AISTATS'26

poster pdf

Duality-based Residual Estimation for Fully Offline Value-based Reinforcement Learning

Kohei Miyaguchi（LY Corporation）
kmiyaguc@lycorp.co.jp

Takeaway

Model-free RL

- learning $\pi^*(a|s)$ without learning $P(s'|s,a)$
- often preferred for high-dimensional states (e.g., images, behavioral user profiles)

has been unsuitable for strictly offline settings.

- bottleneck: lack of established method for offline self-hyperparameter tuning
- hence not ready for safety-first applications (e.g., industry, healthcare)

→ We show that **the bottleneck can be resolved.**

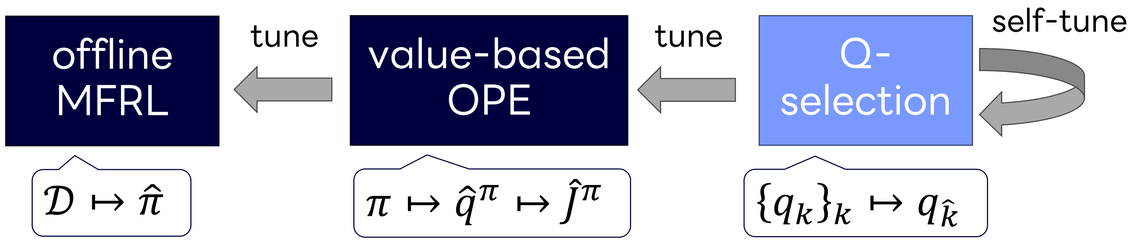
Problem

Problem formulation

Goals:

- Learning policy offline
- No model estimation at any stage
- Capability of offline self-hyperparameter tuning

→ achievable with self-tuning Q-selection:



Weakness of previous methods

They cannot provide clean end-to-end performance guarantee due to

- incompatibility with Bellman residual (common objective of value-based OPE), or
- lack of sample-efficient PAC guarantee

Method	Compatible Q-norm	Error bound
Residual min. [Baird+ '95]	Bellman residual	$O(1)$
BVFT-PE [Zhang & Jiang '21]	L^∞-norm	$O(m^{-1/4})$
LSTD-Tournament [Liu+ '25]	unknown	$O(m^{-1/2})$
DRE [ours]	Bellman residual	$O(m^{-1/2})$

Root cause: double sampling problem

Bellman residual contains squared Bellman operator:

$$R^\pi(q) \coloneqq \sqrt{\mathbb{E}\{q(s,a) - B^\pi q(s,a)\}^2}$$

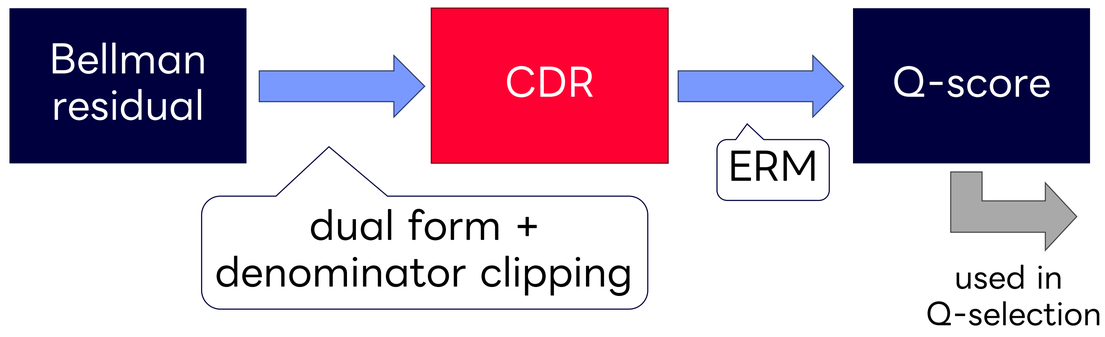
→ Direct estimation requires infamous double sampling, which is infeasible offline ☹

→ Previous workarounds introduce projection-based alternative Q-norms, resulting in compatibility issue.

Proposed Method

Duality-based residual estimation (DRE)

Q-scoring method based on double-sampling-free representation (=CDR) of Bellman residual:



Clipped dual-norm representation (CDR)

A dual form of Bellman residual, stabilized by denominator clipping:

$$R^\pi(q) = \sup_{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \frac{\mathbb{E}\, f(s,a)\,\{q(s,a) - B^\pi q(s,a)\}}{\max\left\{1, \sqrt{\mathbb{E} f^2(s,a)}\right\}}$$

- ✓ Clipping prevents infinite variance without introducing bias
- ✓ "sup" can be handled by empirical risk minimization (ERM), including self-hyperparameter tuning

PAC analysis

When DRE is used as objective of Q-selection, w.h.p.,

$$|J(q_{\hat{k}}) - J^\pi| \le \rho^\pi \left(\min_k R^\pi(q_k) + \min_{\xi\in\Xi}\{\epsilon_\xi + O(m^{-1/2})\} \right)$$

OPE error of selected Q | oracle error | estimation error of DRE

- ✓ This is self-tuning bound: best hyperparameters (k and ξ) are auto selected
- ✓ Oracle term (including coef. ρ^π) is minimax optimal
- ✓ DRE-related term is doubly-robust bias ϵ_ξ + sample-efficient variance $O(m^{-1/2})$

Experiment (proof of concept)

OPE error vs. sample size on FrozenLake-v1

- ✓ DRE-based estimator (red line) successfully tracks best Q among candidates (blue lines).