Off-Policy Evaluation via the Regularized Lagrangian

Sherry Yang*, Ofir Nachum*, Bo Dai*, Lihong Li, Dale Schuurmans

Google Brain

Paper: https://arxiv.org/abs/2007.03438
Code: https://github.com/google-research/dice_rl
Off-policy Evaluation (OPE)

Given $\mathcal{M} = \langle S, A, R, T, \mu_0, \gamma \rangle$ and $\pi(\cdot | s_t)$

Policy value

$$\rho(\pi) = (1 - \gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}[Q^\pi(s_0, a_0)]$$

$$\rho(\pi) = \mathbb{E}_{(s, a) \sim d^\pi}[R(s, a)]$$

Off-policy evaluation via DICE

$$\rho(\pi) = \mathbb{E}_{(s, a) \sim d^\mathcal{D}}[\zeta^*(s, a) \cdot R(s, a)] \text{ where } \zeta^*(s, a) := \frac{d^\pi(s, a)}{d^\mathcal{D}(s, a)}$$

DICE estimators: DualDICE, GenDICE, GradientDICE, ...
\( \rho(\pi) \) as Linear Programs (LPs)

**Primal** Q-LP \( \rho(\pi) = \min_{Q: S \times A \to \mathbb{R}} (1 - \gamma) \mathbb{E}_{\mu_0 \pi} [Q(s, a)] \),

s.t., \( Q(s, a) = R(s, a) + \gamma \cdot \mathcal{P}^\pi Q(s, a) \)

**Dual** d-LP \( \rho(\pi) = \max_{d: S \times A \to \mathbb{R}} \mathbb{E}_d [R(s, a)] \),

s.t., \( d(s, a) = (1 - \gamma) \mu_0(s) \pi(a|s) + \gamma \cdot \mathcal{P}^\pi d(s, a) \)

**Lagrangian** \( \max_d \min_Q L(d, Q) := (1 - \gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)} \mathbb{E}_{s_0 \sim \mu_0} [Q(s_0, a_0)] \\ + \sum_{s, a} d(s, a) \cdot (R(s, a) + \gamma \mathcal{P}^\pi Q(s, a) - Q(s, a)) \)

**Off-policy** \( \max_{\zeta} \min_{Q} L_D(\zeta, Q) := (1 - \gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)} \mathbb{E}_{s_0 \sim \mu_0} [Q(s_0, a_0)] \\ + \mathbb{E}_{(s, a, r, s') \sim d^D} [\zeta(s, a) \cdot (r + \gamma Q(s', a') - Q(s, a))] \)

\[ \zeta = \frac{d}{d^D} \]
Regularized Lagrangian

\[
\max_{\zeta \geq 0} \min_{Q, \lambda} L_D(\zeta, Q, \lambda) := (1 - \gamma) \cdot \mathbb{E}_{s_0 \sim \mu_0}[Q(s_0, a_0)] + \lambda \\
+ \mathbb{E}_{(s, a, r, s') \sim d^p}[\zeta(s, a) \cdot (\alpha_R \cdot R(s, a) + \gamma Q(s', a') - Q(s, a) - \lambda)] \\
+ \alpha_Q \cdot \mathbb{E}_{(s, a) \sim d^p}[f_1(Q(s, a))] - \alpha_\zeta \cdot \mathbb{E}_{(s, a) \sim d^p}[f_2(\zeta(s, a))].
\]

Regularization choices

- **Primal** and **Dual** regularization \( f_1, f_2 \) convex functions
- **Reward** \( \alpha_R \in \{0, 1\} \)
- **Positivity** \( \zeta^*(s, a) = \frac{d^\pi(s, a)}{d^p(s, a)} \geq 0 \)
- **Normalization** \( \mathbb{E}_{d^p}[\zeta(s, a)] = 1 \)
Regularized Lagrangian

\[
\max_{\zeta \geq 0} \min_{Q, \lambda} L_D(\zeta, Q, \lambda) := (1 - \gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}[Q(s_0, a_0)] + \lambda \\
+ \mathbb{E}_{(s, a, r, s') \sim d^\mathcal{D}}[\zeta(s, a) \cdot (\alpha_R \cdot R(s, a) + \gamma Q(s', a') - Q(s, a) - \lambda)] \\
+ \alpha_Q \cdot \mathbb{E}_{(s, a) \sim d^\mathcal{D}}[f_1(Q(s, a))] - \alpha_\zeta \cdot \mathbb{E}_{(s, a) \sim d^\mathcal{D}}[f_2(\zeta(s, a))].
\]

Estimator choices

- **Primal estimator:** \( \hat{\rho}_Q(\pi) := (1 - \gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}[\hat{Q}(s_0, a_0)] + \hat{\lambda}. \)

- **Dual estimator:** \( \hat{\rho}_\zeta(\pi) := \mathbb{E}_{(s, a, r) \sim d^\mathcal{D}}[\hat{\zeta}(s, a) \cdot r]. \)

- **Lagrangian:** \( \hat{\rho}_{Q, \zeta}(\pi) := \hat{\rho}_Q(\pi) + \hat{\rho}_\zeta(\pi) + \mathbb{E}_{(s, a, r, s') \sim d^\mathcal{D}}[\hat{\zeta}(s, a) (\gamma \hat{Q}(s', a') - \hat{Q}(s, a) - \hat{\lambda})]. \)
\[
\max_{\zeta \geq 0} \min_{Q, \lambda} L_D(\zeta, Q, \lambda) := (1 - \gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}[Q(s_0, a_0)] + \lambda \\
+ \mathbb{E}_{(s, a, r, s') \sim d^D} [\zeta(s, a) \cdot (\alpha_R \cdot R(s, a) + \gamma Q(s', a') - Q(s, a) - \lambda)] \\
+ \alpha_Q \cdot \mathbb{E}_{(s, a) \sim d^D} [f_1(Q(s, a))] - \alpha_\zeta \cdot \mathbb{E}_{(s, a) \sim d^D} [f_2(\zeta(s, a))].
\]

### Solution baseness

| Regularization (with or without \(\lambda\)) | \(\hat{\rho}_Q\) | \(\hat{\rho}_\zeta\) | \(\hat{\rho}_{Q, \zeta}\) |
|---------------------------------------------|------------------|------------------|------------------|
| \(\alpha_\zeta = 0\)                      |                  |                  |                  |
| \(\alpha_R = 1\)                          | \(\zeta \geq 0\) | Unbiased         | Biased           |
| \(\alpha_Q > 0\)                          |                  |                  |                  |
| \(\alpha_R = 0\)                          | \(\zeta \geq 0\) | Biased           | Unbiased         |
| \(\alpha_\zeta > 0\)                      |                  |                  |                  |
| \(\alpha_R = 1\)                          | \(\zeta \geq 0\) | Biased           | Unbiased         |
| \(\alpha_Q = 0\)                          |                  |                  |                  |
| \(\alpha_R = 0\)                          | \(\zeta \geq 0\) | Unbiased         | Unbiased         |
Regularized Lagrangian

$$\max_{\zeta \geq 0} \min_{Q, \lambda} L_D(\zeta, Q, \lambda) := (1 - \gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0)}[Q(s_0, a_0)] + \lambda$$

$$+ \mathbb{E}_{(s, a, r, s') \sim \mathcal{D}}[\zeta(s, a) \cdot (\alpha_R \cdot R(s, a) + \gamma Q(s', a') - Q(s, a) - \lambda)]$$

$$+ \alpha_Q \cdot \mathbb{E}_{(s, a) \sim \mathcal{D}}[f_1(Q(s, a))] - \alpha_{\zeta} \cdot \mathbb{E}_{(s, a) \sim \mathcal{D}}[f_2(\zeta(s, a))].$$

---

Recover OPE estimators

- **DualDICE** $\iff$ $(\alpha_Q = 0, \alpha_{\zeta} = 1, \alpha_R = 0)$ without $\zeta \geq 0$ and without $\lambda$.
- **GenDICE** $\iff$ $(\alpha_Q = 1, \alpha_{\zeta} = 0, \alpha_R = 0)$ $\zeta \geq 0$ with $\lambda$.
- **GradientDICE** $\iff$ $(\alpha_Q = 1, \alpha_{\zeta} = 0, \alpha_R = 0)$ without $\zeta \geq 0$ and with $\lambda$.
- **DR-MWQL** $\iff$ $(\alpha_Q = 0, \alpha_{\zeta} = 0, \alpha_R = 1)$ without $\zeta \geq 0$ and without $\lambda$.
- **MWL** $\iff$ $(\alpha_Q = 0, \alpha_{\zeta} = 0, \alpha_R = 0)$ without $\zeta \geq 0$ and without $\lambda$.
- **BestDICE**: $\iff$ $(\alpha_Q = 0, \alpha_{\zeta} = 1, \alpha_R = 0/1)$ with $\zeta \geq 0$ and with $\lambda$. 

Ofir Nachum, Yinlam Chow, Bo Dai, Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In Advances in Neural Information Processing Systems
Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. GenDICE: Generalized offline estimation of stationary values. In International Conference on Learning Representations
Shangtong Zhang, Bo Liu, and Shimon Whiteson. GradientDICE: Rethinking generalized offline estimation of stationary values. arXiv preprint
Masatoshi Uehara and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. arXiv preprint
**BestDICE Performance**

Estimator choice: $\hat{\rho}_\zeta > \hat{\rho}_Q, \hat{\rho}_{Q,\zeta}$

**Reward scale invariance**

Scale=1.0, Shift=0.0  
Scale=0.5  
Scale=10.0  
Shift=5.0  
Shift=10.0  
Exp