Recovering Bandits

Jefferson Huang

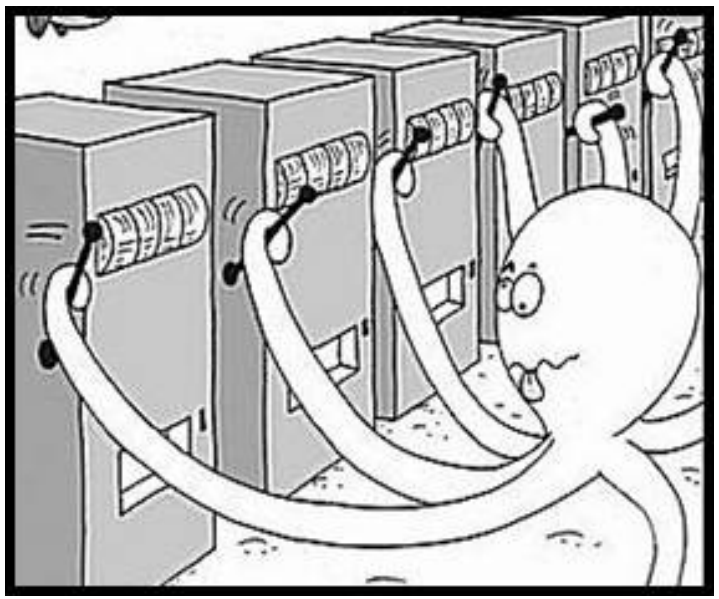
Department of Applied Mathematics and Statistics
Stony Brook University

Algorithms Reading Group
February 26, 2016

Multi-armed bandits (MABs)

A model of sequential decision-making:

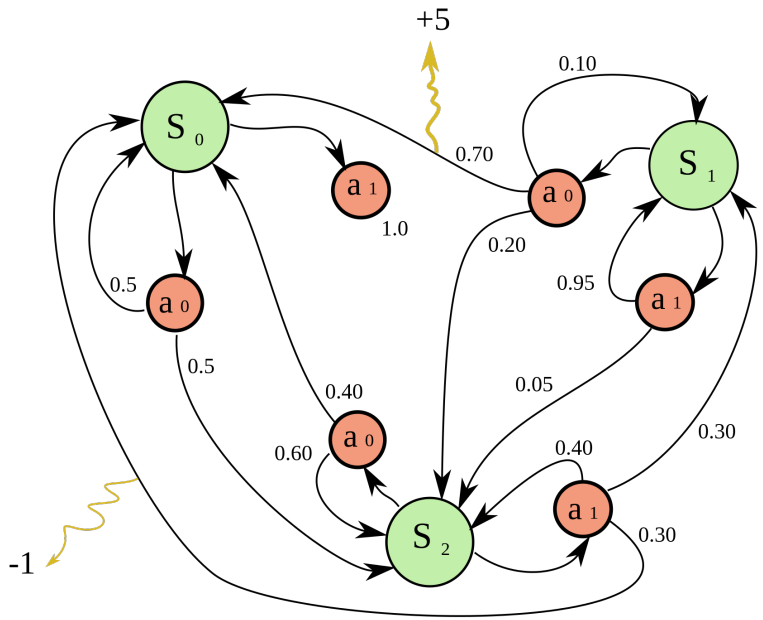
- **Given:** set of K Markov chains.
- Select a chain ("arm") k to "play".
- Chain k undergoes a state transition; the other chains don't move.
- If chain k transitions from state i to state j, a reward $\rho(i, j)$ is earned.
- Select another chain ...



Markov decision processes (MDPs)

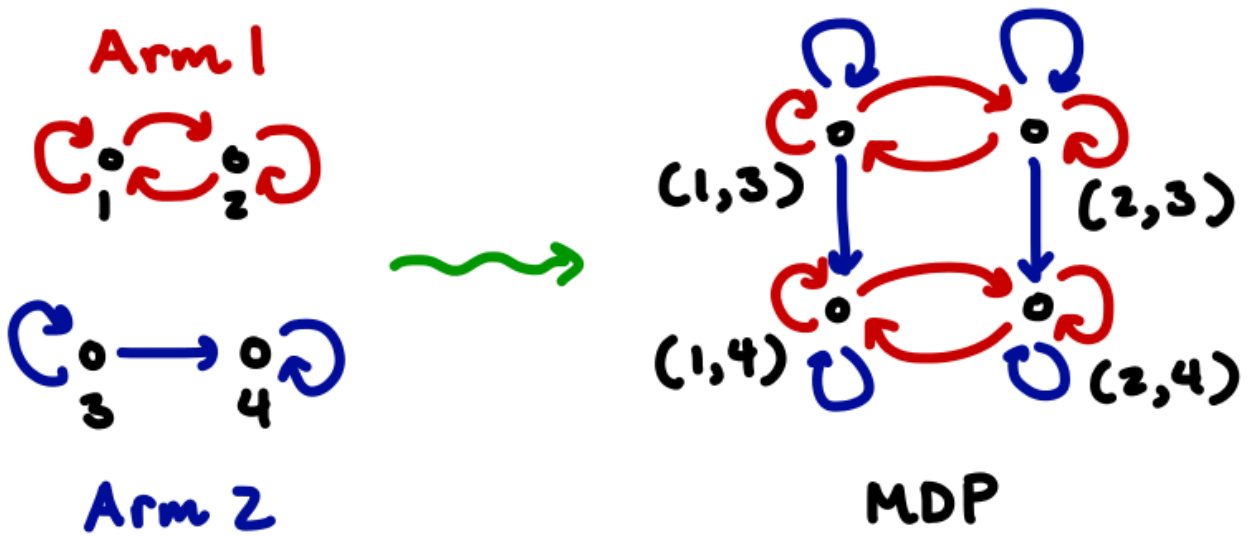
A model of sequential decision-making:

- **Given:** $(\mathcal{S}, \{\mathcal{A}_s\}, r, p)$
- When the system is in state $s \in \mathcal{S}$, select an action $a \in \mathcal{A}_s$.
- A reward $r(s, a)$ is earned.
- The next state is t with probability $p(t|s, a)$.
- Select another action ...



MDP formluation of a multi-armed bandit

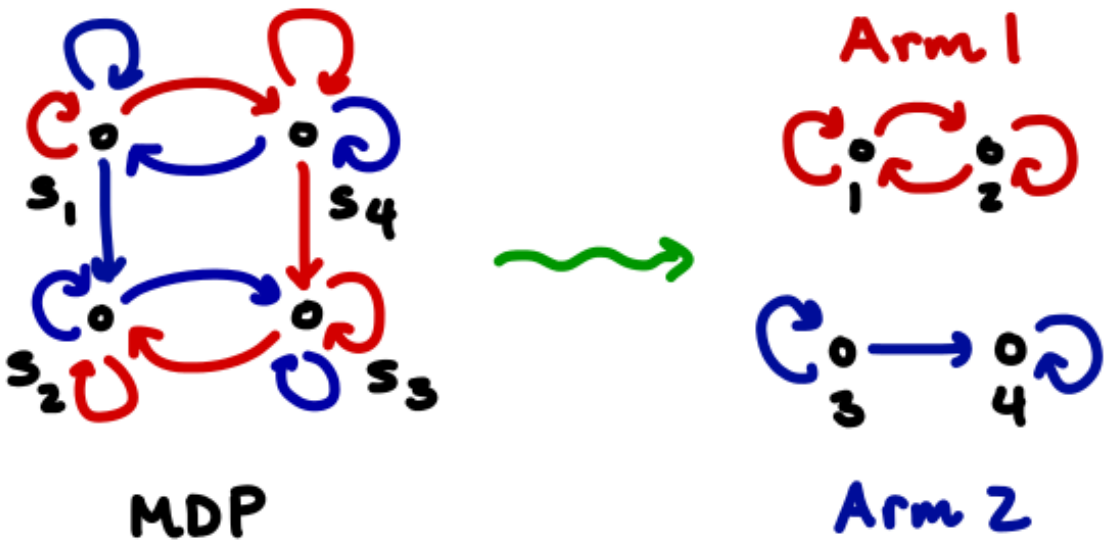
- States: $s = (i_1, \ldots, i_K)$, i_k = state of arm k.
- Actions: $a = 1, \ldots, K$, i.e. which arm to play.
- Transition probabilities & rewards defined according to the arms' transition probabilities & rewards.



Problem

- **Given:** MDP $(\mathcal{S}, \{\mathcal{A}_s\}, r, p)$ generated by a MAB.
- **Question:**

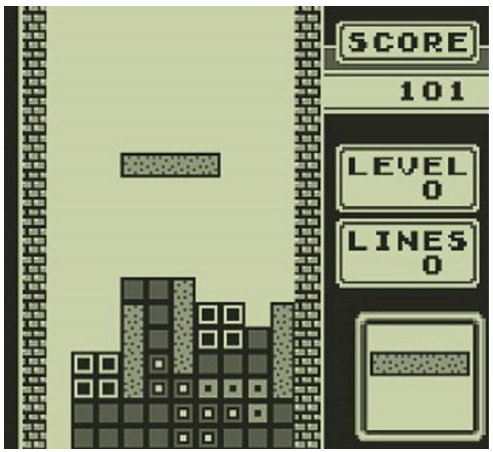
How hard is it to recover the underlying MAB?



Motivation

Benchmarking MDP solution algorithms:

- It can be hard to tell how well an algorithm is doing.
- Example: Tetris.



$\sim 2^{200}$ states

1996: average score $\sim$1,000

2006, 2009: average scores $\sim$900,000

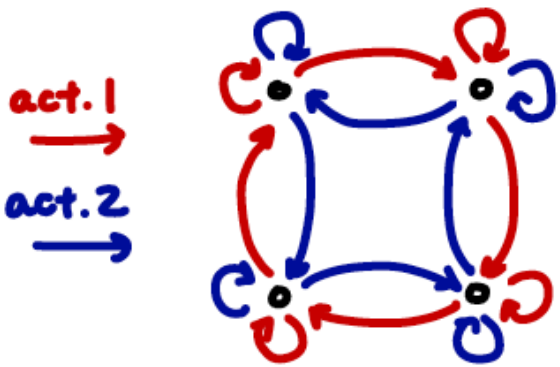
- **Want:** Automatically generate large, difficult-to-solve MDPs whose optimal performance levels are known.

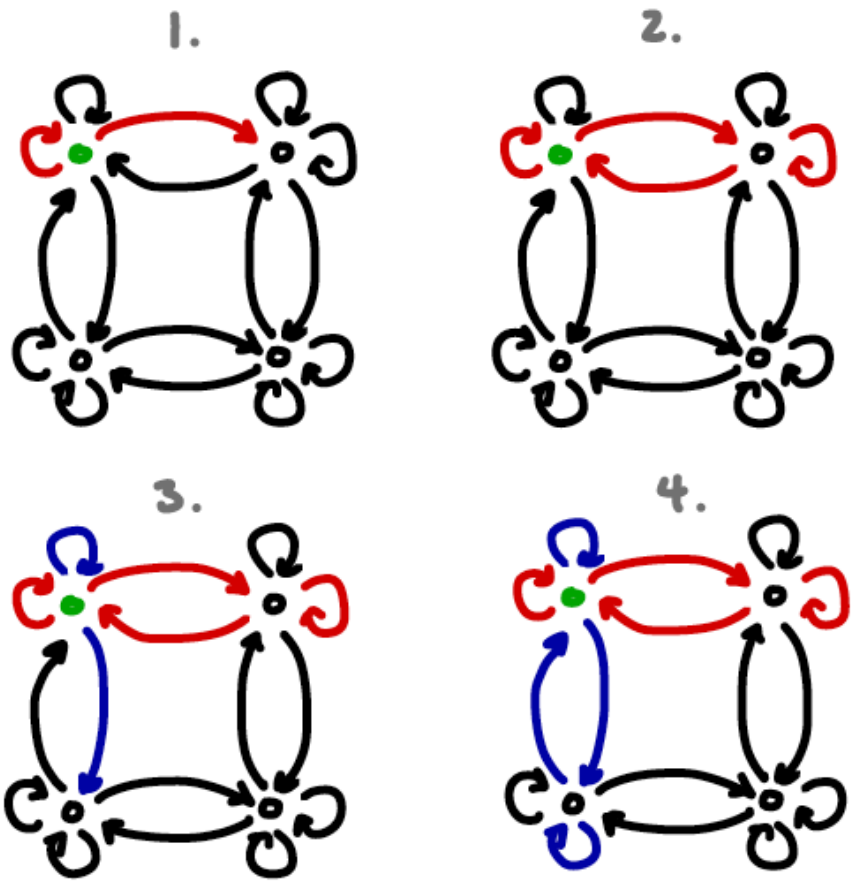
Motivation

- **Idea:** Use MABs.
- Lead to large MDPs:
  - MAB with $K = 100$ arms, where arm k has $N = 10$ states $\implies$ MDP with $N^K = 10^{100}$ states, 100 actions per state.
- Optimal policies can be computed efficiently:
  - Denardo, Feinberg, Rothblum (2013): $O(N^3 K^3)$ arithmetic operations suffice to:
    - compute an optimal policy for the MAB;
    - compute the optimal value earned from a given initial state of the associated MDP.
- **Want:** Ensure the MDP is still difficult to solve.

Easy case: complete transition graphs

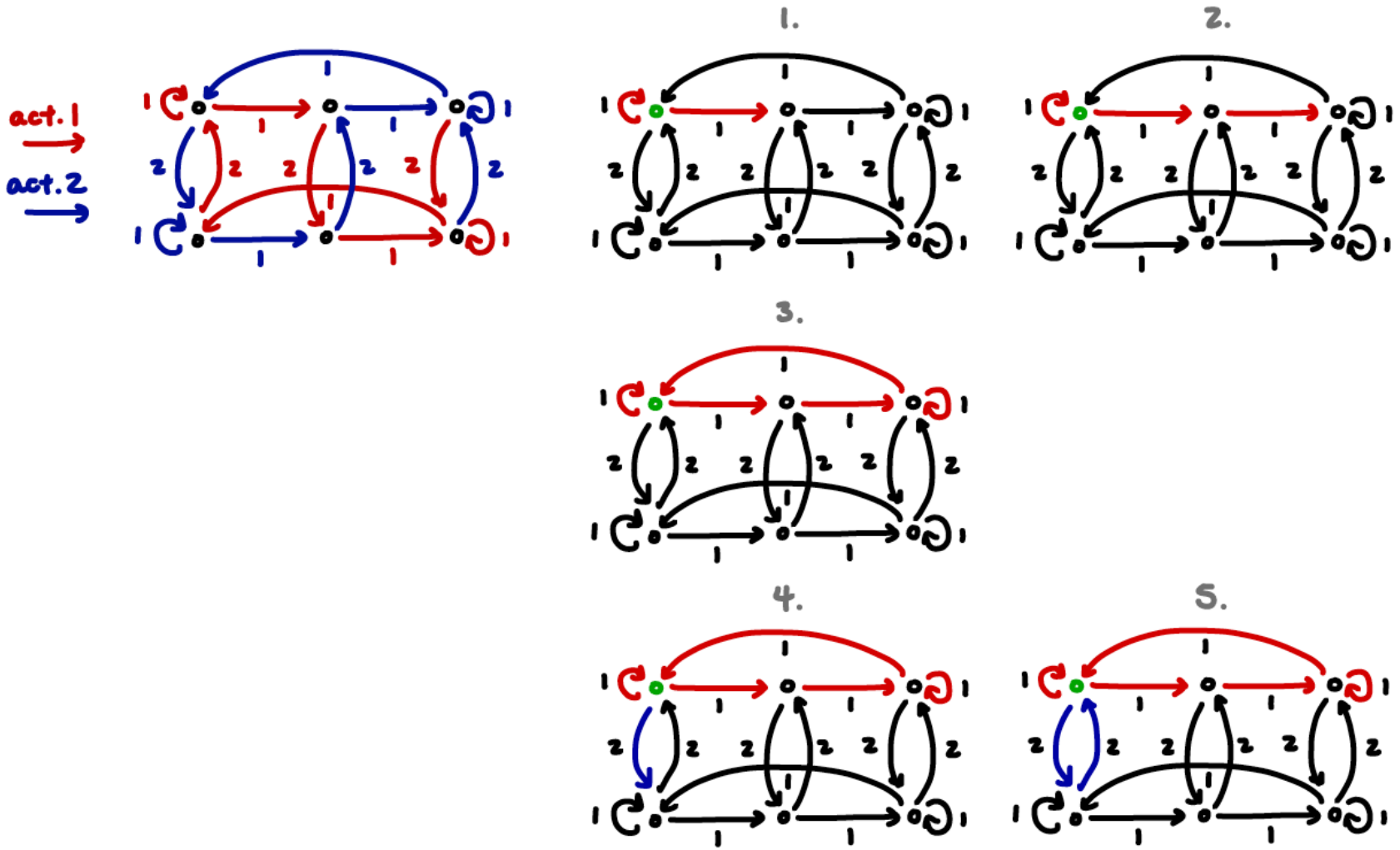
Use transitions to keep track of the current arm:





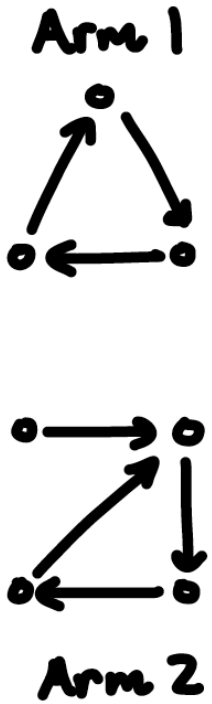
Communicating transition graphs

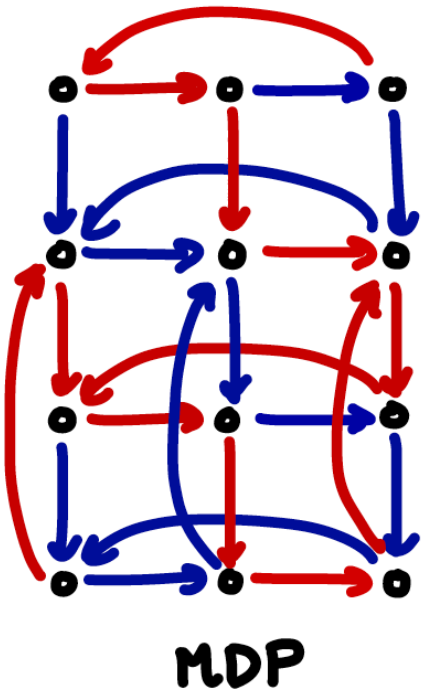
Easy when rewards identify the arm:



Deterministic arms

How to keep track of the current arm?





Other questions

- What other conditions make recovery easy?
- Is NP-hardness relevant to the recovery problem?
- Given a MDP, is it hard in general to decide whether it models a MAB?

Contact email:
jefferson.huang@stonybrook.edu