Regret bounds for kernel-based reinforcement learning

Omar D. Domingues[1], Pierre Ménard[1], Matteo Pirotta[2], Emilie Kaufmann[1,3], Michal Valko[1,4]

[1]Inria Lille - Nord Europe
[2]Facebook AI Research, Paris
[3]CNRS & Université de Lille
[4]DeepMind, Paris

PGMO days 2020

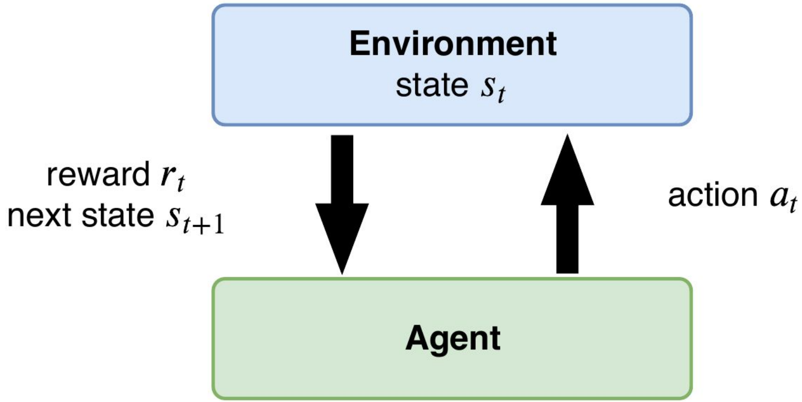
Reinforcement Learning

Framework that **models several learning problems**, for instance

- Controlling robots to reach a goal
- Playing games
- Recommendation systems
- Self-driving cars

Very hard to solve

- Requires a lot of data
- **How to collect data efficiently?**



Mathematical model

- The environment is modeled as a Markov decision process
  - State and action sets $\mathcal{S}, \mathcal{A}$
  - Transition probabilities $P(s'|s, a)$
  - Reward function $r(s, a)$
- The agent follows a policy
  - Action to take in state s at time h: $\pi_h(s)$
- Goal: maximize the sum of rewards

$$\max_{\pi} \mathbf{E}_{\pi} \left[\sum_{h=1}^{H} r(S_h, A_h) \right]$$

- If we have a simulator, use approximate dynamic programming (ADP).

Kernel-based reinforcement learning

- Approximate DP technique introduced by Ormoneit & Sen (2002)*
- Simulate **N independent samples** from the MDP
  - i-th sample = (state, action, reward, next state) = (s_i, a_i, r_i, s_i')
- Estimate a model

$$\widehat{r}(s,a) = \frac{\sum_i w_i(s,a) r_i}{\beta + \sum_i w_i(s,a)}, \qquad \widehat{P}(s'|s,a) = \frac{\sum_i w_i(s,a)\delta_{s_i'}(s')}{\beta + \sum_i w_i(s,a)}$$

- Run dynamic programming on the estimated model, complexity = $O(N^2)$
- Asymptotic convergence guarantee

* Ormoneit, D., & Sen, Ś. (2002). Kernel-based reinforcement learning. Machine learning, 49(2-3), 161-178.

Our contribution: Kernel-based RL with exploration*

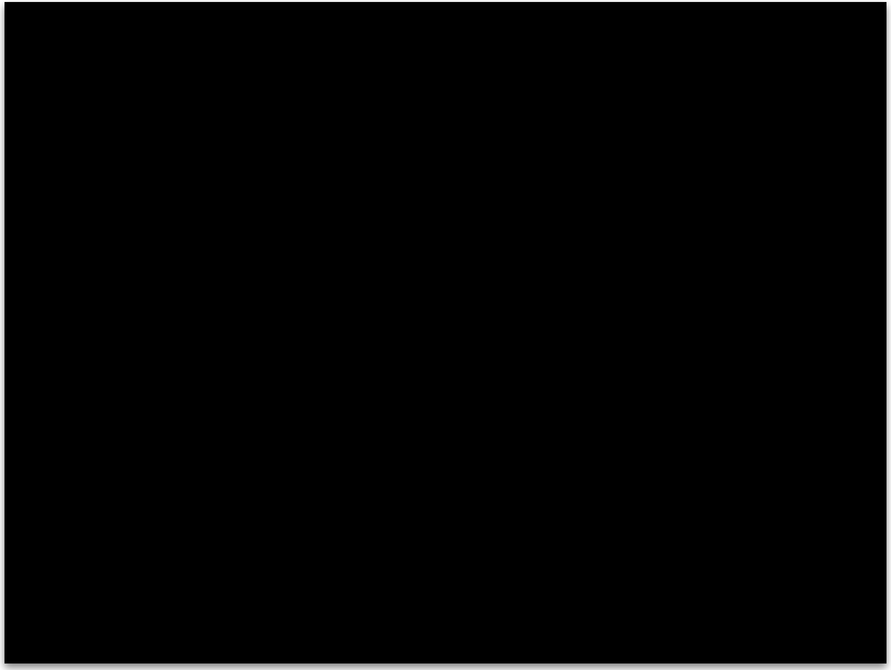
- Exploration strategy to collect data online (not independent)
- Finite time-guarantee, "error per sample" converges to zero

$$\frac{\mathrm{Regret}(N)}{N} \lesssim \left(\frac{1}{N}\right)^{\frac{1}{2d+1}}$$

- Assumption: model is Lipschitz continuous with respect to a given metric
- d = covering dimension of the state-action space

* Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., & Valko, M. (2020). Regret bounds for kernel-based reinforcement learning.

Example



Thank you!

- How to avoid the curse of dimensionality?
- How to improve the runtime?
- Extension to non-stationary MDPs*

* Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., & Valko, M. (2020). A kernel-based approach to non-stationary reinforcement learning in metric spaces.