Research Project Proposal:
Sample complexity of transfer learning in reinforcement learning under a parametrized setting

Riccardo Poiani
riccardo.poiani@mail.polimi.it
CSE Track
Outline

• Motivation

• State of the art
  o Generative setting
  o Non-generative setting

• Research idea and plan
• **Motivation**

• State of the art
  
  o Generative setting
  
  o Non-generative setting

• Research idea and plan
Problems and challenges

• **Superhuman** achievements in some problems but...

• Training costs **money**

• Training is **slow**

• Training can be **dangerous**
Transfer: benefits

Time to Threshold

Threshold Performance

Transfer

No Transfer

Jumpstart
Transfer: an example

Isele et. al [2017]
An **agent** acts in an **environment** in order to maximize a **reward signal**.

The problem is usually formalized as a Markov Decision Process:

- States: $S$
- Actions: $A$
- Initial state distributions
- Reward function
- Transition distribution
- Discount factor: It encodes information about horizon $H$
• A policy $\pi$ is a distribution over the actions, given the state

• The goal is to learn an optimal policy (up to some required accuracy)
  - the policy that maximizes the expected cumulated discounted reward
  - Often expressed in term of $V^\pi(s)$ or $Q^\pi(s, a)$

• Many algorithms exist: SARSA, Q-learning, Delayed Q-learning...
RL: sample complexity

Number of timestamps in which the policy is sub-optimal w.r.t. a fixed quantity $\epsilon$
RL: PAC-MDP efficient algorithm

• **Probabilistic correct** with confidence at least $1 - \delta$

• **Polynomial sample complexity** in the relevant quantities $\left( S, A, \frac{1}{\epsilon}, \frac{1}{\delta}, H \right)$
Setting and goal of the project

• Typical transfer setting

• The agent acts in an environment whose dynamics are characterized by some unknown parameter $\theta \in \Theta$

• Understanding how to exploit transferred knowledge to reduce sample complexity
  
  ○ Generative case

  ○ Non generative case

• Research objective: algorithms with theoretical guarantees; experiments
• Motivation

• State of the art
  o Generative setting
  o Non-generative setting

• Research idea and plan
## RL: Transfer

<table>
<thead>
<tr>
<th>Paper</th>
<th>Allowed Differences</th>
<th>Knowledge Transferred</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abel et. al [2018]</td>
<td>Reward</td>
<td>V(s) / Q(s,a)</td>
<td>Jumpstart and Sample complexity</td>
</tr>
<tr>
<td>Azar et al. [2013]</td>
<td>Transitions and Rewards</td>
<td>Policy</td>
<td>Cumulated reward</td>
</tr>
<tr>
<td>Tirinzoni et al [2019]</td>
<td>Transitions and rewards</td>
<td>Samples</td>
<td>Cumulated reward</td>
</tr>
<tr>
<td>Ammar et al. [2015]</td>
<td>All</td>
<td>Samples</td>
<td>Cumulated reward</td>
</tr>
<tr>
<td>Tirinzoni et al. [2018]</td>
<td>Transitions and rewards</td>
<td>V(s)</td>
<td>Sample complexity</td>
</tr>
</tbody>
</table>

Many others…
• Motivation

• **State of the art**
  
  ○ *Generative setting*
  
  ○ Non-generative setting

• Research idea and plan
Generative settings

• The analysis of the transfer case is currently missing

• Classical RL cases
  
  o A typical lower bound of the problem: $\tilde{O}\left(\frac{|S||A|H^3}{\epsilon^2}\right)$
  
  o **Uniform** sampling approach (Azar et al. [2013])
    
    o match lower bound under some assumptions
  
  o **Variance reduced** approach (Sidford et al. [2019])
• Motivation

• **State of the art**
  - Generative setting
  - **Non-generative setting**

• Research idea and plan
Non-generative setting

- Maximum Exploration Reinforcement Learning (MERL) (Lattimore et al. [2013])
  - \( \tilde{O} \left( \frac{|\Theta| H^3}{\epsilon^2} \right) \) match a lower bound up to a log factor
  - Impractical algorithm

- Parameter elimination method (PEL) (Dyagilev et al. [2008])
  - \( \tilde{O} \left( \frac{|\Theta| H^6}{\epsilon^3} \right) \)
  - Sequential probability ratio test
Non-generative setting

- On the sample complexity of Multi-task RL (Brunskill et. Al [2013])
  - Multi-task setting
  - Clustering approach
  - Theoretical bounds
  - Trade-off between structure exploitation and exploration
Non-generative setting

- **Hidden** parameter MDPs (Killian et. Al [2017])
  - Complex solution that works very well in practice
  - No theoretical guarantees
- **Contextual** MDPs (Modi et. Al [2017])
  - Continuous space for the context
  - Known context
• Motivation

• State of the art
  o Generative setting
  o Non-generative setting

• Research idea and plan
How to tackle the problem

• The problem of sample complexity in the transfer learning setting is hard

• There is little understanding so far in the literature

• We can take advantage of a **generative model** to better understand the problem

• From this simplified case, take insight for more practical algorithms
Desired achievements

- Generative case [65% completed]
  - Online algorithm with theoretical guarantees [85%]
  - Better bounds than the classical RL case by exploiting the structure [95%]
  - Propose a real setting when the algorithm can be used [10%]
Desired achievements

• Non-generative case [0% completed]
  - Online algorithm with theoretical guarantees
  - Experiment to compare against state-of-the-art algorithms
Milestones

- **ICML 2020** 7 February
- **NeurIPS 2020**: around the end of May
Thanks for your attention!