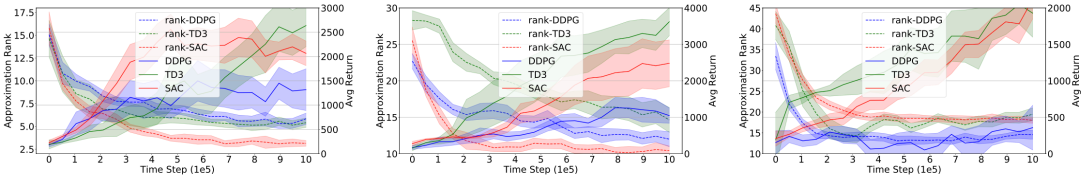


Figure 2: Low-rank structure of Q-matrix in MuJoCo continuous control tasks

Leveraging this low rank structure offers several advantages: low-rank Q-functions allow more efficient approximation of the value landscape, enabling faster learning with better utilization of limited data. The structural assumption also facilitates generalization by guiding the optimization toward smooth, compressible solutions, particularly in settings with high-dimensional state-action spaces or limited coverage.

Method

We investigate the UA-LQE-SAC framework as a representative method to incorporate low-rank structure into offline RL. UA-LQE-SAC builds on the SAC algorithm by augmenting its critic update using uncertainty-guided Q-matrix reconstruction. For each training step, a batch (s, a, r, s') is sampled from the dataset $\mathcal{D}$, and a bootstrapped critic ensemble is used to estimate epistemic uncertainty over the TD targets $r + \gamma Q(s', a')$.

A mask M is then constructed by thresholding the variance across the ensemble, retaining only reliable entries. The masked Q-matrix is completed using a low-rank matrix estimation method (e.g., SoftImpute or SVT) to form $\hat{Q}^{\text{target}}$. This reconstructed matrix is used to compute a second TD loss: $\mathcal{L}_{\text{td}}^{\text{recon}} = E_{(s,a,r,s')}\left[\left(Q(s,a) - (r + \gamma \hat{Q}^{\text{target}}(s',a'))\right)^2\right]$.
The final critic loss is then a weighted combination of the original TD loss and this reconstruction-based loss: $\mathcal{L}_{\text{critic}} = \alpha \cdot \mathcal{L}_{\text{td}}^{\text{orig}} + (1-\alpha) \cdot \mathcal{L}_{\text{td}}^{\text{recon}} + \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}}$, where α controls the interpolation between original and low-rank reconstructed targets, and $\mathcal{L}_{\text{rec}}$ is the matrix reconstruction loss. Actor and entropy parameters are updated using the standard SAC objective, and target networks and the uncertainty model are updated periodically.

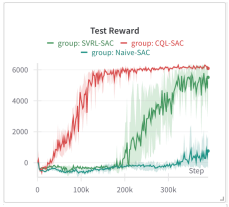
Experiment and Result

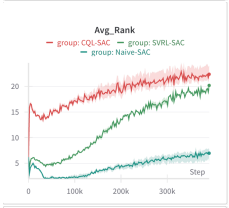
We evaluate our algorithm on the HalfCheetah-medium-v2 benchmark from D4RL, a continuous control task featuring a 17-dimensional state space and 6-dimensional action space. The dataset was collected from a medium-quality SAC-trained policy.
Our empirical evaluation focused on three representative algorithms: naive SAC, CQL-SAC (as strong baselines), and SVRL-SAC (our primary method of interest). SVRL-SAC introduces a structured value estimation framework by applying random masking and low-rank matrix reconstruction to the critic's Q-matrix. Our experimental findings are summarized as follows:

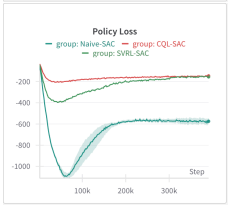
1. **Performance Comparison:** SVRL-SAC consistently outperformed naive SAC and achieved competitive final returns comparable to CQL-SAC across the evaluated benchmarks. This indicates its effectiveness in learning high-quality policies.

2. **Rank Dynamics of the Q-Matrix:** Throughout training, the rank of Q-matrix in SVRL-SAC increased progressively, eventually approaching the similar rank levels observed in CQL-SAC.

3. **Policy Stability and Conservativeness:** The policy loss trajectory of SVRL-SAC converged to similar values as CQL-SAC, indicating similar degrees of stability and conservativeness in policy updates. This supports its capacity to control extrapolation error without excessive pessimism.







Collectively, these results provide empirical support for the efficacy of SVRL-SAC in mitigating extrapolation error and producing more reliable Q-value estimates—especially in unobserved or underrepresented regions of the state-action space. More importantly, the experiments were conducted in a high-dimensional continuous control environment with medium dataset coverage, a regime that aligns closely with the assumptions under which CQL-SAC performs well. We hypothesize that SVRL-SAC may demonstrate even greater robustness in more challenging scenarios, such as sparse-reward problems or low-coverage datasets.

References

[1] Scott Fujimoto, David Meger, and Doina Precup. “Off-Policy Deep Reinforcement Learning without Exploration”. In: *Proceedings of the 36th International Conference on Machine Learning*. 2019. URL: https://proceedings.mlr.press/v97/fujimoto19a.html.

[2] Devavrat Shah et al. *Sample Efficient Reinforcement Learning via Low-Rank Matrix Estimation*. 2020. arXiv: 2006.06135 [cs.LG]. URL: https://arxiv.org/abs/2006.06135.

[3] Xumei Xi, Christina Lee Yu, and Yudong Chen. *Matrix Estimation for Offline Reinforcement Learning with Low-Rank Structure*. 2023. arXiv: 2305.15621 [cs.LG]. URL: https://arxiv.org/abs/2305.15621.

[4] Tong Sang et al. *Uncertainty-aware Low-Rank Q-Matrix Estimation for Deep Reinforcement Learning*. 2021. arXiv: 2111.10103 [cs.LG]. URL: https://arxiv.org/abs/2111.10103.

[5] Yuzhe Yang et al. *Harnessing Structures for Value-Based Planning and Reinforcement Learning*. 2020. arXiv: 1909.12255 [cs.LG]. URL: https://arxiv.org/abs/1909.12255.

[6] Sebastian Dittert. *CQL*. https://github.com/BY571/CQL. 2021.