Logarithmic Smoothing for Pessimistic Off-Policy Evaluation, Selection & Learning

NeurIPS ’24, Spotlight

Otmane Sakhi[1], Imad Aouali[1,2], Pierre Alquier[3], Nicolas Chopin[2]

[1]Criteo AI Lab, [2]CREST-ENSAE, [3]ESSEC

Off-Policy Contextual Bandits

Off-Policy Contextual Bandits

- **Off-Policy (Offline) Contextual Bandit.** A framework that optimizes decision-making by leveraging logged interactions.

Contexts $x \in \mathcal{X}$	Actions $a \in \mathcal{A}$	Logging policy π_0
User features.	Products.	Current RecSys

- **Interactions.** For any $i \in [n]$
  - Observe context $x_i \sim \nu$, where $x_i \in \mathcal{X}$
  - Take action $a_i \sim \pi_0(\cdot \mid x_i)$, where $a_i \in \mathcal{A}$
  - Suffers a cost $c_i \sim p(\cdot \mid x_i, a_i)$. ($c_i \in [-1, 0]$, negative reward)
- **Log** $\mathcal{D}_n = \{x_i, a_i, c_i\}_{i \in [n]}$ and use it to improve the system.

Performance Metric. For $\pi \in \Pi$, the risk is defined as:

$$R(\pi) = \mathbb{E}_{x \sim \nu, a \sim \pi(\cdot|x)} \left[c(x, a) \right] ,$$

where $c(x, a) = \mathbb{E}_{c \sim p(\cdot|x,a)}[c]$ is the expected cost of x and a.

Tasks. Given logged data $\mathcal{D}_n = \{x_i, a_i, c_i\}_{i \in [n]}$ by π_0:

- **Evaluation (OPE).** For a new π, estimate $R(\pi) \approx \hat{R}_n(\pi)$.
- **Selection (OPS).** Given $\{\pi_1, \cdots, \pi_m\}$, select $\arg\min_{i \in [m]} R(\pi_i)$.
- **Learning (OPL).** Find $\pi_* = \arg\min_{\pi \in \Pi} R(\pi)$.

Off-Policy Contextual Bandits

Pessimism is optimal for OPE, OPS & OPL. [1, 2, 3]

- **OPE.** [4, 5] study concentration properties (beyond *MSE*).
- **OPS.** [2, 5] use risk upper bounds (pessimism).
- **OPL.** [1, 3, 6, 7] use risk generalization bounds (pessimism).

Instead of $\hat{R}_n(\pi)$, they use a high-probability bound $\hat{U}_n(\pi)$:

$$R(\pi) \leq \hat{U}_n(\pi) = \hat{R}_n(\pi) + \hat{C}(\pi)\,.$$

Pessimism is optimal for OPE, OPS & OPL. [1, 2, 3]

- **OPE.** [4, 5] study concentration properties (beyond *MSE*).
- **OPS.** [2, 5] use risk upper bounds (pessimism).
- **OPL.** [1, 3, 6, 7] use risk generalization bounds (pessimism).

Instead of $\hat{R}_n(\pi)$, they use a high-probability bound $\hat{U}_n(\pi)$:

$$R(\pi) \leq \hat{U}_n(\pi) = \hat{R}_n(\pi) + \hat{C}(\pi)\,.$$

What we do.

- Derive tight upper bounds for a broad family of estimators.
- Find the estimator (within that family) with the tightest bound.

Novel Concentration Bounds

We focus on the family of regularized IPS estimators:

$$\hat{R}_n^h(\pi) = \frac{1}{n}\sum_{i=1}^{n} h\left(\pi(a_i|x_i), \pi_0(a_i|x_i), c_i\right) = \frac{1}{n}\sum_{i=1}^{n} h_i\,, \tag{1}$$

with h is a transform satisfying **(C1)**: $\frac{p}{q}c \leq h\left(p, q, c\right) \leq 0$.

$$\begin{aligned} h(p,q,c) &= \frac{p}{q}c\,, \implies \text{IPS [8]}\,, \\ h(p,q,c) &= \min\left(\frac{p}{q}, M\right)c\,, M \in \mathbb{R}^+ \implies \text{Clipping [9]}\,, \\ h(p,q,c) &= \left(\frac{p}{q}\right)^{\alpha} c\,, \alpha \in [0,1] \implies \text{Exponential Smoothing [6]}\,, \\ h(p,q,c) &= \frac{p}{q+\gamma}c\,, \gamma \geq 0 \implies \text{Implicit Exploration [5]}\ldots \end{aligned} \tag{2}$$

Novel Concentration Bounds

Let $\pi \in \Pi$, define the empirical ℓ-th moment of $\hat{R}_n^h(\pi)$ as

$$\hat{\mathcal{M}}_n^{h,\ell}(\pi) = \frac{1}{n} \sum_{i=1}^{n} h_i^{\ell} \,. \tag{3}$$

For $\lambda > 0$, we define the function ψ_λ as $\psi_\lambda(x) = \left(1 - \exp(-\lambda x)\right)/\lambda$.

Let $\pi \in \Pi$, $L \geq 1$, h satisfying **(C1)**, $\delta \in (0, 1]$ and $\lambda > 0$. Then it holds with probability at least $1 - \delta$ that

$$R(\pi) \leq \psi_\lambda \left(\hat{R}_n^h(\pi) + \sum_{\ell=2}^{2L} \frac{\lambda^{\ell-1}}{\ell} \hat{\mathcal{M}}_n^{h,\ell}(\pi) + \frac{\ln(1/\delta)}{\lambda n} \right), \tag{4}$$

- L controls the empirical moments, $L \nearrow$ tightens the bound.
- Holds for all h, find the h that minimizes the bound!

Setting $L \to \infty$ and minimizing it w.r.t. h yields a bound:

$$R(\pi) \leq \psi_\lambda \left(\hat{R}_n^\lambda(\pi) + \frac{\ln(1/\delta)}{\lambda n} \right). \tag{5}$$

for a novel estimator, that we call Logarithmic Smoothing (LS):

$$\hat{R}_n^\lambda(\pi) = -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda w_\pi(x_i, a_i) c_i\right), \tag{6}$$

with $w_\pi(x, a) = \pi(a|x)/\pi_0(a|x)$.

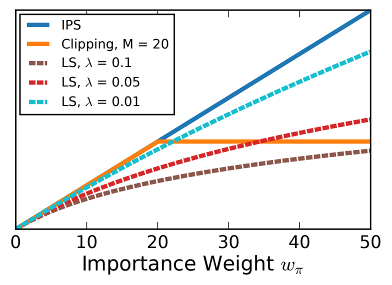
(5) is **provably tighter** than:

- Our bound with $L = 1$.
- cIPS (empirical Bernstein).
- IX bound [5].

Logarithmic Smoothing

$$\forall \lambda \geq 0, \quad \hat{R}_n^{\lambda}(\pi) = -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda w_\pi(x_i, a_i) c_i\right) . \qquad (7)$$

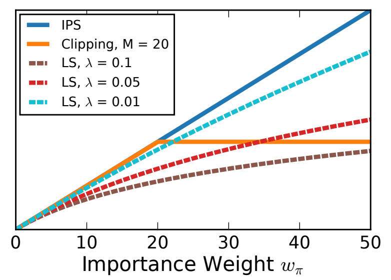
- $\lambda \to 0$ recovers IPS.
- Smoothly corrects the IWs.
- Good bias-variance tradeoff.
- **Unbounded**, with **finite variance!**
- **Sub-Gaussian** concentration:



Logarithmic Smoothing

$$\forall \lambda \geq 0, \quad \hat{R}_n^{\lambda}(\pi) = -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda w_\pi(x_i, a_i) c_i\right) . \qquad (7)$$

- $\lambda \to 0$ recovers IPS.
- Smoothly corrects the IWs.
- Good bias-variance tradeoff.
- **Unbounded**, with **finite variance!**
- **Sub-Gaussian** concentration:



For $\lambda^* = \mathcal{O}(1/\sqrt{n})$, we have with probability at least $1 - \delta$:

$$|R(\pi) - \hat{R}_n^{\lambda^*}(\pi)| \leq \sqrt{2\sigma^2 \ln(2/\delta)}, \quad \text{where } \sigma^2 = 2\mathbb{E}\left[w_\pi(x, a)^2 c^2\right] / n .$$

Logarithmic Smoothing - OPS/OPL

With $\lambda = \mathcal{O}(1/\sqrt{n})$, and by minimizing $\hat{R}_n^\lambda(\pi)$, we reach π_* in:

- $\mathcal{O}\left(\sqrt{\mathbb{E}\left[\left(\frac{\pi_*(a|x)}{\pi_0(a|x)}c\right)^2\right]/n}\right)$ for OPS.
- $\mathcal{O}\left(\sqrt{\left(\mathbb{E}\left[\frac{\pi_*(a|x)c^2}{\pi_0(a|x)^2}\right] + KL(Q^*||P)\right)/n}\right)$ in PAC-Bayes OPL.

→ We identify the best policy with enough n.
→ Faster identification when π_0 is close to π_*.
→ Simple, no additional terms (e.g., Emp. variance in SVP [1])
→ Provably efficient for OPS and OPL.

Experiments

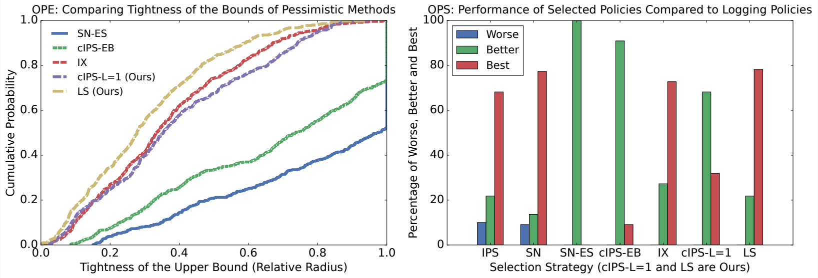


Figure 1: Results for OPE and OPS experiments.

	cIPS	cvcIPS	ES	IX	LS-LIN (Ours)
$rI(U(\hat{\pi}_L))$	14.48%	21.28%	7.78%	<u>24.74%</u>	**26.31%**
$rI(R(\hat{\pi}_L))$	28.13%	<u>33.64%</u>	29.44%	**36.70%**	**36.76%**

Table 1: OPL Improvement of Guaranteed risk *U* and *R* of the bounds.

Conclusion

- Work of theoretical nature with practical implications.
- Principled approach led us to the design of a new estimator.
- A lot more insight can be found in the paper.

- Work of theoretical nature with practical implications.
- Principled approach led us to the design of a new estimator.
- A lot more insight can be found in the paper.
- … Or let’s discuss the work at NeuRIPS, or even by e-mail!

References

[1] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

[2] Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pages 640–648. PMLR, 2021.

[3] Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. PAC-Bayesian Offline Contextual Bandits with Guarantees. In *International Conference on Machine Learning*, pages 29777–29799. PMLR, 2023.

[4] Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in Neural Information Processing Systems*, 34:8119–8132, 2021.

[5] Germano Gabbianelli, Gergely Neu, and Matteo Papini. Importance-weighted offline learning done right. In Claire Vernade and Daniel Hsu, editors, *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 614–634. PMLR, 25–28 Feb 2024.

[6] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential Smoothing for Off-Policy Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 984–1017. PMLR, 2023.

[7] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Unified pac-bayesian study of pessimism for offline policy learning with regularized importance sampling. *UAI 2024*, 2024.

[8] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[9] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.