Transfer and Multi-Task Learning in Reinforcement Learning

Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR
INRIA Lille - Nord Europe, Team SequeL

1. The Context

Reinforcement learning's (RL) (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996) challenging objective is to develop autonomous agents able to learn how to act optimally in an unknown and uncertain environment by trial-and-error and with limited level of supervision (i.e., a reinforcement signal). RL is mostly applied in domains where a precise formalization of the environment and/or the efficient computation of the optimal control policy is particularly difficult (e.g., robotics, human-computer interaction, recommendation systems). An RL problem is formalized as a Markov decision process (MDP) $\mathcal{M}$ characterized by a state space $\mathcal{X}$, an action space $\mathcal{A}$, a (stochastic) dynamics $p : \mathcal{X} \times \mathcal{A} \rightarrow \Delta(\mathcal{X})$ that determines the transition from states to states depending on the action, a reward function $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \rightarrow \mathbb{R}$ that determines the value of a transition x, a, x'. An MDP defines a control **task**. The solution to an MDP/task is an optimal policy $\pi^* : \mathcal{X} \rightarrow \mathcal{A}$ that prescribes the actions to take in each state to maximize the (discounted) sum of rewards measured by the optimal value function $V^* = \max_\pi \mathbb{E}[\sum_t \gamma^t r_t]$ with $\gamma \in (0, 1)$ and $r_t = r(x_t, \pi(x_t), x_{t+1})$. Two of the most difficult challenges in RL are:

1. How to explore the unknown environment so as to maximize the cumulative reward. This requires solving the **exploration-exploitation** problem, well formalized and studied at its core by the multi-armed bandit framework (Bubeck & Cesa-Bianchi, 2012).
2. How to effectively represent the policy and/or the value function. This requires defining an **approximation space** which is well-suited for the specific MDP at hand.

Both previous aspects may greatly benefit from techniques able to define suitable exploration strategies and approximation spaces from past experience or joint experience from other tasks (e.g., designing a intelligent tutoring system for a student and reuse the teaching strategy to other students). The **objective** of my research is to study the problems of transfer learning, multi-task learning, and domain adaptation in the RL (and related) field.

2. The Past

Unlike in supervised learning, transfer learning faces challenges which are specific to field of RL:

- many different things can be transferred (e.g., the MDP parameters, policies, value functions, samples, features),
- the definition of "unsupervised" samples is not clear and thus, domain adaptation methods exploiting target unsupervised samples cannot be easily applied,
- samples are often non-i.i.d. because they are obtained from policies
- tasks may be similar in terms of policies but neither MDPs nor value functions or viceversa.

For this reason, borrowing techniques from "supervised" transfer/multi-task learning is not always trivial or even possible. Early research focused on studying transfer of different kind of solutions from a source to a target task[1]. Later, more sophisticated transfer/multi-task scenarios and algorithms have been developed (e.g., using hierarchical Bayesian solutions to learn "priors" from multiple tasks) to improve the accuracy of the approximation of optimal policies/value functions. The results obtained in the past show a significant sample complexity reduction and an improvement in asymptotic accuracy when transfer/multi-task is applied.

3. The Future

My main interest in the short-term is to study the problem of how transfer/multi-task learning can actually improve exploration-exploitation strategies in multi-armed bandit (MAB) and RL. While the problem of approximation is common in supervised learning as well, the active collection of information is very much specific to RL and MAB.

So far, I have investigated a sequential transfer scenario and investigated two approaches in the linear MAB framework: *(i)* transfer of samples *(under review)*, *(ii)* use of transferred samples to identify the set of possible MAB problems and speed-up the problem identification phase (Gheshlaghi-Azar et al., 2013). In both cases, we proved that the cumulative reward (i.e., reduce the regret) of exploration-exploitation strategies in MAB can be actually improved and that negative transfer can be avoid. Nonetheless, a number of very important questions remain unanswered:

[1] See (Taylor & Stone, 2000; Lazaric, 2011) for a survey.

- Is it possible to incrementally and efficiently estimate the potential bias due to transfer from different tasks? Under which assumptions? *In specific cases, this can be done in supervised learning.*
- What is the measure of similarity between two MDPs that determines the difference in performance of an exploration-exploitation strategy when applied to the two MDPs?
- Is it worth it to explore more in earlier tasks to "unveil" the generative process of the sequence of tasks and exploit it to enhance the transfer? In which scenarios?
- MDPs with different state-action spaces may still be very much similar. Is it possible to map different MDP to an "underlying" common MDP structure in which similar exploration-exploitation solutions can be identified and transferred?

As motivating fields of application, I will focus on *intelligent tutoring systems, recommendation systems, and computer games*.

References

Bertsekas, D. and Tsitsiklis, J. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Gheshlaghi-Azar, M., Lazaric, A., and Brunskill, E. Sequential transfer in multi-arm bandit with finite set of models. In *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS'13)*, 2013.

Lazaric, A. Transfer in reinforcement learning: a framework and a survey. In Wiering, M. and van Otterlo, M. (eds.), *Reinforcement Learning: State of the Art*. Springer, 2011.

Sutton, R. and Barto, A. *Reinforcement Learning, An introduction*. BradFord Book. The MIT Press, 1998.

Taylor, M. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10:1633–1685, 2000.