About this class

Back to MDPs

What happens when we don't have complete knowledge of the environment?

Monte-Carlo Methods

Temporal Difference Methods

Function Approximation

An Example

Blackjack: Goal is to obtain cards whose sum is as great as possible without exceeding 21. All face cards count as 10, and an Ace can be worth either 1 or 11.

Game proceeds as follows: two cards are dealt to both the dealer and the player. One of the dealer's cards is facedown and the other one is faceup. If the player immediately has 21, the game is over, with the player winning if the dealer has less than 21, and the game ending in a draw otherwise.

Otherwise, the player continues by choosing whether to *hit* (get another card) or *stick*. If the total exceeds 21 she goes bust and loses. Otherwise when she sticks, the dealer starts playing using a fixed strategy – she sticks on any sum of 17 or greater. IF the dealer goes

bust the player wins, otherwise the winner is determined by who has a sum closer to 21.

Assume cards are dealt from an infinite deck (i.e. with replacement)

Formulation as an MDP:

1. Episodic, undiscounted
2. Rewards of +1 (winning), 0 (draw), -1 (losing)
3. Actions: hit, stick
4. State space: determined by
   (a) Player's current sum (12-21, because player always hits below 12)
   (b) Presence of a *usable* ace (that doesn't have to be counted as 1)
   (c) Dealer's faceup card

   Total of 200 states

Problem: find the value function for a policy that always hits unless the current total is 20 or 21.

Suppose we wanted to apply a dynamic programming method. We would need to figure out all the transition and reward probabiliies! This is not easy to do for a problem like this.

Monte Carlo methods can work with sample episodes alone!

It's easy to generate sample episodes for our Blackjack example.

In first-visit MC, to evaluate a policy π, we repeatedly generate episodes using π, and then store the return achieved following the *first* occurrence of each state in the episode. Then, averaging these over many simulations gives us the expected value of each state under policy π

The Absence of a Transition Model

We now want to estimate *action* values rather than *state* values. So estimate $Q^{\pi}(s, a)$

Problem? If π is deterministic, we'll never learn the values of taking different actions in particular states...

Must maintain exploration. This is sometimes dealt with through the concept of *exploring starts* – randomize over all actions at the first state in each episode.

Somewhat problematic assumption – nature won't always be so kind – but it should work OK for Blackjack