Data-Efficient Off-Polciy Policy Evaluation for Reinforcement Learning

Xuhui Liu

LAMDA, Nanjing University

April 2, 2021

Table of Contents

NANJING UNIVERSITY

Learning And Mining from DatA

Table of Contents

Off Policy Evaluation

- Goal: To estimate the expected return of the learned policy using data generated by a different policy.
  - Given a dataset $D = \{\tau_i\}_{i=1}^{N}$ of N trajectories, where $\tau_i = s_0^i, a_0^i, s_i^i, \cdots, s_{T-1}^i, a_{T-1}^i, a_t^i$ is generated by a behavior policy π_b
  - We desire to evaluate the policy π_e
  - Off-policy Evaluation (OPE) is to estimate the value:

$$V(\pi_e) = E_\tau[\sum_{t=1}^{T} \gamma^t r_t]$$

where $a_t \sim \pi_e(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$, $r_t \sim R(s_t, a_t)$

Direct Method

- Model-free: Fitted-Q-Evaluation (FQE)
  Use $\hat{Q}(s, a|\theta)$ to estimate $Q^{\pi_e}(s, a)$.

$$\widehat{Q}_k = \min_\theta \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} (\widehat{Q}_{k-1}(x_t^i, a_t^i; \theta) - y_t^i)^2,$$
$$y_t^i \equiv r_t^i + \gamma \mathbb{E}_{\pi_e} \widehat{Q}_{k-1}(x_{t+1}^i, \cdot; \theta), \quad \widehat{Q}_0 \equiv 0.$$

- Model-based: Estimate $\hat{P}$ and $\hat{R}$ from data, and then use the learned MDP to estimate $V(\pi_e)$.

Importance Sampling

- Naive importance sampling

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\beta(\tau)}\left[\frac{\pi_\theta(\tau)}{\pi_\beta(\tau)}\sum_{t=0}^{H}\gamma^t r(\mathbf{s},\mathbf{a})\right]$$

$$= \mathbb{E}_{\tau\sim\pi_\beta(\tau)}\left[\left(\prod_{t=0}^{H}\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\beta(\mathbf{a}_t|\mathbf{s}_t)}\right)\sum_{t=0}^{H}\gamma^t r(\mathbf{s},\mathbf{a})\right] \approx \sum_{i=1}^{n} w_H^i \sum_{t=0}^{H}\gamma^t r_t^i,$$

where $w_t^i = \frac{1}{n}\Pi_{t'=0}^{t}\frac{\pi_\theta(a_{t'}^i|s_{t'}^i)}{\pi_\beta(a_{t'}^i|s_{t'}^i)}$

- Per-decision importance sampling

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\beta(\tau)}\left[\sum_{t=0}^{H}\left(\prod_{t'=0}^{t}\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\beta(\mathbf{a}_t|\mathbf{s}_t)}\right)\gamma^t r(\mathbf{s},\mathbf{a})\right] \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{H} w_t^i\gamma^t r_t^i.$$

Importance Sampling

- Weighted importance sampling

$$w_t^i = \frac{1}{n}\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}^i|s_{t'}^i)}{\pi_\beta(a_{t'}^i|s_{t'}^i)} \longrightarrow w_t^i = \frac{1}{\sum_{i=1}^{n} w_t^i}\prod_{t'=0}^{t}\frac{\pi_\theta(a_{t'}^i|s_{t'}^i)}{\pi_\beta(a_{t'}^i|s_{t'}^i)}$$

Assumptions

Assumption 1

For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, if $\pi_b(a|s) = 0$ then $\pi_e(a|s) = 0$.

Assumption 2

The time horizon L is finite.

Doubly Robust

- Doubly Robust Method was proposed by [Jiang and Li, 2016].

$$\mathrm{DR}(D) := \sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i R_t^{H_i} \tag{2}$$
$$-\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t\left(w_t^i\hat{q}^{\pi_e}\left(S_t^{H_i},A_t^{H_i}\right)-w_{t-1}^i\hat{v}^{\pi_e}\left(S_t^{H_i}\right)\right).$$

Variance Reduction

- Goal: Estimate $\theta := \mathbb{E}[X]$ given a sample of X.
- The estimator will be $\hat{\theta}_1 := X$.
- If we have a sample of another random variable Y, with known expected value, $\mathbb{E}[Y]$.
- We can estimate θ with $\hat{\theta}_2 := X - Y + \mathbb{E}[Y]$.
- $\hat{\theta}_1$ has the same mean with $\hat{\theta}_2$.

Variance Reduction

- $\mathrm{Var}(\hat{\theta}_1) = \mathrm{Var}(X)$.
- $\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)$.
- If $2\mathrm{Cov}(X, Y) > \mathrm{Var}(Y)$, then $\hat{\theta}_2$ has lower variance than $\hat{\theta}_1$.
- Note that the optimal control variate is $Y := X$, since then $\mathrm{Var}(\hat{\theta}_2) = 0$.

Doubly Robust

$$\mathrm{DR}(D) := \underbrace{\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i R_t^{H_i}}_{X}$$
$$\underbrace{-\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t\left(w_t^i \hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) - w_{t-1}^i \hat{v}^{\pi_e}\left(S_t^{H_i}\right)\right)}_{Y}.$$

- Y is mean zero, i.e., $\mathbb{E}[Y] = 0$.

Doubly Robust

$$\mathrm{DR}(D) := \underbrace{\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i R_t^{H_i}}_{X}$$
$$\underbrace{-\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t\left(w_t^i\hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) - w_{t-1}^i\hat{v}^{\pi_e}\left(S_t^{H_i}\right)\right)}_{Y}.$$

$$\hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) \approx R_t^{H_i} + \gamma\hat{v}^{\pi_e}\left(S_{t+1}^{H_i}\right).$$

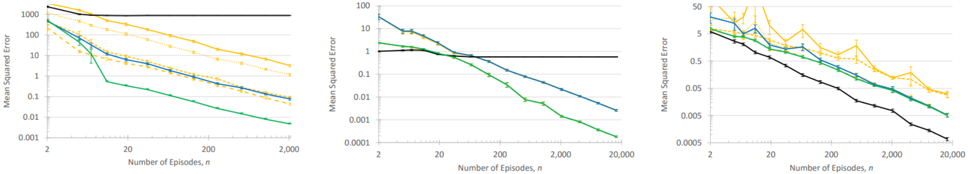
- Y is a decent approximation of X, and therefore DR may have lower variance.

Table of Contents

NANJING UNIVERSITY

Learning And Mining from DatA

Empirical Results



(a) Gridworld (b) ModelFail (c) ModelWin

Figure 1: Empirical results for three different experimental setups. All plots in this paper have the same format: they show the mean squared error of different estimators as n, the number of episodes in D, increases. Both axes always use a logarithmic scale and standard error bars are included from 128 trials. All plots use the following legend:

IS PDIS WIS CWPDIS DR AM WDR

Off-policy j-step return

$$g^{(j)}(D) := \mathrm{IS}^{(j)}(D) + \mathrm{AM}^{(j+1)}(D)$$
$$g^{(\infty)}(D) := \lim_{j\to\infty} g^{(j)}(D).$$

- $\mathrm{IS}^{(j)}(D)$ is an estimate of $\mathbb{E}[\sum_{t=0}^{j} \gamma^t R_t | H \sim \pi_e]$, construted from D using an importance sampling method.
- $\mathrm{AM}^{(j)}(D)$ denote a primarily model-based prediction from D of $\mathbb{E}[\sum_{t=j}^{\infty} \gamma^t R_t | H \sim \pi_e]$.

Blending IS and Model

- Weighting scheme

$$\widehat{\mathbf{x}}^{\star} \in \arg \min_{\mathbf{x} \in \mathbb{R}^{|\mathcal{J}|}} \mathrm{MSE}(\mathbf{x}^{\intercal} \mathbf{g}_{\mathcal{J}}(D), v(\pi_e)),$$

- Bias-variance decomposition

$$\begin{aligned} \widehat{\mathbf{x}}^{\star} &\in \arg \min_{\mathbf{x} \in \Delta^{|\mathcal{J}|}} \mathrm{Bias}(\mathbf{x}^{\intercal} \mathbf{g}_{\mathcal{J}}(D))^2 + \mathrm{Var}(\mathbf{x}^{\intercal} \mathbf{g}_{\mathcal{J}}(D)) \\ &= \arg \min_{\mathbf{x} \in \Delta^{|\mathcal{J}|}} \mathbf{x}^{\intercal} [\Omega_n + \mathbf{b}_n \mathbf{b}_n^{\intercal}] \mathbf{x}, \end{aligned}$$

where $\Omega_n(i,j) = \mathrm{Cov}(\mathbf{g}^{(\mathcal{J}_i)}(D), \mathbf{g}^{(\mathcal{J}_j)}(D))$, and $\mathbf{b}_n(j) = \mathbb{E}[\mathbf{g}^{(\mathcal{J}_j)}(D)] - v(\pi_e)$. And suppose $\sum_{j=1}^{|\mathcal{J}|} x_j = 1$.

Bias-variance Decomposition

Proof.

$$\mathrm{MSE}(x^T g_{\mathcal{J}}(D), v(\pi_e)) \leq \mathrm{MSE}(x^T g_{\mathcal{J}}(D), \mathbb{E}[g_{\mathcal{J}}(D)]) + \mathrm{MSE}(\mathbb{E}[g_{\mathcal{J}}(D)], v(\pi_e))$$
$$= x^T \Omega_n x + (x^T b_n)^2$$

□

Modeling Guided Importance Sampling Combining Estimator

- Variance reduction

$$g^{(j)}(D) := \underbrace{\sum_{i=1}^{n}\sum_{t=0}^{j}\gamma^{t} w_t^i R_t^{H_i}}_{\text{(a)}} + \underbrace{\sum_{i=1}^{n}\gamma^{j+1} w_j^i \hat{v}^{\pi_e}(S_{j+1}^{H_i})}_{\text{(b)}}$$

$$\underbrace{-\sum_{i=1}^{n}\sum_{t=0}^{j}\gamma^{t}\left(w_t^i \hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) - w_{t-1}^i \hat{v}^{\pi_e}\left(S_t^{H_i}\right)\right)}_{\text{(c)}}.$$

Estimating Ω_n

We can write $g^{(j)}(D)$ as the sum of n terms:

$$g^{(j)}(D) = \sum_{i=1}^{n} g_i^{(j)}(D), \tag{24}$$

where

$$g_i^{(j)}(D) := \left(\sum_{t=0}^{j} \gamma^t w_t^i R_t^{H_i} \right) + \gamma^{j+1} w_j^i \hat{v}^{\pi_e}(S_{j+1}^{H_i})$$
$$- \sum_{t=0}^{j} \gamma^t \left(w_t^i \hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) - w_{t-1}^i \hat{v}^{\pi_e}\left(S_t^{H_i}\right)\right).$$

So,

$$\mathrm{Cov}(g^{(i)}(D), g^{(j)}(D)) = \mathrm{Cov}\left(\sum_{k=1}^{n} g_k^{(i)}(D), \sum_{k=1}^{n} g_k^{(j)}(D) \right).$$

Estimating Ω_n

- $g_i^{(j)}(D)$s' distributions are identical.
- Notice that $\omega_t^i = \rho_t^i / \sum_{j=1}^n \rho_t^j$, $g_i^{(j)}(D)$ are not independent.
- But they become less dependent as $n \to \infty$.
- Because the only dependence of $g_i^{(j)}(D)$ comes from the denominator of ω_t^i, which convergence almost surely to n.

Estimating Ω_n

$$
\begin{aligned}
&\mathrm{Cov}(g^{(i)}(D), g^{(j)}(D)) \\
&= \sum_{k\in\{1,\ldots,n\}} \sum_{l\in\{1,\ldots,n\}} \mathrm{Cov}\left(g_k^{(i)}(D), g_l^{(j)}(D)\right) \\
&\overset{(a)}{\approx} \sum_{k\in\{1,\ldots,n\}} \mathrm{Cov}\left(g_k^{(i)}(D), g_k^{(j)}(D)\right) \\
&\overset{(b)}{=} n\,\mathrm{Cov}\left(g_{(\cdot)}^{(i)}(D), g_{(\cdot)}^{(j)}(D)\right),
\end{aligned}
$$

- (a) comes from the assumption that they are independent.
- (b) comes from that they are identical.

Estimating Ω_n

$$\widehat{\Omega}_n(i,j) := \frac{n}{n-1}\sum_{k=1}^{n}\left(g_k^{(\mathcal{J}_i)}(D) - \bar{g}_k^{(\mathcal{J}_i)}(D)\right) \times \left(g_k^{(\mathcal{J}_j)}(D) - \bar{g}_k^{(\mathcal{J}_j)}(D)\right), \tag{25}$$

where

$$\bar{g}_k^{(\mathcal{J}_i)}(D) := \frac{1}{n}\sum_{k=1}^{n} g_k^{(\mathcal{J}_i)}(D).$$

Estimating b_n

- When n, the number of trajectories in D, is small, variance tends to be the root cause of high MSE.
- Proposed an estimator that underestimates the bias initially, but becomes correct as n increase.
- Let $\mathrm{CI}(g^{(\infty)}(D), \delta)$ be a $1-\delta$ confidence interval on the expected value of the random variable $g^{(\infty)}(D)$.
- We estimate $b_n(j)$ as

$$\widehat{\mathbf{b}}_n(j) := \mathrm{dist}\left(g^{(\mathcal{J}_j)}(D), \mathrm{CI}(g^{(\infty)}(D), 0.5)\right)$$

Table of Contents

Consistent estimator

Definition 1 (Almost Sure Convergence). *A sequence of random variables, $(X_n)_{n=1}^{\infty}$, converges almost surely to the random variable X if*

$$\Pr\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

We write $X_n \xrightarrow{\text{a.s.}} X$ to denote that the sequence $(X_n)_{n=1}^{\infty}$ convergences almost surely to X.

Definition 2. *Let θ be a real number and $(\hat{\theta}_n)_{n=1}^{\infty}$ be an infinite sequence of random variables. We call $\hat{\theta}_n$, a (strongly)* **consistent estimator** *of θ if and only if $\hat{\theta}_n \xrightarrow{a.s.} \theta$.*

Law of Large Numbers

Theorem 6 (Khintchine Strong Law of Large Numbers). *Let* $\{X_i\}_{i=1}^{\infty}$ *be independent and identically distributed random variables. Then* $(\frac{1}{n}\sum_{i=1}^{n} X_i)_{n=1}^{\infty}$ *is a sequence of random variables that converges almost surely to* $\mathbf{E}[X_1]$.

Theorem 7 (Kolmogorov Strong Law of Large Numbers). *Let* $\{X_i\}_{i=1}^{\infty}$ *be independent (not necessarily identically distributed) random variables. If all* X_i *have the same mean and bounded variance (i.e., there is a finite constant* b *such that for all* $i \geq 1$, $\mathrm{Var}(X_i) \leq b$*), then* $(\frac{1}{n}\sum_{i=1}^{n} X_i)_{n=1}^{\infty}$ *is a sequence of random variables that converges almost surely to* $\mathbf{E}[X_1]$.

BIM Convergence

Assumption 4 (Bounded importance weight). *There exists a constant $\beta < \infty$ such that for all $(t, i) \in \mathbb{N}_{\geq 0} \times \{1, \ldots, n\}$, $\rho_t^i \leq \beta$ surely.*

Theorem 3. *If Assumption 4 holds, there exists at least one $j \in \mathcal{J}$ such that $g^{(j)}(D)$ is a strongly consistent estimator of $v(\pi_e)$, $\widehat{\mathbf{b}}_n - \mathbf{b}_n \xrightarrow{a.s.} 0$, and $\widehat{\Omega}_n - \Omega_n \xrightarrow{a.s.} 0$, then $\mathrm{BIM}(D, \widehat{\Omega}_n, \widehat{\mathbf{b}}_n) \xrightarrow{a.s.} v(\pi_e)$.* **Proof** *See Appendix E.*

Consistency of BIM

- First assume we have true Ω_n and b_n.
- Let the j^* be an index such that $g^{(j^*)}(D) \xrightarrow{a.s.} v(\pi_e)$, which exits by assumption.
- Let y be a weight vector that places a weight of one on $g^{(j^*)}(D)$ and weight of zero on other returns.
- Then $y^T g(D) = g^{(j^*)}(D) \xrightarrow{a.s.} v(\pi_e)$.

Consistency of BIM

- Remember that
$$x^{\star} \in \arg\min_{x \in \Delta|\mathcal{J}|} \mathrm{MSE}\left(x^{\top} g_{\mathcal{J}}(D), \Omega_n, b_n\right)$$
- $\mathrm{MSE}\left((x^{\star})^{\top} g_{\mathcal{J}}(D), v(\pi_e)\right) \leq \mathrm{MSE}\left(y^{\top} g_{\mathcal{J}}(D), v(\pi_e)\right)$
- $\mathrm{BIM}(D, \Omega_n b_n) \xrightarrow{\text{a.s.}} v(\pi_e)$

Consistency of BIM

Lemma

If f is a continuous function, $f(X_n) \xrightarrow{a.s.} X$, and $Y_n - X_n \xrightarrow{a.s.} 0$, then $f(Y_n) \xrightarrow{a.s.} X$.

- $\hat{b}_n - b_n \xrightarrow{a.s.} 0$ and $\hat{\Omega}_n - \Omega_n \xrightarrow{a.s.} 0$.
- $\Rightarrow \mathrm{BIM}(D, \hat{\Omega}_n, \hat{b}_n) \xrightarrow{a.s.} v(\pi_e)$.

Consistency of BIM

$$
\begin{aligned}
\Pr\left(\lim_{n\to\infty} f(Y_n) = X\right) &= \Pr\left(\lim_{n\to\infty} f(Y_n - X_n + X_n) = X\right) \\
&\overset{(a)}{=} \Pr\left(f\left(\lim_{n\to\infty} Y_n - X_n + X_n\right) = X\right) \\
&\overset{(b)}{\geq} \Pr\left(\left(\lim_{n\to\infty} Y_n - X_n = 0\right) \right. \\
&\qquad \left. \bigcap \left(f\left(\lim_{n\to\infty} X_n\right) = X\right)\right) \\
&= \Pr\left(\left(\lim_{n\to\infty} Y_n - X_n = 0\right) \right. \\
&\qquad \left. \bigcap \left(\lim_{n\to\infty} f(X_n) = X\right)\right) \\
&\overset{(c)}{=} 1,
\end{aligned}
$$

- (a) holds because f is a continuous function.
- (b) holds because it gives sufficient conditions for the event in the line above to hold.
- d (c) holds because under our assumptions the two events both occur with probability one.

NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

References

Jiang, N. and Li, L. (2016).
Doubly robust off-policy value evaluation for reinforcement learning.
In Balcan, M. and Weinberger, K. Q., editors, *Proceedings of the 33nd International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 652–661. JMLR.org.