POLITECNICO MILANO 1863

POIS

Policy Optimization via Importance Sampling

Alberto M. Metelli, Matteo Papini, Francesco Faccio and Marcello Restelli

{albertomaria.metelli, matteo.papini, marcello.restelli}@polimi.it, francesco.faccio@mail.polimi.it

Problem and Motivation

- **Reinforcement Learning (RL):** find optimal policy π^*
- **Policy Search**: search over a class of policies π
  - Every policy induces a distribution $p(\cdot|\pi)$ over **trajectories** τ of the Markov Decision Process (MDP)
  - Every trajectory τ has a **return** $R(\tau)$
- **Goal:** find π^* maximizing $J(\pi)$

$$J(\pi) = \mathop{\mathbb{E}}_{\tau \sim p(\cdot|\pi)}[R(\tau)]$$

- **Using data collected with some policy** π:
  - How can I evaluate proposals $\pi' \neq \pi$?
  - How can I trust counterfactual evaluations?
  - How can I best use my data for optimization?

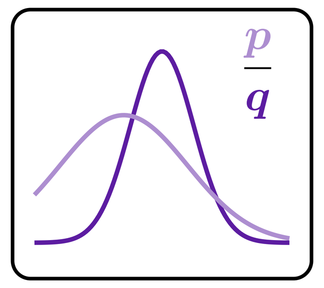
Importance Sampling

How can I evaluate proposals?
With **Importance Sampling (IS)**

- Given a **behavioral** (data-sampling) **distribution** $q(x)$, a **target distribution** $p(x)$, and a function $f(x)$, **estimate**

$$\boldsymbol{\mu} = \mathop{\mathbb{E}}_{x \sim p}[f(x)] \qquad \text{with data from } q, \quad x_i \sim q$$

$$\widehat{\mu}_{\text{IS}} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\frac{p(x_i)}{q(x_i)}}_{w(x_i)} f(x_i)$$



- $w(x) = p(x)/q(x)$ is the **importance weight**
- The estimate is **unbiased**: $\mathbb{E}_q[\widehat{\mu}_{\text{IS}}] = \mu \ldots$
- … **but the variance can be very high!**
- **Rényi divergence**: dissimilarity between p and q:

$$D_2(p\|q) = \log \mathop{\mathbb{E}}_{x\sim q}\left[\left(\frac{p(x)}{q(x)}\right)^2\right] \qquad \underbrace{d_2(p\|q) = \exp\{D_2(p\|q)\}}_{\text{exponentiated Rényi}}$$

- Variance of the weight depends **exponentially** on the distributional divergence (?)

$$\mathbb{V}\text{ar}[\boldsymbol{w}] = d_2(p\|q) - 1$$

- **Effective Sample Size (ESS)**: number of equivalent samples in plain Monte Carlo estimation ($x_i \sim p$)

$$\text{ESS} = \frac{N}{d_2(p\|q)} \approx \frac{\|\boldsymbol{w}\|_1^2}{\|\boldsymbol{w}\|_2^2} = \widehat{ESS}$$

- Variance of the estimator $\widehat{\mu}_{\text{IS}}$ depends **exponentially** on the distributional divergence as well

$$\mathbb{V}\text{ar}[\widehat{\mu}_{\text{IS}}] \leq \frac{1}{N}\|f\|_\infty^2 d_2(p\|q)$$

Off-Distribution Learning

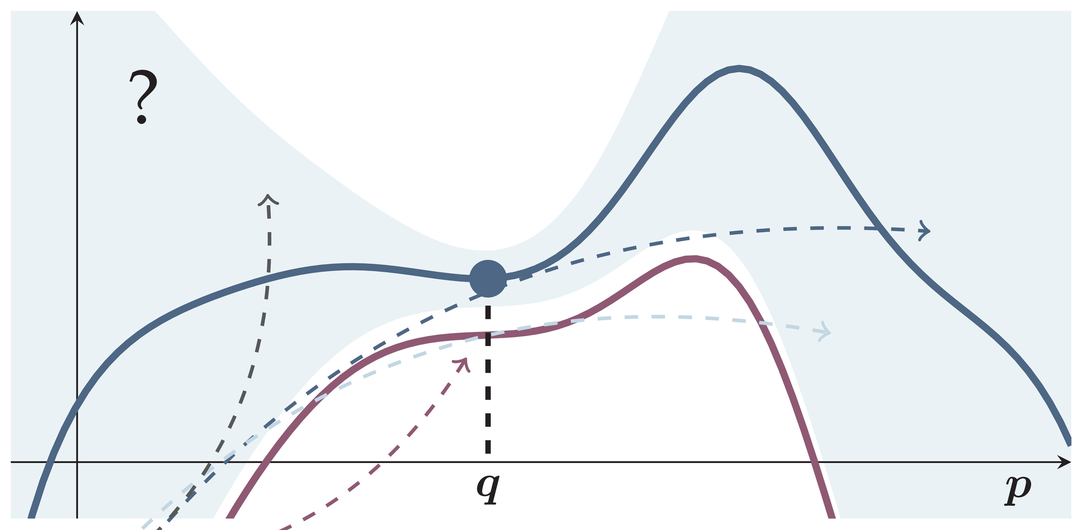
How (far) can I trust conterfactual evaluations?

- Evaluate only close solutions: REPS (?), TRPO (?)
- **Use a lower bound**: EM (??), PPO (?), **POIS**

Given a behavioral $q(x)$, a function $f(x)$ and a proposal $p(x)$, with probability at least $1-\delta$:

True function ≥ Lower Bound = IS Estimator − Variance Bound (Cantelli)

$$\mathop{\mathbb{E}}_{x\sim p}[f(x)] \geq \mathcal{L}_\delta^{\text{POIS}}(p/q) = \frac{1}{N}\sum_{i=1}^{N} w(x_i) f(x_i) - \|f\|_\infty \sqrt{\frac{(1-\delta)d_2(p\|q)}{\delta N}}$$



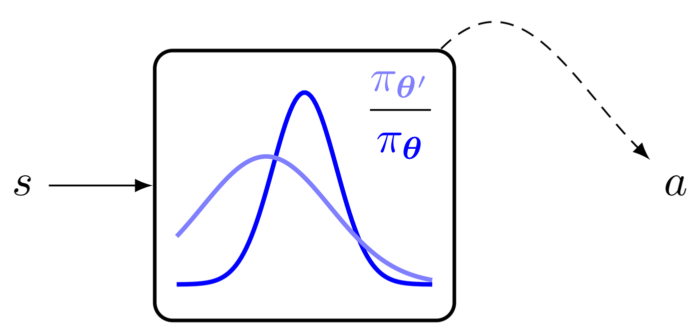
How can I best use my data for optimization?
Given the behavioral q, find p **maximizing** $\mathbb{E}_{x\sim p}[f(x)]$:

1. Collect data with q (expensive in RL)
2. Find p maximizing $\mathcal{L}_\delta^{\text{POIS}}(p/q)$ (offline optimization)
3. Set new behavioral $q \leftarrow p$
4. Repeat until convergence

Action-Based POIS

- Find the **policy** parameters $\boldsymbol{\theta}^*$ that maximize $J(\boldsymbol{\theta}')$ (??)

$$J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\boldsymbol{\theta})}[R(\tau)]$$



- Given a **behavioral policy** $\pi_{\boldsymbol{\theta}}$ we compute a target policy π_{θ} by optimizing:

$$\mathcal{L}_\lambda^{\text{A-POIS}}(\theta'/\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{t=0}^{H-1}\frac{\pi_{\theta'}(a_{\tau_i,t}|s_{\tau_i,t})}{\pi_{\boldsymbol{\theta}}(a_{\tau_i,t}|s_{\tau_i,t})}R(\tau_i) - \lambda\sqrt{\frac{\widehat{d}_2\left(p(\cdot|\theta')\|p(\cdot|\boldsymbol{\theta})\right)}{N}}$$

- The term $d_2\left(p(\cdot|\theta')\|p(\cdot|\boldsymbol{\theta})\right)$ is estimated from samples
- The d_2 grows exponentially with the task horizon H
- λ is a regularization hyperparameter

$$\lambda = \frac{R_{\max}}{1-\gamma}\sqrt{\frac{1-\delta}{\delta}}$$

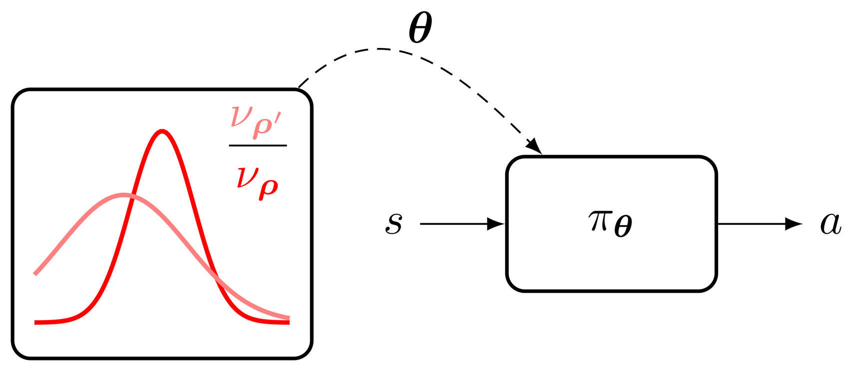
- We consider diagonal Gaussian policies $\pi_{\boldsymbol{\theta}}$

$$a \sim \pi_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\cdot|s) = \mathcal{N}\left(u_{\boldsymbol{\mu}}(s), \text{diag}(\boldsymbol{\sigma}^2)\right)$$

Parameter-Based POIS

- Find the **hyperpolicy** parameters $\boldsymbol{\rho}^*$ that maximize $J(\boldsymbol{\rho})$ (?)

$$J(\boldsymbol{\rho}) = \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim\nu_{\boldsymbol{\rho}}}\mathop{\mathbb{E}}_{\tau\sim p(\cdot|\boldsymbol{\theta})}[R(\tau)]$$



- Given a **behavioral hyperpolicy** $\nu_{\boldsymbol{\rho}}$ we compute a target hyperpolicy $\nu_{\rho'}$ by optimizing:

$$\mathcal{L}_\lambda^{\text{P-POIS}}(\rho'/\boldsymbol{\rho}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\nu_{\rho'}(\boldsymbol{\theta}_i)}{\nu_{\boldsymbol{\rho}}(\boldsymbol{\theta}_i)}R(\tau_i) - \lambda\sqrt{\frac{d_2\left(\nu_{\rho'}\|\nu_{\boldsymbol{\rho}}\right)}{N}}$$

- The term $d_2\left(\nu_{\rho'}\|\nu_{\boldsymbol{\rho}}\right)$ can be computed exactly
- Affected by the parameter space dimension $\dim(\boldsymbol{\theta})$
- λ is a regularization hyperparameter

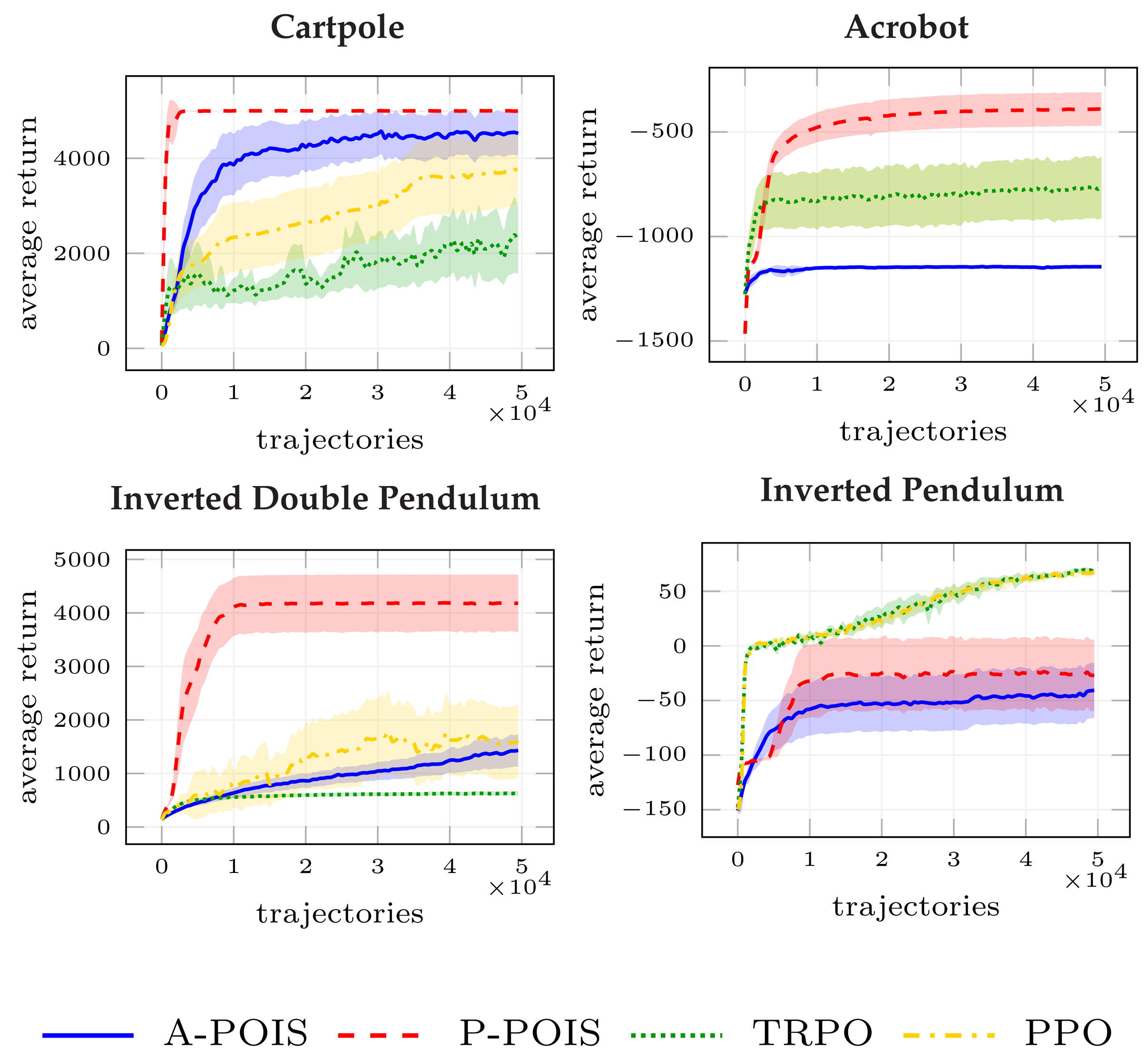
$$\lambda = \frac{R_{\max}}{1-\gamma}\sqrt{\frac{1-\delta}{\delta}}$$

- We consider diagonal Gaussian hyperpolicies $\nu_{\boldsymbol{\rho}}$

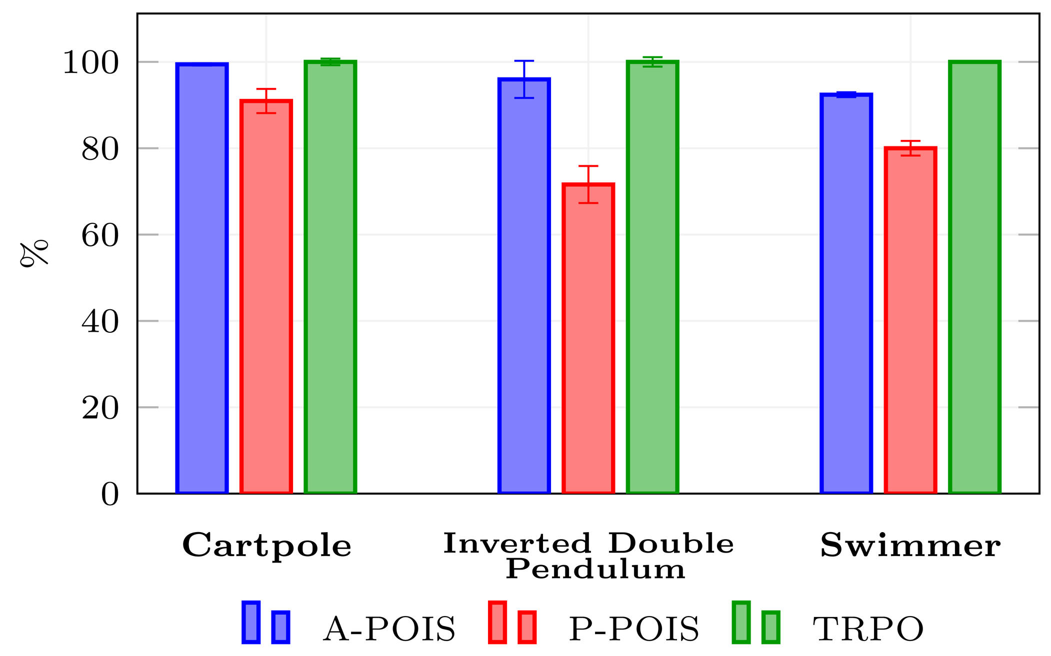
$$\boldsymbol{\theta} \sim \nu_{\boldsymbol{\mu},\boldsymbol{\sigma}} = \mathcal{N}\left(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)\right)$$

Experiments

Linear Policies



Deep Policies



Algorithm Details

- **Self-normalized (SN)** importance sampling (?)

$$\widetilde{\mu}_{\text{SN}} = \frac{\sum_{i=1}^{N} w(x_i) f(x_i)}{\sum_{i=1}^{N} w(x_i)} \qquad x_i \sim q$$

- ESS instead of d_2 as penalization
- Gradient optimization of $\mathcal{L}^{\star-\text{POIS}}$ using *line search*
- Natural gradient for P-POIS

References