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

Table of contents

Motivation

Why Revisit IPS?

Setting

Offline contextual bandit [7, 8, 9, 10] with logged data $\mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^n$ from a known logging policy π_0.

Goal: learn $\hat{\pi} \in \Pi$ maximizing $V(\pi) = \mathbb{E}_{X\sim\nu,\, A\sim\pi(\cdot|X)}\left[r(X, A)\right]$.

Why Revisit IPS?

Setting

Offline contextual bandit [7, 8, 9, 10] with logged data $\mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^n$ from a known logging policy π_0.

Goal: learn $\hat{\pi} \in \Pi$ maximizing $V(\pi) = \mathbb{E}_{X\sim\nu,\, A\sim\pi(\cdot|X)}\left[r(X, A)\right]$.

Problem

IPS is unbiased but has high variance; **IW clipping** reduces variance but: (i) introduces high bias, (ii) is non-differentiable (flat regions), (iii) sensitive to hyperparameter tuning.

Why Revisit IPS?

Setting

Offline contextual bandit [7, 8, 9, 10] with logged data $\mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^n$ from a known logging policy π_0.

Goal: learn $\hat{\pi} \in \Pi$ maximizing $V(\pi) = \mathbb{E}_{X\sim\nu,\, A\sim\pi(\cdot|X)}[r(X, A)]$.

Problem

IPS is unbiased but has high variance; **IW clipping** reduces variance but: (i) introduces high bias, (ii) is non-differentiable (flat regions), (iii) sensitive to hyperparameter tuning.

Our answer

Exponential smoothing (ES): smooth, differentiable IW regularization, and two-sided PAC-Bayes generalization bounds that are *optimizable by SGD*.

Regularized IPS

IPS and Regularized IPS

IPS [8]

$$\hat{V}_{\mathsf{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i\, w(A_i \mid X_i), \quad w(a \mid x) = \frac{\pi(a \mid x)}{\pi_0(a \mid x)}.$$

IPS and Regularized IPS

IPS [8]

$$\hat{V}_{\mathsf{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i\, w(A_i \mid X_i), \quad w(a \mid x) = \frac{\pi(a \mid x)}{\pi_0(a \mid x)}.$$

Regularized IPS [5]

$$\hat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i\, \hat{w}(A_i \mid X_i), \qquad \hat{w} \le w.$$

Hard IW clipping: $\hat{w} = \min\{w,\, M\}$ or $\hat{w} = \frac{\pi}{\max(\pi_0, \tau)}$.

IPS and Regularized IPS

Regularized IPS (generic)

$$\hat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i\,\hat{w}(A_i \mid X_i), \qquad \hat{w} \le w.$$

Hard IW clipping: $\hat{w} = \min\{w,\, M\}$ or $\hat{w} = \frac{\pi}{\max(\pi_0,\tau)}$.

IPS and Regularized IPS

Regularized IPS (generic)

$$\hat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i\,\hat{w}(A_i \mid X_i), \qquad \hat{w} \le w.$$

Hard IW clipping: $\hat{w} = \min\{w,\, M\}$ or $\hat{w} = \frac{\pi}{\max(\pi_0,\tau)}$.

Limitations of hard IW clipping

Non-differentiable (zero gradients beyond M), highly sensitive to M and τ, loses ordering when many $\pi_0(\cdot|x)$ are clipped to the same value.

Exponential Smoothing

Definition and Properties

Smooth variant

$$\texttt{IPS-}\alpha: \quad \hat{V}^{\alpha}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i \, \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)^{\alpha}}, \ \alpha \in [0,1].$$

Definition and Properties

Smooth variant

$$\texttt{IPS-}\alpha : \quad \hat{V}^{\alpha}(\pi) = \frac{1}{n} \sum_{i=1}^{n} R_i \, \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)^{\alpha}}, \ \alpha \in [0, 1].$$

Bias-variance trade-off for α

$$|\mathbb{B}(\hat{V}^{\alpha})| \leq \mathbb{E}_{X, A \sim \pi(\cdot|X)} \left[1 - \pi_0(A|X)^{1-\alpha}\right],$$

$$\mathbb{V}\left[\hat{V}^{\alpha}\right] \leq \frac{1}{n} \mathbb{E}_{X, A \sim \pi(\cdot|X)} \left[\frac{\pi(A|X)}{\pi_0(A|X)^{2\alpha-1}}\right].$$

$\alpha \to 1$: low bias (IPS); $\alpha \to 0$: low variance.

Why ES Beats Clipping in Optimization

- Smooth and everywhere differentiable $\Rightarrow$ stable SGD; no flat regions.
- Preserves ranking induced by π_0: if $\pi_0(a|x) < \pi_0(a'|x)$ then $\pi_0(a|x)^\alpha < \pi_0(a'|x)^\alpha$.
- Single bounded hyperparameter ($\alpha \in [0, 1]$) instead of $M \in [0, \infty)$.

Pessimism via PAC-Bayes

From One-Sided to Two-Sided

Prior pessimistic objectives

One-sided bounds [11, 12] lead to $V(\pi) \geq \hat{V}(\pi) - g(\cdot)$ but cannot certify estimator quality.

From One-Sided to Two-Sided

Prior pessimistic objectives

One-sided bounds [11, 12] lead to $V(\pi) \geq \hat{V}(\pi) - g(\cdot)$ but cannot certify estimator quality.

Our approach

Two-sided, **tractable** PAC-Bayes bound directly optimized by SGD. Works without the bounded-IW assumption and applies to **standard IPS** $(\alpha = 1)$.

Main Theorem (Two-Sided PAC-Bayes, Informal)

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

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.

Adaptive, Data-Driven α

Tuning α

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

$$\underset{\mathbb{Q}\in\mathcal{M}_1(\mathcal{H}),\alpha\in[0,1]}{\arg\min} \; \hat{R}_n^\alpha(\pi_\mathbb{Q}) + \mathcal{O}\left(\frac{D_{\mathsf{KL}}(\mathbb{Q}||\mathbb{P}) + \bar{V}_n^\alpha(\pi_\mathbb{Q})}{\sqrt{n}} + B_n^\alpha(\pi_\mathbb{Q})\right).$$

Experiments

Setup

- Supervised-to-bandit conversion on vision datasets: `MNIST`, `FashionMNIST`, `EMNIST`, `CIFAR100`.
- Policies: Gaussian and Mixed-Logit (PAC-Bayes-friendly); priors tied (optionally) to π_0.
- Optimization: Adam; we compare **two-sided** bound vs. prior one-sided baselines.

ES vs. Clipped-PAC Baselines

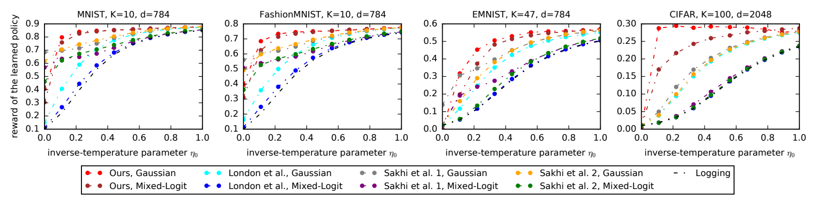

Figure 1: Across logging-quality η_0, ES + two-sided PAC-Bayes outperforms [11]. Gaussian policies typically strongest.

Tuning Sensitivity and Adaptive α

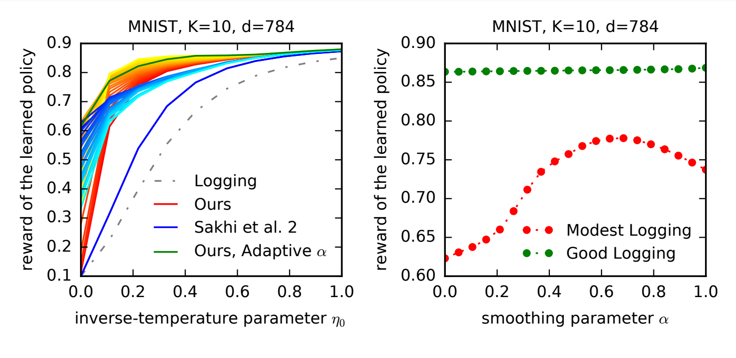

Figure 2: *Left:* grid over τ (clip) and α (ES); adaptive α close to best fixed choice. *Right:* average reward value for varying α using either modest or good logging; IW regularization is much needed for modest logging policies.

Takeaways

Key Takeaways

- **Exponential smoothing**: smooth IW regularization with explicit bias-variance control; better optimization behavior than clipping.
- **Two-sided, tractable PAC-Bayes bounds**: applicable to standard IPS; SGD-friendly.
- **Theory extended to any IW regularization technique** [5].

Limitations: Data-dependent quantities in the bound; symmetric tails may be loose. Performance breaks in large-scale settings [3] where Bayesian direct methods with informative priors [1, 2, 4, 6] perform better when the number of actions is high.

References

[1] Imad Aouali. Linear diffusion models meet contextual bandits with large action spaces. In *NeurIPS 2023 Workshop on Foundation Models for Decision Making*, 2023.

[2] Imad Aouali. Diffusion models meet contextual bandits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

References

[3] Imad Aouali, Achraf Ait Sidi Hammou, Otmane Sakhi, David Rohde, and Flavian Vasile. Probabilistic rank and reward: A scalable model for slate recommendation. *arXiv preprint arXiv:2208.06263*, 2022.

[4] Imad Aouali, Branislav Kveton, and Sumeet Katariya. Mixed-effect thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 2087–2115. PMLR, 2023.

[5] 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.

References

[6] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Bayesian off-policy evaluation and learning for large action spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 136–144. PMLR, 2025.

[7] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

References

[8] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1097–1104, 2011.

[9] Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

[10] Lihong Li, Wei Chu, John Langford, and Robert Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

References

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

[12] 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.