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 $\pi_b$.
- **Question:** What would be mean reward under target policy $\pi_e$?

Why OPE? When too costly/dangerous/unethical to just try out $\pi_e$.

This work:

A novel estimator for OPE in reinforcement learning.
Formalization

\[ S_t : \text{state at } t, \quad A_t : \text{action at } t, \quad R_t : \text{reward at } t, \]
\[ \pi_b : \text{logging/behavior policy}, \quad \pi_e : \text{target policy}, \]
\[ \rho_t := \prod_{t=1}^{T} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} : \text{importance sampling ratio.} \]

Action-value/reward-to-go function:

\[ Q_{\pi_e}^T(s, a) := E_{\pi_e} \left[ \sum_{\tau \geq t} R_{\tau} \mid S_{\tau} = s, A_{\tau} = a \right]. \]

Our estimand: value function

\[ V_{\pi_e}(Q_{\pi_e}) := E_{\pi_e} [Q_{\pi_e}^1(S_1, A_1) \mid S_1 = s_1] \text{ (fix the initial state to } s_1). \]
Our base estimator

Overview of longitudinal TMLE

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

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

LTMLE:

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

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

- 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 $t$:

\[
l_t(\hat{\epsilon}_{t+1})(\epsilon_t) := \rho_t \left\{ \frac{R_t + \hat{V}_{t+1}(\hat{\epsilon}_{t+1}) + \Delta_t}{2\Delta_t} \log \left( \frac{\hat{Q}_t(\epsilon_t) + \Delta_t}{2\Delta_t} \right) \right. \\
\left. + \left( 1 - \frac{R_t + \hat{V}_{t+1}(\hat{\epsilon}_{t+1}) + \Delta_t}{2\Delta_t} \right) \log \left( 1 - \frac{\hat{Q}_t(\epsilon_t) + \Delta_t}{2\Delta_t} \right) \right\}. \]

Recursive fitting: Likelihood for $\epsilon_t$ requires fitted $\hat{\epsilon}_{t+1} \implies$ proceed backwards in time.
Our base estimator
Regularizations

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

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

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

**Penalization.** Add $L_1$-penalty $\lambda |\epsilon_t|$ to each $l_t$. 
Our ensemble estimator

- Make a pool of regularized estimators $g := (g_1, \ldots, g_K)$.
- $\hat{\Omega}_n$: bootstrap estimate of $\text{Cov}(g)$.
- $\hat{b}_n$: bootstrap estimate of bias of $g$.
- Compute

  $$\hat{x} = \arg \min_{0 \leq x \leq 1} \frac{1}{n} x^\top \hat{\Omega}_n x + (x^\top \hat{b}_n)^2.$$  

- Return

  $$\hat{V}^{RLTMLE} = \hat{x}^\top g.$$
Empirical performance

(a) GridWorld

(b) ModelFail

(c) ModelWin low bias

(d) ModelWin high bias

estimator
- MAGIC
- RLTMLE 1
- RLTMLE 2
- WDR

log10(n \times MSE) vs log10(n)