Research Project Proposal:
Policy space identification in Configurable MDPs

Guglielmo Manneschi
guglielmo.manneschi@mail.polimi.it
CSE Track

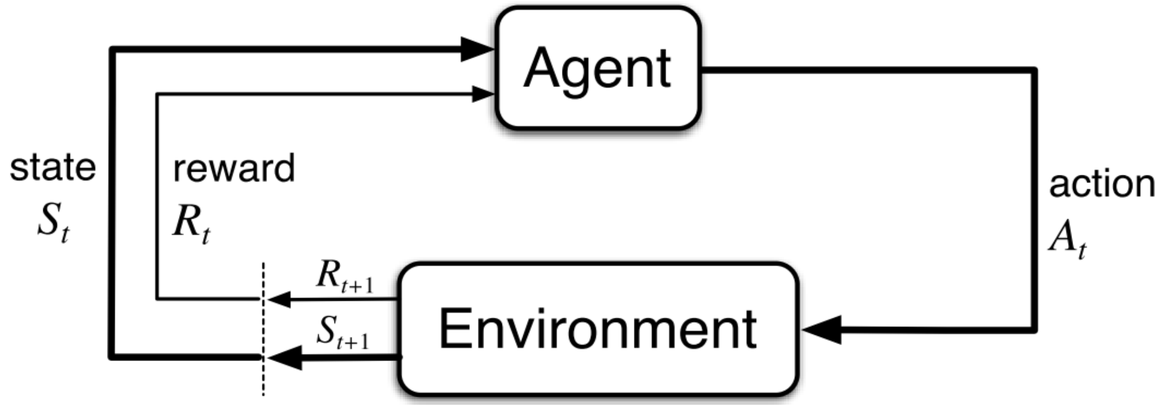
POLITECNICO
MILANO 1863

HONOURS PROGRAMME
HP-SR
in Information Technology

Markov Decision Process

Framework to model **sequential decision-making** problems [1]

- **States**
  - How the agent perceives the world
- **Actions**
- **Rewards**
  - How good the agent behaves
- **Environment dynamics**



[1] *Richard S. Sutton and Andrew G. Barto. 1998. Introduction to Reinforcement Learning (1st ed.). MIT Press, Cambridge, MA, USA.*

Markov Decision Process

- **Policy** $\pi(a|s)$
  - Agent behaviour

- **Expected return** $J_\pi = \mathbb{E}_\pi \left[\sum_{t=1}^{T} R_t \right]$
  - Expected sum of rewards in a complete episode
  - How good is a certain policy

Markov Decision Process

- **Solution** of an MDP
  - Find a policy that maximizes the expected return

$$\pi^* = \arg\max J_\pi$$

  - If the policy is parametric, gradient based approach [2]

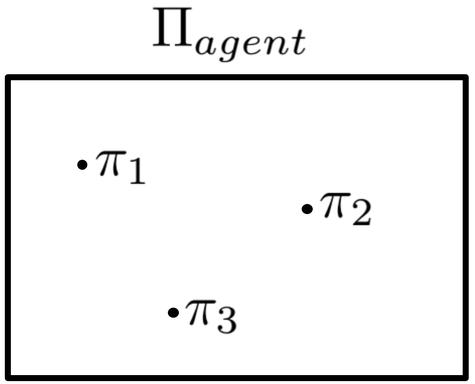
[2] *Peters, Jan, and Stefan Schaal. "Reinforcement learning of motor skills with policy gradients." Neural networks 21.4 (2008): 682-697.*

Policy space

- In case of a large (or infinite) state space, computing the policy for every state is infeasible
  - Instead, represent the state with a **feature vector** $\phi(s)$
  - The policy is a function defined by a **parameter vector** θ: $\pi_\theta(a|\phi(s))$
  - The parameters combine the state features observed by the agent
- To find the best policy, the agent searches inside a space of functions (or parameters) called **policy space**

Policy space identification

- **Goal** of the research project
- If we can only observe an agent, it can be useful to identify its policy space
  - to estimate the agent's capabilities
  - to discover what can the agent perceive
  - to retrieve the reward function (IRL) [3]

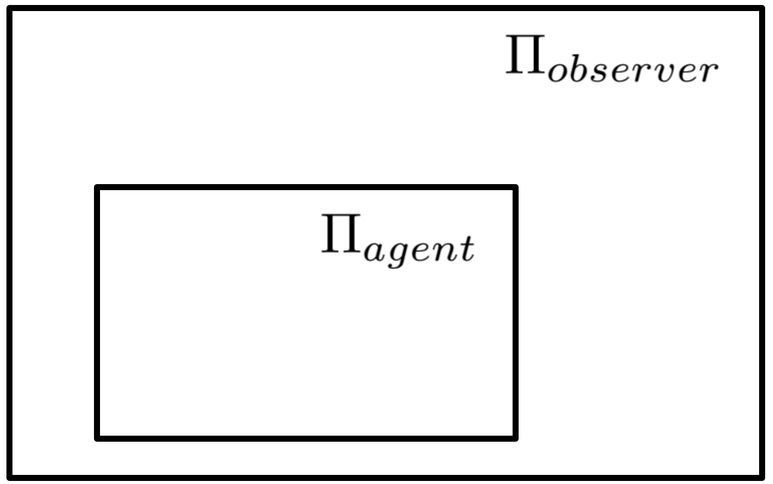


[3] *Ng, Andrew Y., and Stuart J. Russell. "Algorithms for inverse reinforcement learning." Icml. Vol. 1. 2000.*

Policy space identification

- E.g. the policy of an agent is defined by an **unknown parameter vector**
  - The higher a parameter the more an action depends on that feature

- We have a greater set of parameters (and features)
  - Understand which ones are actually used by the agent



Policy space identification

- Limits of the classical MDP framework
  - **Some state features may be useless** for a certain task
  - It may appear that the agent cannot see them

Configurable MDP

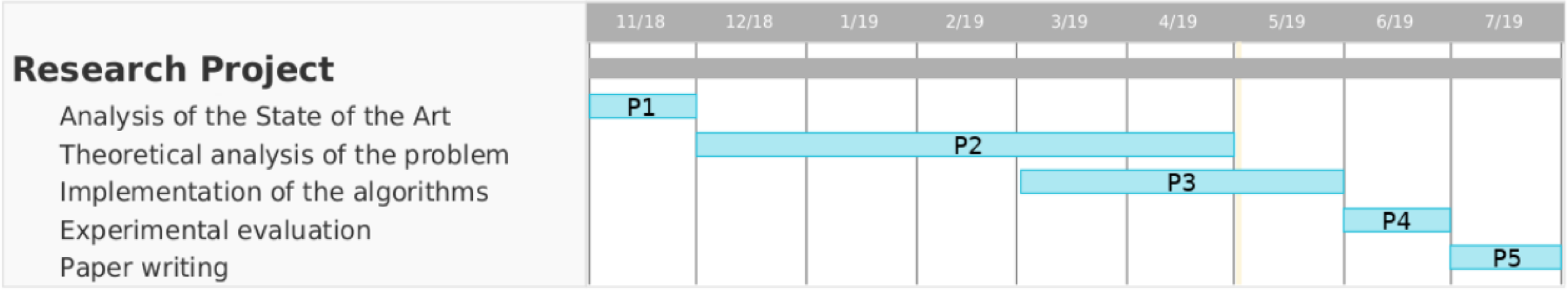
- **Configurable** Markov Decision Process [4]
  - Novel framework
  - Extension of MDP
  - Possibility to configure the environment with a set of parameters

[4] *Metelli, Alberto Maria, Mirco Mutti, and Marcello Restelli. "Configurable Markov Decision Processes." ICML. 2018.*

Configurable MDP

- We want to **solicit the agent to use certain features**
  - **Select an environment** where the task requires having access to those features to be solved
  - Put the agent in this environment and let it **learn** the best policy
  - **Estimate** the value of the parameters (in our policy space)
- Repeat with multiple environments until we have enough confidence

Research project



Gantt diagram of the research project

Thanks for your attention