Transfer of Value Functions via Variational Methods

Andrea Tirinzoni, Rafael Rodriguez, and Marcello Restelli
{andrea.tirinzoni, marcello.restelli}@polimi.it, rafaelalberto.rodriguez@mail.polimi.it
POLITECNICO MILANO 1863

GitHub

Problem

- The agent has solved a finite set of **source tasks** $\mathcal{M}_{\tau_1}, \mathcal{M}_{\tau_2}, \dots, \mathcal{M}_{\tau_M}$ sampled from some **distribution** $\mathcal{D}$
- Each task is an MDP $\mathcal{M}_\tau = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_\tau, \mathcal{R}_\tau, p_0 \rangle$
- A parametric approximation to their **optimal value functions** is available

$$\mathcal{W}_s = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_M\} \;\text{ s.t. }\; Q_{\boldsymbol{w}_j} \simeq Q^*_{\tau_j}$$

- **Assumption**: all tasks share **similarities** in their optimal value functions [4]
- **Goal**: use this knowledge to speed-up the learning process of a new **target task** $\mathcal{M}_\tau$ sampled from $\mathcal{D}$

Motivation

- **Reinforcement learning** algorithms have enjoyed many success stories in complicated tasks
- High **sample complexity** remains a major issue
- Must **adapt** to changing environments and goals
- **Prior knowledge** from related tasks is often available in practice $\rightarrow$ **Transfer learning** [6]
- Need for transfer algorithms that are **general** and widely **applicable**

Contributions

1. **Algorithmic**. We propose a general **framework** for transferring value functions in RL and two **practical algorithms**
   - We learn a **prior** distribution over optimal Q-functions using the given source tasks
   - **Variational** approximation of the corresponding posterior for a new target task
   - Efficient **exploration** via posterior sampling
   - **Any** differentiable Q-function approximator and prior/posterior models could be used
2. **Theoretical**. We provide a theoretical analysis of our practical algorithms offering a better insight into their behavior
3. **Empirical**. We empirically evaluate our algorithms on four different domains with increasing level of difficulty

Variational Transfer Framework

IDEA: use the source weights $\mathcal{W}_s$ to estimate the distribution $p(\boldsymbol{w})$ over optimal Q-functions induced by $\mathcal{D}$

- How to characterize $p(\boldsymbol{w}|D) \propto p(D|\boldsymbol{w})p(\boldsymbol{w})$ given a dataset D of N samples from the target task?
- **PAC-Bayes argument** [3]: the likelihood $p(D|\boldsymbol{w})$ decays exponentially as the TD error of $Q_{\boldsymbol{w}}$ on D increases

$$p(\boldsymbol{w}|D) \simeq \frac{e^{-\Lambda \|B_{\boldsymbol{w}}\|_D^2} p(\boldsymbol{w})}{\int e^{-\Lambda \|B_{\boldsymbol{w}'}\|_D^2} p(d\boldsymbol{w}')}$$

- **Problem**: computing the Gibbs posterior is often intractable $\rightarrow$ **Variational approximation** [1]

$$\min_{\boldsymbol{\xi}\in\Xi} \mathcal{L}(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{w}\sim q_{\boldsymbol{\xi}}}\left[\|B_{\boldsymbol{w}}\|_D^2\right] + \frac{\lambda}{N} KL\left(q_{\boldsymbol{\xi}}(\boldsymbol{w}) \;||\; p(\boldsymbol{w})\right)$$

Main Properties

1. **Prior estimation**: summarize the information to transfer into a single distribution and use it to guide the learning process of the target task
2. **Exploration via posterior sampling** [5, 2]: at each time, the agent guesses the solution of the target task according to the current posterior and acts accordingly
3. **Black-box optimization**: minimizing the variational objective requires only differentiability of the models involved

Algorithm Variational Transfer

Input: Target task $\mathcal{M}_\tau$, source weights $\mathcal{W}_s$

Estimate prior $p(\boldsymbol{w})$ from $\mathcal{W}_s$
$\boldsymbol{\xi} \leftarrow \operatorname{argmin}_{\boldsymbol{\xi}} KL(q_{\boldsymbol{\xi}}||p)$, $D \leftarrow \emptyset$
repeat
  Sample initial state: $s_0 \sim p_0$
  while s_h is not terminal **do**
    $a_h = \operatorname{argmax}_a Q_{\boldsymbol{w}}(s_h, a)$ for $\boldsymbol{w} \sim q_{\boldsymbol{\xi}}(\boldsymbol{w})$
    $s_{h+1} \sim \mathcal{P}_\tau(\cdot|s_h, a_h)$, $r_{h+1} = \mathcal{R}_\tau(s_h, a_h)$
    $D \leftarrow D \cup \langle s_h, a_h, r_{h+1}, s_{h+1}\rangle$
    $\boldsymbol{\xi} \leftarrow \mathit{optimizer}\left(\boldsymbol{\xi}, \nabla_{\boldsymbol{\xi}}\mathcal{L}(\boldsymbol{\xi})\right)$
  end while
until forever

Practical Algorithms

Gaussian Variational Transfer (GVT)

- Prior: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$
- Posterior: $q_{\boldsymbol{\xi}}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Mixture of Gaussian Variational Transfer (MGVT)

- Prior: $p(\boldsymbol{w}) = |\mathcal{W}_s|^{-1} \sum_{\boldsymbol{w}_s \in \mathcal{W}_s} \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_s, \sigma_p^2 \boldsymbol{I})$
- Posterior: $q_{\boldsymbol{\xi}}(\boldsymbol{w}) = C^{-1} \sum_{i=1,\dots,C} \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

Finite-Sample Analysis

Bound the **expected Bellman error** under the optimal variational distribution for a dataset of N samples

$$\mathbb{E}_{q_{\widehat{\boldsymbol{\xi}}}}\left[\left\|\widetilde{B}_{\boldsymbol{w}}\right\|_\nu^2\right] \le 2\left\|\widetilde{B}_{\boldsymbol{w}^*}\right\|_\nu^2 + \upsilon(\boldsymbol{w}^*) + c_1\sqrt{\frac{\log\frac{2}{\delta}}{N}} + \frac{c_2 + \lambda d \log N + \lambda\, \varphi(\mathcal{W}_s)}{N} + \frac{c_3}{N^2}$$

1. **Approximation error** due to the limited hypothesis space
2. **Variance** due to a biased estimation of the Bellman error
3. **Variance** due to the finite samples
4. **Likelihood** of the optimal target weights under the prior

GVT: Distance to the prior mean

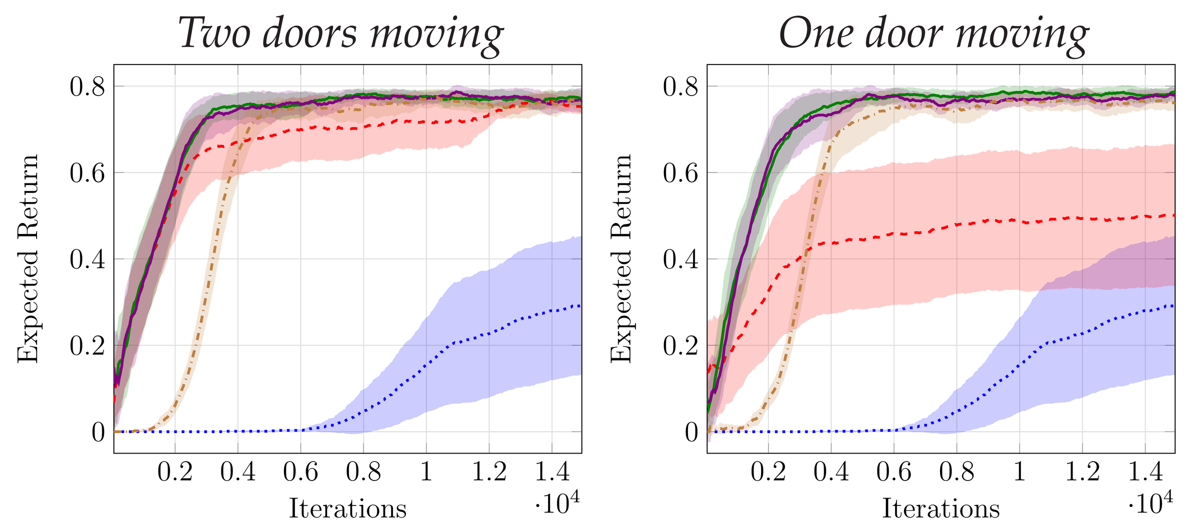
$$\varphi(\mathcal{W}_s) = \|\boldsymbol{w}^* - \boldsymbol{\mu}_p\|_{\boldsymbol{\Sigma}_p^{-1}}$$

MGVT: **Softmin** distance to the sources

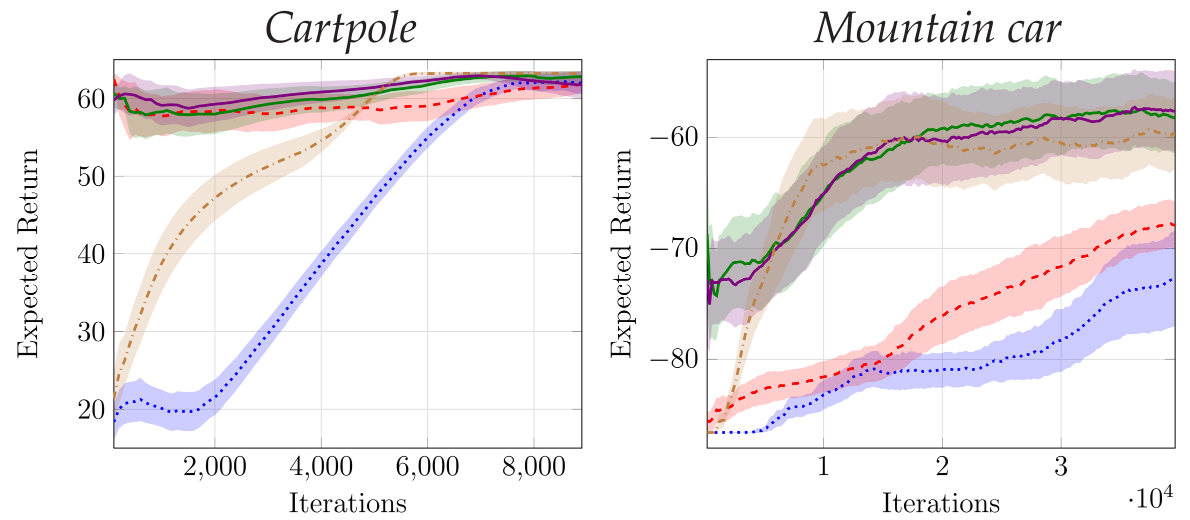
$$\varphi(\mathcal{W}_s) = \operatorname*{softmin}_{\boldsymbol{w}\in\mathcal{W}_s}\left(\|\boldsymbol{w}^* - \boldsymbol{w}\|\right)$$

Empirical Results

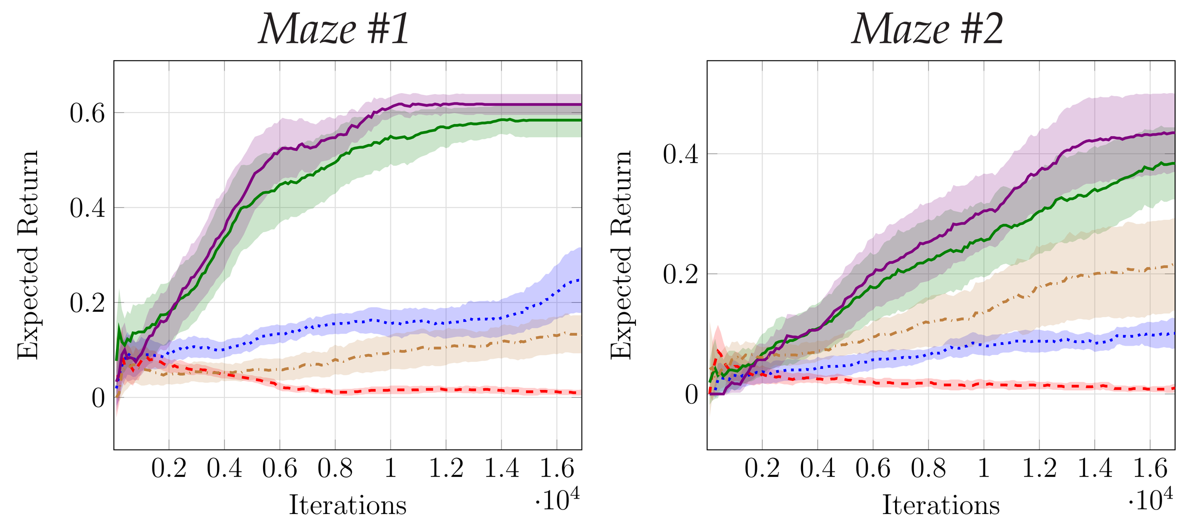
The Rooms Problem



Classic Control



Maze Navigation



- No transfer (DDQN)
- Fine-tuning from random source task
- GVT
- MGVT with 1 component
- MGVT with 3 components

References

[1] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41, 2016.

[2] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. *arXiv preprint arXiv:1802.04412*, 2018.

[3] Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

[4] Alessandro Lazaric and Mohammad Ghavamzadeh. Bayesian multi-task reinforcement learning. In *ICML-27th International Conference on Machine Learning*, pages 599–606. Omnipress, 2010.

[5] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

[6] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 2009.