Motivation

**Policy Search (PS):** very effective RL technique for *continuous control tasks*

- High *sample complexity* remains a major limitation

[Heess et al., 2017]  [OpenAI, 2018]  [Vinyals et al., 2017]
**Motivation**

**Policy Search (PS):** very effective RL technique for continuous control tasks

- High sample complexity remains a major limitation
- Samples available from several sources are discarded
  - Different policies
  - Different environments
Motivation

Policy Search (PS): very effective RL technique for continuous control tasks

- High sample complexity remains a major limitation
- Samples available from several sources are discarded
  - Different policies
  - Different environments

Transfer of Samples

Tirinzoni et al. Transfer of Samples in Policy Search via Multiple Importance Sampling ICML 2019
Transfer of Samples

Source Task $\mathcal{M}_1$

$\tau_{i,1} \sim \pi_{\theta_1}, P_1$

Source Task $\mathcal{M}_2$

$\tau_{i,2} \sim \pi_{\theta_2}, P_2$

Target Task $\mathcal{M}$

$\pi_{\theta}, P$

Source Task $\mathcal{M}_m$

$\tau_{i,m} \sim \pi_{\theta_m}, P_m$


Extension to online PS algorithms not trivial

Tirinzoni et al. Transfer of Samples in Policy Search via Multiple Importance Sampling ICML 2019
Transfer of Samples

- Extension to online PS algorithms **not trivial**
Goal: Transfer source trajectories to improve the target gradient estimation

Multiple Importance Sampling (MIS) Gradient Estimator

\[
\nabla_{\theta}^{\text{MIS}} J(\theta) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} w(\tau_{i,j}) g_{\theta}(\tau_{i,j})
\]

\[
W(\tau) := \frac{p(\tau|\theta, P)}{\sum_{j=1}^{m} \alpha_j p(\tau|\theta_j, P_j)}
\]

Unbiased and bounded weights
Easily combined with other variance reduction techniques
Effective sample size $\equiv \text{Transferable knowledge}$
Adaptive batch size
Provably robust to negative transfer

Tirinzoni et al. Transfer of Samples in Policy Search via Multiple Importance Sampling ICML 2019
**Goal**: Transfer source trajectories to improve the target gradient estimation

### Multiple Importance Sampling (MIS) Gradient Estimator

\[
\nabla_{\theta}^{\text{MIS}} J(\theta) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} w(\tau_{i,j}) g_{\theta}(\tau_{i,j}) \\
\text{weights} \quad \text{gradient} \\
w(\tau) := \frac{p(\tau|\theta, \mathcal{P})}{\sum_{j=1}^{m} \alpha_j p(\tau|\theta_j, \mathcal{P}_j)}
\]

- Unbiased and **bounded** weights
Transferring Samples in Policy Search

**Goal:** Transfer source trajectories to improve the target gradient estimation

Multiple Importance Sampling (MIS) Gradient Estimator

$$\nabla_{\theta}^{\text{MIS}} J(\theta) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left( w(\tau_{i,j}) g_{\theta}(\tau_{i,j}) \right)$$

$$w(\tau) := \frac{p(\tau|\theta, P)}{\sum_{j=1}^{m} \alpha_j p(\tau|\theta_j, P_j)}$$

- Unbiased and **bounded** weights
- Easily combined with other **variance reduction** techniques
**Goal:** Transfer source trajectories to improve the target *gradient estimation*

Multiple Importance Sampling (MIS) Gradient Estimator

\[
\nabla_{\theta}^{\text{MIS}} J(\theta) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} w(\tau_{i,j}) g_{\theta}(\tau_{i,j})
\]

- **Unbiased and bounded weights**
- Easily combined with other *variance reduction* techniques
- **Effective sample size** \(\equiv\) Transferable knowledge \(\rightarrow\) **Adaptive batch size**
**Goal:** Transfer source trajectories to improve the target gradient estimation

Multiple Importance Sampling (MIS) Gradient Estimator

\[
\nabla_{\theta}^{\text{MIS}} J(\theta) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} w(\tau_{i,j}) g_\theta(\tau_{i,j}) \quad w(\tau) := \frac{p(\tau|\theta, P)}{\sum_{j=1}^{m} \alpha_j p(\tau|\theta_j, P_j)}
\]

- Unbiased and **bounded weights**
- Easily combined with other **variance reduction** techniques
- **Effective sample size** $\equiv$ Transferable knowledge $\rightarrow$ **Adaptive batch size**
- Provably **robust to negative transfer**
Problem: $\mathcal{P}$ unknown $\rightarrow$ Importance weights cannot be computed
Solution: Online minimization of an upper-bound to the expected MSE of $\nabla_{\theta}^{\text{MIS}} J(\theta)$
**Problem:** $\mathcal{P}$ unknown $\rightarrow$ Importance weights cannot be computed

**Solution:** Online minimization of an upper-bound to the expected MSE of $\nabla_{\theta}^{\text{MIS}} J(\theta)$

- Obtain principled estimates even **without target samples**
Estimating the Transition Models

**Problem**: $\mathcal{P}$ unknown $\rightarrow$ Importance weights cannot be computed

**Solution**: Online minimization of an upper-bound to the expected MSE of $\nabla_{\theta}^{\text{MIS}} J(\theta)$

- Obtain principled estimates even **without target samples**
- Can be **efficiently optimized** for
  - Discrete set of models
  - **Reproducing Kernel Hilbert Spaces** (RKHS) $\rightarrow$ Closed-form solution
Empirical Results

- **Good performance** with both known and unknown models
- Very effective **sample reuse** from different policies but *same* environment
Thank you!

andrea.tirinzoni@polimi.it

https://github.com/AndreaTirinzoni/

Meet us at poster #118 @ Pacific Ballroom


Tirinzoni et al.

**TRANSFER OF SAMPLES IN POLICY SEARCH VIA MULTIPLE IMPORTANCE SAMPLING**

*ICML 2019*
OpenAI (2018).
Learning dexterous in-hand manipulation.
*CoRR*, abs/1808.00177.

Eligibility traces for off-policy policy evaluation.
*Computer Science Department Faculty Publication Series*, page 80.

Transferring instances for model-based reinforcement learning.

Importance weighted transfer of samples in reinforcement learning.