Near-Optimal Reinforcement Learning in Factored MDPs

Ian Osband and Benjamin Van Roy

Stanford University

Reinforcement Learning

- **Setting:** Learning + decision making + delayed feedback.

  Supervised Learning → *Multi-armed Bandit* → *Reinforcement Learning*

- **Goal:** Maximize cumulative rewards through time.

- **Key tradeoff:** *Exploration vs. Exploitation.*

  "We want algorithms that learn to make good decisions in any unknown environment as efficiently as possible."

- **Measure:** $\mathrm{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T}(r_t^* - r_t)\right]$

  r_t^*: *Rewards of unknown optimal controller*

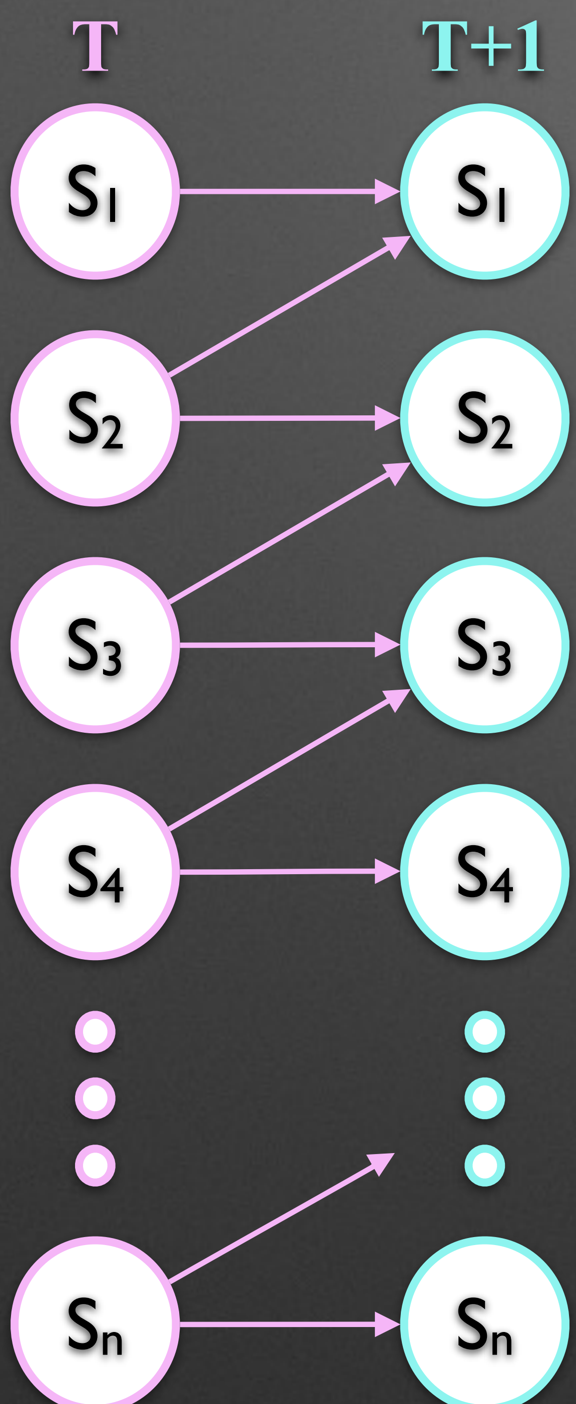
  r_t: *Actual rewards*

- **Theorem:** In a general MDP with S states A actions

$$\mathrm{Regret}(T) = \Omega\left(\sqrt{SAT}\right)$$

- **Problem:** We want low regret even when S, A are huge!

Learning in Factored MDPs



- **Key idea:** Learn quickly via *low-dimensional structure*.
- **Definition:** Factored MDP $\leftrightarrow$ conditional independence.
- **Example:** In a production line, the state of each machine is only directly dependent upon its neighbors.

"We obtain regret bounds that scale with the number of parameters, rather than the number of states."

- **Algorithms:** *Optimism* and *Posterior Sampling*.
- **Result:** For K independent sections in the MDP

Naive Bounds: EXPONENTIAL in K → *New Bounds: LINEAR in K* → *See you at the poster!*