Near-Optimal Reinforcement Learning in Factored MDPs

Ian Osband and Benjamin Van Roy

Stanford University

Reinforcement Learning

- Goal: Maximize long term rewards in an unknown environment.
- Key tradeoff: *Exploration vs Exploitation*

We want algorithms to learn good decisions quickly in any environment.

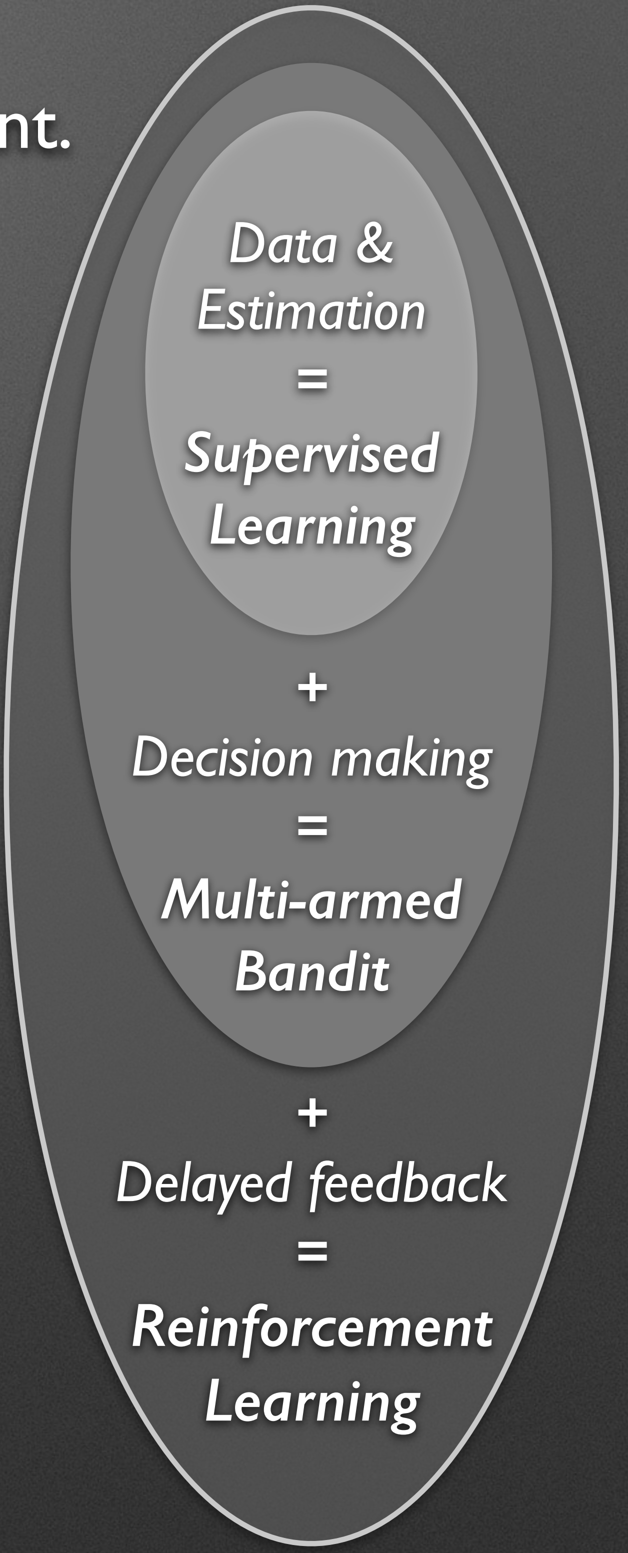
- Measure: $\mathrm{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T}(r_t^* - r_t)\right]$

Rewards of optimal controller *Actual rewards*

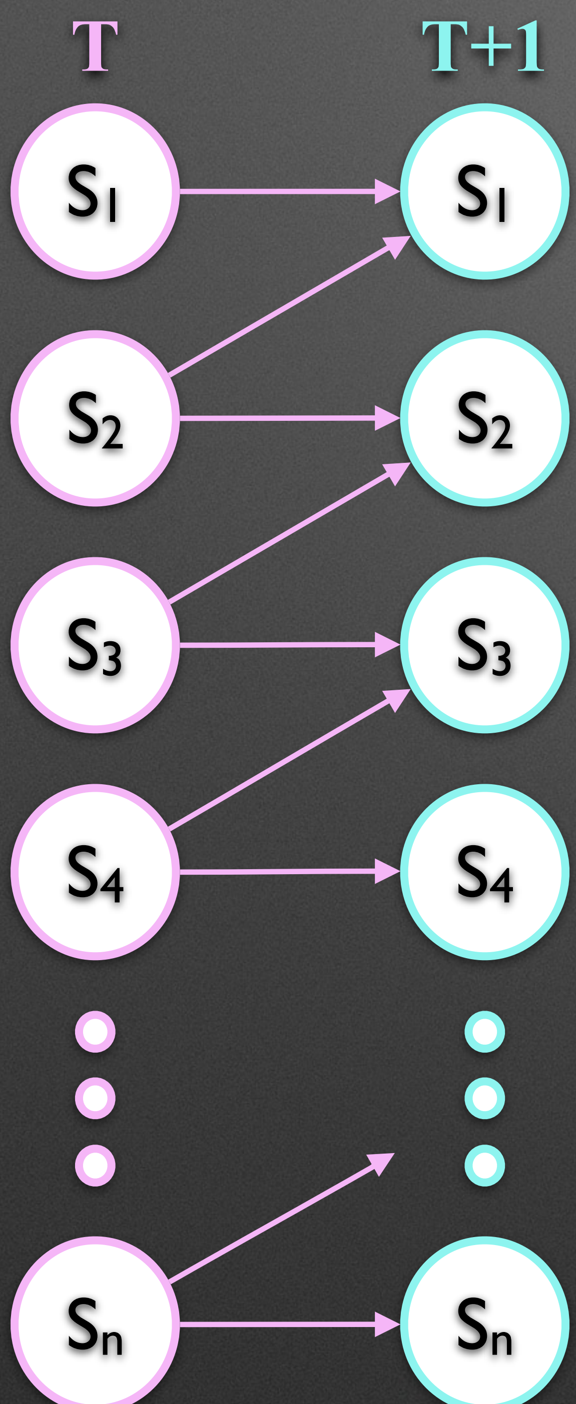
- Theorem: In a general MDP with S states and A actions

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

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



Learning in Factored MDPs



- Key idea: Learn quickly via *low-dimensional structure*.
- Definition: Factored MDP ↔ conditional independence.
- Example: In a production line, each machine's state depends directly only on its neighbors.

Our regret bounds scale with number of parameters rather than number of states.

- Algorithms: *Optimism* and *Posterior Sampling*.
- Bounds: For K independent segments of an MDP

Naive bounds: Exponential in K → *New bounds: Linear in K*