**2/5/18**

Markov decision processes:

* Set of states S
* Start state s0
* Set of actions A
* Transitions P(s’|s,a) (or T(s,a,s’))
* Rewards R(s,a,s’) (and discount γ)

MDP quantities so far:

* Policy = Choice of action for each state
* Utility = sum of (discounted) rewards

The value (utility) of a state s:

* V\*(s) = expected utility starting in s and acting optimally

The value (utility) of a **q-state** (s,a):

* Q\*(s,a) = expected utility starting out having taken action a from state s and (thereafter) acting optimally
* haven’t taken the state yet but have chosen the state
  + up to nature to determine which of the next states you’re going to end up in

The optimal policy:

* π\*(s) = optimal action from state s

Values of States

* Fundamental operation: compute the (expectimax) value of a state
  + Expected utility under optimal action
  + Average sum of (discounted) rewards
  + This is just what expectimax computed!

Bellman Equations

* take correct first action
* keep being optimal
* Definition of “optimal utility” via expectimax recurrence gives a simple onestep lookahead relationship amongst optimal utility values
* characterize optimal values