Transfer of Samples in Policy Search via Multiple Importance Sampling

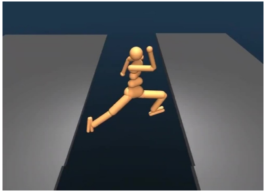
Andrea Tirinzoni, Mattia Salvini, and Marcello Restelli

36th International Conference on Machine Learning, Long Beach, California

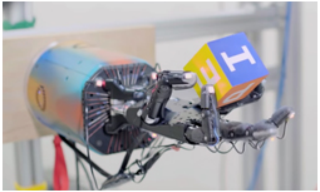
POLITECNICO
MILANO 1863

Motivation

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



[Heess et al., 2017]



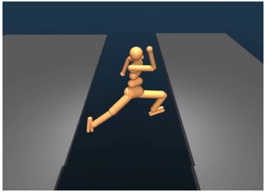
[OpenAI, 2018]

[Vinyals et al., 2017]

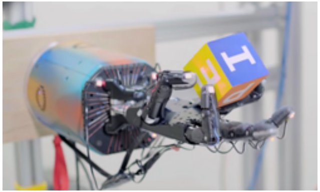
- High **sample complexity** remains a major limitation

Motivation

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



[Heess et al., 2017]



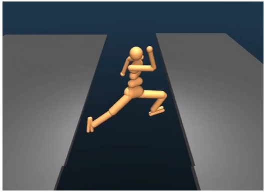
[OpenAI, 2018]

[Vinyals et al., 2017]

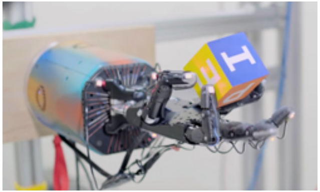
- High **sample complexity** remains a major limitation
- Samples available from several sources are **discarded**
  - Different policies
  - Different environments

Motivation

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



[Heess et al., 2017]



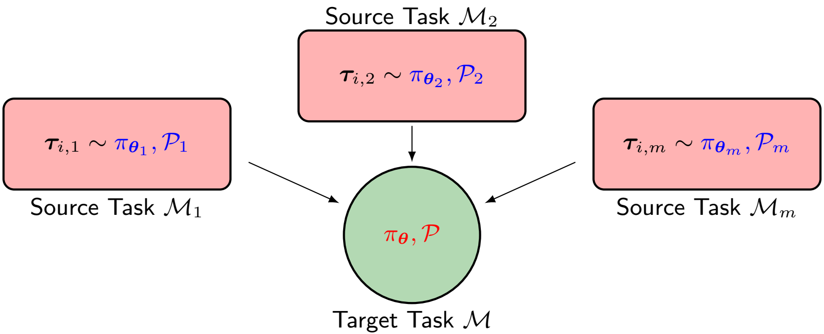
[OpenAI, 2018]

[Vinyals et al., 2017]

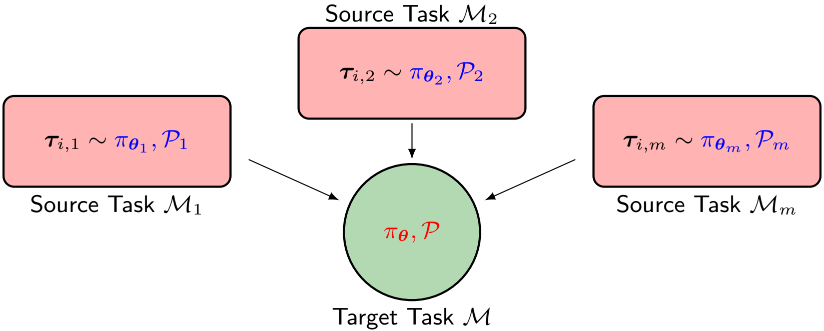
- High **sample complexity** remains a major limitation
- Samples available from several sources are **discarded**
  - Different policies
  - Different environments

Transfer of Samples

Transfer of Samples



Transfer of Samples



- Existing works: **batch value-based** settings [Lazaric et al., 2008, Taylor et al., 2008, Lazaric and Restelli, 2011, Laroche and Barlier, 2017, Tirinzoni et al., 2018]
- Extension to online PS algorithms **not trivial**

Transferring Samples in Policy Search

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

Multiple Importance Sampling (MIS) Gradient Estimator

$$\nabla_{\theta}^{\text{MIS}} J(\boldsymbol{\theta}) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \underbrace{w(\boldsymbol{\tau}_{i,j})}_{\text{weights}} \underbrace{g_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{i,j})}_{\text{gradient}} \qquad w(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, \mathcal{P})}{\sum_{j=1}^{m} \alpha_j p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \mathcal{P}_j)}$$

Transferring Samples in Policy Search

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

Multiple Importance Sampling (MIS) Gradient Estimator

$$\nabla_{\theta}^{\mathsf{MIS}} J(\boldsymbol{\theta}) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \underbrace{w(\boldsymbol{\tau}_{i,j})}_{\text{weights}} \underbrace{g_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{i,j})}_{\text{gradient}} \qquad w(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, \mathcal{P})}{\sum_{j=1}^{m} \alpha_j p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \mathcal{P}_j)}$$

- 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

$$\nabla_{\theta}^{\text{MIS}} J(\boldsymbol{\theta}) := \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} \underbrace{w(\boldsymbol{\tau}_{i,j})}_{\text{weights}}\, \underbrace{g_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{i,j})}_{\text{gradient}} \qquad w(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},\mathcal{P})}{\sum_{j=1}^{m}\alpha_j p(\boldsymbol{\tau}|\boldsymbol{\theta}_j,\mathcal{P}_j)}$$

- Unbiased and **bounded weights**
- Easily combined with other **variance reduction** techniques

Transferring Samples in Policy Search

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

Multiple Importance Sampling (MIS) Gradient Estimator

$$\nabla_\theta^{\text{MIS}} J(\boldsymbol{\theta}) := \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} \underbrace{w(\boldsymbol{\tau}_{i,j})}_{\text{weights}} \underbrace{g_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{i,j})}_{\text{gradient}} \qquad w(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},\mathcal{P})}{\sum_{j=1}^{m}\alpha_j p(\boldsymbol{\tau}|\boldsymbol{\theta}_j,\mathcal{P}_j)}$$

- Unbiased and **bounded weights**
- Easily combined with other **variance reduction** techniques
- *Effective sample size* $\equiv$ Transferable knowledge $\rightarrow$ **Adaptive batch size**

Transferring Samples in Policy Search

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

Multiple Importance Sampling (MIS) Gradient Estimator

$$\nabla_\theta^{\text{MIS}} J(\boldsymbol{\theta}) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \underbrace{w(\boldsymbol{\tau}_{i,j})}_{\text{weights}} \underbrace{g_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{i,j})}_{\text{gradient}} \qquad w(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, \mathcal{P})}{\sum_{j=1}^{m} \alpha_j p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \mathcal{P}_j)}$$

- 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**

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 $\nabla_{\theta}^{\mathsf{MIS}} J(\boldsymbol{\theta})$

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 $\nabla_{\theta}^{\mathsf{MIS}} J(\boldsymbol{\theta})$

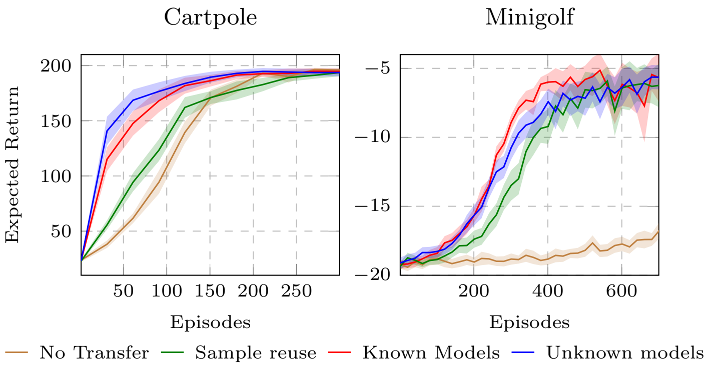
- Obtain principled estimates even **without target samples**

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 $\nabla_{\theta}^{\mathsf{MIS}} J(\boldsymbol{\theta})$

- 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

References

Hammersley, J. and Handscomb, D. (1964).
Monte Carlo Methods.
Methuen's monographs on applied probability and statistics. Methuen.

Heess, N., Sriram, S., Lemmon, J., Merel, J., Wayne, G., Tassa, Y., Erez, T., Wang, Z., Eslami, A., Riedmiller, M., et al. (2017).
Emergence of locomotion behaviours in rich environments.
arXiv preprint arXiv:1707.02286.

Laroche, R. and Barlier, M. (2017).
Transfer reinforcement learning with shared dynamics.
In *AAAI*.

Lazaric, A. and Restelli, M. (2011).
Transfer from multiple mdps.
In *Advances in Neural Information Processing Systems*.

Lazaric, A., Restelli, M., and Bonarini, A. (2008).
Transfer of samples in batch reinforcement learning.
In *Proceedings of the 25th international conference on Machine learning*.

References (cont.)

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

Precup, D. (2000).
Eligibility traces for off-policy policy evaluation.
Computer Science Department Faculty Publication Series, page 80.

Taylor, M. E., Jong, N. K., and Stone, P. (2008).
Transferring instances for model-based reinforcement learning.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 488–505. Springer.

Tirinzoni, A., Sessa, A., Pirotta, M., and Restelli, M. (2018).
Importance weighted transfer of samples in reinforcement learning.
In *International Conference on Machine Learning*, pages 4943–4952.

References (cont.)

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J., et al. (2017).
Starcraft ii: A new challenge for reinforcement learning.
arXiv preprint arXiv:1708.04782.