Subgaussian and Differentiable Importance Sampling for Off-Policy Evaluation and Learning

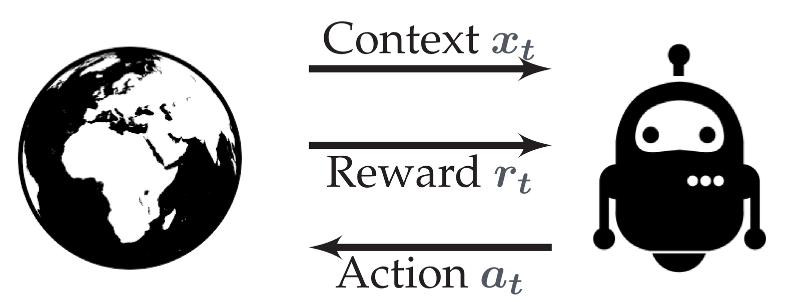
Alberto Maria Metelli, Alessio Russo, and Marcello Restelli

{albertomaria.metelli, alessio.russo, marcello.restelli}@polimi.it

Politecnico Milano 1863 — AIRLAB Artificial Intelligence and Robotics Lab

Neural Information Processing Systems

Contextual Bandits



- Environment samples a context $x_t \sim \rho$
- Agent plays an action $a_t \sim \pi(\cdot|x_t)$
- Environment generates a reward $r_t = r(x_t, a_t)$

Goal: policy π^* maximizing the **expected reward** (Langford and Zhang, 2007)

$$\pi^* \in \arg\max_{\pi} v(\pi) = \mathop{\mathbb{E}}_{\substack{x\sim\rho \\ a\sim\pi(\cdot|x)}} [r(x,a)]$$

Vanilla Importance Sampling

- **Goal**: estimate the expectation μ of a function f under a **target** distribution P having samples collected with a **behavioral** distribution Q (Owen, 2013)

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i\in[n]} \underbrace{\frac{P(y_i)}{Q(y_i)}}_{\omega(y_i)} f(y_i) \qquad y_i \overset{\text{iid}}{\sim} Q, \quad P \ll Q$$

importance weight

☺ **Unbiased**: $\mathop{\mathbb{E}}_{y_i \overset{\text{iid}}{\sim} Q}[\widehat{\mu}_n] = \mathop{\mathbb{E}}_{y\sim P}[f(y)] = \mu$

☹ **Variance**: can be very large! (Metelli et al., 2018)

$$\mathop{\mathrm{Var}}_{y_i \overset{\text{iid}}{\sim} Q}[\widehat{\mu}_n] \leqslant \frac{\|f\|_\infty}{n} \underbrace{I_2(P\|Q)}_{\simeq \text{exp Rényi divergence}} \qquad I_\alpha(P\|Q) = \int_y P(y)^\alpha Q(y)^{1-\alpha}\mathrm{d}y$$

Anticoncentration of Vanilla Importance Sampling

- **Polynomial** (dependence on δ) concentration (Metelli et al., 2018)

$$|\widehat{\mu}_n - \mu| \leqslant O\left(\|f\|_\infty \left(\frac{I_\alpha(P\|Q)}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\right) \quad \text{w.p. } 1-\delta$$

☹ **Anti-concentration** (ours): Polynomial concentration is tight!

$$|\widehat{\mu}_n - \mu| \geqslant \Omega\left(\|f\|_\infty \left(\frac{I_\alpha(P\|Q)-1}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\right) \quad \text{w.p. } \delta$$

Importance Sampling Corrections

- **Self-Normalized** Importance Sampling (SN-IS, Kuzborskij et al., 2021)

$$\omega^{\mathrm{SN}}(y_i) = \frac{n\omega(y_i)}{\sum_{j\in[n]}\omega(y_j)}$$

- Importance Sampling with **TRuncation** (IS-TR, Ionides, 2008)

$$\omega^{\mathrm{TR}}(y_i) = \min\{\omega(y_i), M\}$$

- Importance Sampling with **Optimistic Shrinkage** (IS-OS, Su et al., 2020)

$$\omega^{\mathrm{OS}}(y_i) = \frac{\tau\omega(y_i)}{\omega(y_i)^2 + \tau}$$

Power-Mean Correction of Importance Sampling

- **Idea**: interpolate between **vanilla weight** and **1** in a **smooth** way
- (s, λ)-corrected weight

$$\omega_{\lambda,s}(y) = \left((1-\lambda)\underbrace{\omega(y)}_{\text{vanilla weight}}{}^{s} + \lambda\right)^{\frac{1}{s}}$$

☺ **Unbiased** when $P = Q$ a.s.

☺ If $s < 0$, the weight is **bounded**: $\omega_{\lambda,s}(y) \leqslant \lambda^{\frac{1}{s}}$

We focus on $s = -1$ $\quad\rightarrow\quad$ $\omega_{\lambda,-1}(y) = \dfrac{\omega(y)}{(1-\lambda) + \lambda\omega(y)}$

Concentration Inequalities

- Select λ as a function of $I_\alpha(P\|Q)$ and δ
- **Exponential** (dependence on δ) concentration

$$\widehat{\mu}_{n,\lambda_\alpha^*} - \mu \leqslant \|f\|_\infty(2+\sqrt{3})\left(\frac{2I_\alpha(P\|Q)^{\frac{1}{\alpha-1}}\log\frac{1}{\delta}}{3(\alpha-1)^2 n}\right)^{1-\frac{1}{\alpha}} \quad \text{w.p. } 1-\delta$$

☺ With $\alpha = 2$, we have **Subgaussian** concentration inequality

$$\widehat{\mu}_{n,\lambda_2^*} - \mu \leqslant \|f\|_\infty(2+\sqrt{3})\sqrt{\frac{2I_\alpha(P\|Q)\log\frac{1}{\delta}}{3n}} \quad \text{w.p. } 1-\delta$$

- Method to compute λ_2^* without knowledge of $I_\alpha(P\|Q)$ in the paper

Differentiability

- When the **target** distribution is parametric and differentiable $P_{\boldsymbol{\theta}}$

$$\nabla_{\boldsymbol{\theta}}\omega_\lambda(y) = \frac{(1-\lambda)\omega(y)}{(1-\lambda+\lambda\omega(y))^2}\nabla_{\boldsymbol{\theta}}\log P_{\boldsymbol{\theta}}(y)$$

- **Bounded** gradient when $\lambda > 0$

$$\|\nabla_{\boldsymbol{\theta}}\omega_\lambda(y)\|_\infty \leqslant \frac{1}{4\lambda}\|\nabla_{\boldsymbol{\theta}}\log P_{\boldsymbol{\theta}}(y)\|_\infty$$

Importance Sampling Corrections Comparison

Estimator	Concentration (order O)	Is subgaussian?	Is unbiased when $P = Q$?	Is differentiable?
IS	$\sqrt{\frac{I_2(P\|Q)}{\delta n}}$	☹ (poly)	☺	☺
SN-IS	$B^{\mathrm{SN}} + \sqrt{V^{\mathrm{ES}}\log\frac{1}{\delta}}$	☹ (exp)	☺	☺
IS-TR	$\sqrt{\frac{I_2(P\|Q)\log\frac{1}{\delta}}{n}}$	☺	☹	☹
IS-OS	$\max_{\beta\in\{2,3\}}\sqrt[\beta]{\frac{I_\beta(P\|Q)(\log\frac{1}{\delta})^{\beta-1}}{n^{\beta-1}}}$	☹ (exp)	☹	☺
IS-λ	$\sqrt{\frac{I_2(P\|Q)\log\frac{1}{\delta}}{n}}$	☺	☺	☺

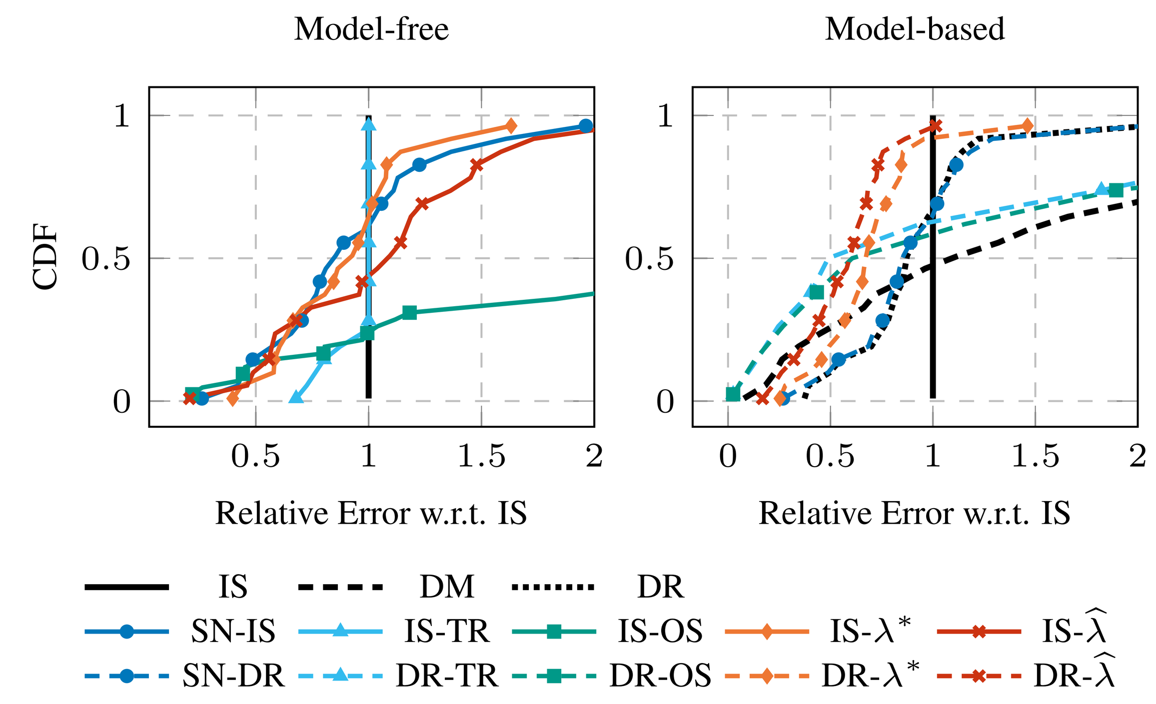
Experiments

Off-Policy Evaluation

1. Synthetic experiment with Gaussian distributions
   - $I_2(P\|Q) \simeq 27.9$ and $f(y) = 100\cos(2\pi y)$
   - MSE (best in **bold** and second best underlined)

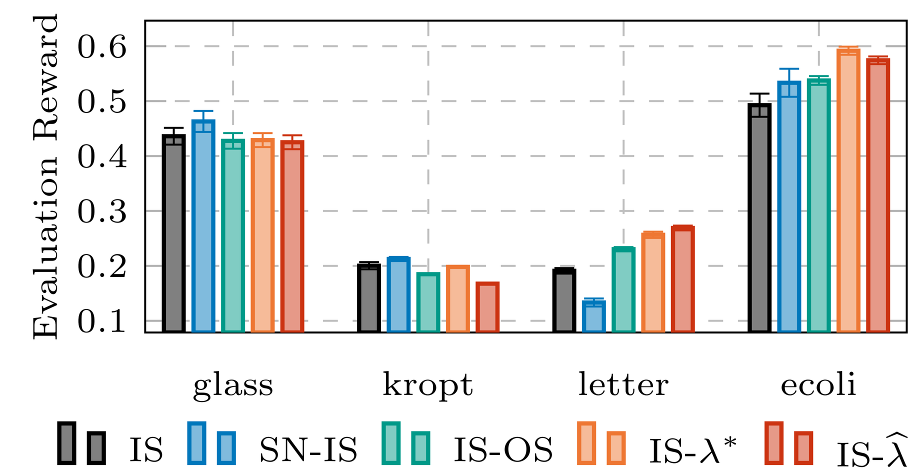
Estimator / n	10	20	50	100	200	500	1000
IS	27.43 ± 13.33	15.70 ± 4.83	10.89 ± 1.81	9.26 ± 0.92	12.41 ± 1.88	9.42 ± 0.68	5.84 ± 0.27
SN-IS	23.89 ± 5.77	15.62 ± 2.62	10.96 ± 1.18	9.53 ± 0.74	8.82 ± 0.62	7.48 ± 0.37	5.14 ± 0.20
IS-TR	23.47 ± 7.52	14.03 ± 2.75	10.32 ± 1.47	8.89 ± 0.79	<u>7.68 ± 0.46</u>	<u>6.21 ± 0.28</u>	4.22 ± 0.15
IS-OS	**19.25 ± 8.68**	**10.93 ± 3.29**	**8.37 ± 1.35**	**7.06 ± 0.61**	8.69 ± 1.44	6.65 ± 0.47	<u>3.97 ± 0.16</u>
IS-λ^*	<u>21.75 ± 6.36</u>	<u>13.17 ± 2.45</u>	<u>9.26 ± 1.19</u>	<u>7.76 ± 0.62</u>	**6.53 ± 0.38**	**5.29 ± 0.23**	**3.52 ± 0.12**

2. Contextual MAB built starting from classification dataset (Dudík et al., 2011)
   - Comparison with Doubly-Robust and Direct Method
   - CDF of the absolute error normalized by IS



Off-Policy Learning

2. Contextual MAB built starting from classification dataset (Dudík et al., 2011)
   - Boltzmann policy
   - Gradient-ascent learning regularized with $I_2(P\|Q)$



References

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1097–1104. Omnipress, 2011.

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

I. Kuzborskij, C. Vernade, A. György, and C. Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 640–648. PMLR, 2021.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 817–824. Curran Associates, Inc., 2007.

A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 5447–5459, 2018.

A. B. Owen. *Monte Carlo theory, methods and examples*. 2013.

Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 9167–9176. PMLR, 2020.