Course 1, Module 1
Sequential Decision Making

K-armed bandit review and discussion

Agenda

- Admin 5 mins
- Review/questions 15 mins
- Worksheets 25 mins
- Answer discussion 5 mins

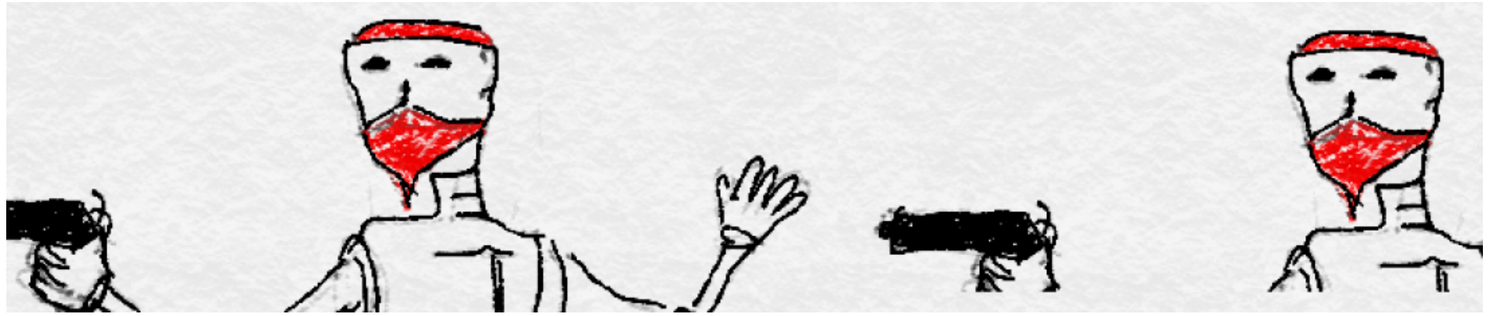
Reminders: Sept 10, 2020

- Schedule with deadlines on github pages (https://marthawhite.github.io/rlcourse/schedule.html)
- Graded Notebook for Course 1, Module 1 (Bandits) due **TODAY**
- Practice Quiz and Discussion Question **due Sunday**, for Course 1, Module 2 (MDPs)
- You should be doing the readings
- Any questions about admin?

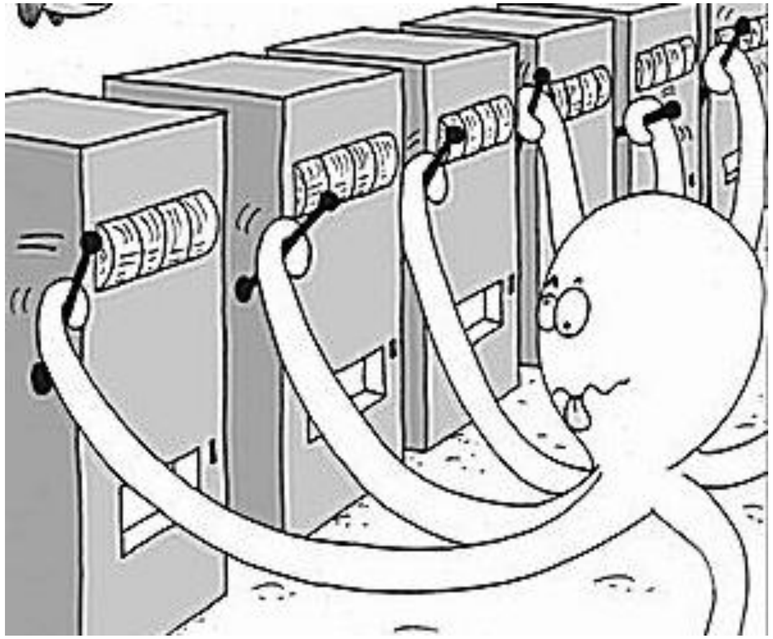
Bad Discussion Questions

1. *email Zero*
2. whis is stepsize means? *Spelling, grammar, ?*
3. When will I be able to plug in my brain to download a better OS?
4. For a stationary problem, what is a possible disadva *Incomplete*
5. In RL you are not told which arm is the best. In supervised learning you have the correct answer and the program modifies it's output based on if its estimate of the correct choice is correct or not. *No question or topic of discussion*
6. Will we have access to the lecture slides notes prior to the actual lecture? *Admin, ask on eclass*
7. The second question of the quiz, I tried to plug the coordinate into the formula, however I can only get 1/7 which closes to 1/8, and I cannot get constant stepsize in that question. *Ask on eclass or go to office hours for help*
8. In Assignment 1, what's thew different between updating self.q_values first and updating current_action first? *Ask on eclass or go to office hours for help*
9. how to identify the estimat is updated by the prediction error?(ex. 1/2 or 1/(t-1))? *Ask on eclass or go to office hours for help*

Quick Review of Bandits



banditalgs.com



Microsoft Research: http://slivkins.com/work/bandits-svc/

“Bandit Algorithms”
by Tor Lattimore and Csaba Szepesvári (page 9)

Review of Course 1, Module 1

- Each week we will give you a chance to ask questions about each topic/video.
- We will not go over the content in the lecture; this is to allow for the questions you would usually ask during lecture.

Video 1: The K-Armed Bandit Problem

- Formalized the problem of decision making under uncertainty using **K-armed bandits.**

- Used this bandit problem to describe fundamental concepts in reinforcement learning, such as **rewards**, **time steps**, and **values (q*).**

Video 2: Estimating Action Values

- Discussed a method for estimating the action-values, called the **sample-average method**.
- Described **greedy** action-selection.
- Introduced the **exploration-exploitation** dilemma in reinforcement learning.

$$Q_T(a) = \frac{\textbf{Sum of Rewards when a was taken}}{\textbf{Number of times a was taken}} = \frac{\sum_{t \in \tau_a} R_t}{N_a}$$

Video 3: Estimating Action Values Incrementally

- Described how action values can be estimated incrementally.
- Identified how the incremental update rule is an instance of a more general learning rule.
- Described how the general learning rule can be used in non-stationary problems.

$$Q_n(a) = Q_n(a) + \frac{1}{n}(R_n - Q_n(a))$$

Video 4-6: The Exploration-Exploitation Trade-off and Exploration Methods

- Defined the exploration-exploitation tradeoff.
- Defined **epsilon-greedy**, as a simple method to balance exploration and exploitation.
- Discussed how optimistic initial values encourage early exploration.
- Described some of the limitations of optimistic initial values as an exploration mechanism.
- Discussed how upper confidence bound action-selection uses uncertainty in the estimates to drive exploration.

Video 4-6: The Exploration-Exploitation Trade-off and Exploration Methods

$$A_t = argmax_{a \in \mathcal{A}} \left[Q_t(a) + c\sqrt{\frac{\ln(t)}{N_a}} \right]$$

Questions:

- If initial estimates of Q(a) = 0, for all a, after first time step if we get a negative reward, will we consider this choice as a greedy choice?
- When we talked about optimistic values, we said that the max reward has to be larger than the actual rewards. So why can't we set the reward to +infinity?
- What should be the max value for the reward? Does the max value affect anything?
- For epsilon-greedy rules, if the epsilon path is taken, is there still a chance to randomly select the greedy action or is the greedy action excluded?
- The lecture said that epsilon 0.1 plateaus after 300 steps while 0.01 improves over time. Why does quiz state that epsilon 0.1 does better than 0.01 over 1000?

Questions:

- What would happen if we used Pessimistic Initial Values, say -5? Would the agent be stuck with whichever action it randomly picked first?
- I keep seeing * in the expected reward function $q^*(a) = E [R | A = a]$. I am not sure if * stands for optimal?

Questions:

- When tracking a non-stationary problem, what is the intuition of using a step size parameter?
- How do we set hyper-parameters? (i.e. α, ϵ, c, etc…)
- Optimistic Initial Values: How do we set the initial estimate values when we don't know what the reward values are?
- In a video, there's a graph comparing an optimistic initial value method with an ε-greedy method to show the former is doing better, but why not combine them?

Questions:

- When we talked about MDPs we said that the agent can be any thing like a wheel controller. Can we have a top level agent and 100's of other agents all talking to the top level agent? I.e in a self driving car one agent checks for lanes another for cars and then a higher agent checks for crossings?

Worksheets

- We will split up into breakout rooms w/ 1 TA per room
- Focus on questions 2 and 3
- You are free to use these rooms or make your own sessions in google meet (or whatever software you wish) to discuss in small groups (TAs will only stay in their breakout rooms)
- We will be back at 1:45 to briefly discuss solutions