OFF-POLICY POLICY EVALUATION

IN FINITE HORIZON

WENHAO LI

2019-11-19

References

- Precup, Doina. "Eligibility traces for off-policy policy evaluation." *Computer Science Department Faculty Publication Series* (2000): 80.
- Precup, Doina, Richard S. Sutton, and Sanjoy Dasgupta. "Off-policy temporal-difference learning with function approximation." *ICML*. 2001.
- Dudík, Miroslav, John Langford, and Lihong Li. "Doubly robust policy evaluation and learning." *arXiv preprint arXiv:1103.4601* (2011).
- Jiang, Nan, and Lihong Li. "Doubly robust off-policy value evaluation for reinforcement learning." *arXiv preprint arXiv:1511.03722* (2015).
- Hanna, Josiah, Peter Stone, and Scott Niekum. "Importance Sampling Policy Evaluation with an Estimated Behavior Policy." *International Conference on Machine Learning*. 2019.

Start From Context Bandit

- Directed Method

$$\hat{V}^{\pi}_{\mathrm{DM}} = \frac{1}{|S|} \sum_{x \in S} \hat{\varrho}_{\pi(x)}(x)$$

- Inverse Propensity Score

$$\hat{V}^{\pi}_{\mathrm{IPS}} = \frac{1}{|S|} \sum_{(x,h,a,r_a) \in S} \frac{r_a \mathbf{I}(\pi(x) = a)}{\hat{p}(a \mid x, h)}$$

Importance Sampling Method

$$
\begin{aligned}
E_d\{x\} &= \int_x x\, d(x)\, dx = \int_x x \frac{d(x)}{d'(x)} d'(x) dx \\
&= E_{d'}\left\{ x \frac{d(x)}{d'(x)} \right\},
\end{aligned}
$$

which leads to the importance sampling estimator,

$$
\approx \frac{1}{n} \sum_{i=1}^{n} x_i \frac{d(x_i)}{d'(x_i)} \tag{1}
$$

Importance Sampling Method (Cont.)

A less well known variant of this technique is *weighted importance sampling*, which performs a weighted average of the samples, with weights $\frac{d(x_i)}{d'(x_i)}$. The weighted importance sampling estimator is:

$$\frac{\sum_{i=1}^{n} x_i \frac{d(x_i)}{d'(x_i)}}{\sum_{i=1}^{n} \frac{d(x_i)}{d'(x_i)}}.$$

Per-Decision Algorithm

$$Q^{IS}(s,a) \stackrel{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} R_m w_m,$$

$$R_m \stackrel{\text{def}}{=} r_{t_m+1} + \gamma r_{t_m+2} + \ldots + \gamma^{T_m - t_m - 1} r_{T_m},$$

$$w_m \stackrel{\text{def}}{=} \frac{\pi_{t_m+1}}{b_{t_m+1}} \frac{\pi_{t_m+2}}{b_{t_m+2}} \cdots \frac{\pi_{T_m-1}}{b_{T_m-1}},$$

$$Q^{ISW}(s,a) \stackrel{\text{def}}{=} \frac{\sum_{m=1}^{M} R_m w_m}{\sum_{m=1}^{M} w_m}.$$

Per-Decision Algorithm (Cont.)

$$Q^{IS}(s,a) \stackrel{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} R_m w_m,$$

$$R_m \stackrel{\text{def}}{=} r_{t_m+1} + \gamma r_{t_m+2} + \ldots + \gamma^{T_m - t_m - 1} r_{T_m},$$

$$w_m \stackrel{\text{def}}{=} \frac{\pi_{t_m+1}}{b_{t_m+1}} \frac{\pi_{t_m+2}}{b_{t_m+2}} \cdots \frac{\pi_{T_m-1}}{b_{T_m-1}},$$

$$R_m w_m = \sum_{i=t_m+1}^{T_m} \gamma^{i-t_m-1} r_i \frac{\pi_{t_m+1}}{b_{t_m+1}} \cdots \frac{\pi_{i-1}}{b_{i-1}} \frac{\pi_i}{b_i} \cdots \frac{\pi_{T_m-1}}{b_{T_m-1}}.$$

Per-Decision Algorithm (Cont.)

$$R_m w_m = \sum_{i=t_m+1}^{T_m} \gamma^{i-t_m-1} r_i \frac{\pi_{t_m+1}}{b_{t_m+1}} \cdots \frac{\pi_{i-1}}{b_{i-1}} \frac{\pi_i}{b_i} \cdots \frac{\pi_{T_m-1}}{b_{T_m-1}}.$$

$$Q^{PD}(s,a) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{T_m-t_m} \gamma^{k-1} r_{t_m+k} \prod_{i=t_m+1}^{t_m+k-1} \frac{\pi_i}{b_i}.$$

$$Q^{PDW}(s,a) \stackrel{\text{def}}{=} \frac{\sum_{m=1}^{M} \sum_{k=1}^{T_m-t_m} \gamma^{k-1} r_{t_m+k} \prod_{i=t_m+1}^{t_m+k-1} \frac{\pi_i}{b_i}}{\sum_{m=1}^{M} \sum_{k=1}^{T_m-t_m} \gamma^{k-1} \prod_{i=t_m+1}^{t_m+k-1} \frac{\pi_i}{b_i}}.$$

Per-Decision Algorithm (Cont.)

Algorithm 1 Online, Eligibility-Trace Version of Per-Decision Importance Sampling

1. Update the eligibility traces for all states:

$$
\begin{aligned}
e_t(s,a) &= e_{t-1}(s,a)\,\gamma\lambda\frac{\pi(s_t,a_t)}{b(s_t,a_t)}, \qquad \forall s,a \\
e_t(s,a) &= 1, \text{iff } t = t_m(s,a),
\end{aligned}
$$

   where $\lambda \in [0,1]$ is an eligibility trace decay factor.

2. Compute the TD error:

$$
\delta_t = r_{t+1} + \gamma\frac{\pi(s_{t+1},a_{t+1})}{b(s_{t+1},a_{t+1})}Q_t(s_{t+1},a_{t+1}) - Q_t(s_t,a_t)
$$

3. Update the action-value function:

$$
Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha e_t(s,a)\,\delta_t, \qquad \forall s,a
$$

Off-Policy TD(λ) with Function Approximation

$$Q^{\pi}(s,a) \approx \theta^T \phi_{sa} = \sum_{i=1}^{m} \theta(i)\phi_{sa}(i),$$

$$\Delta\theta_t = \alpha\left(R_t^{\lambda} - \theta^T\phi_t\right)\phi_t,$$

$$R_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}R_t^{(n)},$$

$$R_t^{(n)} = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n \theta^T \phi_{t+n},$$

$\Rightarrow$

$$\Delta\theta_t = \alpha\left(\bar{R}_t^{\lambda} - \theta^T\phi_t\right)\phi_t\rho_1\rho_2\cdots\rho_t,$$

$$\begin{aligned}\bar{R}_t^{(n)} \quad = \quad & r_{t+1} + \gamma r_{t+2}\rho_{t+1} + \cdots \\ & + \gamma^{n-1} r_{t+n}\rho_{t+1}\cdots\rho_{t+n-1} \\ & + \gamma^n \rho_{t+1}\cdots\rho_{t+n}\theta^T\phi_{t+n}\end{aligned}$$

Off-Policy TD(λ) with Function Approximation (Cont.)

$$\Delta\theta_t = \alpha \left(R_t^\lambda - \theta^T \phi_t \right) \phi_t, \qquad \Delta\theta_t = \alpha \left(\bar{R}_t^\lambda - \theta^T \phi_t \right) \phi_t \rho_1 \rho_2 \cdots \rho_t,$$

Theorem 1 *Let $\Delta\theta$ and $\Delta\bar{\theta}$ be the sum of the parameter increments over an episode under on-policy TD(λ) and importance sampled TD(λ) respectively, assuming that the starting weight vector is θ in both cases. Then*

$$E_b\left\{ \Delta\bar{\theta} \mid s_0, a_0 \right\} = E_\pi\{\Delta\theta \mid s_0, a_0\}, \quad \forall s_0 \in \mathcal{S}, a_0 \in \mathcal{A}.$$

Off-Policy TD(λ) with Function Approximation (Cont.)

$$E_b\{\Delta\bar{\theta}\} = E_b\left\{\sum_{t=0}^{\infty} \alpha\left(\bar{R}_t^{\lambda} - \theta^T\phi_t\right)\phi_t\rho_1\rho_2\cdots\rho_t\right\}$$

$$= E_b\left\{\sum_{t=0}^{\infty}\sum_{n=1}^{\infty} \alpha(1-\lambda)\lambda^{n-1}(\bar{R}_t^{(n)} - \theta^T\phi_t)\phi_t\rho_1\rho_2\cdots\rho_t\right\}.$$

$$\boxed{\begin{aligned} & E_b\left\{\sum_{t=0}^{\infty}\left(\bar{R}_t^{(n)} - \theta^T\phi_t\right)\phi_t\rho_1\rho_2\cdots\rho_t\right\} \\ & \qquad = E_\pi\left\{\sum_{t=0}^{\infty}\left(R_t^{(n)} - \theta^T\phi_t\right)\phi_t\right\}. \end{aligned}}$$

Off-Policy TD(λ) with Function Approximation (Cont.)

$$E_b\left\{\sum_{t=0}^{\infty}\left(\bar{R}_t^{(n)} - \theta^T\phi_t\right)\phi_t\rho_1\rho_2\cdots\rho_t\right\}$$

$$= \sum_{t=0}^{\infty}\sum_{\omega\in\Omega_t} p_b(\omega)\phi_t\prod_{k=1}^{t}\rho_k E_b\left\{\bar{R}_t^{(n)} - \theta^T\phi_t \;\middle|\; s_t, a_t\right\}$$

(given the Markov property)

$$= \sum_{t=0}^{\infty}\sum_{\omega\in\Omega_t}\prod_{j=1}^{t} p_{s_{j-1},s_j}^{a_{j-1}} b(s_j, a_j)\phi_t \prod_{k=1}^{t}\frac{\pi(s_k, a_k)}{b(s_k, a_k)} \cdot \left(E_b\left\{\bar{R}_t^{(n)} \;\middle|\; s_t, a_t\right\} - \theta^T\phi_t\right)$$

$$= \sum_{t=0}^{\infty}\sum_{\omega\in\Omega_t}\prod_{j=1}^{t} p_{s_{j-1},s_j}^{a_{j-1}}\pi(s_j, a_j)\phi_t \cdot \left(E_b\left\{\bar{R}_t^{(n)} \;\middle|\; s_t, a_t\right\} - \theta^T\phi_t\right)$$

$$= \sum_{t=0}^{\infty}\sum_{\omega\in\Omega_t} p_\pi(\omega)\phi_t\left(E_\pi\left\{R_t^{(n)} \;\middle|\; s_t, a_t\right\} - \theta^T\phi_t\right)$$

(using our previous result)

$$= E_\pi\left\{\sum_{t=0}^{\infty}\left(R_t^{(n)} - \theta^T\phi_t\right)\phi_t\right\}. \diamond$$

Revisit Context Bandit

- Directed Method

$$\hat{V}^{\pi}_{\mathrm{DM}} = \frac{1}{|S|} \sum_{x \in S} \hat{\varrho}_{\pi(x)}(x)$$

- Inverse Propensity Score

$$\hat{V}^{\pi}_{\mathrm{IPS}} = \frac{1}{|S|} \sum_{(x,h,a,r_a) \in S} \frac{r_a \mathbf{I}(\pi(x) = a)}{\hat{p}(a \mid x, h)}$$

Doubly Robust Estimator

$$\hat{V}^{\pi}_{\mathrm{DM}} = \frac{1}{|S|} \sum_{x \in S} \hat{\varrho}_{\pi(x)}(x) \qquad \hat{V}^{\pi}_{\mathrm{IPS}} = \frac{1}{|S|} \sum_{(x,h,a,r_a) \in S} \frac{r_a \mathbf{I}(\pi(x) = a)}{\hat{p}(a \mid x, h)}$$

$$\hat{V}^{\pi}_{\mathrm{DR}} = \frac{1}{|S|} \sum_{(x,h,a,r_a) \in S} \left[\frac{(r_a - \hat{\varrho}_a(x)) \mathbf{I}(\pi(x) = a)}{\hat{p}(a \mid x, h)} + \hat{\varrho}_{\pi(x)}(x) \right].$$

Extends to MDP

$$V_{\mathrm{DR}} := \widehat{V}(s) + \rho\left(r - \widehat{R}(s,a)\right), \tag{8}$$

where $\rho := \frac{\pi_1(a|s)}{\pi_0(a|s)}$ and $\widehat{V}(s) := \sum_a \pi_1(a|s)\widehat{R}(s,a)$. It is easy to verify that $\widehat{V}(s) = \mathbb{E}_{a\sim\pi_0}\left[\rho\widehat{R}(s,a)\right]$, as long as $\widehat{R}$ and ρ are independent, which implies the unbiasedness of the estimator. Furthermore, if $\widehat{R}(s,a)$ is a good estimate of r, the magnitude of $r - \widehat{R}(s,a)$ can be much smaller than that of r. Consequently, the variance of $\rho(r - \widehat{R}(s,a))$ *tends to* be smaller than that of ρr, implying that DR often has a lower variance than IS (Dudík et al., 2011).

Doubly Robust Estimator for RL

$$V_{\text{IS}} := \rho_{1:H} \cdot \left(\sum_{t=1}^{H} \gamma^{t-1} r_t\right), \qquad V_{\text{WIS}} = \frac{\rho_{1:H}}{w_H}\left(\sum_{t=1}^{H} \gamma^{t-1} r_t\right),$$

$$V_{\text{step-IS}} := \sum_{t=1}^{H} \gamma^{t-1} \rho_{1:t}\, r_t. \qquad V_{\text{step-WIS}} = \sum_{t=1}^{H} \gamma^{t-1} \frac{\rho_{1:t}}{w_t} r_t\,.$$

$$V_{\text{step-IS}}^{H+1-t} := \rho_t \left(r_t + \gamma V_{\text{step-IS}}^{H-t}\right).$$

$$\boxed{\begin{aligned} V_{\text{DR}}^{H+1-t} &:= \widehat{V}(s_t) + \rho_t\left(r_t + \gamma V_{\text{DR}}^{H-t} - \widehat{Q}(s_t, a_t)\right). \\ V_{\text{DR}} &:= \widehat{V}(s) + \rho\left(r - \widehat{R}(s, a)\right), \end{aligned}}$$

Doubly Robust Estimator for RL (Ext.)

$$V_{\text{DR-v2}}^{H+1-t} = \widehat{V}(s_t) + \rho_t\Big(r_t + \gamma V_{\text{DR-v2}}^{H-t} - \widehat{R}(s_t, a_t) - \gamma \widehat{V}(s_{t+1}) \frac{\widehat{P}(s_{t+1}|s_t,a_t)}{P(s_{t+1}|s_t,a_t)}\Big),$$

$$\boxed{\begin{aligned} &V_{\text{DR}}^{H+1-t} := \widehat{V}(s_t) + \rho_t\Big(r_t + \gamma V_{\text{DR}}^{H-t} - \widehat{Q}(s_t, a_t)\Big). \\ &V_{\text{DR}} := \widehat{V}(s) + \rho\left(r - \widehat{R}(s,a)\right), \end{aligned}}$$

Sampling Error

$$\mathrm{IS}(\pi_e, \mathcal{D}) := \frac{1}{m} \sum_{i=1}^{m} g(H^{(i)}) \prod_{t=0}^{L-1} \frac{\pi_e(A_t^{(i)} | S_t^{(i)})}{\pi_b{}^{(i)}(A_t^{(i)} | S_t^{(i)})}.$$

Regression Importance Sampling

$$\pi_{\mathcal{D}}{}^{(n)} := \operatorname*{argmax}_{\pi \in \Pi^n} \sum_{H \in \mathcal{D}} \sum_{t=0}^{L-1} \log \pi(a|H_{t-n:t}).$$

$$\mathrm{RIS}(n)(\pi_e, \mathcal{D}) := \frac{1}{m} \sum_{i=1}^{m} g(H_i) \prod_{t=0}^{L-1} \frac{\pi_e(A_t|S_t)}{\pi_{\mathcal{D}}{}^{(n)}(A_t|H_{t-n:t})}$$

Regression Importance Sampling

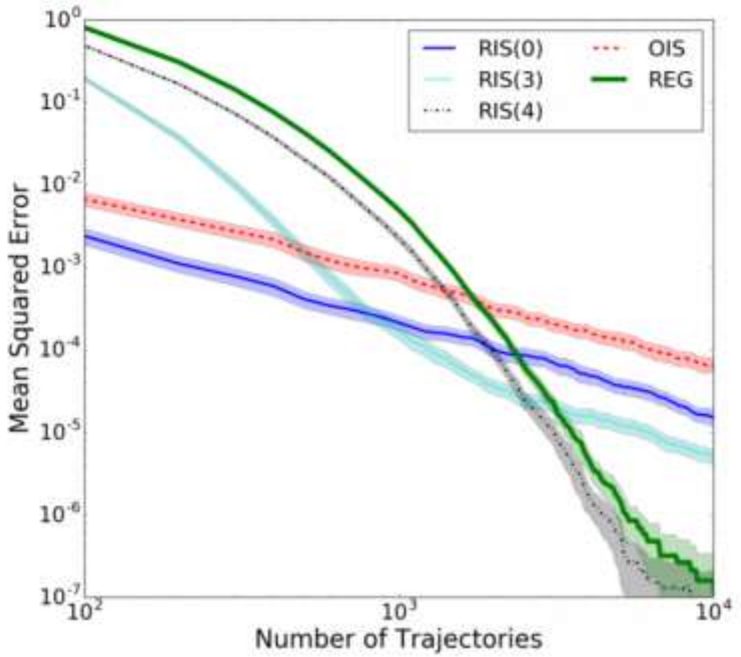

Figure 3: Off-policy evaluation in the SinglePath MDP for various n. The curves for REG and RIS(4) have been cut-off to more clearly show all methods. These methods converge to an MSE value of approximately 1×10^{-31}.