AA228/CS238 Week 7 Problem Session (2/22/23)

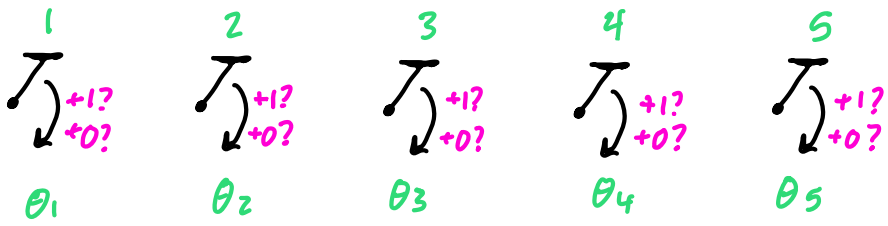
Bandits and Beliefs

Plan for today:

1. Multiarmed bandits
2. Discrete state filter

Multiarmed Bandits

- Common model to study the tradeoff b/w exploration and exploitation.



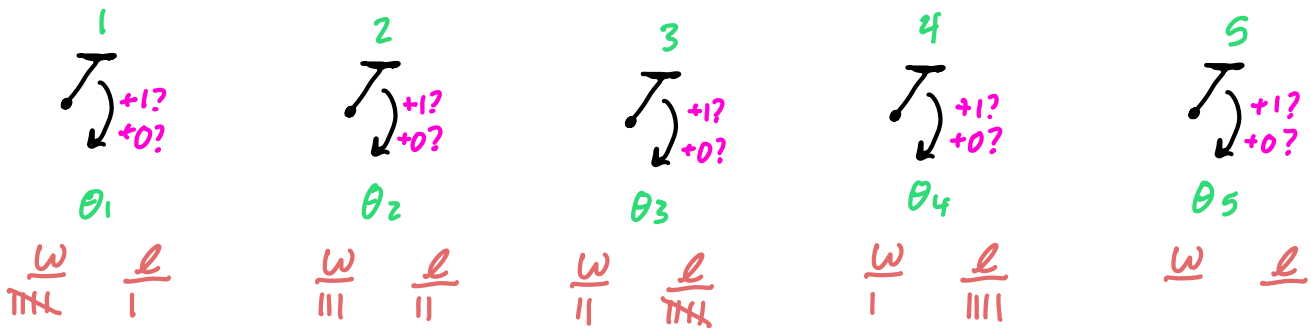
Goal: accumulate as much reward (as many wins) as possible

Challenge: we do not know the θ_a's!

$\Rightarrow$ we can only estimate a belief over their values based on the data we have observed.

Uses:

- Clinical trials
- research fund allocation
- recommender systems
  - * which ads to display
  - * which news articles to show



Which arm should we pull next...

if we want to exploit? Arm 1

if we want to explore? Arm 5

in general? Need a strategy to balance the two

Keeping track of our beliefs (Bayesian Approach)

- Assume a Beta distribution prior for each arm
  - ⇒ if uniform, use $\text{Beta}(1,1)$, else $\text{Beta}(\alpha, \beta)$
- Update belief based on # wins (w_a) and # losses (l_a) observed
  - ⇒ if uniform prior, update to $\text{Beta}(w_a+1, l_a+1)$
  - ⇒ in general, update to $\text{Beta}(w_a+\alpha, l_a+\beta)$
- Posterior probability of winning is then

$$p_a = P(\text{win}_a \mid w_a, l_a) = \int_0^1 \theta \, \text{Beta}(\theta \mid w_a+1, l_a+1)\, d\theta = \frac{w_a}{w_a+l_a+2}$$

Note: for a general prior $\text{Beta}(\alpha, \beta)$

$$p_a = P(\text{win}_a \mid w_a, l_a) = \int_0^1 \theta \, \text{Beta}(\theta \mid w_a+\alpha, l_a+\beta)\, d\theta = \frac{w_a+\alpha}{w_a+l_a+\alpha+\beta}$$

Posterior belief over theta given prior of $\text{Beta}(\alpha, \beta)$:

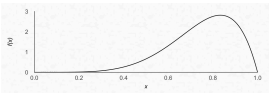
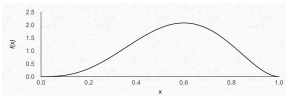
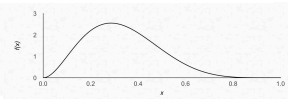
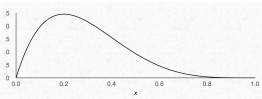
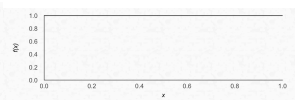
$$\text{Beta}(w_a+\alpha, l_a+\beta)$$

Posterior probability of winning:

$$p_a = \frac{w_a+\alpha}{w_a+l_a+\alpha+\beta}$$

*Assume uniform prior

1	2	3	4	5
+1? +0?	+1? +0?	+1? +0?	+1? +0?	+1? +0?
θ_1	θ_2	θ_3	θ_4	θ_5
w: ~~IIII~~ l: I	w: III l: II	w: II l: ~~IIII~~	w: I l: IIII	w: l:
Beta(5+1, 1+1)	Beta(3+1, 2+1)	Beta(2+1, 5+1)	Beta(1+1, 4+1)	Beta(1, 1)
= Beta(6, 2)	= Beta(4, 3)	= Beta(3, 6)	= Beta(2, 5)	= Beta(1, 1)
$P_1 = \frac{6}{6+2} = 0.75$	$P_2 = \frac{4}{4+3} = 0.57$	$P_3 = \frac{3}{3+6} = 0.33$	$P_4 = \frac{2}{2+5} = 0.29$	$P_5 = \frac{1}{1+1} = 0.5$

Question 1. Multi-Armed Bandits

A student in the Stanford Intelligent Systems Lab has been hard at work trying to write her PhD thesis. To ensure that it is not all left to the last minute, she has adopted the strategy of working on it for at least one hour per day. However, the time of day for her thesis-writing hour varies each day. She has two options:

1. She can write when she first gets to lab in the morning.
2. She can write when she gets home at night.

Each option results in one of the following outcomes:

1. She gets some outlining or 1-2 paragraphs done (reward of 0).
2. She has a burst of inspiration and gets 1-2 pages done (reward of +1).

Working in the morning has some unknown probability of each outcome, while working at night has a different unknown probability of each outcome. She wants to select the time to write each day to maximize her productivity. To procrastinate writing her thesis, she decides to formulate this as a multi-armed bandit problem.

a) What are the "arms" in this bandit problem? What represents a success in this case?

1 (write in morning)

+1? 1-2 pages
+0? 1-2 paragraphs

θ_1

2 (write at night)

+1? 1-2 pages
+0? 1-2 paragraphs

θ_2

b) We want to take a Bayesian approach to modeling our belief over the probability of success for each arm. Let θ_1 represent the probability of success for writing in the morning and θ_2 represent the probability of success for writing at night. How should we represent our belief over these probabilities?

$\theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$ $\quad$ $\theta_2 \sim \text{Beta}(\alpha_2, \beta_2)$

c) If we assume a uniform prior over these outcomes, what should our prior be? What if we believe based on prior knowledge that it is more likely to get 1-2 paragraphs done than to get 1-2 pages. What would a reasonable prior be in this case?

Uniform: Beta(1, 1)

Based on prior knowledge: Beta(1, 5)

d) Assume we start with a Beta(1, 5) prior. We observe data for 4 mornings and 4 nights. One of the nights, the student gets 1-2 pages done. She gets 1-2 paragraphs done in all other writing sessions. What is our posterior belief over the probability of success when writing at in the morning? What about when writing at night?

posterior for arm 1: $\text{Beta}(0+1, 4+5) = \text{Beta}(1, 9)$

posterior for arm 2: $\text{Beta}(1+1, 3+5) = \text{Beta}(2, 8)$

1	2	3	4	5
+1? +0?	+1? +0?	+1? +0?	+1? +0?	+1? +0?
θ_1	θ_2	θ_3	θ_4	θ_5
w: 𝍸 l: I	w: III l: II	w: II l: 𝍸	w: I l: IIII	w: l:
Beta(5+1, 1+1)	Beta(3+1, 2+1)	Beta(2+1, 5+1)	Beta(1+1, 4+1)	Beta(1, 1)
= Beta(6, 2)	= Beta(4, 3)	= Beta(3, 6)	= Beta(2, 5)	= Beta(1, 1)
$P_1 = \frac{6}{6+2} = 0.75$	$P_2 = \frac{4}{4+3} = 0.57$	$P_3 = \frac{3}{3+6} = 0.33$	$P_4 = \frac{2}{2+5} = 0.29$	$P_5 = \frac{1}{1+1} = 0.5$

Given our belief over the probability of success for each arm, which arm should we pull next?

Ad HOC Exploration strategies

ε-greedy: Choose greedy action w/ probability $1-\varepsilon$ and a random arm w/ probability ε

Softmax: Pull arm a w/ probability proportional to $\exp(\lambda P_a)$ where λ is the precision parameter

UCB1: Pull arm that maximizes $P_a + c\sqrt{\frac{\log N}{N(a)}}$

c: exploration parameter N: total # pulls

$N(a)$: # pulls for arm a

Thompson sampling: Sample from each posterior and select action w/ largest sampled value

Question 1. Multi-Armed Bandits

A student in the Stanford Intelligent Systems Lab has been hard at work trying to write her PhD thesis. To ensure that it is not all left to the last minute, she has adopted the strategy of working on it for at least one hour per day. However, the time of day for her thesis-writing hour varies each day. She has two options:

1. She can write when she first gets to lab in the morning.
2. She can write when she gets home at night.

Each option results in one of the following outcomes:

1. She gets some outlining or 1-2 paragraphs done (reward of 0).
2. She has a burst of inspiration and gets 1-2 pages done (reward of +1).

Working in the morning has some unknown probability of each outcome, while working at night has a different unknown probability of each outcome. She wants to select the time to write each day to maximize her productivity. To procrastinate writing her thesis, she decides to formulate this as a multi-armed bandit problem.

e) Assume we are following an ϵ-greedy policy with $\epsilon = 0.2$ to decide when to write tomorrow. Our belief of the probabilities of success is based on the posteriors calculated in part (d). What is the probability that we choose to write in the morning?

posterior for arm 1: $\text{Beta}(0+1, 4+5) = \text{Beta}(1,9)$

posterior for arm 2: $\text{Beta}(1+1, 3+5) = \text{Beta}(2,8)$

calculate posterior probability of success for each arm:

$$p_a = \frac{w_a + \alpha}{w_a + l_a + \alpha + \beta}$$

$$p_1 = \frac{0+1}{0+4+1+5} = \frac{1}{10} = 0.1$$

$$p_2 = \frac{1+1}{1+3+1+5} = \frac{2}{10} = 0.2$$

Greedy Action? arm 2 (night)

Select arm at random w/ probability $\varepsilon = 0.2$

When selecting at random, select morning with 0.5 probability.

$\Rightarrow$ select morning w/ probability $(0.2)(0.5) = \boxed{0.1}$

f) Consider the same scenario as in part (e) but assume instead that we want to use softmax exploration with precision parameter $\lambda = 1$. What is the probability we choose to write in the morning?

Select proportional to $\exp(\lambda \rho_a)$

$\Rightarrow$ normalize to get valid probabilities

$P(\text{morning}) \propto \exp(\rho_1) = \exp(0.1)$

$P(\text{night}) \propto \exp(\rho_2) = \exp(0.2)$

$$P(\text{morning}) = \frac{\exp(0.1)}{\exp(0.2) + \exp(0.1)} = \boxed{0.475}$$

g) Consider the same scenario as in part (e) but assume instead that we want to use the UCB1 strategy for exploration with exploration parameter $c = 1$. What is the probability we choose to write in the morning?

Pull arm that maximizes $\rho_a + c\sqrt{\frac{\log N}{N(a)}}$

$N(1) = 1 + 9 = 10 \quad N(2) = 2 + 8 = 10 \quad N = N(1) + N(2) = 20$

$$\rho_1 + c\sqrt{\frac{\log N}{N(1)}} = 0.1 + (1)\sqrt{\frac{\log 20}{10}} = 0.647$$

$$\rho_2 + c\sqrt{\frac{\log N}{N(2)}} = 0.2 + (1)\sqrt{\frac{\log 20}{10}} = 0.747$$

so UCB1 picks night $\Rightarrow$ prob. selecting morning is 0.0.

Optimal Exploration

Let counts $w_1, l_1, \ldots, w_n, l_n$ represent a belief state.

* Formulate as an MDP where states are counts and actions are which arm to pull.

$Q^*(w_1, l_1, \ldots, w_n, l_n, a)$ is expected payoff for pulling arm a and then acting optimally.

$$U^*(w_1, l_1, \ldots, w_n, l_n) = \max_a Q(w_1, l_1, \ldots, w_n, l_n, a)$$

Can write Q as:

$Q^*(w_1, l_1, \ldots, w_n, l_n, a)$

$= $ (prob. win if pull arm a)(reward if win + utility of new belief state)

$+$ (prob. lose if pull arm a)(reward if lose + utility of new belief state)

$$= \frac{w_a+1}{w_a+l_a+2}\left(1 + U^*(\ldots, w_a+1, l_1, \ldots)\right)$$

$$+ \left(1 - \frac{w_a+1}{w_a+l_a+2}\right) U^*(\ldots, w_a, l_a+1, \ldots)$$

Assume finite horizon h:

- For all states with no pulls left ($\Sigma_a (w_a + l_a) = h$), $U^* = 0$
- work backwards iteratively applying above equation

h) Instead of using the previous ad hoc exploration strategies, we decide to solve for the optimal exploration strategy noting that there is a finite number of days until the thesis is due. Let w_i represent the number of successes when taking action i ($i = 1$ for morning and $i = 2$ for night) and similarly let ℓ_i represent the number of losses. Write down an expression for computing the optimal state-action utility function for the morning action (i.e. $Q^*(w_1, \ell_1, w_2, \ell_2, \text{morning})$). Assume we use the prior from part (d) (Beta(1, 5) for probability of success both in the morning and at night).

$Q^*(w_1, \ell_1, \ldots, w_n, \ell_n, a)$

$=$ (Prob. win if pull arm a)(reward if win + utility of new belief state)

$+$ (Prob. lose if pull arm a)(reward if lose + utility of new belief state)

$$= \frac{w_1+1}{w_1+\ell_1+6}\left(1 + U^*(w_1+1, \ell_1, w_2, \ell_2)\right)$$

$$+ \left(1 - \frac{w_1+1}{w_1+\ell_1+6}\right) U^*(w_1, \ell_1+1, w_2, \ell_2)$$

i) Describe in words how you would calculate the optimal state-action value function if we know that there are 30 days until the thesis is due.

Start w/ all states where $w_1+\ell_1+w_2+\ell_2 = 30$

$\Rightarrow$ assign utility of 0

Apply equation above to all states where $w_1+\ell_1+w_2+\ell_2 = 29$

and continue to work backwards

j) After testing the strategy for a few days, the student realizes there is one more possible outcome of the one hour writing session: she gets distracted by something else and does not make any progress on her thesis for a reward of -2. For each action, we now have three probabilities to estimate. Let $\theta_1 = [\theta_{11}, \theta_{12}, \theta_{13}]$ be the probabilities for each outcome (no progress, 1-2 paragraphs, 1-2 pages respectively) for taking the morning action and let $\theta_2 = [\theta_{21}, \theta_{22}, \theta_{23}]$ be defined similarly. How should we represent our belief over θ_1 and θ_2?

Dirichlet Distribution : $\theta_1 \sim \text{Dir}(\alpha_{11}, \alpha_{12}, \alpha_{13})$

k) Suppose we start with a uniform prior. Let w_{ij} represent the number of times action i resulted in outcome j ($j = 1$ for no progress, $j = 2$ for 1-2 paragraphs, and $j = 3$ for 1-2 pages). Write down the new expression for computing the optimal state-action utility function for the morning action (i.e. $Q^*(w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, \text{morning})$).

$$Q^*(w_{11}, \dots, w_{1N}, \dots, w_{nN}, a)$$

$$= \sum_{i=1}^{N} \sum_{i=1}^{n} (\text{prob outcome } j \text{ if pull arm } i)(\text{reward of outcome } j + \text{utility of new belief state})$$

$$Q^*(w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, \text{morning})$$

$$= \frac{w_{11}+1}{w_{11}+w_{12}+w_{13}+3}\left(-2 + U^*(w_{11}+1, w_{12}, w_{13}, \dots)\right) +$$

$$\frac{w_{12}+1}{w_{11}+w_{12}+w_{13}+3} U^*(w_{11}, w_{12}+1, w_{13}, \dots) +$$

$$\frac{w_{13}+1}{w_{11}+w_{12}+w_{13}+3}\left(1 + U^*(w_{11}, w_{12}, w_{13}+1, \dots)\right)$$

Beliefs

- Probability distribution over the underlying state
- For problems w/ discrete states, can use a belief vector $b \in \mathbb{R}^{|S|}$ which contains the probability of being in each state
- Based on observations of the true state

Question 2. Discrete State Filter

Suppose we have a puppy that we need to care for by making sure we give it proper exercise. When the puppy is feeling restless and wants to run around outside, we should let it go outside. However, we cannot directly tell whether the puppy wants to go outside. Instead, we can only observe whether the puppy scratches at the door. We will model this scenario a POMDP with two states, two actions, and two observations. The state, action, and observation spaces are as follows:

$$\mathcal{S} = \{\text{restless}, \text{relaxed}\}$$

$$\mathcal{A} = \{\text{let outside}, \text{ignore}\}$$

$$\mathcal{O} = \{\text{scratch}, \text{no scratch}\}$$

The transition dynamics are:

$$T(\text{restless} \mid \text{restless}, \text{ignore}) = 1.0$$

$$T(\text{restless} \mid \text{restless}, \text{let outside}) = 0.0$$

$$T(\text{restless} \mid \text{relaxed}, \text{let outside}) = 0.0$$

$$T(\text{restless} \mid \text{relaxed}, \text{ignore}) = 0.25$$

When the puppy is restless, it scratches at the door 90 % of the time. When the puppy is relaxed, it still sometimes scratches at the door (because it sees a squirrel or one of its puppy friends outside). It scratches at the door when relaxed 5 % of the time.

a) Explicitly write down the observation model. In other words, what is $O(\text{scratch} \mid \text{restless}, \text{let outside})$, $O(\text{scratch} \mid \text{restless}, \text{ignore})$, $O(\text{scratch} \mid \text{relaxed}, \text{let outside})$, and $O(\text{scratch} \mid \text{relaxed}, \text{ignore})$.

0.9 0.9 0.05 0.05

Belief Updating

Given our current belief and a new action and observation, how do we update our belief?

Discrete state Filter:

$$
\begin{aligned}
b'(s') &= P(s' \mid b, a, o) \\
&\propto P(o \mid b, a, s') P(s' \mid b, a) \\
&= O(o \mid a, s') P(s' \mid b, a) \\
&= O(o \mid a, s') \sum_s P(s' \mid b, a, s) P(s \mid b, a) \\
&= O(o \mid a, s') \sum_s T(s' \mid s, a) b(s)
\end{aligned}
$$

$$\boxed{b'(s') \propto O(o \mid a, s') \sum_s T(s' \mid s, a) b(s)}$$

* Note, for problems with continuous states, summation becomes an integral $\Rightarrow$ can compute exactly if T and O are linear Gaussian and b is Gaussian (Kalman filter!)

Question 2. Discrete State Filter

Suppose we have a puppy that we need to care for by making sure we give it proper exercise. When the puppy is feeling restless and wants to run around outside, we should let it go outside. However, we cannot directly tell whether the puppy wants to go outside. Instead, we can only observe whether the puppy scratches at the door. We will model this scenario a POMDP with two states, two actions, and two observations. The state, action, and observation spaces are as follows:

$$\mathcal{S} = \{\text{restless}, \text{relaxed}\}$$

$$\mathcal{A} = \{\text{let outside}, \text{ignore}\}$$

$$\mathcal{O} = \{\text{scratch}, \text{no scratch}\}$$

The transition dynamics are:

$$T(\text{restless} \mid \text{restless}, \text{ignore}) = 1.0$$

$$T(\text{restless} \mid \text{restless}, \text{let outside}) = 0.0$$

$$T(\text{restless} \mid \text{relaxed}, \text{let outside}) = 0.0$$

$$T(\text{restless} \mid \text{relaxed}, \text{ignore}) = 0.25$$

When the puppy is restless, it scratches at the door 90 % of the time. When the puppy is relaxed, it still sometimes scratches at the door (because it sees a squirrel or one of its puppy friends outside). It scratches at the door when relaxed 5 % of the time.

b) Suppose we start with a uniform belief over whether the puppy is restless or relaxed. Let the first element of the belief vector represent the probability that the puppy is restless and the second element represent the probability that the puppy is relaxed. What is our initial belief vector **b**?

$b = [0.5, 0.5]$

c) Suppose we do not let the dog out, and we observe that the dog does not scratch at the door. Compute our posterior belief that the dog is restless **up to a proportionality constant**.

$$\boxed{b'(s') \propto O(o \mid a, s') \sum_s T(s' \mid s, a)\, b(s)}$$

$$b'(\text{restless}) \propto O(\text{no scratch} \mid \text{restless}, \text{ignore})\big(T(\text{restless} \mid \text{restless}, \text{ignore})\, b(\text{restless}) + T(\text{restless} \mid \text{relaxed}, \text{ignore})\, b(\text{relaxed})\big)$$

$$= 0.1\,(1.0(0.5) + 0.25(0.5))$$

$$= \boxed{0.0625}$$

d) Compute our posterior belief that the dog is relaxed **up to a proportionality constant**.

$$b'(\text{relaxed}) \propto O(\text{no scratch} \mid \text{relaxed}, \text{ignore}) ($$
$$T(\text{relaxed} \mid \text{restless}, \text{ignore})\, b(\text{restless}) +$$
$$T(\text{relaxed} \mid \text{relaxed}, \text{ignore})\, b(\text{relaxed})$$
$$= 0.95(0.0(0.5) + 0.75(0.5))$$
$$= \boxed{0.3562}$$

e) Based on the results from part (c) and part (d), compute our new belief vector $\mathbf{b}'$.

$$b' = \begin{bmatrix} 0.0625 \\ 0.3562 \end{bmatrix} / (0.0625 + 0.3562)$$

$$\boxed{b' = \begin{bmatrix} 0.1493 \\ 0.8507 \end{bmatrix}}$$