Markov Decision Processes

CS 188: Section Handout

Defining a Markov Decision Process (MDP)

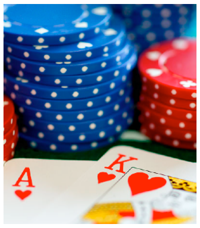
- State Space: $\{S_0, S_1, S_2, ...\}$
- Actions: $\{A_0, A_1, ...\}$
- Initial State: S_0
- Transition Model: $T(s, a, s')$, the probability of going from s to s' with action a.
- Reward Function: $R(s)$, the reward for being in state s.[1]
- Discount Factor: γ, the discount for rewards: a reward r in t steps is worth $r\gamma^t$ now. $0 < \gamma \leq 1$.

A solution to an MDP is called a **policy**, which is a function $\pi(s)$ that maps from states to actions. For a particular policy π, every state has exactly one chosen action.

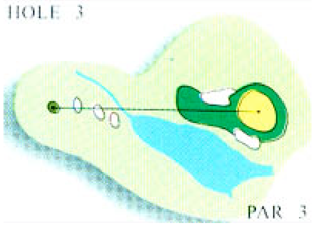
Which of the following are MDPs?

Exercise: For each of the following tasks/games, describe an MDP formulation or state why it is not amenable to the MDP framework.

- Blackjack (21) with no betting
- Rock, Paper, Scissors
- Person trapped in a container ship who can yell for help on sunny days
- Tightrope-walking robot



[1] Sometimes rewards have different structures, such as $R(s, a, s')$: the reward for moving from s to s' via action a.



A Very Simple Example: Golf

- State Space: $\{Tee, Fairway, Sand, Green\}$
- Actions: $\{Conservative, Power\ shot\}$
- Initial State: Tee
- Transition Model: $T(s, a, s')$, the probability of going from s to s' with action a.

s	a	s'	$T(s,a,s')$
Tee	Conservative	Fairway	0.9
Tee	Conservative	Sand	0.1
Tee	Power shot	Green	0.5
Tee	Power shot	Sand	0.5
Fairway	Conservative	Green	0.8
Fairway	Conservative	Sand	0.2
Sand	Conservative	Green	1.0

- Reward Function:

s	$R(s)$
Tee	-1
Fairway	-1
Sand	-2
Green	3

Question: For the *Conservative* policy, what is the utility of being at the *Tee*? What about the *Power shot* policy?

Value iteration: an exact solution to MDPs

The quick and dirty story of value iteration:

- Solves the Bellman equation: $U(s) = R(s) + \gamma \max_a \sum_{s'} T(s,a,s')U(s')$
- Starts with $\hat{U}(s) = 0$ for all s. Iterates through each state many times, updating $\hat{U}(s)$.
- Iteration always converges to the correct answer given infinite time.

Exercise: Compute the value iteration updates for golf.