A Principled Path to Fitted Distributional Evaluation

Sungee Hong, Jiayi Wang, Zhengling Qi, Raymond K. W. Wong

NeurIPS 2025 Spotlight

Link: https://arxiv.org/abs/2402.01900

October 25, 2025

Distributional OPE

- Distributional off-policy evaluation (OPE) is a method of reinforcement learning (RL), by which we evaluate the return distribution for given target policy π (action-choosing rule for each state) by using offline data.
- For each initial state-action pair $s, a \in \mathcal{S} \times \mathcal{A}$, we estimate the distribution of return. This is useful in assessing distributional attributes of π (e.g., risk).

$$\Upsilon_\pi(s,a) := \underbrace{\mathcal{L}\left(\sum_{t=1}^{\infty} \gamma^{t-1} R^{(t)}\right)}_{\text{return distribution at } s,a}, \text{ where } (R^{(t+1)}, S^{(t+1)}) \sim \underbrace{p(\cdot|S^{(t)}, A^{(t)})}_{\text{environment}}, A^{(t+1)} \sim \underbrace{\pi(\cdot|S^{(t+1)})}_{\text{target policy}}$$

- Bellman operator $\mathcal{T}^\pi : \mathcal{P}^{\mathcal{S}\times\mathcal{A}} \to \mathcal{P}^{\mathcal{S}\times\mathcal{A}}$ gives us convolution of distributional with reward R.

$$Z(s,a) \sim \Upsilon(s,a) \Rightarrow \mathcal{L}(R + \gamma Z(S', A') \mid s, a; \pi) \sim \mathcal{T}^\pi \Upsilon(s,a),$$

$$(\mathcal{T}^\pi)^\infty \Upsilon(s,a) = \Upsilon_\pi(s,a) = \mathcal{L}\left(\sum_{t\geq 1} \gamma^{t-1} R_t \mid s, a, \pi\right) \quad \Rightarrow \quad \text{Underlying idea of iterative algorithms.}$$

Fitted distributional evaluation

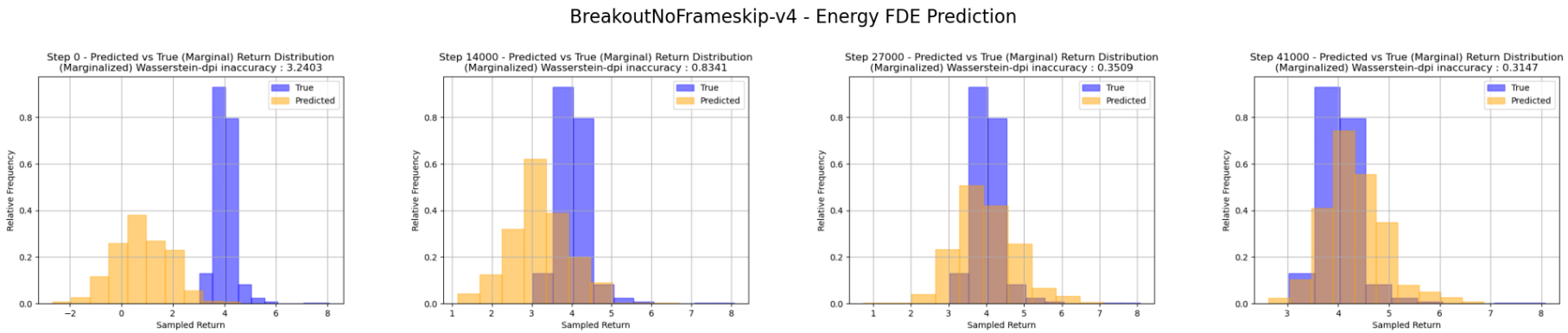
- In traditional OPE, FQE (fitted-Q evaluation) repeats regression to estimate the return expectation $Q_\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,

$$\theta_{t+1} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\left\{Q_\theta(s_i,a_i) - r_i - \gamma\cdot Q_{\theta_t}(s_i',A_i')\mathcal{T}^\pi Q_{\theta_t}\right\}^2 \quad \Rightarrow \quad Q_{\theta_T} \approx (\mathcal{T}^\pi)^T Q_{\theta_0} \approx Q_\pi.$$

- In distributional OPE, FDE (fitted distributional evaluation), we replace MSE with probability divergence $\mathfrak{d} : \mathcal{P}\times\mathcal{P}\to\mathbb{R}$.

$$\theta_{t+1} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\partial\left\{\Upsilon_\theta(s_i,a_i), \Psi(r_i,s_i',\Upsilon_{\theta_t};\pi)\right\} \quad \Rightarrow \quad \Upsilon_{\theta_T} \approx (\mathcal{T}^\pi)^T \Upsilon_{\theta_0} \approx \Upsilon_\pi.$$

- We are the first ones to provide a generalized guideline for FDE methods that can statistically converge.



Principle-1: Contraction-inducing metric

The inaccuracy metric should make $\mathcal{T}^\pi$ contraction.

$$\tilde{\eta}(\mathcal{T}^\pi \Upsilon_1, \mathcal{T}^\pi \Upsilon_2) \leq \zeta \cdot \tilde{\eta}(\Upsilon_1, \Upsilon_2),$$

$$\underbrace{\tilde{\eta}(\Upsilon_T, \Upsilon_\pi)}_{\text{final inaccuracy}} \leq \sum_{t=1}^{T} \zeta^{T-t} \cdot \underbrace{\tilde{\eta}(\Upsilon_t, \mathcal{T}^\pi \Upsilon_{t-1})}_{\text{iteration-level inaccuracy}} + \underbrace{\zeta^T \cdot \tilde{\eta}(\Upsilon_0, \Upsilon_\pi)}_{\text{shrinks to zero as } T \to \infty} .$$

1. Supremum-extended metric (Odin and Charpentier, 2020): for $|\mathcal{S} \times \mathcal{A}| < \infty$

$$\eta_\infty(\Upsilon_1, \Upsilon_2) := \sup_{s,a} \left\{ \eta(\Upsilon_1(s,a), \Upsilon_2(s,a)) \right\} \quad \Rightarrow \quad \zeta = \gamma^c.$$

2. Expectation-extended metric: for $|\mathcal{S} \times \mathcal{A}| = \infty$

$$\bar{\eta}_{d_\pi, q}(\Upsilon_1, \Upsilon_2) := \left[\mathbb{E}_{s,a \sim d_\pi} \left\{ \eta^{2q}(\Upsilon_1(s,a), \Upsilon_2(s,a)) \right\} \right]^{\frac{1}{2q}} \Rightarrow \zeta = \gamma^{c - \frac{1}{2q}},$$

Principle-2: Functional Bregman divergence

Bellman update $\mathcal{T}^\pi \Upsilon_{\theta_t}$ should minimize the population objective function.

$$\mathcal{T}^\pi \Upsilon_{\theta_t} \in \arg \min_{\Upsilon \in \mathcal{P}^{\mathcal{S}\times\mathcal{A}}} \mathbb{E}\left\{ \mathfrak{d}\left(\Upsilon(S,A), \Psi(R,S',\Upsilon_{\theta_t};\pi) \right) \right\}$$

Definition (Functional Bregman divergence)

(Ovcharov, 2018) For a distributional family $\mathcal{P}$ over support $\mathcal{X}$, let $U \subset \text{span}(\mathcal{P})$ be a convex set that contains $\mathcal{P}$. A function $\Phi : U \to \mathbb{R}$ has a subgradient $\Phi^*(P) \in L(\mathcal{P})$ at $P \in U$ if

$$\Phi(Q) \geq \Phi(P) + \Phi^*(P) \cdot (Q - P) \quad \text{for all } Q \in U.$$

Here, $L(\mathcal{P})$ is the collection of $\mathcal{P}$-integrable functions (i.e. $\int_{\mathcal{X}} |f(x)| \mathrm{d}P(x) < \infty$ for every $P \in \mathcal{P}$. For a measurable function $f \in L(\mathcal{P})$, $f \cdot Q := \int_{\mathcal{X}} f(x) \mathrm{d}Q(x)$ for a signed measure Q. A statistical divergence is a functional Bregman divergence if it has such Φ, Φ^* such that

$$D(P,Q) = \Phi(Q) - \Phi(P) - \Phi^*(P) \cdot (Q - P).$$

This opens door to

1. widely-known examples: KLD, squared MMD, squared L2-norm of densities
2. not widely-known examples: survival, spheric, Hyvärinen, Tsallis, Brier divergences.

Principle-3: Association

Following is an example that explains the association between functional Bregman divergence and the inaccuracy metric.

$$\text{Inaccuracy metric}: \bar{\eta}_{d_\pi,q}(\Upsilon_1,\Upsilon_2) := \left[\mathbb{E}_{s,a\sim d_\pi}\left\{\eta^{2q}\left(\Upsilon_1(s,a),\Upsilon_2(s,a)\right)\right\}\right]^{\frac{1}{2q}},$$

$$\text{Objective function of FDE}: \theta_{t+1} = \arg\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\partial\left\{\Upsilon_\theta(s_i,a_i),\Psi(r_i,s_i',\Upsilon_{\theta_t};\pi)\right\}.$$

Theorem (Convergence example)

Let η be a probability metric that satisfies (S-L-C) with $q \geq 1$ and $c > 1/(2q)$. Suppose Assumptions 3.1, 3.2, C.3, and C.4 hold. By letting $T = \lfloor \frac{1}{c-\frac{1}{2q}} \cdot \frac{\delta'}{2(l-1)+\alpha} \cdot \log_{1/\gamma} N \rfloor$ with $\delta' = \min\{\delta, 1/q\}$, the corresponding FDE achieves

$$\bar{\eta}_{d_\pi,q}(\Upsilon_{\theta_T},\Upsilon_\pi) \leq \frac{1}{1-\gamma^{c-\frac{1}{2q}}}\cdot O_P\left\{\left(\frac{N}{\log_{1/\gamma} N}\right)^{\frac{-\delta'}{2(l-1)+\alpha}}\right\} + \frac{1}{1-\gamma^{c-\frac{1}{2q}}}\cdot 2^{1-\frac{1}{2q}}\cdot\epsilon_0.$$

Table of examples

We discovered nine examples.

1. $l_2^2, \mathrm{MMD}_\beta^2, \mathrm{MMD}_{\sigma_{\mathrm{RBF}}}^2$ were previously mentioned (e.g., Nguyen-Tang et al., 2021; Bellemare et al., 2017), but we first proved their statistical convergence.
2. Other examples are first suggested by our work.

Table 2: Comparison of different FDE methods under general state-action space. See Appendix A for definitions of examples of $\mathfrak{d}$ and $\bar{\eta}_{d_\pi,q}$. $p \geq 1$ is an integer, and $r > p$ satisfies $\sup_{\theta\in\Theta} \sup_{s,a} \mathbb{E}_{Z\sim\Upsilon_\theta(s,a)}\{\|Z\|^r\} < \infty$. Convergence rates are displayed up to a logarithmic order.

$\mathfrak{d}$	$\bar{\eta}_{d_\pi,q}$	unbounded	density-free	d	rate	reference
l_2^2	$\overline{\mathbb{W}}_{p,d_\pi,p}$	✗	✓	1	$N^{\frac{-1}{p}\cdot\frac{1}{2+\alpha}}$	Theorem 3.4
$\mathrm{MMD}_{\beta=1}^2$	$\overline{\mathbb{W}}_{1,d_\pi,1}$	✓	✓	≥ 1	$N^{\frac{-1}{\frac{4(d+1)r}{r-1}+\alpha}}$	Theorem 3.4
$\mathrm{MMD}_{\beta>1}^2$	$\overline{\mathrm{MMD}}_{\beta,d_\pi,1}$	✗	✓	≥ 1	$N^{\frac{-1}{2+\alpha}}$	Theorem C.11
$d_{L_2}^2$	$\overline{\mathbb{W}}_{p,d_\pi,p}$	✓	✗	≥ 1	$N^{\frac{-2(r-p)}{(d+2r)p}\cdot\frac{1}{2+\alpha}}$	Theorem 3.4
MMD_ν^2	$\overline{\mathbb{W}}_{p,d_\pi,p}$	✓	✗	≥ 1	$N^{\frac{-(r-p)}{(d+2r)p}\cdot\frac{1}{2+\alpha}}$	Theorem 3.4
$\mathrm{MMD}_{\sigma_{\mathrm{RBF}}}^2$	$\overline{\mathbb{W}}_{p,d_\pi,p}$	✓	✓	≥ 1	–	Theorem 3.4
$\mathrm{MMD}_{\sigma,f}^2$	$\overline{\mathbb{W}}_{p,d_\pi,p}$	✗	✓	≥ 1	–	Theorem 3.4
MMD_{cou}^2	$\overline{\mathbb{W}}_{p,d_\pi,p}$	✗	✗	$\neq 1$	$N^{\frac{-1}{p}\cdot\frac{(6-\alpha)}{16}}$	Theorem C.11
KL	$\overline{\mathbb{W}}_{p,d_\pi,p}$	✗	✗	≥ 1	$N^{\frac{-1}{p}\cdot\frac{1}{2+\alpha}}$	Corollary B.9

Simulation

In experiments on Atari games, we have following conclusions.

1. Our methods (KL, Energy, Hyv) outperform baseline methods (FLE, QRD, IQN) in mean rank values.
2. TVD (not a functional Bregman divergence) malfunctions. This corroborates our claim to use functional Bregman divergence.

Table 1: Mean of rank values (highest: 1, lowest: 9) of inaccuracy in Atari games

Category		Method								
Reward	Cover	FLE	QRD	IQN	KL	Energy	PDF	RBF	TVD	Hyv
Deterministic	Strong	3.60	6.19	8.25	**1.77**	**3.36**	4.38	5.03	8.11	**2.78**
Deterministic	Weak	**3.00**	6.39	8.30	**1.84**	3.14	4.69	5.47	8.09	**2.42**
Random	Strong	4.42	5.30	7.00	**3.03**	**3.15**	4.07	4.07	8.63	**3.59**
Random	Weak	3.76	6.09	7.31	**2.34**	**3.61**	4.46	4.23	8.61	**2.40**

References

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017). The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*.

Nguyen-Tang, T., Gupta, S., and Venkatesh, S. (2021). Distributional reinforcement learning via moment matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9144–9152.

Odin, E. and Charpentier, A. (2020). *Dynamic Programming in Distributional Reinforcement Learning*. PhD thesis, Université du Québec à Montréal.

Ovcharov, E. Y. (2018). Proper scoring rules and bregman divergence. *Bernoulli*, 24(1):53–79.