Confident Off-Policy Evaluation and Selection through Self-Normalized Importance Weighting

Csaba Szepesvári

June 25, 2020

University of Alberta and DeepMind
IST, Lisbon, June 25, 2020

Collaborators

- Ilja Kuzborskij
- Claire Vernade
- András György

Off-Policy Contextual Bandit Model

Model: $(P_X, P_{R|X,A}, \pi_b)$

- P_X – prob. measure over context space $\mathcal{X}$
- $P_{R|X,A}$ – prob. kernel producing reward dist. given $X \in \mathcal{X}$ and action $A \in [K]$
- π_b – behaviour policy, e.g. $\pi_b(\cdot|X)$

Off-Policy Contextual Bandit Model

Model: $(P_X, P_{R|X,A}, \pi_b)$

- P_X – prob. measure over context space $\mathcal{X}$
- $P_{R|X,A}$ – prob. kernel producing reward dist. given $X \in \mathcal{X}$ and action $A \in [K]$
- π_b – behaviour policy, e.g. $\pi_b(\cdot|X)$

Contextual off-policy evaluation problem

- An agent observes indep. $S = ((X_1, A_1, R_1), \dots, (X_n, A_n, R_n))$
  $A_i \sim \pi_b(\cdot|X_i)$, $X_i \sim P_X$, $R_i \sim P_{R|X,A}$
- An agent follows a randomized *target policy* π

Goal: estimate the value $v(\pi)$ of that policy:

$$v(\pi) = \int_{\mathcal{X}} \sum_{a \in [K]} \pi(a|x) r(x,a) \, \mathrm{d}P_X(x)$$

where $r(x,a) = \int u \, \mathrm{d}P_{R|X,A}(u|x,a)$.

Value estimation through Importance Sampling

Many ways to do that...

At the core of many is to use *importance weights*

$$W_i = \frac{\pi(A_i|X_i)}{\pi_b(A_i|X_i)} \qquad i \in [n] \ .$$

For example, (unbiased) *importance sampling* estimator

$$\hat{v}^{\mathrm{IS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} W_i R_i \ .$$

Value estimation through Importance Sampling

Many ways to do that...

At the core of many is to use *importance weights*

$$W_i = \frac{\pi(A_i|X_i)}{\pi_b(A_i|X_i)} \qquad i \in [n] \ .$$

For example, (unbiased) *importance sampling* estimator

$$\hat{v}^{\mathrm{IS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} W_i R_i \ .$$

High variance!

For example, $W_i \sim p$, where p is heavy-tailed (disagreeing policies)

Value estimation through DR

Another popular estimator is *Doubly-Robust* estimator

$$\hat{v}^{\mathrm{DR}}(\pi) = \frac{1}{n}\sum_i \pi(A_i|X_i)\hat{\eta}(X_i, A_i) + \frac{1}{n}\sum_i W_i(R_i - \hat{\eta}(X_i, A_i)),$$

for some fixed $\hat{\eta} : (x, a) \to [0, 1]$ (typically a reward estimator learned on a held-out dataset).

- Unbiased
- Reduces variance, but we need a reward modeling (training, tuning, dataset splitting)...

Value estimation through Importance Sampling

Something simpler — a *weighted importance sampling* estimator

$$\hat{v}^{\mathrm{WIS}}(\pi) = \frac{\sum_{i=1}^{n} W_i R_i}{\sum_{i=1}^{n} W_i} \ .$$

- *Biased* (asymptotically unbiased (IID))
- In practice, low variance (self-normalization)

Some intuition: $\mathrm{Var}(\hat{v}^{\mathrm{WIS}}(\pi)) \leq \mathbb{E}\left[\sum_k \frac{W_k^2}{\left(\sum_i W_i\right)^2}\right]$

What about $v(\pi)$?

- Of course, estimator alone is not enough. We want:

$$1 - e^{-x} \le \mathbb{P}\Big(\hat{v}(\pi) + \varepsilon(x, S, \pi, \pi_b) \le v(\pi)\Big) \qquad x > 0 \ .$$

Some challenges:

- Even for basic importance sampling $(W_1 R_1 + \cdots + W_n R_n)/n$ it's non-trivial: unbiased, but W_i are **unbounded**
  - Excludes Hoeffding/Bernstein/McDiarmid

What about $v(\pi)$?

- Of course, estimator alone is not enough. We want:

$$1 - e^{-x} \le \mathbb{P}\Big(\hat{v}(\pi) + \varepsilon(x, S, \pi, \pi_b) \le v(\pi)\Big) \qquad x > 0 \ .$$

Some challenges:

- Even for basic importance sampling $(W_1 R_1 + \cdots + W_n R_n)/n$ it's non-trivial: unbiased, but W_i are **unbounded**
  - Excludes Hoeffding/Bernstein/McDiarmid
  - We can "truncate", e.g. $W_i^\lambda = \pi(A_i|X_i)/(\pi_b(A_i|X_i) + \lambda)$ for some h.p. $\lambda > 0$.
  - Ugly! Needs tuning, doesn't always work...

What about $v(\pi)$?

- Of course, estimator alone is not enough. We want:

$$1 - e^{-x} \leq \mathbb{P}\Big(\hat{v}(\pi) + \varepsilon(x, S, \pi, \pi_b) \leq v(\pi)\Big) \qquad x > 0 \ .$$

Some challenges:

- Even for basic importance sampling $(W_1 R_1 + \cdots + W_n R_n)/n$ it's non-trivial: unbiased, but W_i are **unbounded**
  - Excludes Hoeffding/Bernstein/McDiarmid
  - We can "truncate", e.g. $W_i^\lambda = \pi(A_i|X_i)/(\pi_b(A_i|X_i) + \lambda)$ for some h.p. $\lambda > 0$.
  - Ugly! Needs tuning, doesn't always work...
- Variance is important: need bounds with empirical variance.
- Sometimes, estimator is not a sum of indep. elements (self-normalization).

Semi-empirical Efron-Stein Bound for WIS

Let's go back and pick WIS:

$$\hat{v}^{\mathrm{WIS}}(\pi) = \frac{1}{Z}\sum_{i=1}^{n} W_i R_i \ , \qquad Z = \sum_{i=1}^{n} W_i \ .$$

Theorem W.h.p.,

$$v(\pi) \overset{\widetilde{\Omega}}{=} \left(B \cdot \left(\hat{v}^{\mathrm{WIS}}(\pi) - \sqrt{V^{\mathrm{WIS}} + \frac{1}{n}} \right) - \frac{1}{\sqrt{n}} \right)_+$$

$$V^{\mathrm{WIS}} = \sum_{k=1}^{n} \mathbb{E}\left[\left(\frac{W_k}{Z} + \frac{W_k'}{Z^{(k)}} \right)^2 \,\middle|\, W_1^k, X_1^n \right] \qquad (\text{"variance"})$$

$$B = \min\left(\mathbb{E}\left[\frac{n}{Z} \,\middle|\, X_1^n \right]^{-1}, 1 \right) \ , \qquad (\text{bias})$$

where $Z^{(k)} = Z + (W_k' - W_k)$, and W_k' indep. dist. as W_k.

Semi-empirical Efron-Stein Bound for WIS

Theorem W.h.p.,

$$v(\pi) \overset{\widetilde{\Omega}}{=} \left(B \cdot \left(\hat{v}^{\mathrm{WIS}}(\pi) - \sqrt{V^{\mathrm{WIS}} + \frac{1}{n}} \right) - \frac{1}{\sqrt{n}} \right)_+$$

$$V^{\mathrm{WIS}} = \sum_{k=1}^{n} \mathbb{E}\left[\left(\frac{W_k}{Z} + \frac{W_k'}{Z^{(k)}} \right)^2 \,\middle|\, W_1^k, X_1^n \right] \qquad (\text{"variance"})$$

$$B = \min\left(\mathbb{E}\left[\frac{n}{Z} \,\middle|\, X_1^n \right]^{-1}, 1 \right) \,, \qquad (\text{bias})$$

where $Z^{(k)} = Z + (W_k' - W_k)$, and W_k' indep. dist. as W_k.

- No truncation! No hyperparameters.
- Contexts are fixed.
- Needs knowledge of π_b — only partly empirical:

  V^{WIS} and B can be computed exactly. Cost: n^K :-(

  Can approximate using Monte-Carlo simulation! :-)

Semi-empirical Efron-Stein Bound for WIS

Theorem W.h.p.,

$$v(\pi) \overset{\widetilde{\Omega}}{=} \left(B \cdot \left(\hat{v}^{\mathrm{WIS}}(\pi) - \sqrt{V^{\mathrm{WIS}} + \frac{1}{n}} \right) - \frac{1}{\sqrt{n}} \right)_+$$

$$V^{\mathrm{WIS}} = \sum_{k=1}^{n} \mathbb{E}\left[\left(\frac{W_k}{Z} + \frac{W'_k}{Z^{(k)}} \right)^2 \middle| W_1^k, X_1^n \right] \qquad \text{("variance")}$$

$$B = \min\left(\mathbb{E}\left[\frac{n}{Z} \middle| X_1^n \right]^{-1}, 1 \right) \ , \qquad \text{(bias)}$$

where $Z^{(k)} = Z + (W'_k - W_k)$, and W'_k indep. dist. as W_k.

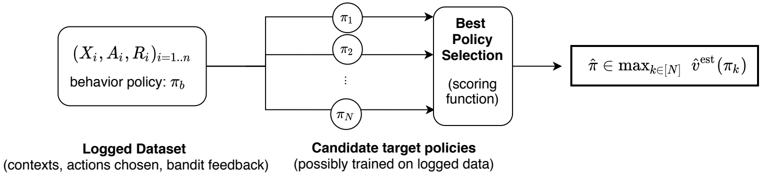
- No truncation! No hyperparameters.
- Contexts are fixed.

Recall some intuition: $\mathrm{Var}(\hat{v}^{\mathrm{WIS}}(\pi)) \le \mathbb{E}\left[\sum_k \left(\frac{W_k^2}{Z} \right)^2 \right]$

Is it any good?

The Best Policy Identification problem

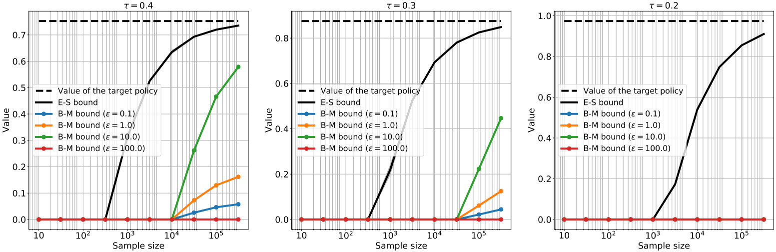
- We have a finite set of target policies Π.
- We do $\hat{\pi} \in \arg\max_{\pi \in \Pi} \hat{v}^{\text{est}}(\pi)$.
- We want to maximize $v(\hat{\pi})$
  — we'll use confidence bounds as $\hat{v}^{\text{est}}$.



Synthetic Experiments – Setup

- Fix $K > 0$
- $\pi_b(a) \propto e^{\frac{1}{\tau}\mathbb{I}\{a=1\}}$
- $\pi(a) \propto e^{\frac{1}{\tau}\mathbb{I}\{a=1\}}$
- $R_i = \mathbb{I}\{A_i = k\}$, $A_i \sim \pi_b(\cdot)$
- As $\tau \to 0$, π_b and π become increasingly misaligned

Results



τ = 0.4
τ = 0.3
τ = 0.2
Value
Sample size
Value of the target policy
E-S bound
B-M bound (ε = 0.1)
B-M bound (ε = 1.0)
B-M bound (ε = 10.0)
B-M bound (ε = 100.0)

Nonsynthetic Experiments – Setup

Target policies are $\left\{\pi^{\mathrm{ideal}}, \pi^{\hat{\mathbf{\Theta}}_{\mathsf{IS}}}, \pi^{\hat{\mathbf{\Theta}}_{\mathsf{WIS}}}\right\}$ where

$$\pi^{\mathbf{\Theta}}(y = k \mid \boldsymbol{x}) \propto e^{\frac{1}{\tau}\boldsymbol{x}^\top \boldsymbol{\theta}_k}$$

with two choices of parameters given by the optimization problems:

$$\hat{\mathbf{\Theta}}_{\mathsf{IS}} \in \operatorname*{arg\,min}_{\mathbf{\Theta} \in \mathbb{R}^{d \times K}} \hat{v}^{\mathrm{IS}}(\pi^{\mathbf{\Theta}}) \ , \qquad \hat{\mathbf{\Theta}}_{\mathsf{WIS}} \in \operatorname*{arg\,min}_{\mathbf{\Theta} \in \mathbb{R}^{d \times K}} \hat{v}^{\mathrm{WIS}}(\pi^{\mathbf{\Theta}}) \ .$$

- Trained by GD with $\eta = 0.01$, $T = 10^5$.
- $\tau = 0.1$ — cold! Almost deterministic.

Table: Average test rewards of the target policy when chosen by each method of the benchmark.

name Size	Ecoli 336	Vehicle 846	Yeast 1484	PageBlok 5473	OptDigits 5620	SatImage 6435	PenDigits 10992
ESLB	**0.913 ± 0.263**	**0.716 ± 0.389**	**0.912 ± 0.267**	**0.910 ± 0.270**	**0.843 ± 0.325**	**0.910 ± 0.270**	**0.910 ± 0.270**
DR	0.656 ± 0.410	0.610 ± 0.443	0.563 ± 0.392	0.888 ± 0.291	0.616 ± 0.344	0.423 ± 0.361	0.565 ± 0.382
IS (trunc+Bern)	$-\infty$	$-\infty$	**0.916 ± 0.262**	**0.910 ± 0.270**	0.748 ± 0.404	0.658 ± 0.413	0.810 ± 0.345
Chebyshev-WIS	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
Emp.Lik.	0.511 ± 0.298	0.455 ± 0.405	0.312 ± 0.325	0.669 ± 0.409	0.285 ± 0.359	0.634 ± 0.409	0.549 ± 0.426

Proof sketch

$$\underbrace{v(\pi) - \mathbb{E}\left[v(\pi) \,|\, X_1^n\right]}_{\textbf{Concentration of contexts}} + \underbrace{\mathbb{E}\left[v(\pi) \,|\, X_1^n\right] - \mathbb{E}\left[\hat{v}^{\textbf{wis}}(\pi) \mid X_1^n\right]}_{\textbf{Bias}} + \underbrace{\mathbb{E}\left[\hat{v}^{\textbf{wis}}(\pi) \mid X_1^n\right] - \hat{v}^{\textbf{wis}}(\pi)}_{\textbf{Concentration}}$$

1. Concentration of contexts – Hoeffding since X_1^n are IID. $\mathbb{E}\left[v(\pi) \,|\, X_1^n\right] = \mathbb{E}\left[\frac{1}{n}\sum_i W_i R_i \,|\, X_1^n\right]$.
2. Bias – IS is unbiased, let's try to "split" WIS into IS and denominator.

Proof sketch

$$\underbrace{v(\pi) - \mathbb{E}\left[v(\pi) \mid X_1^n\right]}_{\textbf{Concentration of contexts}} + \underbrace{\mathbb{E}\left[v(\pi) \mid X_1^n\right] - \mathbb{E}\left[\hat{v}^{\textbf{wis}}(\pi) \mid X_1^n\right]}_{\textbf{Bias}} + \underbrace{\mathbb{E}\left[\hat{v}^{\textbf{wis}}(\pi) \mid X_1^n\right] - \hat{v}^{\textbf{wis}}(\pi)}_{\textbf{Concentration}}$$

1. Concentration of contexts – Hoeffding since X_1^n are IID.
   $\mathbb{E}\left[v(\pi) \mid X_1^n\right] = \mathbb{E}\left[\frac{1}{n}\sum_i W_i R_i \mid X_1^n\right]$.
2. Bias – IS is unbiased, let's try to "split" WIS into IS and denominator.
   Harris' inequality. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a non-increasing and $g : \mathbb{R}^n \to \mathbb{R}$ be a non-decreasing function. Then for real-valued random variables $(X_1, \ldots, X_n)$ independent from each other, we have

   $$\mathbb{E}[f(X_1, \ldots, X_n) g(X_1, \ldots, X_n)] \leq \mathbb{E}[f(X_1, \ldots, X_n)]\,\mathbb{E}[g(X_1, \ldots, X_n)]\,.$$

   This gives us:

   $$\mathbb{E}\left[\frac{\sum_{k=1}^n W_k R_k}{\sum_{k=1}^n W_k} \;\middle|\; X_1^n\right] \leq \mathbb{E}\left[\frac{1}{\sum_{k=1}^n W_k} \;\middle|\; X_1^n\right] \mathbb{E}\left[\sum_{k=1}^n W_k R_k \;\middle|\; X_1^n\right]$$

Proof sketch

$$\underbrace{v(\pi) - \mathbb{E}\left[v(\pi) \mid X_1^n\right]}_{\textbf{Concentration of contexts}} + \underbrace{\mathbb{E}\left[v(\pi) \mid X_1^n\right] - \mathbb{E}\left[\hat{v}^{\textbf{wis}}(\pi) \mid X_1^n\right]}_{\textbf{Bias}} + \underbrace{\mathbb{E}\left[\hat{v}^{\textbf{wis}}(\pi) \mid X_1^n\right] - \hat{v}^{\textbf{wis}}(\pi)}_{\textbf{Concentration}}$$

Concentration... (Remember) Some challenges:

- Even for basic importance sampling $(W_1 R_1 + \cdots + W_n R_n)/n$ it's non-trivial: unbiased, but W_i are **unbounded**
  - Excludes Hoeffding/Bernstein/McDiarmid
  - We can "truncate", e.g. $W_i^\lambda = \pi(A_i|X_i)/(\pi_b(A_i|X_i) + \lambda)$ for some h.p. $\lambda > 0$.
  - Ugly! Needs tuning, doesn't always work...

Proof sketch

$$\underbrace{v(\pi) - \mathbb{E}\left[v(\pi) \mid X_1^n\right]}_{\textbf{Concentration of contexts}} + \underbrace{\mathbb{E}\left[v(\pi) \mid X_1^n\right] - \mathbb{E}\left[\hat{v}^{\mathbf{wis}}(\pi) \mid X_1^n\right]}_{\textbf{Bias}} + \underbrace{\mathbb{E}\left[\hat{v}^{\mathbf{wis}}(\pi) \mid X_1^n\right] - \hat{v}^{\mathbf{wis}}(\pi)}_{\textbf{Concentration}}$$

Concentration... (Remember) Some challenges:

- Even for basic importance sampling $(W_1 R_1 + \cdots + W_n R_n)/n$ it's non-trivial: unbiased, but W_i are **unbounded**
  - Excludes Hoeffding/Bernstein/McDiarmid
  - We can "truncate", e.g. $W_i^\lambda = \pi(A_i|X_i)/(\pi_b(A_i|X_i) + \lambda)$ for some h.p. $\lambda > 0$.
  - Ugly! Needs tuning, doesn't always work...
- Variance is important: need bounds with empirical variance.
- Sometimes, estimator is not a sum of indep. elements (self-normalization).

Concentration of $\hat{v}^{\mathbf{wis}}$

Goal: lower bound on $\mathbb{E}\left[\hat{v}^{\mathrm{WIS}}(\pi) \mid X_1^n\right] - \hat{v}^{\mathrm{WIS}}(\pi)$.

Theorem Assume elements of $S = (X_1, X_2, \ldots, X_n)$ are independent, and let

$$\Delta = f(S) - \mathbb{E}[f(S)] \;, \quad V = \sum_{k=1}^{n} \mathbb{E}\left[(f(S) - f(S^{(k)}))^2 \,\middle|\, X_1, \ldots, X_k \right] \;.$$

Then, for any $x \geq 2, \; y > 0$,

$$\mathbb{P}\left(|\Delta| \geq \sqrt{(V + y)\,(2 + \ln(1 + V/y))\, x} \right) \geq e^{-x} \;.$$

Take $f = \hat{v}^{\mathrm{WIS}}$, condition on X_1^n, and choose $y = 1/n$. Algebra gives that V obeys

$$V \leq \sum_{k=1}^{n} \mathbb{E}\left[\left(\frac{W_k}{Z} + \frac{W_k'}{Z^{(k)}} \right)^2 \,\middle|\, W_1^k, X_1^n \right] \;.$$

Canonical Pairs – [dlPLS08]

We call (A, B) a canonical pair if $B \geq 0$ and

$$\sup_{\lambda \in \mathbb{R}} \mathbb{E}\left[\exp\left(\lambda A - \frac{\lambda^2}{2} B^2\right)\right] \leq 1 \ .$$

Theorem 12.4 of [dlPLS08]

Theorem

Let (A, B) be a canonical pair. Then, for any $t > 0$,

$$\mathbb{P}\left(\frac{|A|}{\sqrt{B^2 + (\mathbb{E}[B])^2}} \geq t\right) \leq \sqrt{2}e^{-\frac{t^2}{4}} \ .$$

In addition, for all $t \geq \sqrt{2}$ and $y > 0$,

$$\mathbb{P}\left(\frac{|A|}{(B^2 + y)\left(1 + \frac{1}{2}\ln\left(1 + \frac{B^2}{y}\right)\right)} \geq t\right) \leq e^{-\frac{t^2}{2}} \ .$$

Recall

$$\Delta = f(S) - \mathbb{E}[f(S)] \ , \quad V = \sum_{k=1}^{n} \mathbb{E}\left[(f(S) - f(S^{(k)}))^2 \,\middle|\, X_1, \ldots, X_k \right] \ .$$

Lemma

$(\Delta, \sqrt{V})$ *is a canonical pair.*

Proof.

Let $\mathbb{E}_k[\cdot]$ stand for $\mathbb{E}[\cdot \mid X_1, \ldots, X_k]$. The Doob martingale decomposition of $f(S) - \mathbb{E}[f(S)]$ gives

$$f(S) - \mathbb{E}[f(S)] = \sum_{k=1}^{n} D_k \ ,$$

where $D_k = \mathbb{E}_k[f(S)] - \mathbb{E}_{k-1}[f(S)] = \mathbb{E}_k[f(S) - f(S^{(k)})]$ and the last equality follows from the elementary identity $\mathbb{E}_{k-1}[f(S)] = \mathbb{E}_k[f(S^{(k)})]$. □

Conclusions

- Confident off-policy estimation
- Self-normalized importance weighting estimator
- Harris-inequality + Efron-Stein: Value lower bound
- Appears to be tighter than alternatives
- Where is the limit? Bootstrapping? Honest coverage?

[dlPLS08] V. H. de la Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.