TEXAS
The University of Texas at Austin

TOPIC 2 STOCHASTIC PROGRAMMING

Bandit Problems

- Bandit problems fit nicely with stochastic programming
- We already have the SP homework due tonight
- There will be bandit problems on the RL homework
- Conceptual bandit questions are fair game for the midterm
- Please read Chapter 2 of Sutton and Bartow

Stochastic Programming

- Every problem we have worked on so far has been solvable

- In much of the rest of the semester we will deal with problems that can't actually be solved to optimality
- We must rely on heuristics

Bandits

- One such example is called the multi-armed bandit problem
  - This name comes from the fact that some people call slot machines one-armed bandits – because they steal your money
- Imagine you're at a casino with several slot machines
- They all have the same cost to play but they all have different distributions of payoffs
- You don't know the payoff distributions
- If you can only play a fixed number of times, how should you pick when to play each slot, to maximize the payout?

Bandits

- There are tons of ‘solutions’ to the multi-armed bandit problem
- They almost all rely on some sort of assumptions about the payoff distribution
  - Payoffs follow a normal distribution with unknown mean
  - Beliefs about the mean follow a normal distribution
- How do we know if any of these assumptions are valid?
  - We don’t!!!
- It’s therefore necessary to rely on heuristic solutions

Bandits

- The basic concept of solving the bandit problem is based on the idea of exploration vs exploitation
- We must play all the slots several times to get a sense of each payoff distribution
  - Exploration
- Once we are confident about which one pays the best, we play that one repeatedly
  - Exploitation
- This concept will be very important in reinforcement learning later in the semester

Exploration vs Exploitation

- The concept is simple enough, but the details can be tricky
  - How much exploring should we do before exploiting?
  - Should we ever go back to exploring?
- We'll learn about 2 methods to 'solve' the problem

Bandits

- Since we want to play the slot with the highest expected return, we should keep track of each slot's average payoff when we play them
- One way to attack this problem is to initialize each slot's average payoff as 0 and then each time we pull an arm, pick the one with the highest average so far
  - Ties are broken randomly
  - This is called a greedy algorithm
  - Very little exploration and lots of exploitation

Class Participation

- This greedy algorithm is quite terrible!

- Talk about how you could modify this greedy algorithm to have more exploration

Bandits

- A simple change to the greedy algorithm makes it surprisingly good!
- Each time we pull an arm behave greedily, except with a probability of ϵ, choose the arm randomly
  - This is called an ϵ-greedy algorithm
  - Lots of exploration and exploitation
- Picking ϵ is similar to a hyperparameter tuning problem
- Let's try it out!

Bandits

- Another possibility is to always behave greedily, but change our initialization of distributions
- If we initialize the mean for each slot to be larger than is reasonable then each time we play a slot we'll decrease the mean
- This will naturally lead to exploration at the beginning and exploitation later
- This is called Optimistic Initial Values
- ϵ-greedy methods are widely useful, but OIV is really only useful for bandit problems ☹