Transferring Value Functions via Variational Methods

Andrea Tirinzoni, Rafael A. Rodriguez, and Marcello Restelli

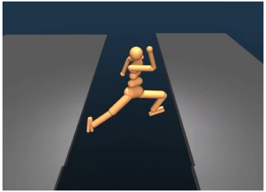
14th European Workshop on Reinforcement Learning, Lille, France

POLITECNICO
MILANO 1863

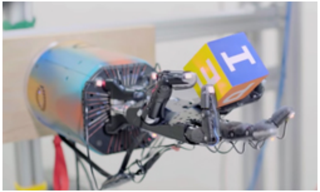
Motivation

Why do we need transfer?

Reinforcement Learning (RL) has been successfully applied to many **complex tasks**



[Heess et al., 2017]



[OpenAI, 2018]

[Vinyals et al., 2017]

- High **sample complexity** remains a major limitation
- Both humans and artificial agents repeatedly face several related tasks
  - Changing environments
  - New goals

Transfer

Our Settings

- The agent has solved a *finite* set of **source tasks** sampled from a **distribution** $\mathcal{D}$

$$\mathcal{M}_{\tau_1}, \mathcal{M}_{\tau_2}, \dots, \mathcal{M}_{\tau_M} \ \text{ s.t. } \ \mathcal{M}_\tau = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_\tau, \mathcal{R}_\tau, p_0 \rangle \sim \mathcal{D}$$

- A parametric approximation to their **optimal value functions** is available

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

- **Goal**: use this knowledge to speed-up the learning process of a new **target task** $\mathcal{M}_\tau$ sampled from $\mathcal{D}$

Related Works

- Use the source Q-functions as **initializers**
  [Tanaka and Yamamura, 2003, Taylor and Stone, 2009, Abel et al., 2018]
- **Bayesian** methods
  [Lazaric and Ghavamzadeh, 2010]
- Transfer via **successor features**
  [Barreto et al., 2017, Barreto et al., 2018]
- **Fast adaptation** / Meta-learning
  [Finn et al., 2017, Grant et al., 2018, Amit and Meir, 2018]

Our Objectives

What we would like to have

Our transfer algorithm should

- *not* make strong **assumptions** that limit its applicability
- dynamically use information from the source tasks during the learning process
- drive the **exploration** of the target task based on transferred knowledge

Transfer via Variational Methods

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 $p(\boldsymbol{w}|D) \propto \boxed{p(D|\boldsymbol{w})} p(\boldsymbol{w})$ given a dataset D of N samples from the target task?
- **PAC-Bayes argument** [Catoni, 2007]: the likelihood $p(D|\boldsymbol{w})$ decays exponentially as the *TD error* $\|B_{\boldsymbol{w}}\|_D^2$ of $Q_{\boldsymbol{w}}$ on D increases

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

Transfer via Variational Methods

Problem: computing the Gibbs posterior $q(\boldsymbol{w})$ is often intractable

- **Variational approximation** [Alquier et al., 2016] $\rightarrow \operatorname{argmin}_{\boldsymbol{\xi}} KL(q_{\boldsymbol{\xi}}(\boldsymbol{w}) \mid\mid q(\boldsymbol{w}))$

$$\underbrace{\min_{\boldsymbol{\xi}\in\Xi} \mathcal{L}(\boldsymbol{\xi})}_{\text{Variational objective}} = \underbrace{\mathbb{E}_{\boldsymbol{w}\sim q_{\boldsymbol{\xi}}}\left[\|B_{\boldsymbol{w}}\|_D^2\right]}_{\text{Expected TD error}} + \frac{\lambda}{N} \underbrace{KL\left(q_{\boldsymbol{\xi}}(\boldsymbol{w}) \mid\mid p(\boldsymbol{w})\right)}_{\text{Divergence w.r.t. the prior}}$$

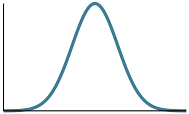
Algorithm 1 Variational Transfer

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

- Summarize transferred information into the **prior** distribution
- Exploration via **posterior sampling** [Osband et al., 2014]
- Requires only **differentiable** models

Practical Algorithms

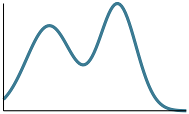
Gaussian Variational Transfer (GVT)



- Prior/posterior models are multivariate **Gaussians**

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \quad q_{\boldsymbol{\xi}}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Mixture of Gaussian Variational Transfer (MGVT)



- **Kernel density estimator** for the prior

$$p(\boldsymbol{w}) = \frac{1}{|\mathcal{W}_s|} \sum_{\boldsymbol{w}_s \in \mathcal{W}_s} \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_s, \sigma_p^2 \boldsymbol{I})$$

- C-component **mixture of Gaussian** model for the posterior

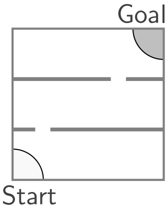
$$q_{\boldsymbol{\xi}}(\boldsymbol{w}) = \frac{1}{C} \sum_{i=1}^{C} \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

Theoretical Properties

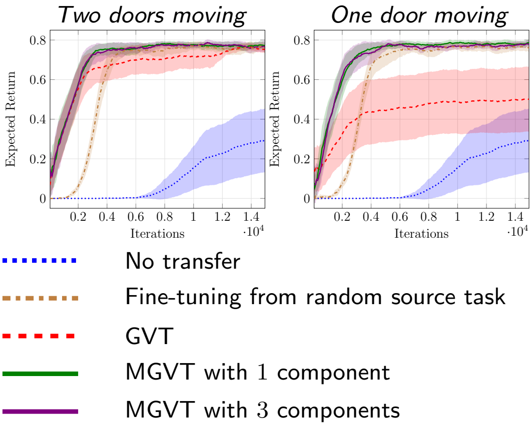
We provide a **finite-sample** analysis of both our practical algorithms

$$\underbrace{\mathbb{E}_{q_{\widehat{\boldsymbol{\xi}}}}\left[\left\|\widetilde{B}_{\boldsymbol{w}}\right\|_{\nu}^{2}\right]}_{\text{Expected Bellman error}} \leq \underbrace{2\left\|\widetilde{B}_{\boldsymbol{w}^*}\right\|_{\nu}^{2}}_{\text{Approximation error}} + \underbrace{v(\boldsymbol{w}^*)}_{\text{Variance}} + \underbrace{\frac{\lambda}{N}\boxed{\varphi\left(\mathcal{W}_s\right)}}_{\text{Distance to the prior}} + \mathcal{O}(1/N)$$

- Bound the expected Bellman error under the *optimal variational distribution* $q_{\widehat{\boldsymbol{\xi}}}$
- Same bounds, different distances to 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 source weights $\varphi(\mathcal{W}_s) = \operatorname{softmin}_{\boldsymbol{w}\in\mathcal{W}_s}\left(\|\boldsymbol{w}^* - \boldsymbol{w}\|\right)$

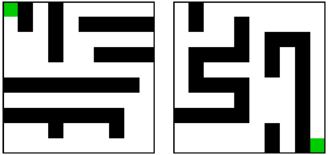


- Linear value functions with RBFs
- Different door positions
- $M = 10$ source tasks

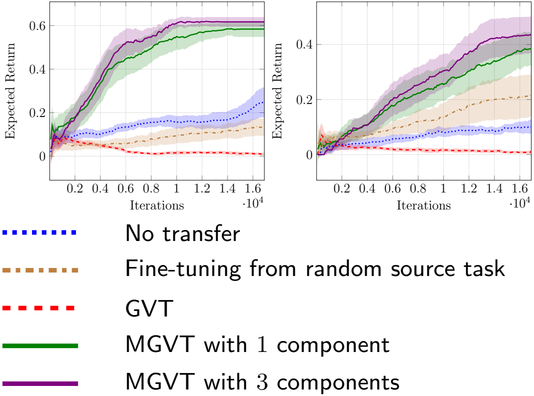


Empirical Evaluation

Maze Navigation



- NN value function approximators
- 20 different mazes
- $M = 5$ source tasks
- Target *not* in the source tasks



Conclusion

We presented a general approach for transferring value functions in RL

- *No strong assumptions* on the approximators/distributions involved
- Exploration of the target task via *posterior sampling*
- Two practical and efficient algorithms

Future works

- transfer parameterized *policies*
- *active* exploration

Contacts

andrea.tirinzoni@polimi.it

https://github.com/AndreaTirinzoni/

References

Abel, D., Jinnai, Y., Guo, S. Y., Konidaris, G., and Littman, M. (2018).
Policy and value transfer in lifelong reinforcement learning.
In *International Conference on Machine Learning*, pages 20–29.

Alquier, P., Ridgway, J., and Chopin, N. (2016).
On the properties of variational approximations of gibbs posteriors.
Journal of Machine Learning Research, 17(239):1–41.

Amit, R. and Meir, R. (2018).
Meta-learning by adjusting priors based on extended PAC-Bayes theory.
In *Proceedings of the 35th International Conference on Machine Learning*.

Barreto, A., Borsa, D., Quan, J., Schaul, T., Silver, D., Hessel, M., Mankowitz, D., Zidek, A., and Munos, R. (2018).
Transfer in deep reinforcement learning using successor features and generalised policy improvement.
In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 501–510, Stockholmsmässan, Stockholm Sweden. PMLR.

References (cont.)

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. (2017).
Successor features for transfer in reinforcement learning.
In *Advances in neural information processing systems*, pages 4055–4065.

Catoni, O. (2007).
Pac-bayesian supervised classification: the thermodynamics of statistical learning.
arXiv preprint arXiv:0712.0248.

Finn, C., Abbeel, P., and Levine, S. (2017).
Model-agnostic meta-learning for fast adaptation of deep networks.
arXiv preprint arXiv:1703.03400.

Grant, E., Finn, C., Levine, S., Darrell, T., and Griffiths, T. (2018).
Recasting gradient-based meta-learning as hierarchical bayes.
arXiv preprint arXiv:1801.08930.

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.

References (cont.)

Lazaric, A. and Ghavamzadeh, M. (2010).
Bayesian multi-task reinforcement learning.
In *ICML-27th International Conference on Machine Learning*, pages 599–606. Omnipress.

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

Osband, I., Van Roy, B., and Wen, Z. (2014).
Generalization and exploration via randomized value functions.
arXiv preprint arXiv:1402.0635.

Tanaka, F. and Yamamura, M. (2003).
Multitask reinforcement learning on the distribution of mdps.
In *Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on*, volume 3, pages 1108–1113. IEEE.

Taylor, M. E. and Stone, P. (2009).
Transfer learning for reinforcement learning domains: A survey.
Journal of Machine Learning Research, 10(Jul):1633–1685.

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.