Importance Weighted Transfer of Samples in Reinforcement Learning

A. Tirinzoni, A. Sessa, M. Pirotta, and M. Restelli
{andrea.tirinzoni, andrea.sessa, marcello.restelli}@polimi.it, {matteo.pirotta}@inria.fr

POLITECNICO MILANO 1863

Problem

- Transfer experience samples $\langle s, a, s', r \rangle$ from a set of m **source tasks** $\tau_1, \tau_2, \dots, \tau_m$ to speed-up the learning process in a given **target task** τ_0
- Each task is an MDP $\tau_j = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_j, \mathcal{R}_j, \gamma \rangle$ with shared state-action space but different transitions and rewards
- Tasks are different → Many challenges:
  - Which samples should be transferred?
  - How should they be transferred?

Motivation

- Transfer learning → Reduce **sample complexity** of RL
- Why **transferring samples**?
  - Samples are the most **basic** pieces of information available to RL agents
  - Does not require source tasks to be **solved**
  - No dependency on the **learning algorithm**
- Limitations of most current approaches:
  - Strong assumptions on the **similarities between tasks**
  - Time-consuming sample selection process. Bad samples selected → **Negative transfer**
  - Transferred samples are used to learn the target task without considering the differences in the task models → **Asymptotic bias**

Contributions

1. We propose **Importance Weighted Fitted Q-Iteration** (IWFQI):
   - IWFQI transfers **all** source samples into a modified version of FQI
   - **Implicit** sample selection via **importance weighting**
   - IWFQI **decouples** rewards and transitions to maximize transferred information
2. We provide a **finite-sample analysis** showing the correctness of our approach
3. We **empirically** evaluate IWFQI on two synthetic tasks and a real-world domain, proving:
   - Better results than competitive methods [Lazaric et al., 2008, Laroche and Barlier, 2017]
   - **Robustness** to negative transfer

Importance Weighted Fitted Q-Iteration

Fitted Q-Iteration - Sequence of supervised learning problems:

$$Q_{k+1} = \arg\inf_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=0}^{N} |h(x_i) - y_i|^2 \qquad y_i = \widehat{T} Q_k(x_i) = r_i + \gamma \max_{a'} Q_k(s_i', a') \qquad x_i = (s_i, a_i)$$

- In transfer settings, we have *sample selection bias* → Use **Importance weighting**

Reward-Transition Decoupling

1. **Reward fitting**: use the importance weighted reward samples from all the tasks to fit a model of the target reward function.

$$\widehat{R} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_r} \sum_{j=0}^{m} \sum_{i=0}^{N_j} w_{r,i,j} \, |h(x_{i,j}) - r_{i,j}|^2$$

2. **Modified Bellman operator**: replace the reward samples in the empirical Bellman operator with the function $\widehat{R}$ fitted at step 1.

$$\widetilde{T} Q(s, a) = \widehat{R}(s, a) + \gamma \max_{a'} Q(s', a')$$

3. **Iterated Q-function fitting**: use the modified Bellman operator and the importance weighted transition samples to iteratively fit the target Q-function.

$$Q_{k+1} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_p} \sum_{j=0}^{m} \sum_{i=0}^{N_j} w_{p,i,j} \, |h(x_{i,j}) - y_{i,j}|^2$$

Algorithm

Algorithm IWFQI

Input: Number of iterations K, dataset $\mathcal{D}^+ = \{s_{i,j}, a_{i,j}, s'_{i,j}, r_{i,j}, w_{r,i,j}, w_{p,i,j}\}$, hypothesis space $\mathcal{H}$

Output: Greedy policy π_K

$\widehat{R} \leftarrow$ FIT-REWARD$(\mathcal{D}, \mathcal{H})$
$Q_0 \leftarrow \widehat{R}$
for $k = 0, \dots, K-1$ **do**
  $y_{i,j} \leftarrow \widetilde{T} Q_k(s_{i,j}, a_{i,j})$
  $Q_{k+1} \leftarrow$ FIT-Q$(\mathcal{D}, \mathcal{H}, y)$
end for
return $\pi_K(s) \leftarrow \arg\max_{a \in \mathcal{A}} Q_K(s, a)$

Importance Weights

$$w_{r,i,j} = \frac{\mathcal{R}_0(r_{i,j}|x_{i,j})}{\mathcal{R}_j(r_{i,j}|x_{i,j})} \qquad w_{p,i,j} = \frac{\mathcal{P}_0(s'_{i,j}|x_{i,j})}{\mathcal{P}_j(s'_{i,j}|x_{i,j})}$$

Problem: the task models $\mathcal{R}$ and $\mathcal{P}$ are **unknown** → The importance weights **cannot** be computed exactly
Solution: Fit **Gaussian processes** for the models $\mathcal{R}$ and $\mathcal{P}$ of each task

- Try to characterize the resulting weight distribution $\mathcal{G}$
- Gaussian models → **Closed-form** for the mean weights

$$\mathbb{E}_{\mathcal{G}}[w_r(x)] = C \frac{\mathcal{N}\left(r | \mu_{GP_0}(x), \sigma_0^2(x) + \sigma_{GP_0}^2(x)\right)}{\mathcal{N}\left(r | \mu_{GP_j}(x), \sigma_j^2(x) - \sigma_{GP_j}^2(x)\right)}$$

Theoretical Analysis

Finite-Sample Analysis of AVI [Farahmand et al., 2010]

$$\|Q^* - Q^{\pi_K}\|_{1,\rho} \le \frac{2\gamma}{(1-\gamma)^2} \left[2\gamma^K Q_{\max} + \inf_{b \in [0,1]} \sqrt{C_{\rho,\mu}(K; b) \sum_{k=0}^{K-1} \alpha_k^{2b} \boxed{\|\epsilon_k\|_\mu^2}} \right]$$

Challenges

- Importance weighted regression
- Biased estimators
- Modified Bellman operator

Error Bound for IWFQI

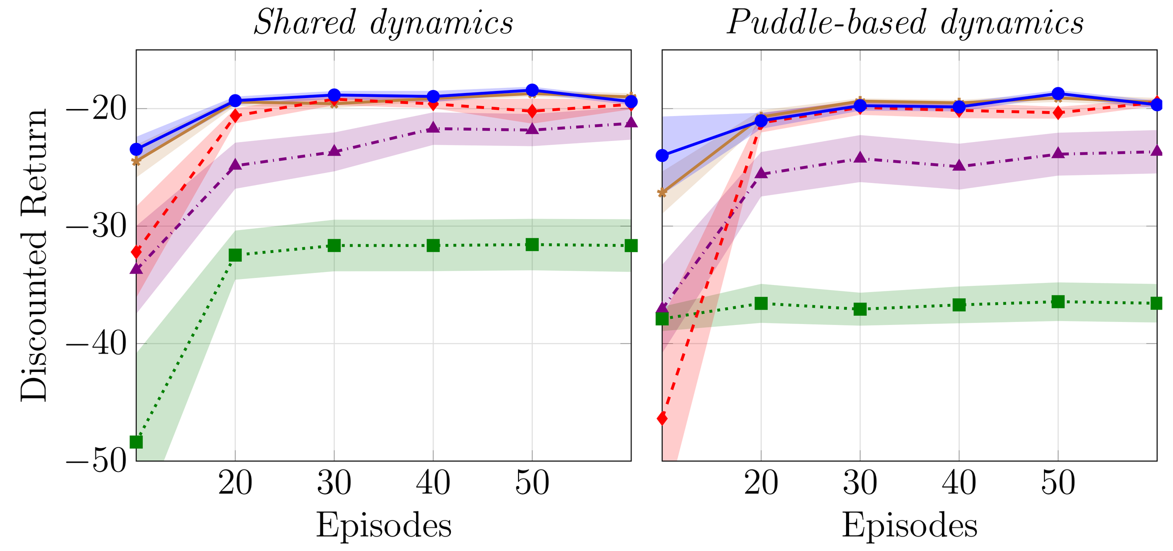
$$\begin{aligned}
\boxed{\|T^* Q_k - Q_{k+1}\|_\mu} \le\ & Q_{\max} \sqrt{\|g_p\|_{1,\mu}} + 2R_{\max} \sqrt{\|g_r\|_{1,\mu}} \\
&+ 2Q_{\max} \|\widetilde{w}_p - w_p\|_{\phi_{\mathcal{S}}^P} + 4R_{\max} \|\widetilde{w}_r - w_r\|_{\phi_{\mathcal{S}}^R} \\
&+ \inf_{f \in \mathcal{H}} \|f - (T^*)^{k+1} Q_0\|_\mu + 2 \inf_{f \in \mathcal{H}} \|f - R\|_\mu \\
&+ 2^{\frac{13}{8}} Q_{\max} \left(\sqrt{M(\widetilde{w}_p)} + 2\sqrt{M(\widetilde{w}_r)} \right) \left(\frac{d \log \frac{2Ne}{d} + \log \frac{4}{\delta}}{N} \right)^{\frac{3}{16}} \\
&+ \sum_{i=0}^{k-1} (\gamma C_{\mathrm{AE}}(\mu))^{k-i} \|T^* Q_i - Q_{i+1}\|_\mu
\end{aligned}$$

Error Decomposition

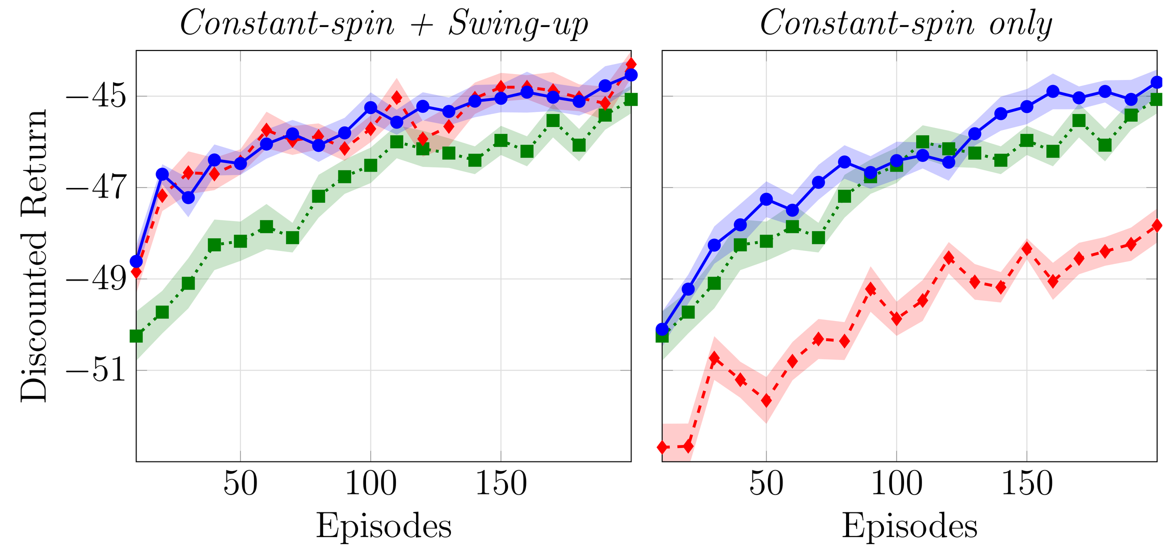
1. **Bias** due to the estimated importance weights $\widetilde{w}_p$ and $\widetilde{w}_r$
2. **Approximation** error due to the functional spaces of limited capacity
3. **Estimation** error due to the limited samples and the variance of the importance weights
4. **Propagation** error due to repeated iterations

Results

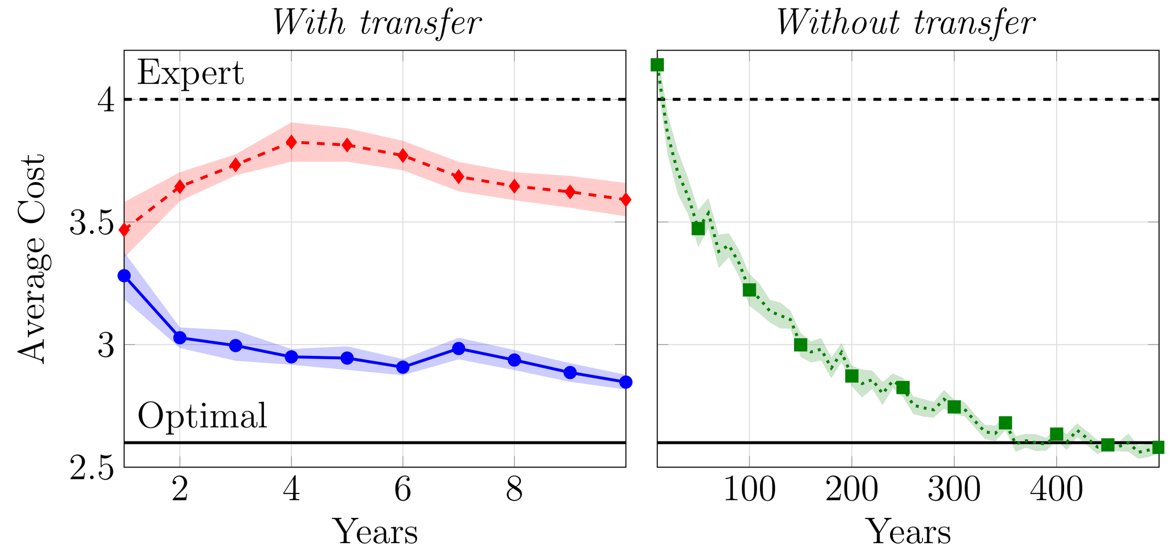
Puddle World



Acrobot



Water Reservoir Control



- IWFQI with estimated importance weights
- IWFQI with ideal importance weights
- Relevance-based [Lazaric et al., 2008]
- Shared-dynamics [Laroche and Barlier, 2017]
- Fitted Q-Iteration [Ernst et al., 2005]

References

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 2005.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

Romain Laroche and Merwan Barlier. Transfer reinforcement learning with shared dynamics. In *AAAI*, 2017.

Alessandro Lazaric, Marcello Restelli, and Andrea Bonarini. Transfer of samples in batch reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, 2008.

GitHub