**Problem and Motivation**
- Reinforcement Learning (RL): find optimal policy \( \pi^* \)
- Policy Search: search over a class of policies \( \pi \)
  - Every policy induces a distribution \( \pi(\tau) \) over trajectories \( \tau \) of the Markov Decision Process (MDP)
  - Every trajectory \( \tau \) has a return \( R(\tau) \)
- Goal: find \( \pi^* \) maximizing \( J(\pi) = \mathbb{E}_{\tau \sim \pi}[R(\tau)] \)
- Using data collected with some policy \( \pi \):
  - 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 \( q(\tau) \), a target distribution \( p(\tau) \), estimate
\[
\hat{\pi}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{p(\tau_i)}{q(\tau_i)} f(\tau_i)
\]
- \( w(x) = \frac{p(x)}{q(x)} \) is the importance weight
- The estimate is unbiased: \( \mathbb{E}[\hat{\pi}_{IS}] = \mu \ldots \) but the variance can be very high!
- Rényi divergence: dissimilarity between \( p \) and \( q \):
\[
D_\alpha(p||q) = \log \mathbb{E} \left[ \left( \frac{p(x)}{q(x)} \right)^\alpha \right] = D_2(p||q) = \exp(D_2(p||q))
\]
- Variance of the weight depends exponentially on the distributional divergence (Cortes et al., 2010)
\[
\text{Var}[w] = d_2(p||q) - 1
\]
- Effective Sample Size (ESS): number of equivalent samples in plain Monte Carlo estimation (\( x_i \sim p \))
\[
\text{ESS} = \frac{N}{d_2(p||q)} = \frac{\sum_i w_i^2}{\sum_i w_i^2} = \text{ESS}
\]
- Variance of the estimator \( \hat{\pi}_{IS} \) depends exponentially on the distributional divergence as well
\[
\text{Var}[\hat{\pi}_{IS}] \leq \frac{1}{N} \sum_i^N d_2(p||q)
\]

**Off-Distribution Learning**

How far can I trust counterfactual evaluations?
- Evaluate only close solutions: REPS (Peters et al., 2010), TRPO (Schulman et al., 2015)
- Use a lower bound: EM (Dayan and Hinton, 1997; Kobert et al., 2011), PPO (Schulman et al., 2017), POIS

Given a behavioral \( q(\tau) \), a function \( f(\tau) \) and a proposal \( p(\tau) \), with probability at least \( 1 - \delta \):
\[
\mathbb{E}_{\tau \sim q}[f(\tau)] \geq \mathbb{E}_{\tau \sim p}[f(\tau)] - \frac{1}{N} \sum_{i=1}^{N} w(x_i)f(x_i) - \left\| f \right\|_\infty \sqrt{(1 - \delta)d_2(p||q)} \frac{1}{\delta N}
\]

How can I best use my data for optimization?
- Given the behavioral \( q \), find \( p \) maximizing \( \mathbb{E}_{\tau \sim p}[f(\tau)] \):
  1. Collect data with \( q \) (expensive in RL)
  2. Find \( p \) maximizing \( \mathcal{L}_{\lambda}^{\text{POIS}}(p||q) \) (offline optimization)
  3. Set new behavioral \( q \leftarrow p \)
  4. Repeat until convergence

**Action-Based POIS**
- Find the policy parameters \( \theta' \) that maximize \( J(\theta') \) (Sutton et al., 2000; Peters and Schaal, 2008)

\[
J(\theta) = \mathbb{E}_{\tau \sim q}[R(\tau)]
\]

- Given a behavioral policy \( \tau_q \) we compute a target policy \( \tau_p \) by optimizing:
\[
\mathcal{L}_{\lambda}^{\text{POIS}}(\theta' || \theta) = \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \tau_p(x_t, a_t, s_t) \mathbb{E}[R(t)]
- \lambda \left( \frac{d_p\left(\mu(\tau' || \theta)\right)}{d_2\left(q(\tau || \theta)\right)} \right)\left( \frac{N}{\sum_{i=1}^{N} w_i} \right)
\]
- The term \( d_p(\mu(\tau' || \theta)) \) is estimated from samples
- The \( d_2 \) grows exponentially with the task horizon \( H \)
- \( \lambda \) is a regularization hyperparameter
- \( \lambda = \frac{R_{\text{max}}}{1 - \gamma} \sqrt{\frac{1}{1 - \delta}} \)
- We consider diagonal Gaussian policies \( \tau_p \)
\[
\alpha \sim \mathcal{N}_\mu(\sigma) = N(\mu, \sigma) \text{diag}(\sigma^2)
\]

**Parameter-Based POIS**
- Find the hyperpolicy parameters \( \rho' \) that maximize \( J(\rho) \) (Sehnke et al., 2008)
\[
J(\rho) = \mathbb{E}_{\theta \sim p}[\mathbb{E}_{\tau \sim q(\theta)}[R(\tau)]]
\]

- Given a behavioral hyperpolicy \( \tau_p \) we compute a target hyperpolicy \( \tau_p \) by optimizing:
\[
\mathcal{L}_{\lambda}^{\text{POIS}}(\rho' || \rho) = \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \tau_p(x_t, a_t, s_t) \mathbb{E}[R(t)]
- \lambda \left( \frac{d_p\left(\mu(\tau' || \rho)\right)}{d_2\left(q(\tau || \rho)\right)} \right)\left( \frac{N}{\sum_{i=1}^{N} w_i} \right)
\]
- The term \( d_p(\mu(\tau' || \rho)) \) can be computed exactly
- Affected by the parameter space dimension \( \text{dim}(\theta) \)
- \( \lambda \) is a regularization hyperparameter
- \( \lambda = \frac{R_{\text{max}}}{1 - \gamma} \sqrt{\frac{1}{1 - \delta}} \)
- We consider diagonal Gaussian hyperpolicies \( \tau_p \)
\[
\theta \sim \mathcal{N}_\mu(\sigma) = N(\mu, \sigma) \text{diag}(\sigma^2)
\]

**Experiments**

- Linear Policies
- Cartpole
- Acrobot
- Inverted Double Pendulum
- Swimmer

**Algorithm Details**
- Self-normalized (SN) importance sampling (Owen, 2013)
- ESS instead of \( d_2 \) as penalization
- Gradient optimization of \( \mathcal{L}_{\lambda}^{\text{POIS}} \) using line search
- Natural gradient for P-POIS

**REFERENCES**