Near-Optimal Reinforcement Learning in Factored MDPs

Ian Osband and Benjamin Van Roy

Stanford University

Reinforcement Learning

- **Setting:** Decision agent in unknown environment.
- **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:**

$$\mathrm{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_t^* - r(s_t, a_t)\right]$$

Rewards of unknown optimal controller

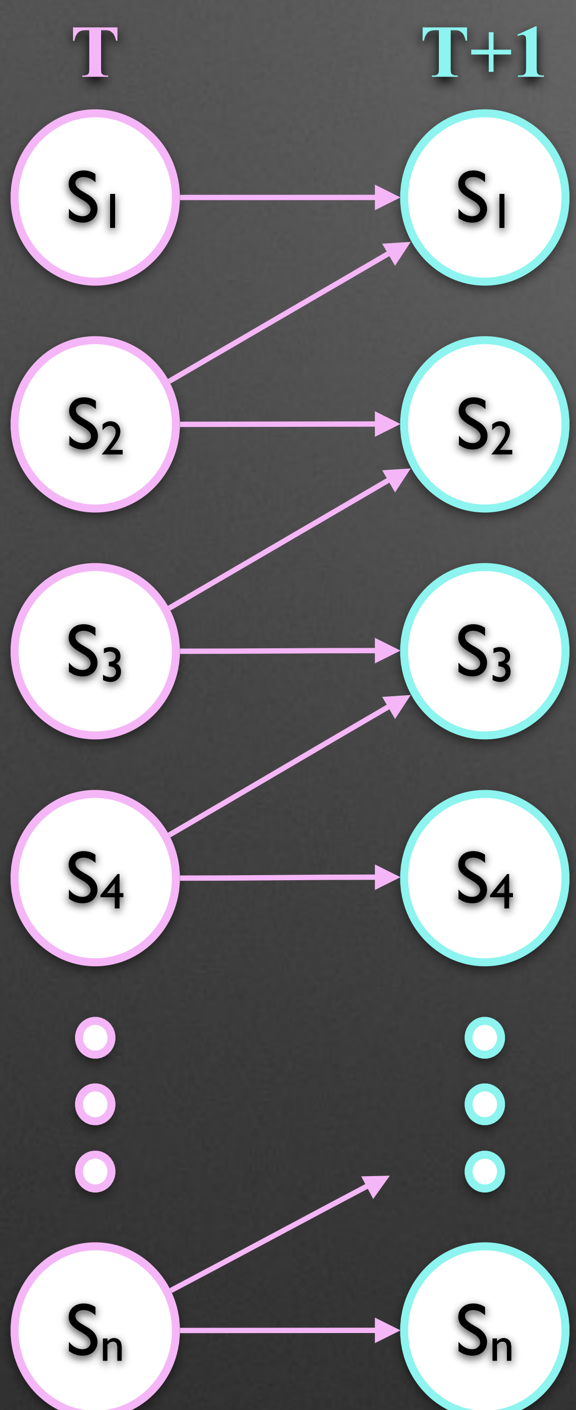
Actual rewards

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

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

- **Problem:** We want to learn even when S, A are huge!

Learning in Factored MDPs



- **Key idea:** Learn quickly via low-dimensional structure.
- **Definition:** Factored MDP $\leftrightarrow$ conditional independence.

"We obtain regret bounds that scale with the number of parameters, rather than the cardinality, of the MDP"

- **Algorithms:** *Optimism* and *Posterior Sampling*.
- **Result:** For m independent sections S states, A actions:

$$\mathrm{Regret}(T) = \tilde{O}\left(m\sqrt{S^2AT}\right) \ll \tilde{O}\left(\sqrt{(S^2A)^mT}\right)$$

- **Example:** Production line with 100 machines, 3 states, 3 actions, each only depends directly on its neighbors

Without factored structure:
Regret bound $\approx 10^{120}\sqrt{T}$

→

Using graphical knowledge:
Regret bound $\approx 10^3\sqrt{T}$