Uncertainty-Aware Instance Reweighting for Off-Policy Learning

Xiaoying Zhang, Junpu Chen, Hongning Wang, Hong Xie, Yang Liu, John C.S. Lui, Hang Li

香港中文大學 The Chinese University of Hong Kong

ByteDance

Background : Off-Policy Learning

Input: A logged dataset $D = \{(x_n, a_n, r_{x_n,a_n})\}_{n=1}^N$ generated by logging policy $\beta^*(a|x)$.

Goal: Learning a policy $\pi(a|x)$ that maximize

$$V(\pi) = E_\pi[r_{x,a}] = E_{\beta^*}[\frac{\pi(a|x)}{\beta^*(a|x)} r_{x,a}]$$

Inverse Propensity Score (IPS)

$$\hat{V}_{IPS}(\pi) = \frac{1}{N}\sum_{n=1}^{N} \frac{\pi(a_n|x_n)}{\beta^*(a_n|x_n)} r_{x_n,a_n}$$

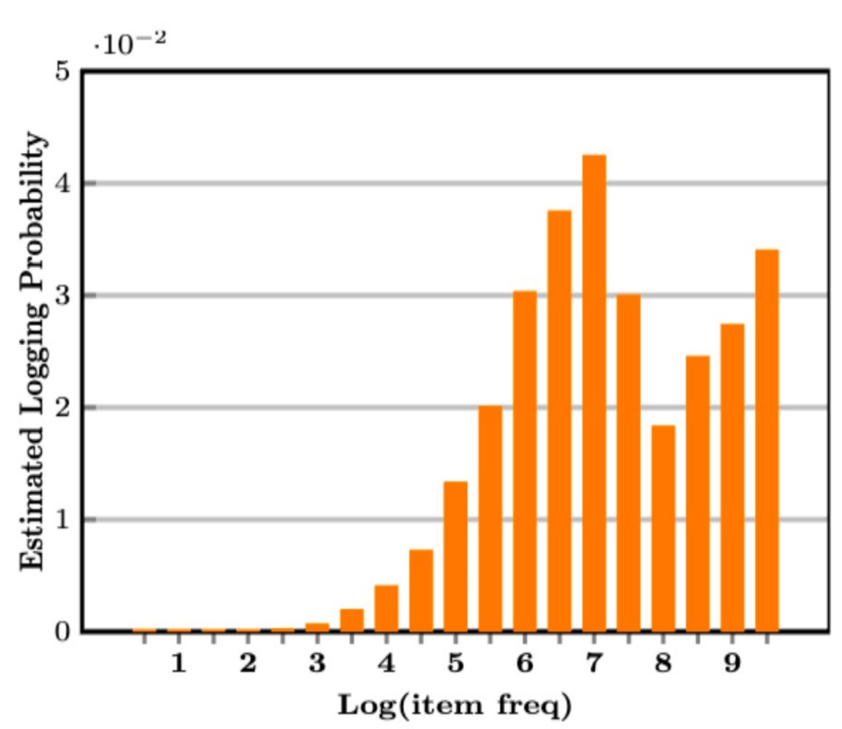
Approximation when β^* is unknown → BIPS

- In practice, $\beta^*$is usually unknown and approximated by its estimate $\hat{\beta}$.

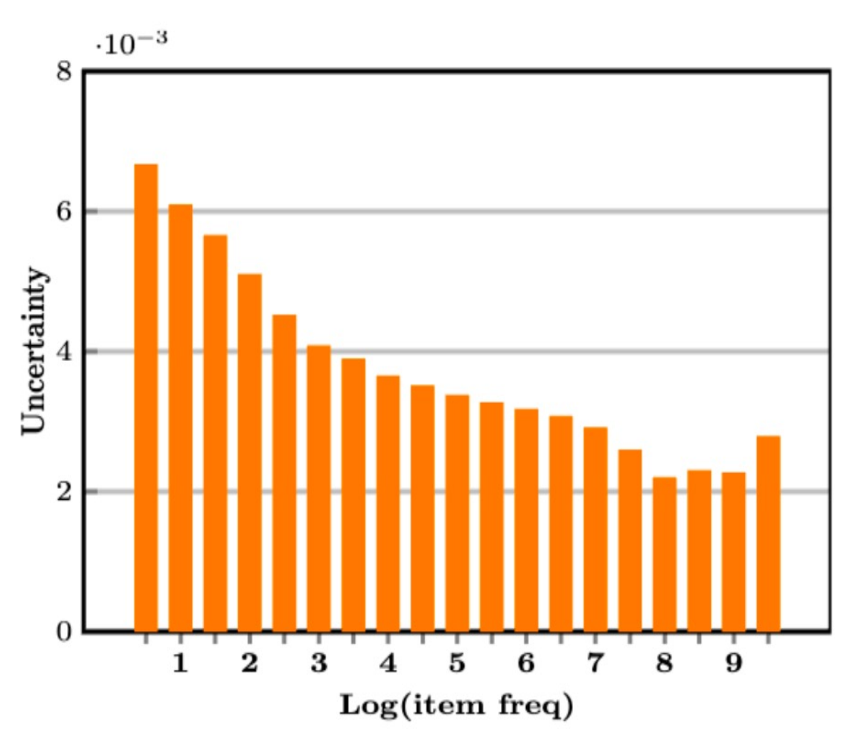
$$\hat{V}_{BIPS}(\pi) = \frac{1}{N}\sum_{n=1}^{N} \frac{\pi(a_n|x_n)}{\hat{\beta}(a_n|x_n)} r_{x_n,a_n}$$

BIPS suffers high bias and variance with inaccurate $\hat{\beta}$

- Theoretically, inaccurate and small $\hat{\beta}$(a|x) → high bias and variance of $\hat{V}_{BIPS}$.
- However, smaller $\hat{\beta}$(a|x) → more likely to be inaccurate (i.e., high uncertainty in estimation).



(a) Estimated Logging Probability



(b) Uncertainty of Estimation

Uncertainty-Aware Off-Policy Learning (UIPS)

$$\hat{V}_{UIPS}(\pi_\vartheta) = \frac{1}{N}\sum_{n=1}^{N} \frac{\pi_\vartheta(a_n|x_n)}{\hat{\beta}(a_n|x_n)} \cdot \phi_{x_n,a_n} \cdot r_{x_n,a_n}$$

- Per-sample weight $\phi_{x,a}$: small value when $\hat{\beta}$(a|x) is small and far from β^*(a|x).

Iterate between two steps

➢ Step 1: Deriving Optimal $\phi^*_{x,a}$: $\hat{V}_{UIPS}(\pi_\vartheta) \rightarrow V(\pi_\vartheta)$

Minimize the upper bound of Mean Square Error (MSE) of $\hat{V}_{UIPS}(\pi_\vartheta)$ to its ground-truth $V(\pi_\vartheta)$ → Per-sample optimization:

$$\min_{\phi_{x,a}} \lambda\left(\frac{\beta^*(a|x)}{\hat{\beta}(a|x)}\phi_{x,a} - 1\right)^2 + \frac{\pi_\vartheta(a|x)^2}{\hat{\beta}(a|x)^2}\phi_{x,a}^2$$

Direct optimization is infeasible due to the unknown $\beta^*(a|x)$, but uncertainty estimation provides its confidence interval $\beta^*(a|x) \in B_{x,a}$:

$$\min_{\phi_{x,a}} \max_{\beta_{x,a}} \lambda\left(\frac{\beta_{x,a}}{\hat{\beta}(a|x)}\phi_{x,a} - 1\right)^2 + \frac{\pi_\vartheta(a|x)^2}{\hat{\beta}(a|x)^2}\phi_{x,a}^2$$

→ Closed-form solution of $\phi^*_{x,a}$!

➢ Step 2: Policy Improvement:

$$\nabla_\vartheta \hat{V}_{UIPS}(\pi_\vartheta) = \frac{1}{N}\sum_{n=1}^{N} \frac{\pi_\vartheta(a_n|x_n)}{\beta^*(a_n|x_n)} \cdot \phi^*_{x_n,a_n} \cdot r_{x_n,a_n} \nabla_\vartheta \log(\pi_\vartheta(a_n|x_n))$$

Theoretical Convergence of UIPS

UIPS converges to a stationary point where the true policy gradient is zero, while convergence of policy learning under BIPS is not guaranteed!

Empirical Results

➢ Improved performance on both synthetic datasets and three unbiased recommendation datasets.

	Yahoo			Coat			KuaiRec		
Algorithm	P@5	R@5	NDCG@5	P@5	R@5	NDCG@5	P@50	R@50	NDCG@50
CE	$0.2819 \pm 2e^{-3}$	$0.7594 \pm 6e^{-3}$	$0.6073 \pm 7e^{-3}$	$0.2799 \pm 5e^{-3}$	$0.4618 \pm 1e^{-2}$	$0.4529 \pm 7e^{-3}$	$0.8802 \pm 2e^{-3}$	$0.0240 \pm 8e^{-5}$	$0.8810 \pm 6e^{-3}$
BIPS-Cap	$0.2808 \pm 2e^{-3}$	$0.7576 \pm 5e^{-3}$	$0.6099 \pm 8e^{-3}$	$0.2758 \pm 6e^{-3}$	$0.4582 \pm 7e^{-3}$	$0.4399 \pm 9e^{-3}$	$0.8750 \pm 3e^{-3}$	$0.0238 \pm 7e^{-5}$	$0.8788 \pm 5e^{-3}$
MinVar	$0.2843 \pm 4e^{-3}$	$0.7685 \pm 1e^{-2}$	$0.6168 \pm 1e^{-2}$	$0.2813 \pm 3e^{-3}$	$0.4668 \pm 9e^{-3}$	$0.4414 \pm 8e^{-3}$	$0.8827 \pm 1e^{-3}$	$0.0240 \pm 5e^{-5}$	$0.8886 \pm 2e^{-3}$
stableVar	$0.2787 \pm 2e^{-3}$	$0.7499 \pm 7e^{-3}$	$0.5919 \pm 7e^{-3}$	$0.2840 \pm 3e^{-3}$	$0.4662 \pm 5e^{-3}$	$0.4393 \pm 7e^{-3}$	$0.8524 \pm 7e^{-3}$	$0.0231 \pm 2e^{-4}$	$0.8570 \pm 4e^{-3}$
Shrinkage	$0.2843 \pm 3e^{-3}$	$0.7654 \pm 8e^{-3}$	$0.6204 \pm 7e^{-3}$	$0.2790 \pm 5e^{-3}$	$0.4636 \pm 4e^{-3}$	$0.4464 \pm 1e^{-2}$	$0.8744 \pm 3e^{-3}$	$0.0238 \pm 9e^{-5}$	$0.8771 \pm 6e^{-3}$
SNIPS	$0.2222 \pm 4e^{-3}$	$0.5828 \pm 1e^{-2}$	$0.4357 \pm 1e^{-2}$	$0.2643 \pm 7e^{-3}$	$0.4287 \pm 1e^{-2}$	$0.4009 \pm 9e^{-3}$	$0.8411 \pm 6e^{-3}$	$0.0228 \pm 2e^{-4}$	$0.8431 \pm 6e^{-3}$
BanditNet	$0.2413 \pm 8e^{-3}$	$0.6442 \pm 2e^{-2}$	$0.4988 \pm 2e^{-2}$	$0.2781 \pm 8e^{-3}$	$0.4527 \pm 1e^{-2}$	$0.4251 \pm 1e^{-2}$	$0.8758 \pm 5e^{-3}$	$0.0239 \pm 2e^{-4}$	$0.8810 \pm 4e^{-3}$
POEM	$0.2732 \pm 3e^{-3}$	$0.7357 \pm 1e^{-2}$	$0.5880 \pm 1e^{-2}$	$0.2791 \pm 4e^{-3}$	$0.4566 \pm 6e^{-3}$	$0.4375 \pm 6e^{-3}$	$0.7785 \pm 1e^{-2}$	$0.0210 \pm 2e^{-4}$	$0.7779 \pm 6e^{-3}$
POXM	$0.2250 \pm 5e^{-3}$	$0.5940 \pm 1e^{-2}$	$0.4542 \pm 2e^{-2}$	$0.2663 \pm 6e^{-3}$	$0.4308 \pm 9e^{-3}$	$0.4006 \pm 1e^{-2}$	$0.8962 \pm 1e^{-2}$	$0.0245 \pm 4e^{-4}$	$0.9041 \pm 1e^{-2}$
Adaptive	$0.2762 \pm 3e^{-3}$	$0.7451 \pm 9e^{-3}$	$0.5919 \pm 8e^{-3}$	$0.2830 \pm 3e^{-3}$	$0.4634 \pm 5e^{-3}$	$0.4217 \pm 5e^{-3}$	$0.8375 \pm 1e^{-2}$	$0.0227 \pm 4e^{-4}$	$0.8460 \pm 1e^{-2}$
ApproxKNN	$0.2697 \pm 2e^{-3}$	$0.7225 \pm 5e^{-3}$	$0.5760 \pm 6e^{-3}$	$0.2755 \pm 2e^{-3}$	$0.4594 \pm 5e^{-3}$	$0.4490 \pm 4e^{-3}$	$0.8839 \pm 2e^{-6}$	$0.0240 \pm 5e^{-5}$	$0.8895 \pm 2e^{-3}$
IPS-C-TS	$0.2816 \pm 2e^{-3}$	$0.7582 \pm 5e^{-3}$	$0.6114 \pm 5e^{-3}$	$0.2799 \pm 3e^{-3}$	$0.4625 \pm 7e^{-3}$	$0.4462 \pm 6e^{-3}$	$0.8781 \pm 3e^{-3}$	$0.0239 \pm 1e^{-4}$	$0.8749 \pm 3e^{-3}$
UIPS-P	$0.1829 \pm 8e^{-3}$	$0.4560 \pm 3e^{-2}$	$0.3300 \pm 1e^{-2}$	$0.2685 \pm 7e^{-3}$	$0.4364 \pm 9e^{-3}$	$0.4087 \pm 7e^{-3}$	$0.8638 \pm 8e^{-3}$	$0.0235 \pm 3e^{-4}$	$0.8685 \pm 7e^{-3}$
UIPS-O	$0.1947 \pm 3e^{-3}$	$0.4959 \pm 1e^{-2}$	$0.3600 \pm 8e^{-3}$	$0.2657 \pm 5e^{-3}$	$0.4306 \pm 9e^{-3}$	$0.4146 \pm 9e^{-3}$	$0.8651 \pm 8e^{-3}$	$0.0235 \pm 2e^{-4}$	$0.8697 \pm 7e^{-3}$
UIPS	$\mathbf{0.2868} \pm 2e^{-3}$	$\mathbf{0.7742} \pm 5e^{-3}$	$\mathbf{0.6274} \pm 5e^{-3}$	$\mathbf{0.2877} \pm 3e^{-3}$	$\mathbf{0.4757} \pm 5e^{-3}$	$\mathbf{0.4576} \pm 8e^{-3}$	$\mathbf{0.9120} \pm 1e^{-3}$	$\mathbf{0.0250} \pm 5e^{-5}$	$\mathbf{0.9174} \pm 7e^{-4}$
p-value	$4e^{-2}$	$1e^{-2}$	$3e^{-2}$	$2e^{-2}$	$6e^{-4}$	$5e^{-5}$	$6e^{-4}$	$6e^{-4}$	$1e^{-3}$

➢ More accurate Off-policy Evaluation

Table 3: MSE of different off-policy estimators. A lower MSE indicates a more accurate estimator.

Algorithm	IPS-GT	BIPS	minVar	stableVar	Shrinage	UIPS
$\tau = 0.5$	$0.0875 \pm 4e^{-4}$	15.786 ± 1.51	$0.9021 \pm 7e^{-13}$	$0.8612 \pm 5e^{-8}$	$0.0718 \pm 5e^{-6}$	$\mathbf{0.0210} \pm 2e^{-6}$
$\tau = 1.0$	$0.0209 \pm 8e^{-5}$	0.5510 ± 0.388	$0.9019 \pm 8e^{-12}$	$0.8578 \pm 2e^{-7}$	$0.1978 \pm 2e^{-5}$	$\mathbf{0.0093} \pm 1e^{-6}$
$\tau = 2.0$	$0.0020 \pm 6e^{-6}$	0.5669 ± 0.013	$0.9015 \pm 5e^{-15}$	$0.8342 \pm 5e^{-7}$	$0.2952 \pm 3e^{-5}$	$\mathbf{0.0043} \pm 4e^{-7}$

➢ Performance under different uncertainties

- the only off-policy algorithm that outperforms CE on test samples with high uncertainty.

	Actions on Samples with High Uncertainty			Actions on Samples with Low Uncertainty		
Algorithm	P@5(RI)	R@5(RI)	NDCG@5(RI)	P@5(RI)	R@5(RI)	NDCG@5(RI)
CE	0.5190	0.1231	0.5526	0.5913	0.1915	0.6549
BIPS-Cap	0.5117 (-1.41%)	0.1202 (-2.33%)	0.5488 (-0.68%)	0.5913 (+0.00%)	0.1903 (-0.64%)	0.6574 (+0.39%)
Shrinkage	0.5158 (-0.62%)	0.1217 (-1.11%)	0.5505 (-0.37%)	0.5892 (-0.35%)	0.1905 (-0.55%)	0.6546 (-0.05%)
UIPS	0.5222 (+0.61%)	0.1237 (+0.50%)	0.5568 (+0.77%)	0.5994 (+1.38%)	0.1940 (+1.28%)	0.6658 (+1.66%)