New Topic: Infinite Horizon MDPs

All of the dynamic programs we have discussed so far have involved a **finite number of stages**

Infinite horizon dynamic programs represent an unbounded decision process

We will discuss the following aspects of infinite horizon MDPs:

- Conditions under which they are well defined
- Optimality equations
- Solution methods
- Applications

Motivation

Why study infinite horizon MDPs?

- Sometimes they are **easier to solve** numerically than finite horizon counterparts
- In some cases the optimal policy can be deduced without solving the MDP

Puterman Chapter 5 is an introduction to infinite horizon MDPs

Assumptions

We will assume the following when discussing infinite horizon MDPs:

- The set of actions, A, and states, S, are finite
- The decision maker's goal can be represented by additive rewards
- The rewards, $r(s,a)$, are bounded, i.e., $r(s,a) \leq M$ and $\lambda < 1$
- All problem data (transition probability matrix, rewards, state and action sets) are **stationary**

Total Expected Discounted Rewards

Given some policy π the total expected discounted reward for a finite horizon MDP is:

state action

$$u_\lambda^\pi(s) = \mathrm{E}_s^\pi\left[\sum_{t=1}^{N} \lambda^{t-1} r_t(X_t, Y_t)\right]$$

current state

optional discount factor

For an infinite horizon MDP:

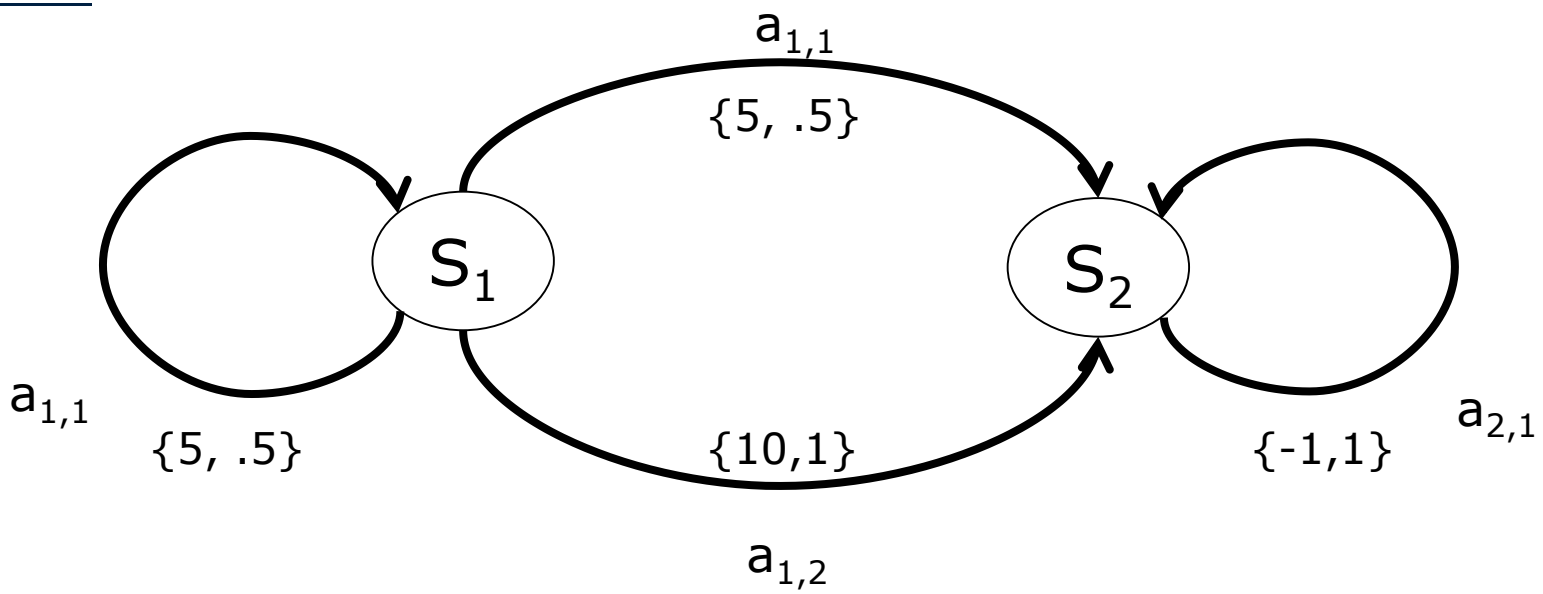
$$u_\lambda^\pi(s) = Lim_{N\to\infty} E_s^\pi\left[\sum_{t=1}^{N} \lambda^{t-1} r_t(X_t, Y_t)\right]$$

where $0 < \lambda < 1$ is required. The limit exists if

$$\max_{s\in S, a\in A}\{r(s,a)\} = M < \infty$$

Example: 2 State MDP

In state S_1 actions $a_{1,1}$ and $a_{1,2}$ are available; in state S_2 only $a_{2,1}$ is available. Rewards and transition probabilities are defined below as {r,p}. At each stage the associated reward is received and then the transition occurs.



Exercise: Find $u_\lambda^\pi(s)$ given policy (a) $d^a(s_1) = a_{1,2}$ and $d^a(s_2) = a_{2,1}$ and policy (b) $\mathrm{d}^\mathrm{b}(s_1) = a_{1,1}$ and $d^b(s_2) = a_{2,1}$

Vector Notation

Given some decision rule $d(s)$ that is applied over an infinite horizon, let $r_d(s) \equiv r(s, d(s))$ and $\mathrm{p_d(j|s)} \equiv p(j|s, d(s))$

Let $\mathrm{r_d}$ denote the $|S|$ vector with $\mathrm{s^{th}}$ component $\mathrm{r_d(s)}$, referred to as the **reward vector**

Let $\mathrm{P_d}$ be the $|\mathrm{S}| \times |S|$ **transition probability matrix** with $(\mathrm{s}, j)^{\mathrm{th}}$ entry $\mathrm{p_d(j|s)}$.

The vector of expected values of the policy for each state is:

$$u_\lambda^\pi = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} r_d$$

Policy Evaluation

Assuming a **stationary policy** $\pi = (d, d, \dots)$, the expected value can be expressed as:

$$u_\lambda^\pi = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} r_d$$

$$= r_d + \lambda P_d r_d + \lambda^2 P_d P_d r_d + \cdots$$

$$= r_d + \lambda P_d \left(r_d + \lambda P_d r_d + \lambda^2 P_d^2 r_d \ + \cdots \right)$$

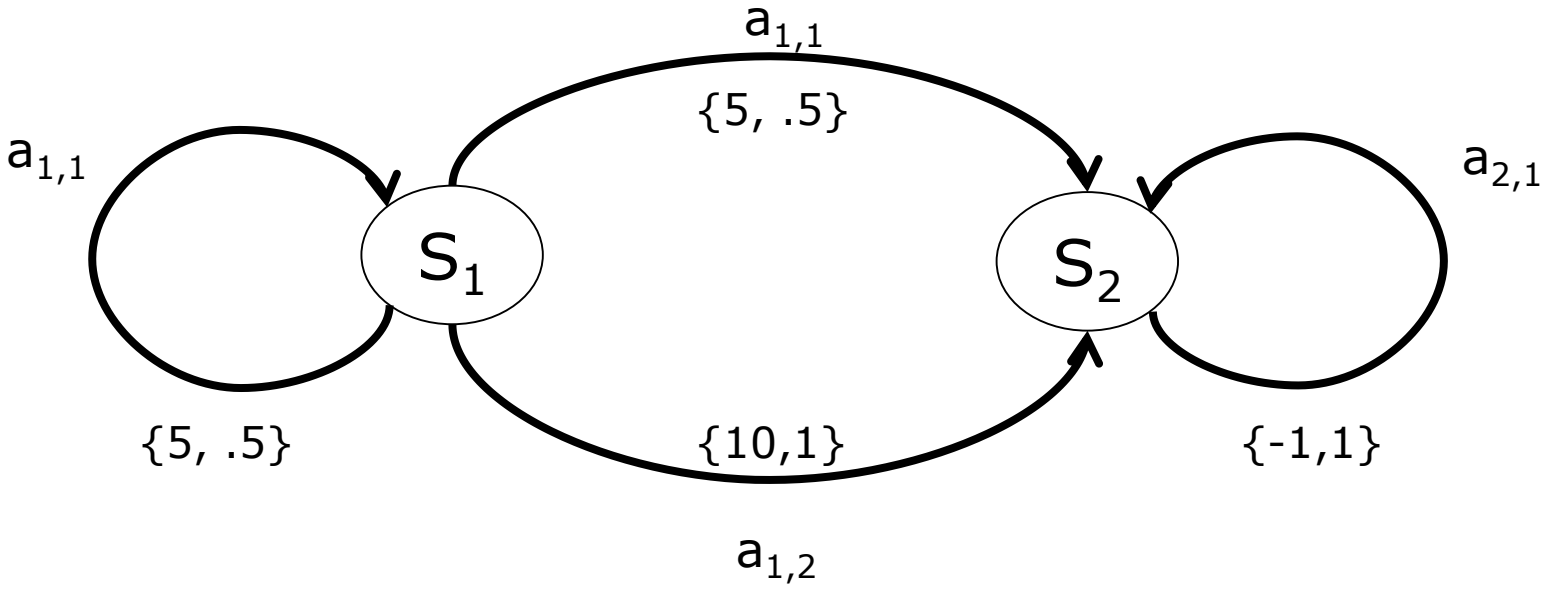
$$= r_d + \lambda P_d u_\lambda^\pi$$

Thus the expected value of a policy is

$$u_\lambda^\pi = (I - \lambda P_d)^{-1} r_d$$

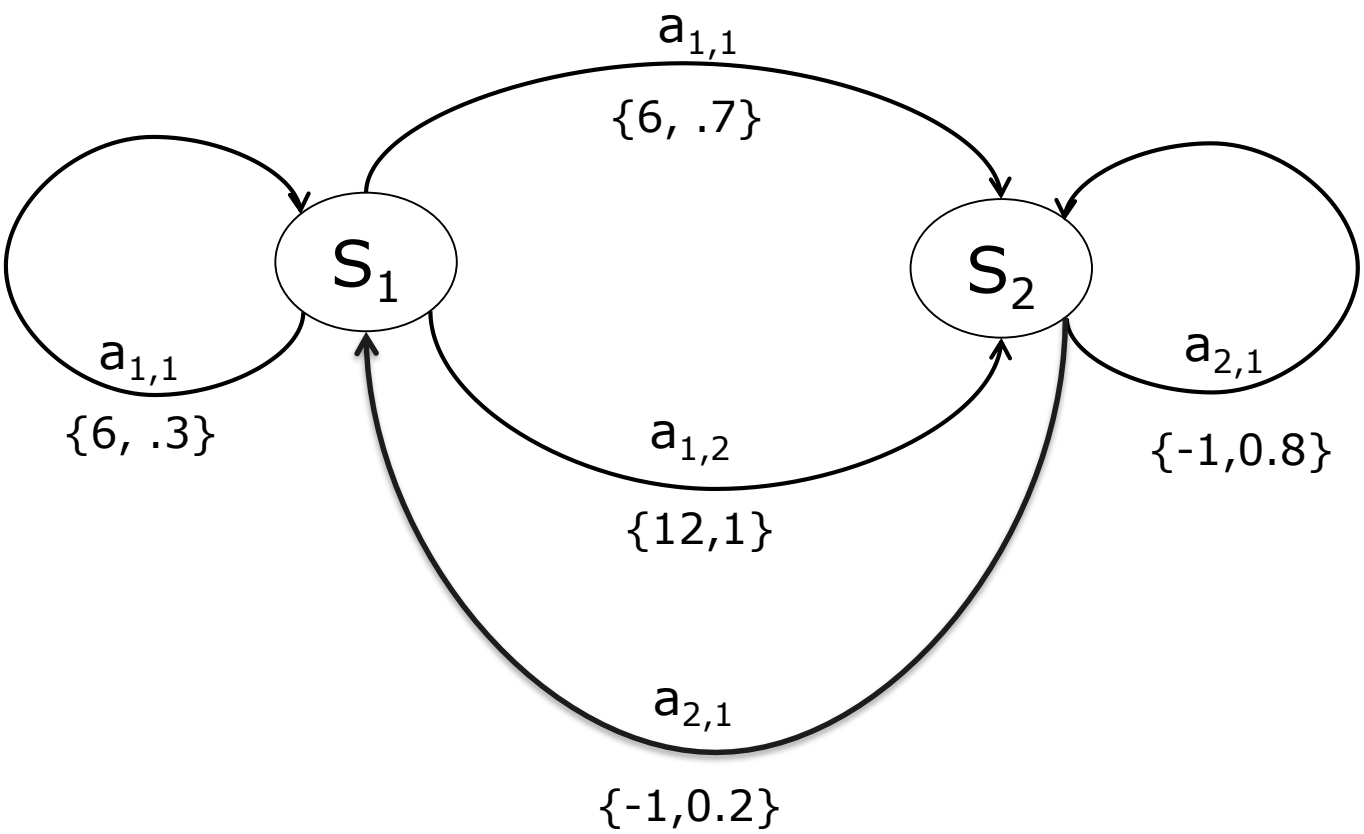
Example: 2 State MDP

In state S_1 actions $a_{1,1}$ and $a_{1,2}$ are available; in state S_2 only $a_{2,1}$ is available. Rewards and transition probabilities are defined below as {r,p} At each stage the associated reward is received and then the transition occurs.



Exercise: Find $u_\lambda^\pi(s)$ for policies (a) $d^a(s_1) = a_{1,2}$ and $d^a(s_2) = a_{2,1}$ and (b) $\mathrm{d}^b(s_1) = a_{1,1}$ and $d^b(s_2) = a_{2,1}$ using: $u_\lambda^\pi = (I - \lambda P_d)^{-1} r_d$

Example: In-Class Assignment



Exercise: Find $u_\lambda^\pi(s)$ for all policies.