Intrinsically Efficient, Stable, and Bounded Off-Policy Evaluation for RL

Nathan Kallus , Masatoshi Uehara

Cornell University and Cornell Tech, Harvard University

Summary

We propose new estimators for OPE based on empirical likelihood that are always more efficient than IS (importance sampling), SNIS (self normalized importance sampling), and DR (doubly robust) and satisfy the same stability and boundedness properties as SNIS.

Introduction

Contextual Bandit Setting

- State space $\mathcal{X}$, Action space $\mathcal{A}$
- Reward space $[0, R_{\max}]$
- Evaluation Policy; $\pi_e : \mathcal{X} \times \mathcal{A} \to [0, 1]$
- Behavior Policy; $\pi_b : \mathcal{X} \times \mathcal{A} \to [0, 1]$

We want to estimate $\beta_T^{\pi_e} = \mathrm{E}_{\pi_e}[r]$ from n observation from the behavior policy $\{x^{(i)}, a^{(i)}, r^{(i)}\}_{i=1}^n$.

Existing approaches

Define

$$\omega = \pi_e(a|x)/\pi_b(a|x),\ q(a,x) = \mathrm{E}[r|a,x]$$
$$v(x) = \mathrm{E}_{a\sim\pi_e(x)}[q(a,x)].$$

- **IS**; $\hat{\beta}_{\mathrm{IS}} = \mathrm{E}_n[\omega r]$
- **SIS**; $\hat{\beta}_{\mathrm{SIS}} = \mathrm{E}_n[\omega r]/\mathrm{E}_n[\omega]$
- **DR**; $\hat{\beta}_{\mathrm{dr}} = \hat{\beta}_d(\hat{q}(x,a))$, where $\hat{\beta}_d(m) = \mathrm{E}_n\left[wr - \mathcal{F}(m)\right]$, $\mathcal{F}(m(x,a)) = wm(x,a) - \left\{\Sigma_{a\in A}\, m(x,a)\pi_e(a|x)\right\}$.

Definition

- Local efficiency - When the model $q(x,a;\tau)$ is well-specified, the estimator achieves the efficiency bound.
- α-Boundedness - Bounded by $\alpha R_{\max}$.
- Stability - If the conditional variance of $r^{(i)}$, given $\mathcal{D}_{x,a}$, is bounded by σ^2, then the conditional variance of the estimator, given $\mathcal{D}_{x,a}$, is also bounded by σ^2 ($\mathcal{D}_{x,a} = \{(x^{(i)}, a^{(i)}) : i \le n\}$).
- Intrinsic efficiency — The asymptotic MSE of the estimator is smaller than that of any of $\hat{\beta}_{\mathrm{sis}}, \hat{\beta}_{\mathrm{is}}, \hat{\beta}_{\mathrm{sis}}, \hat{\beta}_{\mathrm{dr}}$, irrespective of model specification.

REG

We let the REG estimator be $\hat{\beta}_{\mathrm{reg}} = \hat{\beta}_d(\hat{\zeta}_1 + \hat{\zeta}_2 q(x,a;\hat{\tau}))$ where we choose the parameters by minimizing the estimated variance:

$$(\hat{\zeta}, \hat{\tau}) = \underset{\zeta\in\mathbb{R}^2, \tau\in\Theta_\tau}{\arg\min}\ \mathrm{E}_n\left[\left\{wr - \mathcal{F}(\zeta_1 + \zeta_2 q(x,a;\tau))\right\}^2\right].$$

- When $\zeta_1 = 0, \zeta_2 = 1$, more doubly robust estimator (MDR).
- When $\zeta_2 = 1, \zeta_2 = 0$, IS estimator
- When $\zeta_2 = \beta^*, \zeta_2 = 0$, SIS estimator

EMP

We let the EMP estimator be

$$\hat{\beta}_{\mathrm{emp}} = \mathrm{E}_n\left[\hat{c}^{-1}\hat{\kappa}(x,a)\pi_e(a|x)r\right], \quad \text{where}$$
$$\hat{\kappa}(x,a) = \{\pi_b(a|x)[1 + \mathcal{F}(m(x,a;\hat{\xi},\hat{\tau}))]\}^{-1},$$
$$\hat{c} = \mathrm{E}_n\left[\{1 + \mathcal{F}(m(x,a;\hat{\xi},\hat{\tau}))\}^{-1}\right],$$
$$\hat{\xi}, \hat{\tau} = \underset{\xi\in\mathbb{R}, \tau\in\Theta_\tau}{\arg\max}\ \mathrm{E}_n[\log\{1 + \mathcal{F}(m(x,a;\xi,\tau))\}].$$

This is based on empirical likelihood

$$\max_{\kappa} \sum_{i=1}^n \log \kappa^{(i)},\ \text{s.t.} \sum_{i=1}^n \kappa^{(i)}\pi_b(a^{(i)}|x^{(i)}) = 1,$$
$$\sum_{i=1}^n \kappa^{(i)}\pi_b(a^{(i)}|x^{(i)})\mathcal{F}(m(x^{(i)}, a^{(i)};\xi,\tau)) = 0.$$

The dual formulation derives the estimator.

Practical REG and EMP

We estimate a parameter τ in $q(x,a;\tau)$ as in DM to obtain $\hat{\tau}$, which we assume has a limit, $\hat{\tau} \xrightarrow{p} \tau^\dagger$. Then, we consider solving the following optimization problems for REG and EMP, respectively

$$\hat{\zeta} = \underset{\zeta\in\mathbb{R}^2}{\arg\min}\ \mathrm{E}_n\left[\left\{wr - \mathcal{F}(m(x,a;\zeta,\hat{\tau}))\right\}^2\right],$$
$$\hat{\xi} = \underset{\xi\in\mathbb{R}^2}{\arg\max}\ \mathrm{E}_n[\log\{1 + \mathcal{F}(m(x,a;\xi,\hat{\tau}))\}],$$

where $m(x,a;\zeta,\hat{\tau}) = \zeta_1 + \zeta_2 q(x,a;\hat{\tau})$ or $m(x,a;\xi,\hat{\tau}) = \xi_1 + \xi_2 q(x,a;\hat{\tau})$.

- Optimization becomes easier
- Desires properties are still kept

Theoretical results

Theorem; The estimators $\hat{\beta}_{\mathrm{emp}}$, $\hat{\beta}_{\mathrm{reg}}$ have local and intrinsic efficiency, and $\mathrm{Asmse}[\hat{\beta}_{\mathrm{reg}}] = \mathrm{Asmse}[\hat{\beta}_{\mathrm{emp}}]$

$$n^{-1} \min_{\zeta\in\mathbb{R}, \tau\in\mathbb{R}^{d_\tau}} \mathrm{E}\left[\left\{wr - \mathcal{F}(\zeta + q(x,a;\tau))\right\}^2 - \beta^{*2}\right].$$

The estimator $\hat{\beta}_{\mathrm{emp}}$ satisfies 1-boundedness and partial stability.

Theorem; The plug-in-τ versions of $\hat{\beta}_{\mathrm{reg}}$ and $\hat{\beta}_{\mathrm{emp}}$ still satisfy local and intrinsic efficiency, and $\hat{\beta}_{\mathrm{emp}}$ satisfies 1-boundedness and partial stability. Their asymptotic MSEs are

$$n^{-1} \min_{\zeta\in\mathbb{R}^2} \mathrm{E}\left[\left\{wr - \mathcal{F}(\zeta_1 + \zeta_2 q(x,a;\tau^\dagger))\right\}^2 - \beta^{*2}\right]. \quad (1)$$

Extension to RL

In RL, the goal is evaluating $\mathrm{E}_{\pi_e}[\Sigma\gamma^t r_t]$. We can do the same technique. The intrinsic efficiency is defined as a superiority over

$$\hat{\beta}_{\mathrm{snis}} = \frac{\mathrm{E}_n\left[\omega_{0:T-1}\,\Sigma_{t=0}^{T-1}\gamma^t r_t\right]}{\mathrm{E}_n[\omega_{0:T-1}]},\ \hat{\beta}_{\mathrm{snsis}} = \mathrm{E}_n\left[\sum_{t=0}^{T-1}\frac{\omega_{0:t}}{\mathrm{E}_n[\omega_{0:t}]}\gamma^t r_t\right]$$

$\hat{\beta}_{\mathrm{dr}} = \hat{\beta}_d(\{q(x,a;\hat{\tau})\}_{t=0}^{T-1})$, where $\hat{\beta}_d(\{m_t\}_{t=0}^{T-1})$is

$$\mathrm{E}_n\left[\sum_{t=0}^{T-1}\gamma^t\omega_{0:t}r_t - \gamma^t\left(\omega_{0:t}m_t(x_t,a_t) - m_t(x_t,\pi_e)\right)\right].$$

To derive REG, consider the class of estimators $\hat{\beta}_d(\{m_t\}_{t=0}^{T-1})$ where

$m_t(x_t,a_t;\zeta) = \zeta_{1t} + \zeta_{2t}q(x_t,a_t;\hat{\tau})\,(0 \le t \le T-1)$.

Then, we define an estimator $\hat{\zeta}$ as

$$\hat{\zeta} = \underset{\zeta\in\mathbb{R}^2}{\arg\min}\ \mathrm{E}_n[v(\{m_t(x_t,a_t;\zeta)\}_{t=0}^{T-1})]. \quad (2)$$

Same as EMP.

Experiment

We compare common methods with the proposed methods.

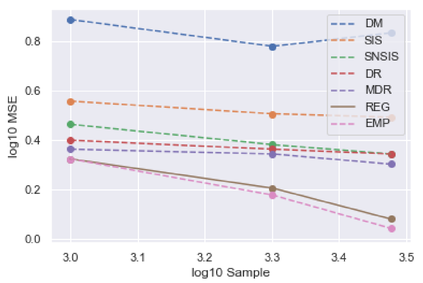

Figure 1:Cliff Walking

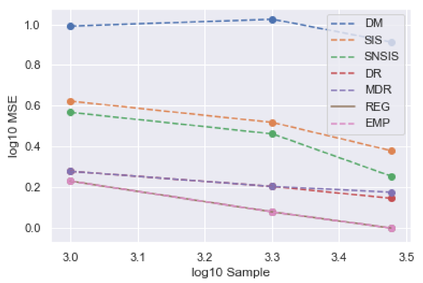

Figure 2:Mountain Car

Important Comparison

Table 1:Comparison of policy evaluation methods. The notation (*) means proposed estimator. The notation # means partially satisfied, as discussed in the text. (S)IS and SN(S)IS refer either to stepwise or non-stepwise.

	DM	(S)IS	SN(S)IS	DR	SNDR	MDR	REG(*)	SNREG(*)	EMP(*)
Consistency		✓	✓	✓	✓	✓	✓	✓	✓
Local efficiency	✓			✓	✓	✓	✓	✓	✓
Intrinsic efficiency						#	✓	#	✓
Boundedness	1		1		2			2	1
Stability	#		✓		#				✓