More efficient Off-Policy Evaluation through Regularized Targeted Learning

Aurelien F. Bibaut, Ivana Malenica, Nikos Vlassis, Mark J. van der Laan

University of California, Berkeley
Netflix, Los Gatos, CA

aurelien.bibaut@berkeley.edu

June 8, 2019

Problem statement

What is Off-Policy Evaluation?

- Data: MDP trajectories collected under behavior policy π_b.
- Question: What would be mean reward under target policy π_e?

Why OPE? When too costly/dangerous/unethical to just try out π_e.

This work:

A novel estimator for OPE in reinforcement learning.

Formalization

S_t : state at t, A_t : action at t, R_t : reward at t,

π_b : logging/behavior policy, π_e : target policy,

$$\rho_t := \prod_{t=1}^{T} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} \text{ : importance sampling ratio.}$$

Action-value/reward-to-go function:

$$Q_t^{\pi_e}(s,a) := E_{\pi_e}\left[\sum_{\tau \geq t} R_t \Big| S_t = s, A_t = a\right].$$

Our estimand: value function

$$V^{\pi_e}(\boldsymbol{Q}^{\pi_e}) := E_{\pi_e}\left[Q_1^{\pi_e}(S_1, A_1)|S_1 = s_1\right] \text{ (fix the initial state to } s_1\text{.)}$$

Our base estimator

Overview of longitudinal TMLE

Say we have an estimator $\hat{\boldsymbol{Q}} = (\hat{Q}_1, ..., \hat{Q}_T)$ of $\boldsymbol{Q}^{\pi_e} = (Q_1^{\pi_e}, ..., Q_T^{\pi_e})$ (e.g. SARSA or dynamics estimators).
m

Traditional Direct Model estimator: $\hat{V} := V_1^{\pi_e}(\hat{\boldsymbol{Q}})$

LTMLE:

- Define, for $t = 1, ..., T$, **logistic intercept model**,

$$\hat{Q}_t(\epsilon_t)(s,a) = 2 \underbrace{\Delta_t}_{\substack{\text{max} \\ \text{r.t.g.}}} \left(\underbrace{\sigma}_{\substack{\text{logit} \\ \text{link}}} \left(\sigma^{-1} \left(\frac{\hat{Q}_t(s,a) + \Delta_t}{2\Delta_t} \right) + \epsilon \right) - 0.5 \right).$$

- Fit $\hat{\epsilon_t}$ by maximum weighted likelihood
- Define $\hat{V}^{LTMLE} := V_1^{\pi_e}(\hat{Q}_1(\hat{\epsilon_1}))$

Our base estimator

Loss and recursive fitting

Log likelihood of for logistic intercept at t:

$$l_t(\hat{\epsilon}_{t+1})(\epsilon_t) := \rho_t \Bigg\{ \underbrace{\frac{R_t + \hat{V}_{t+1}(\hat{\epsilon}_{t+1}) + \Delta_t}{2\Delta_t}}_{\text{normalized r.t.g.}} \log \underbrace{\left(\frac{\hat{Q}_t(\epsilon_t) + \Delta_t}{2\Delta_t} \right)}_{\substack{\text{normalized} \\ \text{predicted r.t.g.}}} + \left(1 - \frac{R_t + \hat{V}_{t+1}(\hat{\epsilon}_{t+1}) + \Delta_t}{2\Delta_t} \right) \log \left(1 - \frac{\hat{Q}_t(\epsilon_t) + \Delta_t}{2\Delta_t} \right) \Bigg\}.$$

Recursive fitting: Likelihood for ϵ_t requires fitted $\hat{\epsilon}_{t+1} \implies$ proceed backwards in time.

Our base estimator

Regularizations

Softening. Trajectories $i = 1, ..., n$ with IS ratios $\rho_t^{(1)}, ..., \rho_t^{(n)}$. For $0 < \alpha < 1$, replace IS ratios by

$$\frac{(\rho_t^{(i)})^\alpha}{\sum_j (\rho_t^{(j)})^\alpha}.$$

Partialing. For some τ, set $\hat{\epsilon}_\tau = ...\hat{\epsilon}_T = 0$.

Penalization. Add L_1-penalty $\lambda|\epsilon_t|$ to each l_t.

Our ensemble estimator

- Make a pool of regularized estimators $\boldsymbol{g} := (g_1, ...g_K)$.
- $\hat{\boldsymbol{\Omega}}_n$: bootstrap estimate of $Cov(\boldsymbol{g})$.
- $\hat{\boldsymbol{b}}_n$: bootstrap estimate of bias of $\boldsymbol{g}$.
- Compute
$$\hat{\boldsymbol{x}} = \arg \min_{\substack{\mathbf{0} \leq \boldsymbol{x} \leq \mathbf{1} \\ \boldsymbol{x}^\top \mathbf{1} = 1}} \frac{1}{n} \boldsymbol{x}^\top \hat{\boldsymbol{\Omega}}_n \boldsymbol{x} + (\boldsymbol{x}^\top \hat{\boldsymbol{b}}_n)^2.$$
- Return
$$\hat{V}^{RLTMLE} = \hat{\boldsymbol{x}}^\top \boldsymbol{g}.$$

Empirical performance