Exponential Smoothing for Off-Policy Learning

Imad Aouali[1,2], Victor-Emmanuel Brunel[2], David Rohde[1], Anna Korba[2]

[1]Criteo AI Lab
[2]CREST, ENSAE, Institut Polytechnique de Paris

International Conference on Machine Learning 2023

Off-Policy Contextual Bandits

Interactions: For any $i \in [n]$

- Observe context $x_i \sim \nu$
- Take action $a_i \sim \pi_0(\cdot \mid x_i)$, π_0 is the **logging policy**
- Receive cost $c_i \sim p(\cdot \mid x_i, a_i)$.

Off-Policy Contextual Bandits

Interactions: For any $i \in [n]$

- Observe context $x_i \sim \nu$
- Take action $a_i \sim \pi_0(\cdot \mid x_i)$, π_0 is the **logging policy**
- Receive cost $c_i \sim p(\cdot \mid x_i, a_i)$.

Performance metric: The risk of a policy π is defined as

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

where $x \in \mathcal{X}$ is a context, $a \in \mathcal{A}$ is an action, and
$c(x, a) = -r(x, a)$ is the expected cost (negative reward) of (x, a).

Off-Policy Contextual Bandits

Tasks: Given $\mathcal{D}_n = (x_i, a_i, c_i)_{i \in [n]}$, where (x_i, a_i, c_i) are i.i.d.

- **Off-Policy Evaluation (OPE):** Build an estimator of $R(\pi)$

$$\hat{R}_n(\pi) = f(\pi, \mathcal{D}_n) \approx R(\pi)\,.$$

- **Off-Policy Learning (OPL):** Find $\hat{\pi}_n$, $R(\hat{\pi}_n) \approx \min_{\pi \in \Pi} R(\pi)$

$$\hat{\pi}_n = \arg\min_{\pi} \hat{R}_n(\pi) + \mathrm{pen}(\pi) \approx \pi_*\,,$$

where $\pi_* = \arg\min_\pi R(\pi)$.

Exponential Smoothing for OPE

Inverse Propensity Scoring (IPS) estimates the risk $R(\pi)$ such as

$$\hat{R}_n^{\mathrm{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} w_\pi(a_i|x_i) c_i \,,$$

where $w_\pi(a|x) = \frac{\pi(a|x)}{\pi_0(a|x)}$ are the **importance weights**.

Exponential Smoothing for OPE

Inverse Propensity Scoring (IPS) estimates the risk $R(\pi)$ such as

$$\hat{R}_n^{\mathrm{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} w_\pi(a_i|x_i)c_i \,,$$

where $w_\pi(a|x) = \frac{\pi(a|x)}{\pi_0(a|x)}$ are the **importance weights**.

Problem: Large variance when π is different from π_0.

Exponential Smoothing for OPE

Inverse Propensity Scoring (IPS) estimates the risk $R(\pi)$ such as

$$\hat{R}_n^{\mathrm{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} w_\pi(a_i|x_i)c_i \,,$$

where $w_\pi(a|x) = \frac{\pi(a|x)}{\pi_0(a|x)}$ are the **importance weights**.

Problem: Large variance when π is different from π_0.

Common Solution: Hard clipping with $\tau \in [0, \infty)$,

$$w_\pi(a|x) \leftarrow \min\left(\tau, \frac{\pi(a|x)}{\pi_0(a|x)}\right).$$

Our Proposal: Exponential smoothing with $\alpha \in [0, 1]$,

$$w_\pi(a|x) \leftarrow \frac{\pi(a|x)}{\pi_0(a|x)^\alpha}\,.$$

Exponential Smoothing for OPE

$\min(\tau, \frac{\pi(\mathrm{a}|\mathrm{x})}{\pi_0(\mathrm{a}|\mathrm{x})})$, $\tau \in \mathbb{R}^+$ $\dashrightarrow$ $\frac{\pi(a|x)}{\pi_0(a|x)^\alpha}$, $\alpha \in [0, 1]$

(1) τ in an unbounded domain $\mathbb{R}^+$
(2) $\min(\tau, \frac{\pi(\mathrm{a}|\mathrm{x})}{\pi_0(\mathrm{a}|\mathrm{x})})$ is non-differentiable in π
(3) $\min(\tau, \frac{\pi(\mathrm{a}|\mathrm{x})}{\pi_0(\mathrm{a}|\mathrm{x})})$ is bounded

(1) α in a bounded domain $[0, 1]$
(2) $\frac{\pi(a|x)}{\pi_0(a|x)^\alpha}$ is differentiable and linear in π
(3) $\frac{\pi(a|x)}{\pi_0(a|x)^\alpha}$ is unbounded

Other corrections were proposed, but ours *simultaneously* allows

(1) easier tuning of $\alpha \in [0, 1]$, **(2)** differentiable objectives,
(3) smaller bias as the corrected importance weights are not constrained to be bounded.

Exponential Smoothing for OPL

PAC-Bayes formulation: $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{A}\}$ is a hypothesis space. Then, policies are defined as[a]

$$\pi(a|x) = \pi_{\mathbb{Q}}(a|x) = \mathbb{P}_{h\sim\mathbb{Q}}\big(h(x) = a\big) = \mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{h(x)=a}\right].$$

[a] Ben London and Ted Sandler. "Bayesian counterfactual risk minimization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4125–4133.

[1] O. Sakhi, N. Chopin, and P. Alquier. "PAC-Bayesian Offline Contextual Bandits With Guarantees". In: *ICML* (2023).

Exponential Smoothing for OPL

PAC-Bayes formulation: $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{A}\}$ is a hypothesis space. Then, policies are defined as[a]

$$\pi(a|x) = \pi_{\mathbb{Q}}(a|x) = \mathbb{P}_{h\sim\mathbb{Q}}\big(h(x) = a\big) = \mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{h(x)=a}\right].$$

[a] Ben London and Ted Sandler. "Bayesian counterfactual risk minimization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4125–4133.

- This is not an assumption[1].
- Softmax, mixed-logit, and Gaussian policies have this form.
- **Suitable for PAC-Bayes:**
  - We control $|\mathbb{E}_{h\sim\mathbb{Q}}[\hat{R}_n(h) - R(h)]|$.
  - Given a prior $\mathbb{P}$ (e.g., $\pi_0 = \pi_{\mathbb{P}}$), learn a posterior $\mathbb{Q}$ that minimizes the expected risk $\mathbb{E}_{h\sim\mathbb{Q}}[R(h)]$.

[1] O. Sakhi, N. Chopin, and P. Alquier. "PAC-Bayesian Offline Contextual Bandits With Guarantees". In: *ICML* (2023).

Exponential Smoothing for OPL

We derive tight and tractable PAC-Bayesian bounds under our estimator:

$$|R(\pi_{\mathbb{Q}}) - \hat{R}_n^\alpha(\pi_{\mathbb{Q}})| \leq \mathcal{O}\Big(\frac{D_{\mathrm{KL}}(\mathbb{Q}||\mathbb{P}) + \bar{V}_n^\alpha(\pi_{\mathbb{Q}})}{\sqrt{n}} + B_n^\alpha(\pi_{\mathbb{Q}})\Big)\,,$$

where

- $\hat{R}_n^\alpha(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^n \frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^\alpha} c_i\,, \qquad \forall \alpha \in [0,1]\,.$
- $\pi_0 = \pi_{\mathbb{P}}$.
- $B_n^\alpha(\pi_{\mathbb{Q}})$ is a bias term.
- $\bar{V}_n^\alpha(\pi_{\mathbb{Q}})$ is a variance term.

Exponential Smoothing for OPL

We derive tight and tractable PAC-Bayesian bounds under our estimator:

$$|R(\pi_{\mathbb{Q}}) - \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \mathcal{O}\Big(\frac{D_{\mathrm{KL}}(\mathbb{Q}||\mathbb{P}) + \bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})}{\sqrt{n}} + B_n^{\alpha}(\pi_{\mathbb{Q}})\Big),$$

where

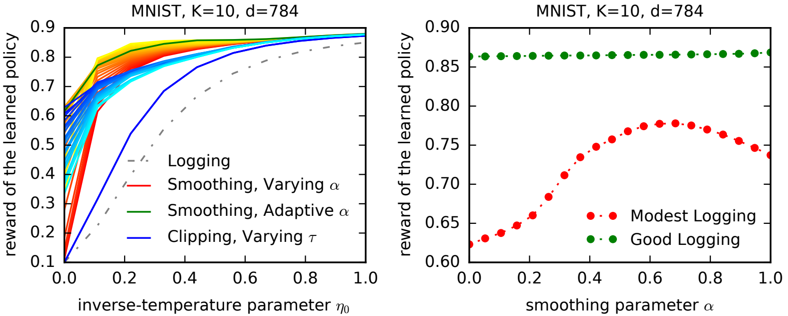
- $\hat{R}_n^{\alpha}(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{\alpha}} c_i\,, \qquad \forall \alpha \in [0,1]\,.$
- $\pi_0 = \pi_{\mathbb{P}}$.
- $B_n^{\alpha}(\pi_{\mathbb{Q}})$ is a bias term.
- $\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})$ is a variance term.

Grounded and data-**adaptive** principle to simultaneously optimize $\alpha \in [0,1]$ and $\mathbb{Q} \in \mathcal{M}_1(\mathcal{H})$ as

$$\operatorname*{arg\,min}_{\mathbb{Q}\in\mathcal{M}_1(\mathcal{H}),\alpha\in[0,1]} \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}}) + \mathcal{O}\Big(\frac{D_{\mathrm{KL}}(\mathbb{Q}||\mathbb{P}) + \bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})}{\sqrt{n}} + B_n^{\alpha}(\pi_{\mathbb{Q}})\Big)\,.$$

Experiments

Below, η_0 represents the quality of the logging policy (the higher the better). We perform better than the most competitive baseline[1].



POSTER: Exhibit Hall 1 #309, Wed 26 Jul 11 a.m. - 12:30 p.m.

Thank you!

[1] O. Sakhi, N. Chopin, and P. Alquier. "PAC-Bayesian Offline Contextual Bandits With Guarantees". In: *ICML* (2023).