Importance Weighted Transfer of Samples in Reinforcement Learning

Andrea Tirinzoni, Andrea Sessa, Matteo Pirotta, and Marcello Restelli

35th International Conference on Machine Learning, Stockholm, Sweden

POLITECNICO MILANO 1863

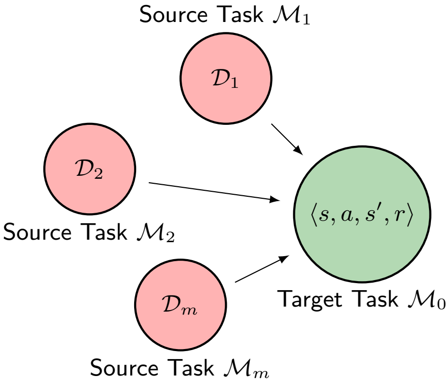
inventeurs du monde numérique

Motivating Example

Optimal control of a **water reservoir**

- Learn per-day water release decisions
- **1 sample = 1 day** $\Rightarrow$ Impractical to learn in the real world
- Lots of historical data might be available from different reservoirs $\rightarrow$ **Transfer**

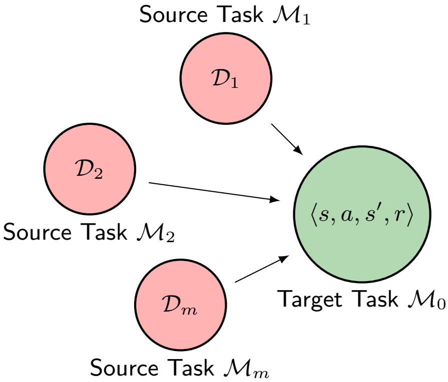
Transfer of Samples



Tasks are **MDPs** $\mathcal{M}_j = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_j, \mathcal{R}_j \rangle$

- Shared state-action space ($\mathcal{S} \times \mathcal{A}$)
- Different reward ($\mathcal{R}_j$) and transition ($\mathcal{P}_j$) models

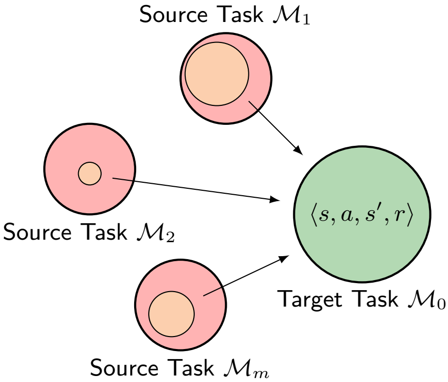
Transfer of Samples



Why transferring samples?

- Decoupled from the **learning algorithm**
- Does not require source tasks to be **solved**
- Data can come from **any distribution**

Previous Works

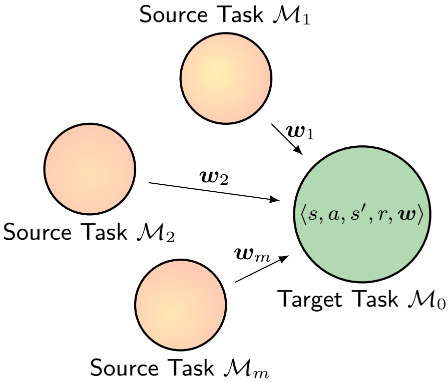


Mostly focus on **sample selection**

- **Intuition**: We are willing to introduce some **bias** to greatly reduce the **variance**
- Bias > variance $\Rightarrow$ **Negative transfer**
- Non-trivial task

[Lazaric et al., 2008, Taylor et al., 2008, Lazaric and Restelli, 2011, Laroche and Barlier, 2017]

Our Proposal



Transfer **all samples** available

- Assign **weights** proportional to their **importance** in solving the target task
- Reduce variance while ideally **unbiased**

Transfer via Importance Weighting

Fitted Q-Iteration [Ernst et al., 2005] $\rightarrow$ Sequence of supervised learning problems:

$$Q_{k+1} = \arg\inf_{h \in \mathcal{H}} \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \left| h(s,a) - \widehat{T} Q_k(s,a) \right|^2 \qquad \widehat{T} Q_k(s,a) = r + \gamma \max_{a'} Q_k(s', a')$$

- Different tasks $\Rightarrow$ *sample-selection bias* $\rightarrow$ Use **importance weighting**

Importance Weighted Fitted Q-Iteration (IWFQI)

1. Fit the target **reward function**

$$\widehat{R} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_r} \sum_{j=0}^{m} \sum_{\mathcal{D}_j} \boxed{w_r} \left| h(s,a) - r \right|^2 \qquad \boxed{w_r} = \frac{\mathcal{R}_0(r|s,a)}{\mathcal{R}_j(r|s,a)}$$

Importance Weighted Fitted Q-Iteration (IWFQI)

1. Fit the target **reward function**

$$\widehat{R} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_r} \sum_{j=0}^{m} \sum_{\mathcal{D}_j} \boxed{w_r} \left| h(s,a) - r \right|^2 \qquad \boxed{w_r} = \frac{\mathcal{R}_0(r|s,a)}{\mathcal{R}_j(r|s,a)}$$

2. Replace the empirical Bellman operator with:

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

Importance Weighted Fitted Q-Iteration (IWFQI)

1. Fit the target **reward function**

$$\widehat{R} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_r} \sum_{j=0}^{m} \sum_{\mathcal{D}_j} \boxed{w_r} \left|h(s,a) - r\right|^2 \qquad \boxed{w_r} = \frac{\mathcal{R}_0(r|s,a)}{\mathcal{R}_j(r|s,a)}$$

2. Replace the empirical Bellman operator with:

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

3. Iteratively fit the **value function**:

$$Q_{k+1} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_p} \sum_{j=0}^{m} \sum_{\mathcal{D}_j} \boxed{w_p} \left|h(s,a) - \widetilde{T}Q_k(s,a)\right|^2 \qquad \boxed{w_p} = \frac{\mathcal{P}_0(s'|s,a)}{\mathcal{P}_j(s'|s,a)}$$

Importance Weighted Fitted Q-Iteration (IWFQI)

1 Fit the target **reward function**

$$\widehat{R} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_r} \sum_{j=0}^{m} \sum_{\mathcal{D}_j} w_r \left|h(s,a) - r\right|^2 \qquad w_r = \frac{\mathcal{R}_0(r|s,a)}{\mathcal{R}_j(r|s,a)}$$

2 Replace the empirical Bellman operator with:

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

3 Iteratively fit the **value function**:

$$Q_{k+1} = \arg\inf_{h \in \mathcal{H}} \frac{1}{Z_p} \sum_{j=0}^{m} \sum_{\mathcal{D}_j} w_p \left|h(s,a) - \widetilde{T}Q_k(s,a)\right|^2 \qquad w_p = \frac{\mathcal{P}_0(s'|s,a)}{\mathcal{P}_j(s'|s,a)}$$

- Weights have to be **estimated** $\rightarrow$ We use Gaussian processes

Theoretical Analysis

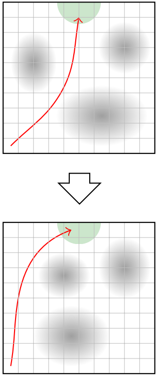
Error bound for IWFQI
(extends [Munos and Szepesvári, 2008, Farahmand et al., 2010, Cortes et al., 2010])

$$\|Q^* - Q^{\pi_K}\|_{1,\rho} \leq f\left(\text{approximation} + \text{estimation} + \text{bias} + \text{propagation}\right)$$

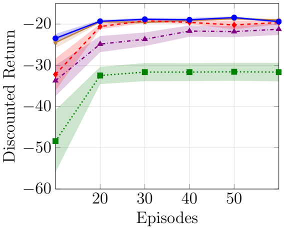
Differently from previous works [Lazaric and Restelli, 2011]:

- Bias does **not** depend on the differences between tasks
- Estimation error depends on the number of **effectively transferred** samples

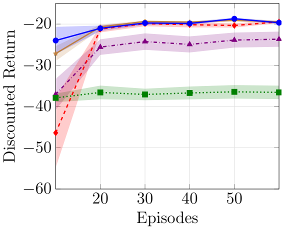
Empirical Evaluation - Puddle World



Shared Dynamics



Puddle-based Dynamics



IWFQI | IWFQI (ideal weights)

Relevance-based [Lazaric et al., 2008]

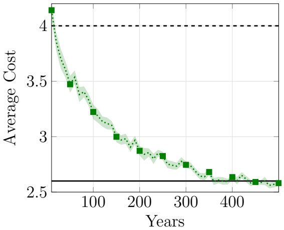
Shared-dynamics [Laroche and Barlier, 2017]

FQI [Ernst et al., 2005]

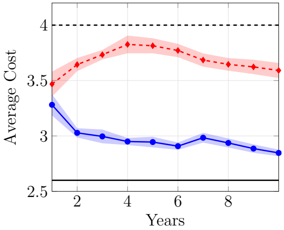
Empirical Evaluation - Water Reservoir Control

No Transfer

- $200k$ samples $\approx$ **500 years!**



Transfer



IWFQI

Optimal　　Expert

Relevance-based [Lazaric et al., 2008]

FQI [Ernst et al., 2005]

Conclusion

We presented **Importance Weighted Fitted Q-iteration**

- Transfer all samples via importance weighting
- Decouple rewards and transitions
- Theoretically well-grounded
- Better empirical performance than existing methods

Contacts

andrea.tirinzoni@polimi.it

https://github.com/AndreaTirinzoni/

Please visit us at poster #207 @ Hall B

References

Cortes, C., Mansour, Y., and Mohri, M. (2010).
Learning bounds for importance weighting.
In *Advances in neural information processing systems*, pages 442–450.

Ernst, D., Geurts, P., and Wehenkel, L. (2005).
Tree-based batch mode reinforcement learning.
Journal of Machine Learning Research.

Farahmand, A.-m., Szepesvári, C., and Munos, R. (2010).
Error propagation for approximate policy and value iteration.
In *Advances in Neural Information Processing Systems*.

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*.

References (cont.)

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*.

Munos, R. and Szepesvári, C. (2008).
Finite-time bounds for fitted value iteration.
Journal of Machine Learning Research, 9(May):815–857.

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.

Theoretical Analysis

L_p-norm bounds for AVI [Munos and Szepesvári, 2008, Farahmand et al., 2010]

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

$$\epsilon_k = T^* Q_k - Q_{k+1}$$

Theoretical Analysis

Error bound for IWFQI

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

- Irreducible **approximation** error of value and **reward** functions
- Error **propagation** through iterations
- **Estimation** error: finite samples & importance-weight **variance** [Cortes et al., 2010]
- Importance weights must be estimated → **bias**

Importance Weight Estimation

Problem: the task models $\mathcal{R}$ and $\mathcal{P}$ are **unknown** $\rightarrow$ The importance weights **cannot** be computed exactly

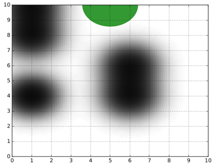
Solution: Fit **Gaussian processes** for the models $\mathcal{R}$ and $\mathcal{P}$ of each task

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

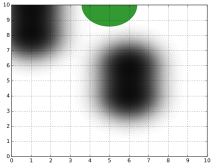
$$\mathbb{E}_{\mathcal{G}}\left[w_r(s,a)\right] = C\,\frac{\mathcal{N}\left(r \middle| \mu_{GP_0}(s,a), \sigma_0^2(s,a) + \sigma_{GP_0}^2(s,a)\right)}{\mathcal{N}\left(r \middle| \mu_{GP_j}(s,a), \sigma_j^2(s,a) - \sigma_{GP_j}^2(s,a)\right)}$$

Experimental Evaluation - Puddle World

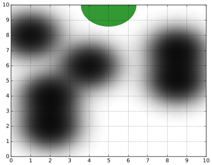
Target Task



Source Task 1



Source Task 2



Source Task 3