Course 1, Module 1
Sequential Decision Making

K-armed bandit review and discussion

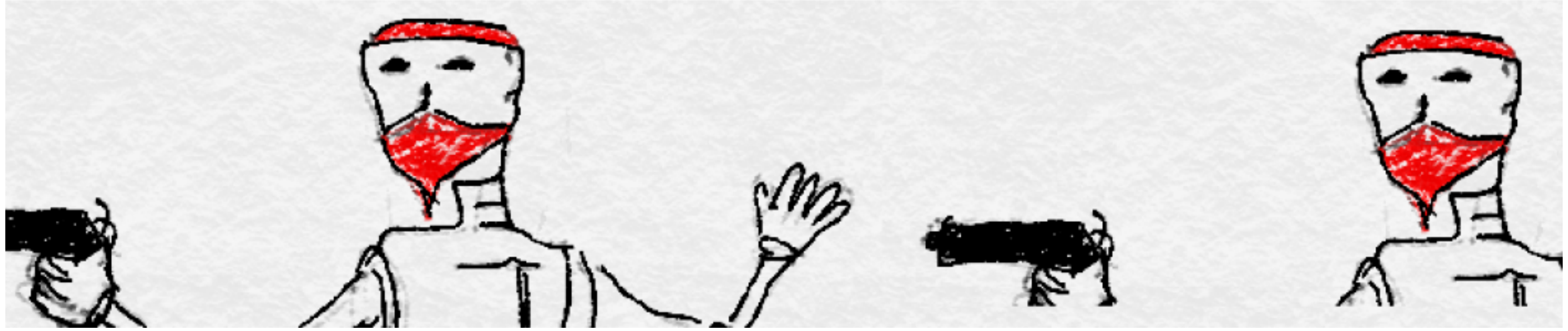
Agenda

- Admin 5 mins
- Review/questions 20 mins
- Lab session with TAs 25 mins

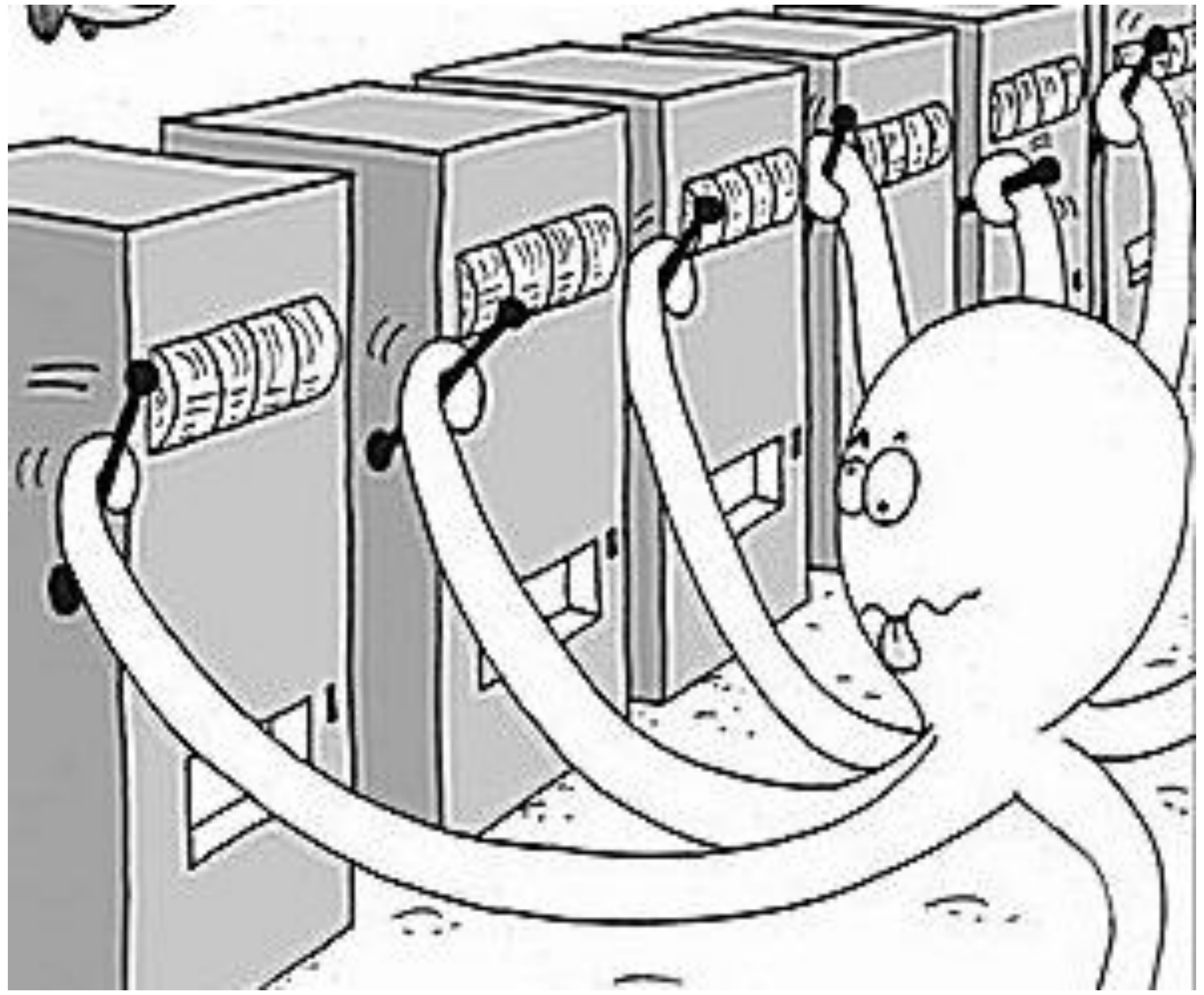
Reminders: Sept 8, 2021

- Schedule with deadlines on github pages (https://docs.google.com/spreadsheets/d/1ooFqttGCklw7rsst9xwL77_SA84LszLvZwpWo06Ltas)
- Graded Notebook for Course 1, Module 1 (Bandits) due **Friday Noon**
- Next practice Quiz **due Sunday**, for Course 1, Module 2 (MDPs)
- You should be doing the readings
- TAs have posted office hours (likely different times next week). Over zoom/meet for now
- Any questions about admin?

Quick Review of Bandits



banditalgs.com



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

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

Demo of Bandits

- https://www.coursera.org/learn/fundamentals-of-reinforcement-learning/ungradedWidget/44Z9R/lets-play-a-game
- https://www.coursera.org/learn/fundamentals-of-reinforcement-learning/ungradedWidget/jEYTO/whats-underneath

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

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

In class questions

- In future lectures these questions will come from Discord
- This week I will review questions from last year that may be helpful to you
- Feel free to put up your hand and ask additional questions

Questions:

- When we talked about optimistic values, we said that the max value 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 choice 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 equation q*(a) = E [R | A = a]. I am not sure if * stands for optimal?
  - In bandits we define two key things:
    - the **true action-value function q***: this defines the problem mathematically
    - The **agents estimate** of the true action-value function which we call **Q**
  - **q*** is the true expected value of the rewards generated by each arm.
  - **Q** is something the agent updates from data as it chooses arms and gets rewards

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?

Lab time

- Raise your hand and the TA can come help you
- Feel free to discuss with classmates
  - No pair coding!