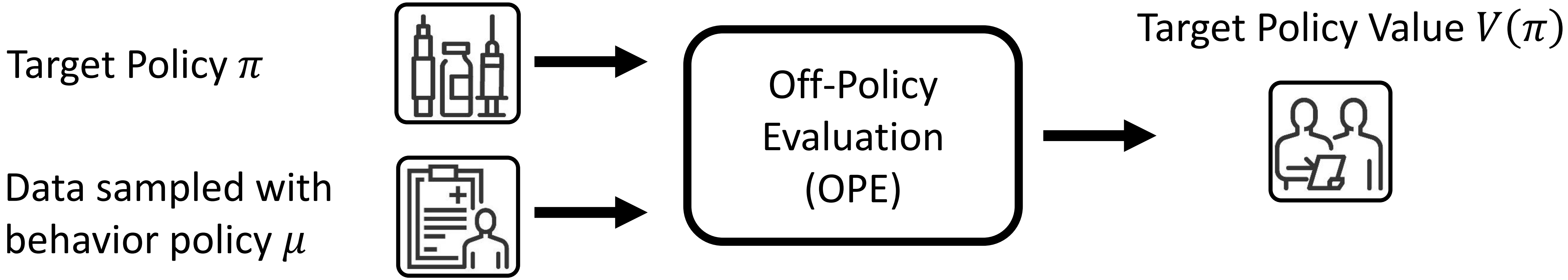
Kernel Metric Learning for In-Sample Off-Policy Evaluation of Deterministic RL Policies

ICLR 2024 | Spotlight

Haanvid Lee[1], Tri Wahyu Guntara[1], Jongmin Lee[2], Yung-Kyun Noh[3,4], Kee-Eung Kim[1]

1 KAIST
2 Berkeley UNIVERSITY OF CALIFORNIA
3 HANYANG UNIVERSITY 한양 1939
4 KIAS KOREA INSTITUTE FOR ADVANCED STUDY

Off-Policy Evaluation of Deterministic Policies



- OPE is used when interaction with an environment is expensive or dangerous

  For example, OPE can be used to predict the effects of

  - Medical drug prescribing policies
  - Policies controlling the durations or intensities of users' exposure to interventions
  - Dynamic pricing policies

- We focus on the OPE of deterministic policy since greedy (deterministic) policies are used in real-world applications

Extrapolation Error

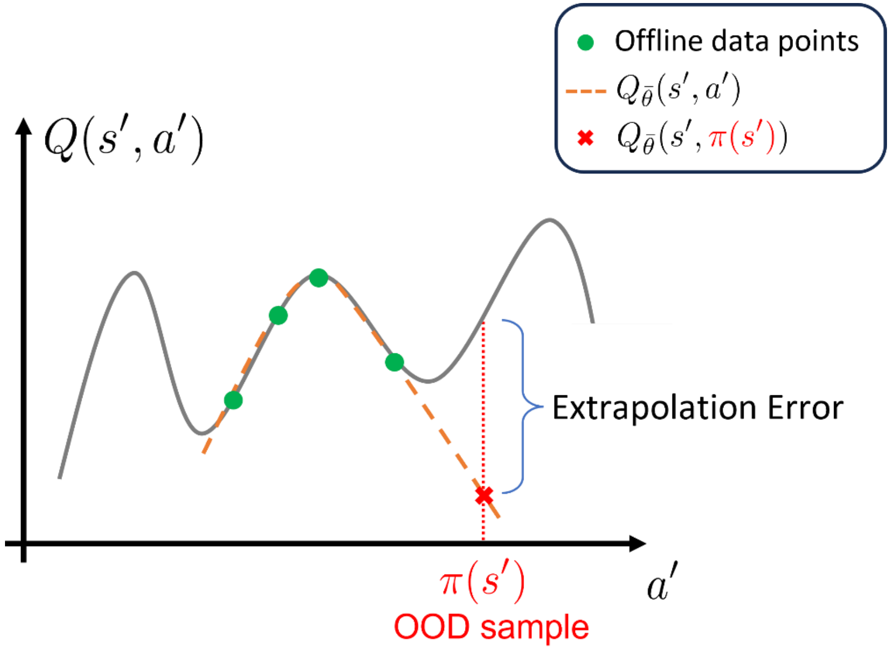
Extrapolation error occurs when using bootstrapping with out-of-distribution (OOD) samples to estimate functions such as Q-functions for OPE.

- Deterministic policy π can be evaluated as $\mathbb{E}\left[Q_\theta(s_0, \pi(s_0))\right]$.

- Fitted Q Evaluation (FQE) objective

$$\min_\theta \mathbb{E}_{(s,a,r,s')\sim p_\mu}\left[\left(Q_\theta(s,a) - (r + \gamma Q_{\bar{\theta}}(s', \pi(s')))\right)^2\right]$$

- Extrapolation error due to querying $Q_{\bar{\theta}}(s', \pi(s'))$ to fit $Q_\theta(s,a)$.

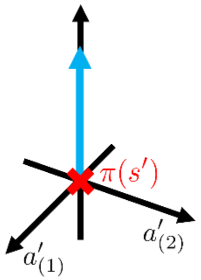


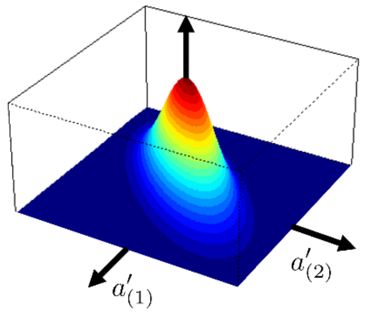
In-Sample FQE with Kernel-Based Importance Resampling

Importance resampling can be applied on FQE to avoid using OOD samples.

- Kernel relaxation

$$\delta\left(a' - \pi(s')\right) \longrightarrow K(a', \pi(s'))$$





In-Sample FQE with Kernel-Based Importance Resampling

Importance resampling can be applied on FQE to avoid using OOD samples.

- Kernel relaxation

$$\delta\left(a' - \pi(s')\right) \longrightarrow K(a', \pi(s'))$$

- Importance resampling probability

$$\rho_j^K := \frac{w^K\left(s_j', a_j'\right)}{\sum_{i=1}^n w^K\left(s_i', a_i'\right)}, \text{ where } w^K(s', a') := \frac{K(a', \pi(s'))}{\mu\left(a' \mid s'\right)}$$

- FQE with importance resampling

$$\min_\theta \mathbb{E}_{D\sim p_\mu, (s,a,r,s',a')\sim\rho^K(\cdot|D)}\left[\frac{1}{n}\sum_{i=1}^n w^K\left(s_i', a_i'\right)\left(Q_\theta(s,a) - \left(r + \gamma Q_{\bar\theta}\left(s', a'\right)\right)\right)^2\right]$$

- Kernel metrics can be learned to assign high importance resampling probabilities on the transitions that are helpful in fitting a Q-function.

Kernel Metric Learning for In-Sample FQE (KMIFQE)

- For convenience, the scale of the kernel metric is referred to as bandwidth h, and the shape of the metric is referred to as metric $A(s)$.
- Gaussian kernel with metrics $A(s)$ and bandwidths h is used.

$$K\left(a', \pi(s')\right) := \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\left(a' - \pi(s')\right)^\top A(s') \left(a' - \pi(s')\right)}{2h^2}\right),$$

$$|A(s')| = 1, A(s')^\top = A(s'), A(s') \succ 0$$

- We aim to accurately estimate the Q-function update vector from the bootstrapping objective by kernel-based importance resampling with learned kernel metrics and bandwidths.

$$\widehat{\Delta} := \frac{1}{nk} \sum_{i=1}^{n} w^K(s_i', a_i') \sum_{j=1}^{k} \left(r_j + \gamma Q_{\bar{\theta}}\left(s_j', a_j'\right) - Q_\theta(s_j, a_j)\right) \nabla_\theta Q_\theta(s_j, a_j),$$

$$\text{where } (s_j, a_j, r_j, s_j', a_j') \overset{\rho_j^K}{\sim} D$$

Optimal Bandwidth

MSE with h and $A(s) = I$

$$\mathrm{MSE}(h,n,k,d) = \underbrace{h^4\|\mathbf{b}\|_2^2}_{(\text{leading-order bias})^2} + \underbrace{\frac{v}{nh^d}}_{(\text{leading-order variance})} + O\left(h^6\right) + O\left(\frac{1}{nh^{d-2}}\right) + O\left(\frac{1}{k}\right),$$

$$\mathbf{b} := \frac{\gamma}{2}\mathbb{E}_{\mathcal{D}\sim p_\mu}\left[\nabla_{a'}^2 Q_{\bar\theta}(s',a')|_{a'=\pi(s')}\nabla_\theta Q_\theta(s,a)\right],$$

$$v := (4\pi)^{-\frac{d}{2}}\mathbb{E}_{p_\mu}\left[\frac{\left(r+\gamma Q_{\bar\theta}\left(s',\pi\left(s'\right)\right)-Q_\theta(s,a)\right)^2\|\nabla_\theta Q_\theta(s,a)\|_2^2}{\mu\left(\pi\left(s'\right)\mid s'\right)}\right]$$

- Leading-order MSE minimizing optimal bandwidth h^*

$$h^* = \left(\frac{vd}{4n\|\mathbf{b}\|_2^2}\right)^{\frac{1}{d+4}}$$

Optimal Metric

With the optimal bandwidth h^*, bias becomes dominant in the MSE in high dimensional action spaces.

$$\lim_{d\to\infty} \mathrm{MSE}\left(h^*, n, k, d\right) \approx \Big\| \underbrace{\frac{\gamma}{2}\mathbb{E}_{\mathcal{D}\sim p_\mu}\left[\nabla^2_{a'} Q_{\bar{\theta}}(s', a')|_{a'=\pi(s')} \nabla_\theta Q_\theta(s, a)\right]}_{=\mathbf{b}} \Big\|_2^2$$

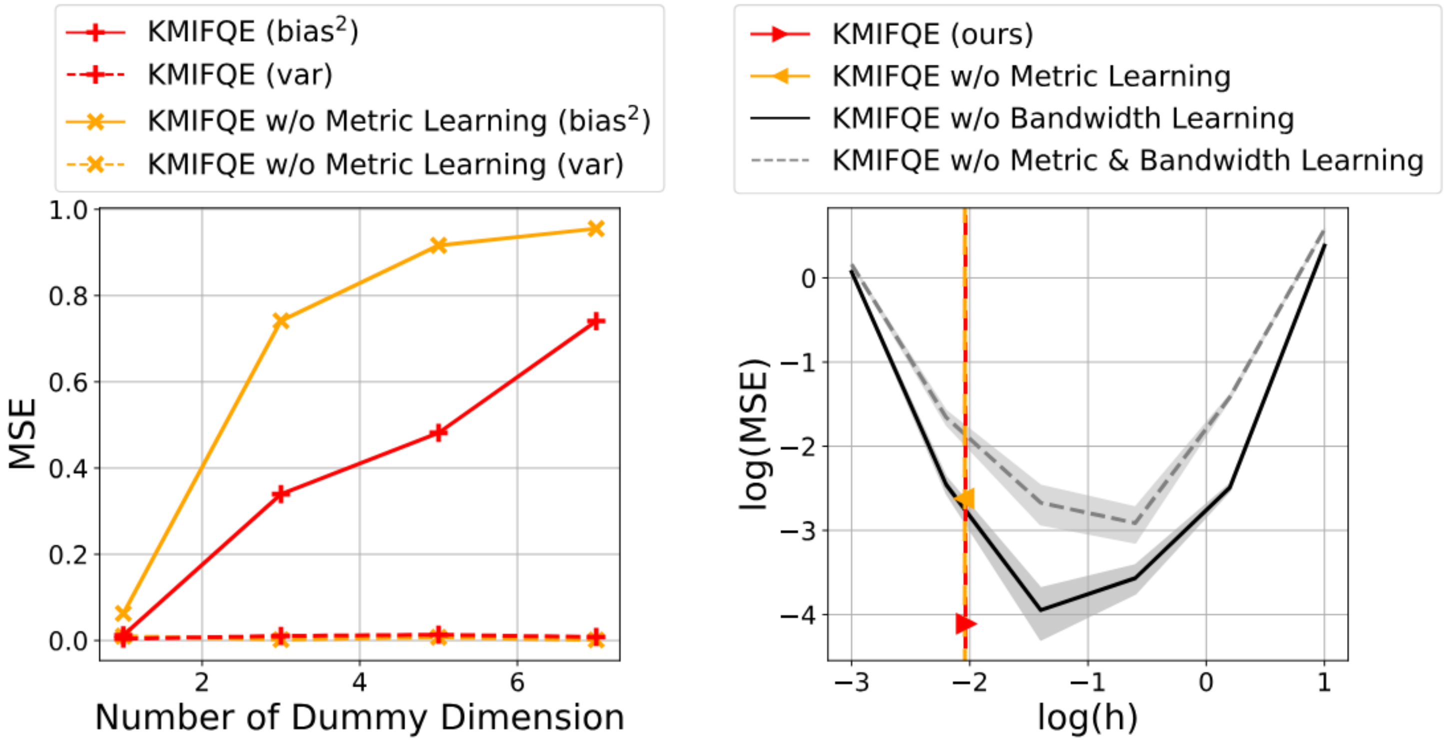
- Bias ($\|\boldsymbol{b}\|_2^2$ with $A(s)$) minimizing optimal metric $A^*(s)$ is the closed-form solution of the objective,

$$\min_{\substack{A:\ A(s')\succ 0,\\ A(s')=A(s')^\top, |A(s')|=1\ \forall s'}} \mathrm{tr}\left(A(s')^{-1} \mathbf{H}_{a'} Q_{\bar{\theta}}(s', a')\Big|_{a'=\pi(s')} \right)^2$$

- As both the optimal metric and bandwidth are dependent on the fitted Q-function, KMIFQE iteratively updates bandwidth, metric, and Q-function until the Q-function converges.

Experiments: Synthetic Data

Synthetic data is generated from the modified inverted pendulum environment with additional dummy action dimensions irrelevant to rewards and next state transitions.



Experiments: MuJoCo Control Tasks

(RMSE)

Dataset	known μ	KMIFQE	KMIFQE w/o Metric	SR-DICE	FQE
Hopper-v2	O	**0.023 ± 0.006**	0.034 ± 0.009	0.129 ± 0.023	0.083 ± 0.011
HalfCheetah-v2	O	2.080 ± 0.010	2.549 ± 0.017	2.784 ± 0.030	**1.637 ± 0.051**
Walker2d-v2	O	**0.032 ± 0.008**	0.048 ± 0.009	0.273 ± 0.054	241.319 ± 49.248
Ant-v2	O	**1.800 ± 0.013**	2.255 ± 0.014	1.996 ± 0.030	3.219 ± 0.736
Humanoid-v2	O	**0.246 ± 0.010**	0.293 ± 0.021	1.285 ± 0.050	8.860 ± 8.196
hopper-m-e-v2	X	**0.019 ± 0.003**	**0.020 ± 0.005**	0.045 ± 0.007	0.033 ± 0.010
halfcheetah-m-e-v2	X	0.418 ± 0.016	0.457 ± 0.007	0.239 ± 0.025	**0.080 ± 0.007**
walker2d-m-e-v2	X	**0.036 ± 0.006**	**0.038 ± 0.006**	0.115 ± 0.017	1.051 ± 0.633
hopper-m-r-v2	X	**0.536 ± 0.099**	**0.517 ± 0.120**	0.849 ± 0.052	**0.561 ± 0.118**
halfcheetah-m-r-v2	X	**4.698 ± 0.044**	4.765 ± 0.026	5.048 ± 0.090	6.394 ± 1.769
walker2d-m-r-v2	X	**1.364 ± 0.052**	**1.360 ± 0.025**	1.523 ± 0.061	86.315 ± 29.206

Thank You!

- Poster | Halle B, Tue 7 May 4:30 p.m. CEST — 6:30 p.m. CEST
- Paper |