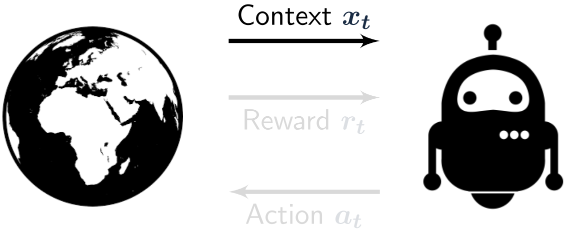
Subgaussian and Differentiable Importance Sampling for Off-Policy Evaluation and Learning

Alberto Maria Metelli   Alessio Russo   Marcello Restelli

December 2021
Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)

Contextual Multi-Armed Bandits



- Environment samples a **context**

$$x_t \sim \rho$$

- Agent plays an **action**

$$a_t \sim \pi(\cdot|x_t)$$

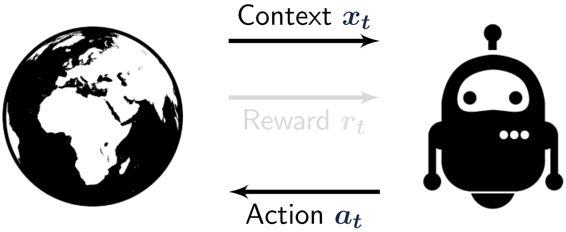
- Environment generates a **reward**

$$r_t = r(x_t, a_t)$$

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

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

Contextual Multi-Armed Bandits



- Environment samples a **context**

$$x_t \sim \rho$$

- Agent plays an **action**

$$a_t \sim \pi(\cdot|x_t)$$

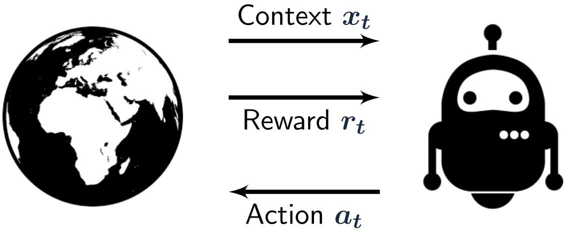
- Environment generates a **reward**

$$r_t = r(x_t, a_t)$$

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

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

Contextual Multi-Armed Bandits

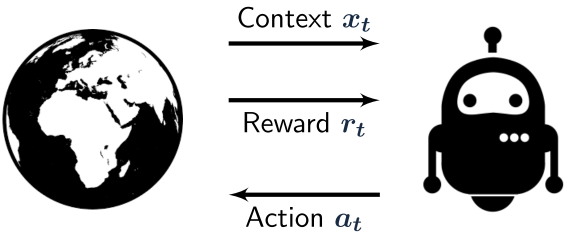


- Environment samples a **context**
$$x_t \sim \rho$$
- Agent plays an **action**
$$a_t \sim \pi(\cdot|x_t)$$
- Environment generates a **reward**
$$r_t = r(x_t, a_t)$$

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

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

Contextual Multi-Armed 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: find a policy π^* maximizing the **expected reward** (Langford and Zhang, 2007)

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

Off-Policy Evaluation and Learning

Input: $\mathcal{D} = \{(x_t, a_t, r_t)\}_{t \in [n]}$ samples collected with a **behavioral policy** π_b

Off-Policy Evaluation (Off-PE)
evaluate a given **target** policy π_e

Off-Policy Learning (Off-PL)
learn an **optimal** policy π_e

How to estimate the expected reward under π_e having samples collected with π_b?

Off-Policy Evaluation and Learning

Input: $\mathcal{D} = \{(x_t, a_t, r_t)\}_{t \in [n]}$ samples collected with a **behavioral policy** π_b

Off-Policy Evaluation (Off-PE)
evaluate a given **target** policy π_e

Off-Policy Learning (Off-PL)
learn an **optimal** policy π_e

How to estimate the expected reward under π_e having samples collected with π_b?

Input: $\mathcal{D} = \{(x_t, a_t, r_t)\}_{t \in [n]}$ samples collected with a **behavioral policy** π_b

Off-Policy Evaluation (Off-PE)
evaluate a given **target** policy π_e

Off-Policy Learning (Off-PL)
learn an **optimal** policy π_e

How to estimate the expected reward under π_e having samples collected with π_b?

Input: $\mathcal{D} = \{(x_t, a_t, r_t)\}_{t \in [n]}$ samples collected with a **behavioral policy** π_b

Off-Policy Evaluation (Off-PE)
evaluate a given **target** policy π_e

Off-Policy Learning (Off-PL)
learn an **optimal** policy π_e

How to estimate the expected reward under π_e having samples collected with π_b?

Vanilla Importance Sampling (IS)

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

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

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

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

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

Vanilla Importance Sampling (IS)

- **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**: $\mathbb{E}_{y_i \overset{\text{iid}}{\sim} Q}[\widehat{\mu}_n] = \mathbb{E}_{y \sim P}[f(y)] = \mu$
- **Variance**: can be very large! (Metelli et al., 2018)

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

Vanilla Importance Sampling (IS)

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

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

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

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

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

Vanilla Importance Sampling (IS)

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

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

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

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

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

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!

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

How to cope with this behavior?

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!

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

How to cope with this behavior?

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!

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

How to cope with this behavior?

Importance Sampling Corrections

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

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

- Importance Sampling **with TRuncation** (**IS-TR**, Ionides, 2008; Papini et al., 2019)

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

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

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

Importance Sampling Corrections

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

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

- Importance Sampling **with TRuncation** (**IS-TR**, Ionides, 2008; Papini et al., 2019)

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

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

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

Importance Sampling Corrections

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

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

- Importance Sampling **with TRuncation** (**IS-TR**, Ionides, 2008; Papini et al., 2019)

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

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

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

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	$B^{\mathsf{SN}} + \sqrt{V^{\mathsf{ES}} \log \frac{1}{\delta}}$	☹ (exp)	☺	☺
IS-TR	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	☺	☹	☹
IS-OS	$\max_{\beta \in \{2,3\}} \sqrt[\beta]{\frac{I_\beta(P\|Q)\left(\log \frac{1}{\delta}\right)^{\beta-1}}{n^{\beta-1}}}$	☹ (exp)	☹	☺

Goal: design an estimator that fulfills **all the three** properties!

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	$\max_{\beta \in \{2,3\}} \sqrt[\beta]{\frac{I_\beta(P\|Q)\left(\log \frac{1}{\delta}\right)^{\beta-1}}{n^{\beta-1}}}$	☹ (exp)	☹	☺

Goal: design an estimator that fulfills **all the three** properties!

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	$B^{\mathsf{SN}} + \sqrt{V^{\mathsf{ES}} \log \frac{1}{\delta}}$	☹ (exp)	☺	☺
IS-TR	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	☺	☹	☹
IS-OS	$\max_{\beta \in \{2,3\}} \sqrt[\beta]{\frac{I_\beta(P\|Q)\left(\log \frac{1}{\delta}\right)^{\beta-1}}{n^{\beta-1}}}$	☹ (exp)	☹	☺

Goal: design an estimator that fulfills **all the three** properties!

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	$B^{\mathsf{SN}} + \sqrt{V^{\mathsf{ES}} \log \frac{1}{\delta}}$	☹ (exp)	☺	☺
IS-TR	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	☺	☹	☹
IS-OS	$\max_{\beta \in \{2,3\}} \sqrt[\beta]{\frac{I_\beta(P\|Q)\left(\log \frac{1}{\delta}\right)^{\beta-1}}{n^{\beta-1}}}$	☹ (exp)	☹	☺

Goal: design an estimator that fulfills **all the three** properties!

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	$B^{\text{SN}} + \sqrt{V^{\text{ES}} \log \frac{1}{\delta}}$	☹ (exp)	☺	☺
IS-TR	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	☺	☹	☹
IS-OS	$\max\limits_{\beta \in \{2,3\}} \sqrt[\beta]{\frac{I_\beta(P\|Q)\left(\log \frac{1}{\delta}\right)^{\beta-1}}{n^{\beta-1}}}$	☹ (exp)	☹	☺

Goal: design an estimator that fulfills **all the three** properties!

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

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

☺ **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 $s = -1$

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 $s = -1$

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 $s = -1$

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 $s = -1$

- **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 $s = -1$

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

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

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

$$\widehat{\mu}_{n,\lambda^*_2} - \mu \leqslant \|f\|_\infty (2+\sqrt{3}) \sqrt{\frac{2 I_\alpha(P\|Q) \log \frac{1}{\delta}}{3n}} \quad \text{w.p. } 1-\delta$$

- Method to compute λ^*_2 without knowledge of $I_\alpha(P\|Q)$ in the paper

Concentration Inequalities

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

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

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

$$\widehat{\mu}_{n,\lambda^*_2} - \mu \leqslant \|f\|_\infty (2+\sqrt{3}) \sqrt{\frac{2 I_\alpha(P\|Q) \log \frac{1}{\delta}}{3n}} \quad \text{w.p. } 1-\delta$$

- Method to compute λ^*_2 without knowledge of $I_\alpha(P\|Q)$ in the paper

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

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

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

$$\widehat{\mu}_{n,\lambda^*_2} - \mu \leqslant \|f\|_\infty (2+\sqrt{3}) \sqrt{\frac{2 I_\alpha(P\|Q) \log \frac{1}{\delta}}{3n}} \quad \text{w.p. } 1-\delta$$

- 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_θ

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

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

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

Differentiability

- When the **target** distribution is parametric and differentiable P_{θ}

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

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

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

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	$B^{\mathsf{SN}} + \sqrt{V^{\mathsf{ES}} \log \frac{1}{\delta}}$	☹ (exp)	☺	☺
IS-TR	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	☺	☹	☹
IS-OS	$\max_{\beta \in \{2,3\}} \sqrt[\beta]{\frac{I_\beta(P\|Q)\left(\log \frac{1}{\delta}\right)^{\beta-1}}{n^{\beta-1}}}$	☹ (exp)	☹	☺
IS-λ	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	☺	☺	☺

Off-Policy Evaluation Experiments

- Synthetic experiment with Gaussian distributions
- $I_2(P\|Q) \simeq 27.9$ and $f(y) = 100\cos(2\pi y)$

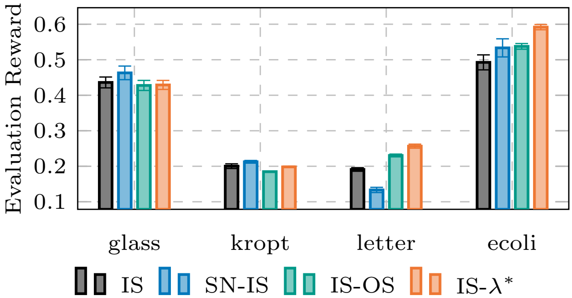
(best in **bold** and second best <u>underlined</u>)

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**

- Other experiments in contextual MABs in the paper

Off-Policy Learning Experiments

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



Discussion and Conclusions

Contributions

- **Anti-concentration** bound proving that vanilla IS has **polynomial** concentration
- **First** importance sampling correction that ensures:
  - **Subgaussian** concentration
  - **Differentiability** in the target distribution
- Experimental evaluation showing promising results

Future Works

- Study different values of s
- Extend to **reinforcement learning**

Contributions

- **Anti-concentration** bound proving that vanilla IS has **polynomial** concentration
- **First** importance sampling correction that ensures:
  - **Subgaussian** concentration
  - **Differentiability** in the target distribution
- Experimental evaluation showing promising results

Future Works

- Study different values of s
- Extend to **reinforcement learning**

Discussion and Conclusions

Contributions

- **Anti-concentration** bound proving that vanilla IS has **polynomial** concentration
- **First** importance sampling correction that ensures:
  - **Subgaussian** concentration
  - **Differentiability** in the target distribution
- Experimental evaluation showing promising results

Future Works

- Study different values of s
- Extend to **reinforcement learning**

Discussion and Conclusions

Contributions

- **Anti-concentration** bound proving that vanilla IS has **polynomial** concentration
- **First** importance sampling correction that ensures:
  - **Subgaussian** concentration
  - **Differentiability** in the target distribution
- Experimental evaluation showing promising results

Future Works

- Study different values of s
- Extend to **reinforcement learning**

Discussion and Conclusions

Contributions

- **Anti-concentration** bound proving that vanilla IS has **polynomial** concentration
- **First** importance sampling correction that ensures:
  - **Subgaussian** concentration
  - **Differentiability** in the target distribution
- Experimental evaluation showing promising results

Future Works

- Study different values of s
- Extend to **reinforcement learning**

Discussion and Conclusions

Contributions

- **Anti-concentration** bound proving that vanilla IS has **polynomial** concentration
- **First** importance sampling correction that ensures:
  - **Subgaussian** concentration
  - **Differentiability** in the target distribution
- Experimental evaluation showing promising results

Future Works

- Study different values of s
- Extend to **reinforcement learning**

Discussion and Conclusions

Contributions

- **Anti-concentration** bound proving that vanilla IS has **polynomial** concentration
- **First** importance sampling correction that ensures:
  - **Subgaussian** concentration
  - **Differentiability** in the target distribution
- Experimental evaluation showing promising results

Future Works

- Study different values of s
- Extend to **reinforcement learning**

Thank You for Your Attention!

Code: `github.com/albertometelli/subgaussian-is`

Contact: `albertomaria.metelli@polimi.it`

References

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 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. 130:640–648, 2021.

J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 817–824. Citeseer, 2007.

A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5447–5459, 2018.

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

M. Papini, A. M. Metelli, L. Lupo, and M. Restelli. Optimistic policy optimization via multiple importance sampling. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4989–4999. PMLR, 2019.

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 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9167–9176. PMLR, 2020.