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

Ilja Kuzborskij

DeepMind
Simons Institute, November 30, 2020

Joint work with

- Claire Vernade
- András György
- Csaba Szepesvári

Contents

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 context $X \in \mathcal{X}$ 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 context $X \in \mathcal{X}$ and action $A \in [K]$
- π_b — behaviour policy, e.g. $\pi_b(\cdot|X)$

Contextual off-policy evaluation problem

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

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 context $X \in \mathcal{X}$ and action $A \in [K]$
- π_b — behaviour policy, e.g. $\pi_b(\cdot|X)$

Contextual off-policy evaluation problem

- An agent observes $S = ((X_1, A_1, R_1), \ldots, (X_n, A_n, R_n))$
  $A_i \overset{\text{ind.}}{\sim} \pi_b(\cdot|X_i)$, $X_i \overset{\text{ind.}}{\sim} P_X$, $R_i \overset{\text{ind.}}{\sim} P_{R|X,A}$
- 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 $\quad r(x,a) = \int u \, \mathrm{d}P_{R|X,A}(u|x,a).$

Value estimation through Importance Weighting

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 weighting* estimator

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

Indeed,

$$\mathbb{E}[\hat{v}^{\mathrm{IW}}(\pi)] = v(\pi)$$

Value estimation through Importance Weighting

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 weighting* estimator

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

Indeed,

$$\mathbb{E}[\hat{v}^{\mathrm{IW}}(\pi)] = v(\pi)$$

High variance!

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

Value estimation through Doubly-robust Estimator

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 fitted on a held-out dataset).

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

Value estimation through Self-normalized Estimator

Something simpler, a *self-normalized importance weighting*:

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

- *Biased* (asymptotically unbiased (IID))
- In practice, low variance (self-normalization)
  - Some intuition: $\mathrm{Var}(\hat{v}^{\mathrm{SN}}(\pi)) \leq \mathbb{E}\left[\sum_k \frac{W_k^2}{\left(\sum_i W_i\right)^2}\right]$
  - $\frac{W_k}{\sum_i W_i} \sim \frac{1}{n^{\alpha}}$ for $\alpha \in [0, 1]$
  - (depending on "niceness" of the weight distribution)

What about $v(\pi)$?

Estimator alone is not enough. We want confidence intervals.

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

How to do that? General decomposition:

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

- *Concentration of texts*: standard concentration (X_1^n are IID)
- *Bias*: sometimes estimator is unbiased, we'll skip this for now..
- *Concentration of estimator* ...

Contents

What is concentration?

- We have $S = (Z_1, Z_2, \ldots, Z_n) \sim \mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$ *independent* r.v. taking values in some $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$.
- We have $f : \mathcal{Z} \to \mathbb{R}$ – a fixed measurable function.

How large are typical deviations $\Delta = f(S) - \mathbb{E}[f(S)]$?

We care about bounds on the *tail probability*

$$\mathbb{P}(\Delta < t) \quad \text{and} \quad \mathbb{P}(\Delta > t) \qquad t > 0\ .$$

What is concentration?

Some classical examples:

$$\mathbb{P}\left(|\Delta| \leq \mathbb{E}[|\Delta|]/x\right) \geq 1 - x, \quad x \in (0,1)\ . \qquad \text{(Markov)}$$

Typically we are after bounds which decay “quickly” in $x > 0$.

What is concentration?

Some classical examples:

$$\mathbb{P}\left(|\Delta| \le \mathbb{E}[|\Delta|]/x\right) \ge 1 - x, \quad x \in (0,1)\ . \qquad \text{(Markov)}$$

Typically we are after bounds which decay "quickly" in $x > 0$.

Assume that indep. $Z_1, \ldots, Z_n \in [0,1]$ and $f(S) = \frac{1}{n}\sum_k Z_k$,

$$\mathbb{P}\left(\Delta \le \sqrt{\frac{x}{2n}}\right) \ge 1 - e^{-x} \qquad \text{(Hoeffding)}$$

$$\mathbb{P}\left(\Delta \le \frac{1}{n}\left(\sqrt{2x\sum_k \mathbb{E}[Z_k^2]} + \frac{2}{3}x\right)\right) \ge 1 - e^{-x}$$

(Bernstein)

What about $v(\pi)$ again?

Estimator alone is not enough. We want confidence intervals.

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

How to do that? General structure:

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

Some challenges for concentration of $\hat{v}$:

- Even for basic importance weighting $\hat{v}^{\mathrm{IW}}$ it's non-trivial: $W_i = \frac{\pi(A_i \mid X_i)}{\pi_b(A_i \mid X_i)}$ are **unbounded**
  - Excludes standard concentration inequalities (moments of $\hat{v}(\pi)$ can't be easily controlled)

What about $v(\pi)$ again?

Estimator alone is not enough. We want confidence intervals.

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

How to do that? General structure:

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

Some challenges for concentration of $\hat{v}$:

- Even for basic importance weighting $\hat{v}^{\mathrm{IW}}$ it's non-trivial: $W_i = \frac{\pi(A_i \,|\, X_i)}{\pi_b(A_i \,|\, X_i)}$ are **unbounded**
  - Excludes standard concentration inequalities (moments of $\hat{v}(\pi)$ can't be easily controlled)
  - We can "truncate", e.g. $W_i^\lambda = \frac{\pi(A_i|X_i)}{\pi_b(A_i|X_i)+\lambda}$ for some $\lambda > 0$.
  - Ugly! In practice needs tuning of λ, doesn't always work...

What about $v(\pi)$ again?

Estimator alone is not enough. We want confidence intervals.

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

How to do that? General structure:

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

Some challenges for concentration of $\hat{v}$:

- Even for basic importance weighting $\hat{v}^{\mathrm{IW}}$ it's non-trivial: $W_i = \frac{\pi(A_i \,|\, X_i)}{\pi_b(A_i \,|\, X_i)}$ are **unbounded**
  - Excludes standard concentration inequalities (moments of $\hat{v}(\pi)$ can't be easily controlled)
  - We can "truncate", e.g. $W_i^\lambda = \frac{\pi(A_i|X_i)}{\pi_b(A_i|X_i)+\lambda}$ for some $\lambda > 0$.
  - Ugly! In practice needs tuning of λ, doesn't always work...
- **Variance is important**: need bounds with empirical variance.
- Sometimes estimator is not a sum IID elements.

What about $v(\pi)$ again?

Estimator alone is not enough. We want confidence intervals.

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

Let's go back and pick Self-normalized Estimator (SN):

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

- $(W_i)_i$ are unbounded
- $\hat{v}^{\mathrm{SN}}$ is not a sum of IID elements (self-normalization)
- We really want CI to be controlled by the variance of $\hat{v}^{\mathrm{SN}}$.

Contents

Concentration of general functions

- Going beyond "simple" functions: f is not necessarily a sum, possibly non-linear.
- One possible way: sensitivity of f to *"small perturbations"* controls concentration.

Concentration of general functions

- Going beyond "simple" functions: f is not necessarily a sum, possibly non-linear.
- One possible way: sensitivity of f to *"small perturbations"* controls concentration.

Let:

- $S' = (Z'_1, Z'_2, \ldots, Z'_n)$ be an independent copy of S
- $S^{(k)} = (Z_1, \ldots, Z_{k-1}, Z'_k, Z_{k+1}, \ldots, Z_n)$

Concentration of general functions

- Going beyond "simple" functions: f is not necessarily a sum, possibly non-linear.
- One possible way: sensitivity of f to *"small perturbations"* controls concentration.

Let:

- $S' = (Z'_1, Z'_2, \ldots, Z'_n)$ be an independent copy of S
- $S^{(k)} = (Z_1, \ldots, Z_{k-1}, Z'_k, Z_{k+1}, \ldots, Z_n)$

Classical Efron-Stein (ES) inequality:

$$\mathrm{Var}(f) \leq \frac{1}{2} \sum_{k=1}^{n} \mathbb{E}\left[\left(f(S) - f(S^{(k)}) \right)^2 \right] .$$

Tail Bounds through Bounded differences

The same idea extended to tail bounds.

Introduce Efron-Stein *variance proxy*

$$V^{\mathrm{ES}}(S, S') = \sum_{k=1}^{n} (f(S) - f(S^{(k)}))_+^2 \,,$$

Bounded Differences

Assume: $\sup_{s,s' \in \mathcal{Z}} V^{\mathrm{ES}}(s, s') \le c$ a.s. for some $c > 0$. Then:

$$\mathbb{P}\left(|\Delta| \le \sqrt{2cx}\right) \ge 1 - e^{-x} \,, \qquad x \ge 0 \,.$$

For averages, $V^{\mathrm{ES}}(S, S') \lesssim 1/n$ recovers Hoeffding's inequality.

Beyond Bounded Differences

Bounded Differences

Assume: $\sup_{s,s' \in \mathcal{Z}} V^{\mathrm{ES}}(s,s') \leq c$ a.s. for some $c > 0$. Then:

$$\mathbb{P}\left(|\Delta| \leq \sqrt{2cx}\right) \geq 1 - e^{-x}\,, \qquad x \geq 0\;.$$

- Powerful, but pessimistic...
- ...neglects information about moments of Δ.

Beyond Bounded Differences

Bounded Differences

Assume: $\sup_{s,s' \in \mathcal{Z}} V^{\mathrm{ES}}(s, s') \le c$ a.s. for some $c > 0$. Then:

$$\mathbb{P}\left(|\Delta| \le \sqrt{2cx}\right) \ge 1 - e^{-x}, \qquad x \ge 0 .$$

- Powerful, but pessimistic...
- ...neglects information about moments of Δ.

Exponential Efron-Stein Inequality [BLM03]

Let $\lambda \in (0, 1)$. Then:

$$\ln \mathbb{E}[e^{\lambda\Delta}] \le \frac{\lambda}{1-\lambda} \ln \mathbb{E}\left[e^{\mathbb{E}[V^{\mathrm{ES}}(S,S') \mid S]}\right]$$

Beyond Bounded Differences

Bounded Differences

Assume: $\sup_{s,s' \in \mathcal{Z}} V^{\mathrm{ES}}(s,s') \leq c$ a.s. for some $c > 0$. Then:

$$\mathbb{P}\left(|\Delta| \leq \sqrt{2cx}\right) \geq 1 - e^{-x}\,, \qquad x \geq 0\,.$$

- Powerful, but pessimistic...
- ...neglects information about moments of Δ.

Exponential Efron-Stein Inequality [BLM03]

Let $\lambda \in (0,1)$. Then:

$$\ln \mathbb{E}[e^{\lambda \Delta}] \leq \frac{\lambda}{1-\lambda} \ln \mathbb{E}\left[e^{\mathbb{E}[V^{\mathrm{ES}}(S,S') \mid S]}\right]$$

Chernoff bound gives us a tail bound:

$\mathbb{P}(\Delta \geq x) \leq \inf_{\lambda \in (0,1)} \mathbb{E}[\exp(\lambda\Delta - \lambda x)]$

Beyond Bounded Differences

Exponential ES

$$\ln \mathbb{E}[e^{\lambda\Delta}] \leq \frac{\lambda}{1-\lambda} \ln \mathbb{E}\left[e^{\lambda\,\mathbb{E}[V^{\mathrm{ES}}(S,S')\,|\,S]}\right] \qquad \lambda \in (0,1)$$

- Control of exponential moment of $V^{\mathrm{ES}} \Rightarrow$ concentration of Δ.
- Can we get something more user-friendly?

Beyond Bounded Differences

Exponential ES

$$\ln \mathbb{E}[e^{\lambda\Delta}] \leq \frac{\lambda}{1-\lambda} \ln \mathbb{E}\left[e^{\lambda\, \mathbb{E}[V^{\mathrm{ES}}(S,S') \mid S]}\right] \qquad \lambda \in (0,1)$$

- Control of exponential moment of $V^{\mathrm{ES}} \Rightarrow$ concentration of Δ.
- Can we get something more user-friendly?

Assume that f satisfies **second-order bounded differences** [Mau19, MP18]: for any $\mathcal{D}$, some $a, b > 0$,

$$\sup_{s,s' \in \mathcal{Z}} \sum_{k,j : k \neq j} \left((f(s) - f(s^{(k)})) - (f(s^{(j)}) - f(s^{(k,j)})) \right)^2 \leq a^2/2 \ ,$$

$$\max_{k \in [n]} f(S) - \mathbb{E}[f(S) \mid Z_1, \dots Z_{k-1}, Z_k, \dots Z_n] \leq b \ .$$

Then, for any $x \geq 0$,

$$\mathbb{P}\left(\Delta \leq \sqrt{2\mathbb{E}[V^{\mathrm{ES}}(S,S')]\, x} + (a + 2/3b)x\right) \geq 1 - e^{-x} \ .$$

Limitations

- All of these inequalities implicitly control moments of $V^{\mathrm{ES}}(S, S')$
- Constants a, b are data-independent
- ... typically we need boundedness of f or its domain to easily get a finite pair a, b.

Semi-Empirical Inequalities

Introduce *Semi-Empirical ES* variance proxy

$$V = \sum_{k=1}^{n} \mathbb{E}\left[(f(S) - f(S^{(k)}))^2 \,\middle|\, Z_1, \dots, Z_k \right] .$$

Semi-Empirical Inequalities

Introduce *Semi-Empirical ES* variance proxy

$$V = \sum_{k=1}^{n} \mathbb{E}\left[(f(S) - f(S^{(k)}))^2 \,\middle|\, Z_1, \dots, Z_k \right] .$$

Semi-empirical Efron-Stein (ES) [KS19]

For any $x \geq 2,\ y > 0$,

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

Semi-Empirical Inequalities

Introduce *Semi-Empirical ES* variance proxy

$$V = \sum_{k=1}^{n} \mathbb{E}\left[(f(S) - f(S^{(k)}))^2 \,\middle|\, Z_1, \dots, Z_k \right] .$$

Semi-empirical Efron-Stein (ES) [KS19]

For any $x \geq 2,\ y > 0$,

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

- Does not require boundedness of RVs, nor of co-domain of f.
- Essentially depends on V and a free parameter $y > 0$ (selected by the user). E.g. $y = 1/n^2$ w.h.p. gives

$$|\Delta| \lessapprox \sqrt{V} + \frac{1}{n} . \qquad \text{(Bernstein-type behavior)}$$

Contents

What about $v(\pi)$ again?

Estimator alone is not enough. We want confidence intervals.

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

Let's go back and pick SN:

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

- $(W_i)_i$ are unbounded
- $\hat{v}^{\mathrm{SN}}$ is not a sum of IID elements (self-normalization)
- We really want CI to be controlled by the variance of $\hat{v}^{\mathrm{SN}}$.

Semi-empirical Efron-Stein Bound for SN

Theorem. [KVGS20] W.h.p.,

$$v(\pi) \geq \boxed{B} \cdot \left(\hat{v}^{\mathrm{SN}}(\pi) - \sqrt{c \cdot \left(\boxed{V^{\mathrm{SN}}} + \frac{1}{n} \right)} \right) - \frac{c'}{\sqrt{n}} \, ,$$

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

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

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: $\mathrm{Var}(\hat{v}^{\mathrm{SN}}(\pi)) \leq \mathbb{E}\left[\sum_k \left(\frac{W_k^2}{Z} \right)^2 \right] \approx \boxed{V^{\mathrm{SN}}}$

Bias $\boxed{B}$ is multiplicative, ≈ 1 for "easy" distributions of W_i

Semi-empirical Efron-Stein Bound for SN

Theorem. [KVGS20] W.h.p.,

$$v(\pi) \geq \boxed{B} \cdot \left(\hat{v}^{\mathrm{SN}}(\pi) - \sqrt{c \cdot \left(\boxed{V^{\mathrm{SN}}} + \frac{1}{n} \right)} \right) - \frac{c'}{\sqrt{n}} ,$$

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

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

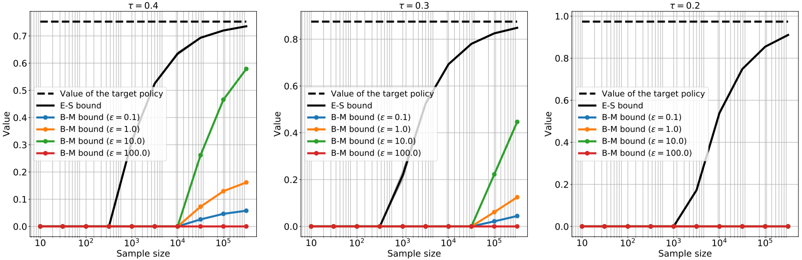
where $Z^{(k)} = Z + (W_k' - W_k)$, and W_k' indep. dist. as W_k.

- No truncation! No hyperparameters.
- Contexts are fixed.
- Needs knowledge of π_b — only partly empirical:
  V^{SN} and B can be computed exactly. Cost: n^K :-(
  Can approximate using Monte-Carlo simulation! :-)

Is it any good? Synthetic Experiments

- Fix $K > 0$, $\tau > 0$
- $\pi_b(a) \propto e^{\frac{1}{\tau}\mathbb{I}\{a=1\}}$
- $\pi(a) \propto e^{\frac{1}{\tau}\mathbb{I}\{a=2\}}$
- $R_i = \mathbb{I}\{A_i = k\}$, $A_i \sim \pi_b(\cdot)$
- As $\tau \to 0$, π_b and π become increasingly misaligned

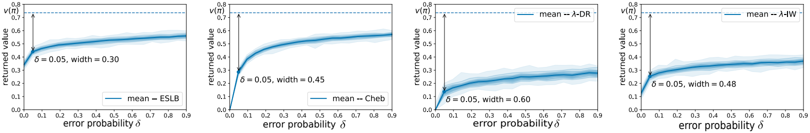
Numerical tightness in sample size



E-S — Our bound

B-M — Empirical Bernstein's bound with ε-truncated weights

Numerical tightness in error probability



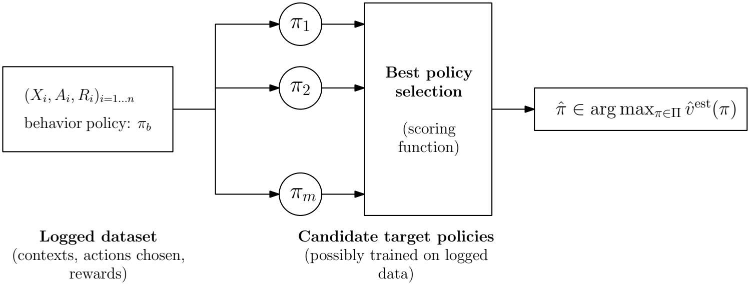
Similar setup as before, sample size $= 10^4$, left to right:

- E-S — our bound.
- Chebyshev's ineq.-based CI for SN.
- Empirical Bernstein's ineq.-based CI for Doubly-robust Estimator (DR) with $W_i^\lambda = \frac{\pi(A_i|X_i)}{\pi_b(A_i|X_i)+\lambda}$ for some $\lambda = 1/\sqrt{n}$.
- Empirical Bernstein's ineq.-based CI for Importance Weighting (IW) with W_i^λ.

Is it any good? Nonsynthetic Experiments

The Best Policy Selection problem

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



Nonsynthetic Experiments – Setup

Target policies are $\left\{\pi^{\mathrm{ideal}}, \pi^{\hat{\Theta}_{\mathsf{IW}}}, \pi^{\hat{\Theta}_{\mathsf{SN}}}\right\}$ where

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

with two choices of parameters given by the optimization problems:

$$\hat{\Theta}_{\mathsf{IW}} \in \operatorname*{arg\,max}_{\Theta \in \mathbb{R}^{d \times K}} \hat{v}^{\mathrm{IW}}(\pi^{\Theta}) \ , \qquad \hat{\Theta}_{\mathsf{SN}} \in \operatorname*{arg\,max}_{\Theta \in \mathbb{R}^{d \times K}} \hat{v}^{\mathrm{SN}}(\pi^{\Theta}) \ .$$

- 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	Yeast	PageBlok	OptDigits	SatImage	isolet	PenDigits	Letter	kropt
Size	1484	5473	5620	6435	7797	10992	20000	28056
Efron-Stein LB	**0.90 $\pm$ 0.27**	**0.90 $\pm$ 0.27**	**0.90 $\pm$ 0.27**	**0.91 $\pm$ 0.26**	**0.90 $\pm$ 0.27**	**0.91 $\pm$ 0.27**	**0.91 $\pm$ 0.27**	**0.91 $\pm$ 0.27**
Trunc-IW + Bern.	**0.91 $\pm$ 0.26**	**0.91 $\pm$ 0.27**	0.74 $\pm$ 0.40	0.79 $\pm$ 0.33	0.74 $\pm$ 0.40	0.81 $\pm$ 0.34	**0.90 $\pm$ 0.27**	**0.90 $\pm$ 0.27**
Trunc-DR + Bern.	$-\infty$	**0.91 $\pm$ 0.27**	0.77 $\pm$ 0.37	$-\infty$	0.74 $\pm$ 0.40	**0.91 $\pm$ 0.26**	**0.91 $\pm$ 0.27**	**0.91 $\pm$ 0.27**
SN + Cheb.	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	**0.90 $\pm$ 0.27**	$-\infty$
DR	0.52 $\pm$ 0.31	0.77 $\pm$ 0.35	0.51 $\pm$ 0.33	0.75 $\pm$ 0.35	0.21 $\pm$ 0.29	0.79 $\pm$ 0.31	0.77 $\pm$ 0.28	**0.91 $\pm$ 0.27**

Contents

Proof sketch

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

1. Concentration of contexts – Hoeffding since X_1^n are IID.
   $\mathbb{E}\left[v(\pi) \,|\, X_1^n\right] = \frac{1}{n}\sum_i \mathbb{E}\left[W_i R_i \,|\, X_i\right]$.
2. Bias – IW is unbiased: "split" SN into IW and denominator.

Proof sketch

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

1. Concentration of contexts – Hoeffding since X_1^n are IID. $\mathbb{E}\left[v(\pi) \,|\, X_1^n\right] = \frac{1}{n}\sum_i \mathbb{E}\left[W_i R_i \,|\, X_i\right]$.
2. Bias – IW is unbiased: "split" SN into IW 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 $(Z_1, \ldots, Z_n)$ independent from each other, we have

   $$\mathbb{E}[f(Z_1, \ldots, Z_n) g(Z_1, \ldots, Z_n)] \le \mathbb{E}[f(Z_1, \ldots, Z_n)]\, \mathbb{E}[g(Z_1, \ldots, Z_n)] \;.$$

   This gives us:

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

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

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

Semi-empirical Efron-Stein (ES) [KS19]

For any $x \geq 2,\ y > 0$,

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

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

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

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

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 $\mathbb{E}_{k-1}[f(S)] = \mathbb{E}_k[f(S^{(k)})]$. □

Take-home message

- Tighter off-policy evaluation bounds for contextual bandits
- Tighter CIs for Self-normalized Estimator
- New high-probability user-friendly variance-dependent concentration inequalities for general functions

Some limitations / future challenges:

- Requires knowledge of π_b
- Requires π_b to be static, observations are IID
  — in many practical cases this is not a problem!
- Policy optimization (learning)
  - Extension of the about to the PAC-Bayes setting [KS19]

https://arxiv.org/abs/2006.10460

https://arxiv.org/abs/1909.01931

[BLM03] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

[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.

[KS19] I. Kuzborskij and Cs. Szepesvári. Efron-Stein PAC-Bayesian Inequalities. arXiv:1909.01931, 2019.

[KVGS20] I. Kuzborskij, C. Vernade, A. György, and Cs. Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. arXiv:2006.10460, 2020.

[Mau19] A. Maurer. A bernstein-type inequality for functions of bounded interaction. *Bernoulli*, 25(2):1451–1471, 2019.

[MP18] A. Maurer and M. Pontil. Empirical bounds for functions with weak interactions. In *Conference on Computational Learning Theory (COLT)*, pages 987–1010, 2018.