Notes on Minimax Weight and Q-function Learning for Off-policy Evaluation

Nan Jiang

October 31, 2019

This note attempts to explain the simple intuitions behind the marginalized importance sampling methods for off-policy evaluation. We will use the algorithms in our recent paper (Uehara and Jiang, 2019) as examples. Everything will be kept minimal (and sometimes informal).

1 MWL

Goal Estimate $R_{\pi_e} := (1-\gamma)\mathbb{E}_{\pi_e}[\sum_{t=0}^{\infty}\gamma^t r_t | s_0 \sim d_0]$. d_0 and π_e are known. Let $d_{\pi_e,\gamma}$ be the normalized discounted state-action occupancy induced by π_e from d_0.

Data Each data point looks like $(s,a) \sim \mu$, $r \sim R(s,a)$, $s' \sim P(s,a)$. Assume μ covers $d_{\pi_e,\gamma}$. We will use exact expectations instead of sample-based approximation for simplicity.

Format of Estimator We want to find $w : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that

$$\mathbb{E}_\mu[w(s,a)\cdot r] = R_{\pi_e}. \tag{1}$$

Define

$$w_{\pi_e/\mu}(s,a) := \frac{d_{\pi_e,\gamma}(s,a)}{\mu(s,a)}.$$

Note that choosing $w = w_{\pi_e/\mu}$ satisfies Eq.(1), but how to learn this function?

Derivation Eq.(1) is equivalent to

$$\mathbb{E}_\mu[w(s,a)\cdot r] = (1-\gamma)\mathbb{E}_{s\sim d_0}[Q^{\pi_e}(s,\pi_e)]. \tag{2}$$

Key observation: by Bellman equation,

$$\mathbb{E}[r|s,a] = \mathbb{E}[Q^{\pi_e}(s,a) - \gamma Q^{\pi_e}(s',\pi_e)|s,a].$$

Use the "$Q - \gamma Q$" term on the RHS to replace r, and Eq.(2) is equivalent to

$$\mathbb{E}_\mu[w(s,a)\cdot(Q^{\pi_e}(s,a) - \gamma Q^{\pi_e}(s',\pi_e))] = (1-\gamma)\mathbb{E}_{s\sim d_0}[Q^{\pi_e}(s,\pi_e)]. \tag{3}$$

Of course we do not know Q^{π_e}, and this is where function approximation is useful: Use $\mathcal{F} \subset (\mathcal{S}\times\mathcal{A} \to \mathbb{R})$ to model Q^{π_e}, and let's find w such that

$$\mathbb{E}_\mu[w(s,a)\cdot(f(s,a) - \gamma f(s',\pi_e))] = (1-\gamma)\mathbb{E}_{s\sim d_0}[f(s,\pi_e)], \ \ \forall f \in \mathcal{F}. \tag{4}$$

Note that adding other functions to $\mathcal{F}$ does not hurt, because Eq.(4) always holds for any f when $w = w_{\pi_e/\mu}$: note that $\mathbb{E}_\mu[w_{\pi_e/\mu}(s,a)\cdot(\cdot)] = \mathbb{E}_{d_{\pi_e,\gamma}}[(\cdot)]$, and under $d_{\pi_e,\gamma}$ the terms in Eq.(4) telescope and cancel out nicely.

And that's it! All you need to do is to use another function class $\mathcal{W}$ to model $w_{\pi_e/\mu}$, and find the best $w \in \mathcal{W}$ that minimizes the worst (over $f \in \mathcal{F}$) violation of Eq.(4).

The realizability conditions on both $\mathcal{F}$ and $\mathcal{W}$ can be relaxed, which we do not further discuss here. See Uehara and Jiang (2019) for details.

The original algorithm by Liu et al. (2018) MWL is inspired by and extends the algorithm by Liu et al. (2018), who use importance weights and value functions of states (instead of state-action pairs) and requires knowledge of the behavior policy ($\mu(a|s)$ in our notation). Their algorithm can be derived similarly, e.g., by noting that reward r can be replaced by $V^{\pi_e}(s) - \gamma\frac{\pi_e(a|s)}{\mu(a|s)}V^{\pi_e}(s')$. We leave the derivation as an exercise to the readers.

2 MQL

In Uehara and Jiang (2019) we find that things become more interesting when we flip w and f around. Consider the same setting as in the previous section, and now let's try to learn a Q-function for OPE.

Format of Estimator We want to find $q : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ such that

$$(1-\gamma)\mathbb{E}_{s\sim d_0}[q(s,\pi_e)] = R_{\pi_e}. \tag{5}$$

Note that $q = Q^{\pi_e}$ satisfies the equation.

Derivation Key observation: for any q,[1]

$$(1-\gamma)\mathbb{E}_{s\sim d_0}[q(s,\pi_e)] - R_{\pi_e} = \mathbb{E}_{(s,a)\sim d_{\pi_e,\gamma}}[q(s,a) - r - \gamma q(s',\pi_e)]. \tag{6}$$

To prove this identity, note that $\mathbb{E}_{(s,a)\sim d_{\pi_e,\gamma}}[r]$ on the RHS is just R_{π_e}, and the remaining terms telescope and cancel out once we rewrite $d_{\pi_e,\gamma}$ as the discounted sum of distributions.

We can use a function class $\mathcal{Q}$ to model Q^{π_e} and try to minimize the RHS of Eq.(6). The trouble is that we do not have access to $(s,a)\sim d_{\pi_e,\gamma}$, but this can be resolved if we have the importance weight $w_{\pi_e/\mu}$, since the RHS of Eq.(6) is equal to

$$\mathbb{E}_{(s,a)\sim\mu}[w_{\pi_e/\mu}(s,a)\cdot(q(s,a) - r - \gamma q(s',\pi_e))].$$

While we do not know $w_{\pi_e/\mu}$ either, we can use another rich function class to model $w_{\pi_e/\mu}$, and find $q \in \mathcal{Q}$ that satisfies the equation for all weights in the function class. The rest of the derivation is similar to MWL.

References

Jiang, N., A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*.

Liu, Q., L. Li, Z. Tang, and D. Zhou (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371.

Uehara, M. and N. Jiang (2019). Minimax Weight and Q-Function Learning for Off-Policy Evaluation. *arXiv:1910.12809*.

[1] The finite-horizon version of this identity has been given in Jiang et al. (2017, Lemma 1).