

---

The preamble is an abbreviation of the lecture notes

# Markov Decision Processes

A Markov Decision Process is defined by several properties:

- A set of states  $S$
- A set of actions  $A$ .
- A start state.
- Possibly one or more terminal states.
- Possibly a **discount factor**  $\gamma$ .
- A **transition function**  $T(s, a, s')$ .
- A **reward function**  $R(s, a, s')$ .

## The Bellman Equation

- $V^*(s)$  – the optimal value of  $s$  is the expected value of the utility an optimally-behaving agent that starts in  $s$  will receive, over the rest of the agent's lifetime.
- $Q^*(s, a)$  - the optimal value of  $(s, a)$  is the expected value of the utility an agent receives after starting in  $s$ , taking  $a$ , and acting optimally henceforth.

Using these two new quantities and the other MDP quantities discussed earlier, the Bellman equation is defined as follows:

$$V^*(s) = \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V^*(s')]$$

We can also define the equation for the optimal value of a q-state (more commonly known as an optimal **q-value**):

$$Q^*(s, a) = \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V^*(s')]$$

which allows us to reexpress the Bellman equation as

$$V^*(s) = \max_a Q^*(s, a).$$

# Value Iteration

The **time-limited value** for a state  $s$  with a time-limit of  $k$  timesteps is denoted  $V_k(s)$ , and represents the maximum expected utility attainable from  $s$  given that the Markov decision process under consideration terminates in  $k$  timesteps. Equivalently, this is what a depth- $k$  expectimax run on the search tree for a MDP returns.

**Value iteration** is a **dynamic programming algorithm** that uses an iteratively longer time limit to compute time-limited values until convergence (that is, until the  $V$  values are the same for each state as they were in the past iteration:  $\forall s, V_{k+1}(s) = V_k(s)$ ). It operates as follows:

1.  $\forall s \in S$ , initialize  $V_0(s) = 0$ . This should be intuitive, since setting a time limit of 0 timesteps means no actions can be taken before termination, and so no rewards can be acquired.
2. Repeat the following update rule until convergence:

$$\forall s \in S, V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V_k(s')]$$

At iteration  $k$  of value iteration, we use the time-limited values for with limit  $k$  for each state to generate the time-limited values with limit  $(k + 1)$ . In essence, we use computed solutions to subproblems (all the  $V_k(s)$ ) to iteratively build up solutions to larger subproblems (all the  $V_{k+1}(s)$ ); this is what makes value iteration a dynamic programming algorithm.

# 1 MDPs: Micro-Blackjack

In micro-blackjack, you repeatedly draw a card (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw or Stop if the total score of the cards you have drawn is less than 6. If your total score is 6 or higher, the game ends, and you receive a utility of 0. When you Stop, your utility is equal to your total score (up to 5), and the game ends. When you Draw, you receive no utility. There is no discount ( $\gamma = 1$ ). Let's formulate this problem as an MDP with the following states: 0, 2, 3, 4, 5 and a *Done* state, for when the game ends.

(a) What is the transition function and the reward function for this MDP?

(b) Fill in the following table of value iteration values for the first 4 iterations.

States	0	2	3	4	5
$V_0$					
$V_1$					
$V_2$					
$V_3$					
$V_4$					

(c) You should have noticed that value iteration converged above. What is the optimal policy for the MDP?

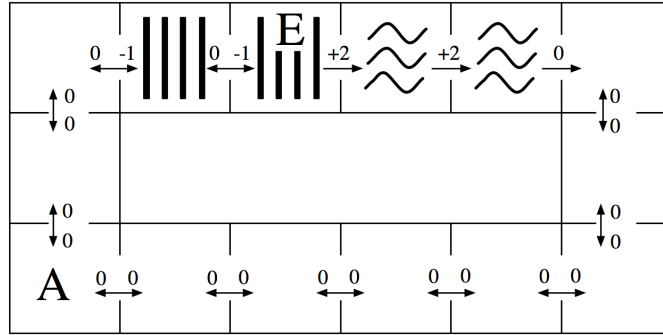
States	0	2	3	4	5
$\pi^*$					

(d) Perform one iteration of policy iteration for one step of this MDP, starting from the fixed policy below:

States	0	2	3	4	5
$\pi_i$	Draw	Stop	Draw	Stop	Draw
$V^{\pi_i}$					
$\pi_{i+1}$					

## 2 MDPs: Grid-World Water Park

Consider the MDP drawn below. The state space consists of all squares in a grid-world water park. There is a single waterslide that is composed of two ladder squares and two slide squares (marked with vertical bars and squiggly lines respectively). An agent in this water park can move from any square to any neighboring square, unless the current square is a slide in which case it must move forward one square along the slide. The actions are denoted by arrows between squares on the map and all deterministically move the agent in the given direction. The agent cannot stand still: it must move on each time step. Rewards are also shown below: the agent feels great pleasure as it slides down the water slide (+2), a certain amount of discomfort as it climbs the rungs of the ladder (-1), and receives rewards of 0 otherwise. The time horizon is infinite; this MDP goes on forever.

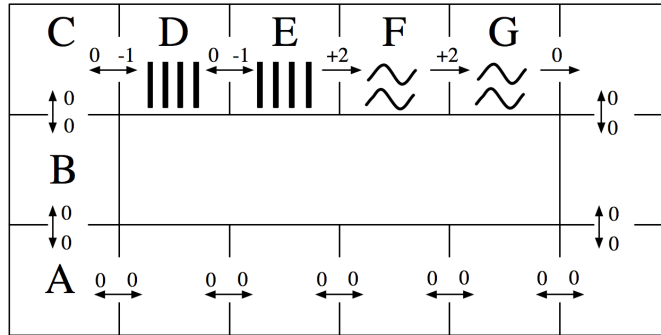


(a) How many (deterministic) policies  $\pi$  are possible for this MDP?

(b) Fill in the blank cells of this table with values that are correct for the corresponding function, discount, and state. *Hint: You should not need to do substantial calculation here.*

	$\gamma$	$s = A$	$s = E$
$V_3^*(s)$	1.0		
$V_{10}^*(s)$	1.0		
$V_{10}^*(s)$	0.1		
$Q_1^*(s, \text{west})$	1.0	—	
$Q_{10}^*(s, \text{west})$	1.0	—	
$V^*(s)$	1.0		
$V^*(s)$	0.1		

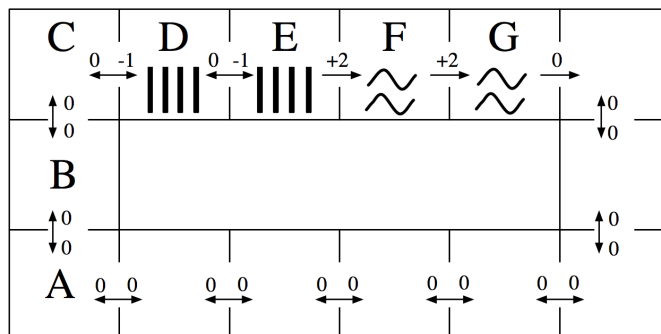
Use this labeling of the state space to complete the remaining subproblems:



- (c) Fill in the blank cells of this table with the Q-values that result from applying the Q-update for the transition specified on each row. You may leave Q-values that are unaffected by the current update blank. Use discount  $\gamma = 1.0$  and learning rate  $\alpha = 0.5$ . Assume all Q-values are initialized to 0. (Note: the specified transitions would not arise from a single episode.)

	$Q(D, \text{west})$	$Q(D, \text{east})$	$Q(E, \text{west})$	$Q(E, \text{east})$
Initial:	0	0	0	0
Transition 1: $(s = D, a = \text{east}, r = -1, s' = E)$				
Transition 2: $(s = E, a = \text{east}, r = +2, s' = F)$				
Transition 3: $(s = E, a = \text{west}, r = 0, s' = D)$				
Transition 4: $(s = D, a = \text{east}, r = -1, s' = E)$				

The agent is still at the water park MDP, but now we're going to use function approximation to represent Q-values. Recall that a policy  $\pi$  is *greedy* with respect to a set of Q-values as long as  $\forall a, s \ Q(s, \pi(s)) \geq Q(s, a)$  (so ties may be broken in any way).



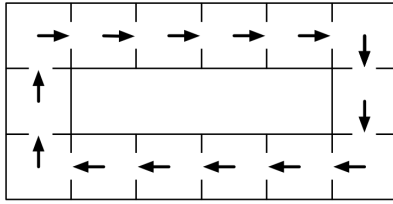
For the next subproblem, consider the following feature functions:

$$f(s, a) = \begin{cases} 1 & \text{if } a = \text{east,} \\ 0 & \text{otherwise.} \end{cases}$$

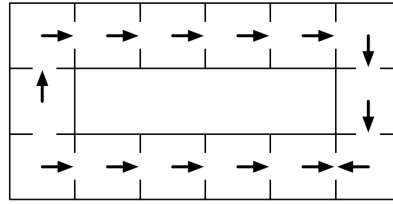
$$f'(s, a) = \begin{cases} 1 & \text{if } (a = \text{east}) \wedge \text{isSlide}(s), \\ 0 & \text{otherwise.} \end{cases}$$

(Note:  $\text{isSlide}(s)$  is true iff the state  $s$  is a slide square, i.e. either  $F$  or  $G$ .)

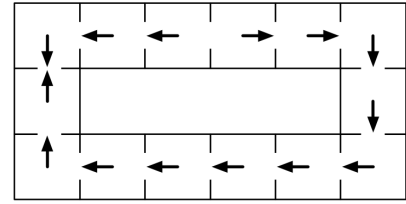
Also consider the following policies:



$\pi_1$



$\pi_2$



$\pi_3$

- (d) Which are greedy policies with respect to the Q-value approximation function obtained by running the single Q-update for the transition  $(s = F, a = \text{east}, r = +2, s' = G)$  while using the specified feature function? You may assume that all feature weights are zero before the update. Use discount  $\gamma = 1.0$  and learning rate  $\alpha = 1.0$ . Circle all that apply.

$f$	$\pi_1$	$\pi_2$	$\pi_3$
$f'$	$\pi_1$	$\pi_2$	$\pi_3$