6.4110
Representation, Inference and Reasoning in AI

Quiz 3

Solutions

April 27, 2026

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.

You are permitted to use a single sheet of paper with notes on (both sides). You may not use a calculator. **Box all answers** for free response questions.

Name: _____

MIT email: _____

Question	Points	Score
1	23	
2	20	
3	22	
4	26	
5	9	
Total:	100	

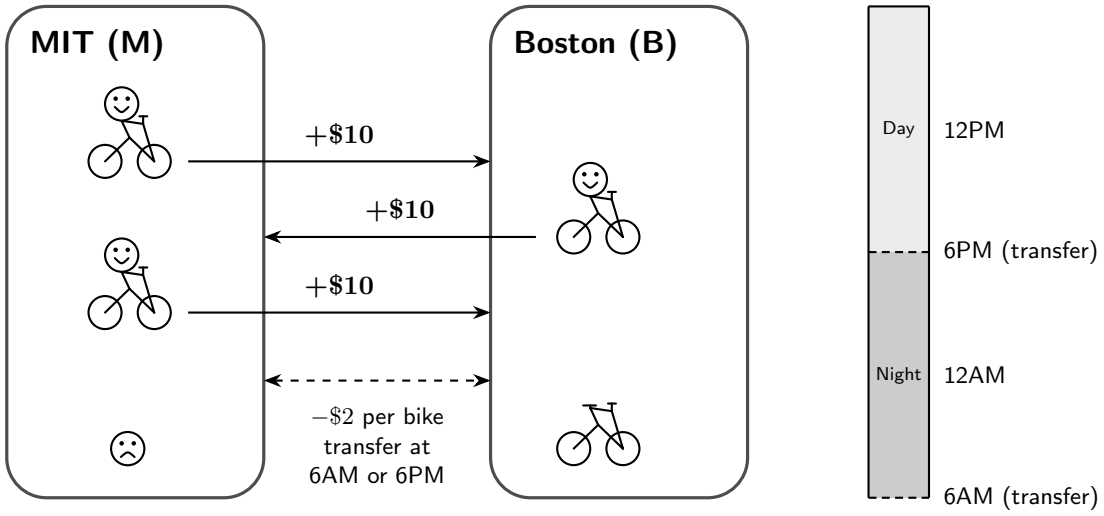
6.4110
Quiz 3, April 27, 2026

1 Blue Bikes

1. You are in charge of a small-city deployment of a bike-share system.

   - There are four (4) bikes.
   - There are two locations: MIT (M) and Boston (B). Each location has a dock that can contain at most four (4) bikes.
   - At 6AM and at 6PM, you can (instantaneously!) move as many bikes as you want between locations.
   - During the day (between 6AM and 6PM), and again during the night (between 6PM and 6AM), some number between 0 and 4 (inclusive) people want to ride between the two locations: the number is drawn from a distribution P_{ijk} where i and j are locations (M or B) and $k = D$ for daytime and $k = N$ for nighttime. So, for example, $P_{MBD}(3)$ is the probability that three people want to ride from MIT to Boston during the day. Each bike can be ridden at most once per day and once per night. No person can bike from MIT to MIT or from Boston to Boston directly.
   - If someone successfully completes a ride, you earn +10. This can only happen if there is a bike available at the passenger's starting location when they want to ride.
   - It costs −2 for the bike-sharing company to move a bike between locations at 6AM or 6PM.

   The diagram below illustrates a sample arrangement (not necessarily describing any of the following problems) in which there are two bikes at MIT and two bikes in Boston. Over the course of a day, three people at MIT wish to travel to Boston, while one person in Boston wants to go to MIT. However, there are only enough bicycles for two of the three people at MIT to ride to Boston since the bike from the person in Boston can not be used again until the next night. During the night, assuming no transfers have taken place, there will be one bike at MIT and three bikes in Boston.



   Your goal is to move bikes around to maximize your profit! Since you're a 6.4110 student, you decide to formulate this as an MDP.

(a) (5 points) What is the smallest state space we could use to represent this problem? How big is it?

Solution:
Since the number of bikes at MIT + the number of bikes at Boston = 4, we only need to store the number of bikes at one location as our state.

Thus, we can define our state space as # of bikes at MIT (or Boston) which can take values $\{0, 1, 2, 3, 4\}$ and thus has size 5.

(b) (5 points) What is the action space (the set of actions that could possibly be taken in any state)?

Solution: The number of actions are the net number of bikes moved from MIT. This can be any of the following numbers: $\{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ thus there are 9 possible actions.

(c) (5 points) Assume that today, all the bikes start at MIT at 6AM. If you don't take any action, what is the probability that the system transitions to a state in which there are two bikes at MIT and two at Boston at 6PM? Answer in terms of the appropriate P_{ijk} values.

Solution: With all 4 bikes at MIT and no action taken, no Boston→MIT rides are possible (no bikes at Boston). For the state to be 2 at MIT and 2 at Boston at 6PM, exactly 2 MIT→Boston rides must complete:

$$P = P_{MBD}(2).$$

(d) For each question below, indicate whether it is best modeled as an MDP, a POMDP, a deterministic reward-maximization problem, or another type of problem. If you select **"Other," explain in the box why it cannot be modeled as an MDP, POMDP, or Deterministic problem**; if you select any of the other options, you may leave the box blank.

i. (2 points) The same 4 students bike from MIT to Boston during the day, and then bike back from Boston to MIT at midnight.

◯ MDP ◯ POMDP √ **Deterministic** ◯ Other

Solution: Transitions are deterministic.

ii. (2 points) Your bike-dock system is kind of flakey and it sometimes mis-reports the locations of the bikes.

◯ MDP √ **POMDP** ◯ Deterministic ◯ Other

Solution: You no longer have full state observability, so this becomes a POMDP.

iii. (2 points) Sometimes when a rider tries to take a bike out of the dock, it fails. Assume each rider only tries once before giving up, and failure probabilities are independent.

√ **MDP** ◯ POMDP ◯ Deterministic ◯ Other

Solution: Dock failure adds noise, but bike locations remain fully observable, so it is a MDP with modified transition probabilities.

iv. (2 points) Riders tell their friends about where they have had success finding a bike, and their friends change their commuting habits accordingly.

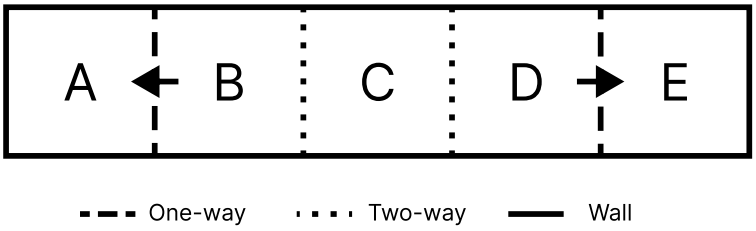
◯ MDP ◯ POMDP ◯ Deterministic √ **Other**

Solution: The demand distribution evolves with history, violating the Markov property.

2 MDP Decision-Making

2. Consider the following problem with 5 states $\{A, B, C, D, E\}$ and two actions $\{L, R\}$. States A and E are terminal states; after either one is reached, it cannot be escaped, and all actions taken from those states have 0 reward. In the other states, each action has a 5/6 chance of moving in the targeted direction; otherwise, you move in the opposite direction. You are given rewards for taking specific actions in specific states (regardless of the actual direction you move). A visualization of the setup is shown on the left below, and a table of the rewards is shown on the right. Use a discount factor of $\gamma = 0.9$.

Problem setup



Reward table $R(s, a)$

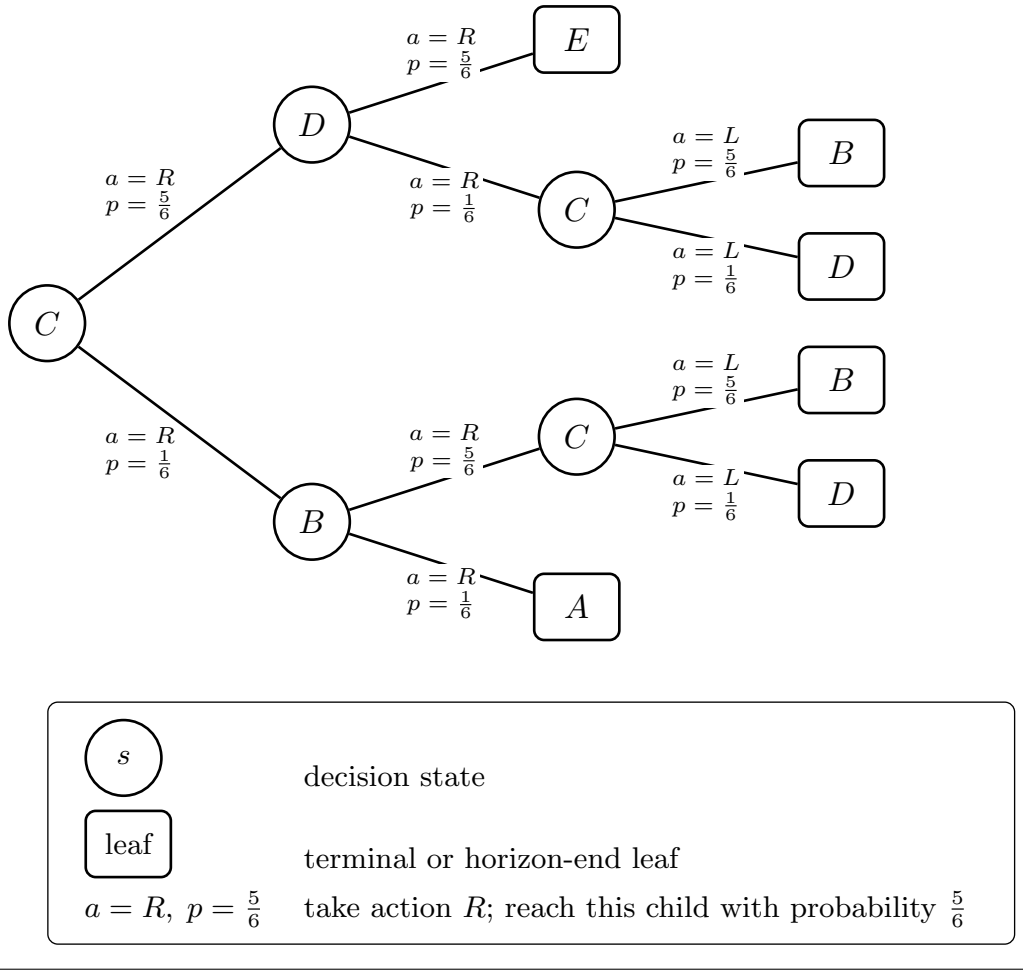
State	L	R
A	0	0
B	-4	0
C	$+1$	-1
D	0	$+4$
E	0	0

(a) (10 points) Assuming a discount factor of 0.9, what are the horizon 0, 1 and 2 value functions? **Fill in the missing rows.** We left some space for calculations at the bottom of this page.

	A	B	C	D	E
$V_0(s)$	0	0	0	0	0
$V_1(s)$	0	0	1	4	0
$V_2(s)$	0	$\frac{15}{20}$	$\frac{40}{20}$	$\frac{83}{20}$	0
$V_3(s)$	0	$\frac{60}{40}$	$\frac{89}{40}$	$\frac{172}{40}$	0

(b) (5 points) Draw the *policy tree* for the optimal policy when starting in state 3 using the horizon 3 values.

Solution:



(c) (5 points) Provide two discount factors $0 < \gamma < 1$ that induce different optimal policies. Explain your answer.

Solution:

Small γ (0.1) : favors immediate reward, so C prefers L.

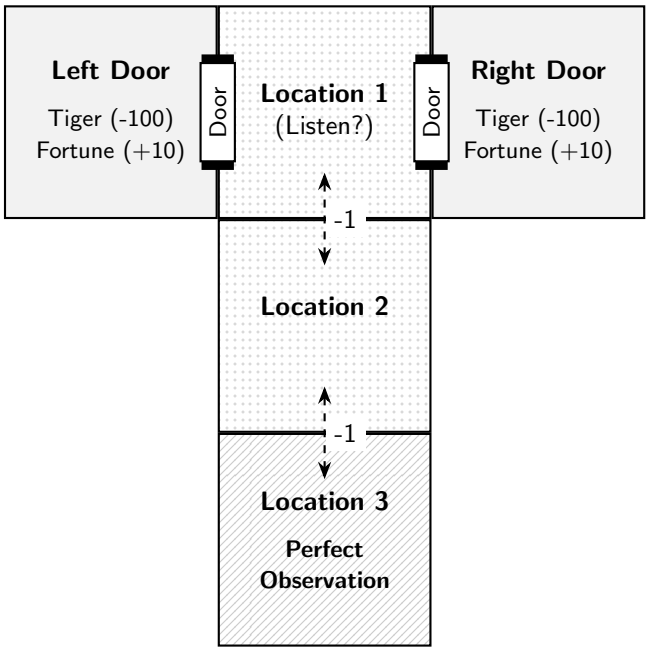
Large γ (0.99) : favors reaching D, so C prefers R.

$\gamma = 0.25$ is the cut off, so long as you had one above and one below, and explained, you should have gotten points.

3 Tiger or Ruin?

3. Leslie mixed up her notes and somehow ended up with a combination of the tiger problem and the fortune-or-ruin problem. The situation is:

   - There are 3 locations in a hallway: 1, 2, 3
   - In location 1, you can open the door on the left or the door on the right.
   - There is a tiger behind one of the doors and a fortune behind the other.
   - If a tiger is behind the door you open you get -100, otherwise you get +10. In either case, the game ends (you get 0 reward for the infinite future).
   - You may stand at location 1 and listen. If you do this and detect a tiger on either the left or the right, then the chance that the tiger truly is on the side where you heard it is 0.85. Listening has reward -1.
   - You can also move to a neighboring location for a reward of -1.
   - In location 3, you will get a perfect observation of where the tiger is.
   - In the following, we will only consider finite-horizon values so you can assume $\gamma = 1$.



(a) (4 points) What is the belief space of this problem?

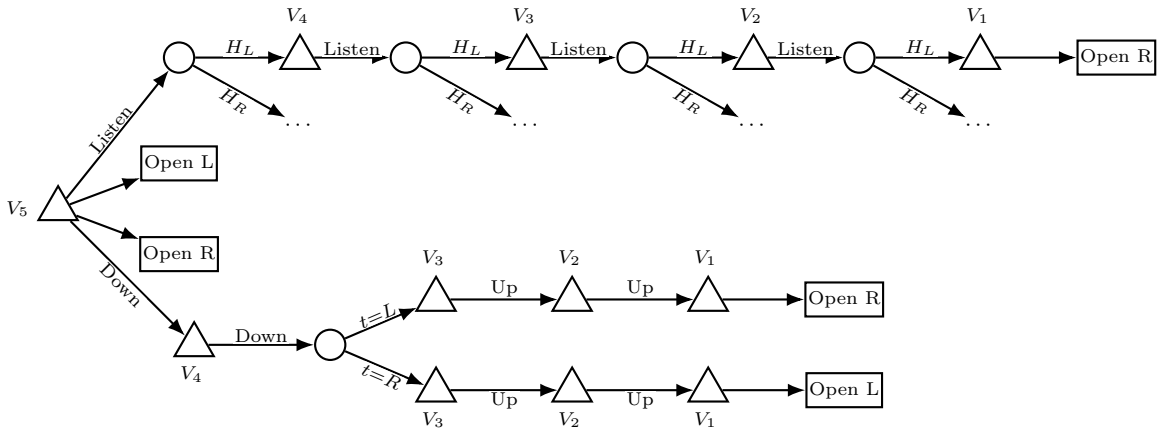
Solution:
The belief space is over the location of the tiger and the location of the agent. The tiger can be behind the left door or the right door, and the agent can be in locations $1, 2, 3$.
Since the agent's own location is fully observable, the belief state can be written as

$$b = \big(\Pr(t = L), \Pr(t = R), a\big),$$

where

$$\Pr(t = L) + \Pr(t = R) = 1, \qquad a \in \{1, 2, 3\}.$$

(b) (8 points) Here is is a partial horizon 5 expectimax search tree for the combined problem.



Assume that, in b_0, the agent is in location 1 and believes that tiger is on the left with probability 0.5.

What is the horizon 5 value of selecting *Down* as your first action in the combined problem, and continuing optimally, $Q^5(b_0, \mathit{Down})$?

Explain your reasoning and write down a numerical expression (but you don't need to evaluate it.)

Solution: If the agent chooses DOWN first, then the optimal plan is to move to location 3, observe the tiger perfectly, return to location 1, and open the correct door:

DOWN, DOWN, UP, UP, OPEN CORRECT DOOR.

The four movement actions each have reward -1, and opening the correct door has reward $+10$. Thus,

$$Q^5(b_0, \text{DOWN}) = -1 - 1 - 1 - 1 + 10 = -4 + 10 = 6.$$

(c) (5 points) Assume the optimal horizon 5 value of being in belief state b_0 in the original tiger problem is $V^5_{\text{tiger}}(b_0)$. What must be true of the value of $Q^5(b_0, \textit{Down})$ in order for *Down* to be the optimal first action in the combined problem in belief state b_0?

Solution: $Q^5(b_0, \textit{Down}) \geq V^5_{\text{tiger}}(b_0)$, or equivalently,
$Q^5(b_0, \textit{Down}) \geq \max\{Q^5(b_0, \textit{Listen}), Q^5(b_0, \textit{Open L}), Q^5(b_0, \textit{Open R})\}$.

(d) (5 points) Recall that the optimal horizon 3 action from belief $b = (0.5, 0.5)$ in the original tiger problem is *Listen*. What is the optimal horizon 3 action in the combined problem? Explain why your answer does or does not differ from the answer for the horizon 5 case.

Solution: Listen, because with horizon 3 there is not enough time to go down and come back and open the correct door.

4 Ternary Bandits

4. You find yourself in a casino with two slot machines, each of which has a hidden payoff probability p_m (where m is 1 or 2, indicating the machine) which is an element of $\{0.1, 0.5, 0.9\}$.

   Unlike a typical bandit problem: You have to pay \$0.50 each time you play, and you have the option of just not playing anymore–taking the nop action.

   Furthermore, you can also pay \$.10 to call a guy named Vinnie, and give him the serial number of the machine you're playing. Vinnie will tell you the payoff probability of that machine (that is, 0.1, 0.5, or 0.9) but will only give you the correct answer with probability 0.99. Also, Vinnie has no memory, so each time you call, it's an independent chance the answer is correct.

   Let's focus on **a single machine**, to start with.

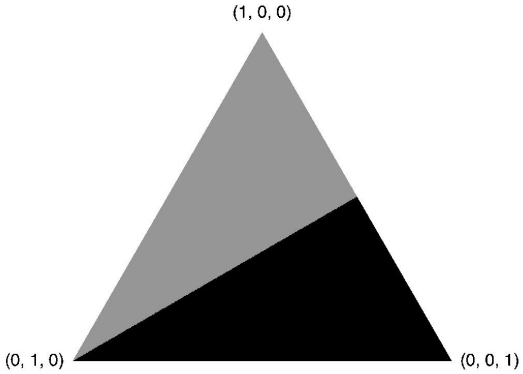
   (a) (3 points) If you start with a belief $(0.1, 0.1, 0.8)$ on p_1, what is your belief about the probability of winning if you play that machine?

   Solution:

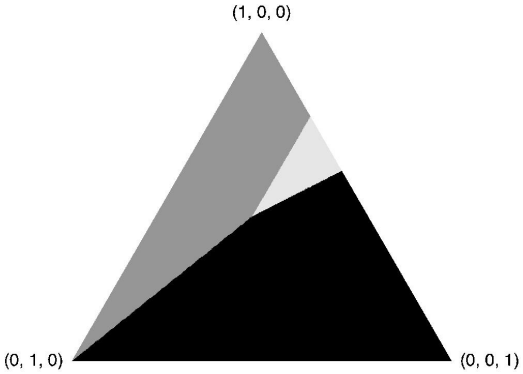
   $$P(\text{win}) = 0.1(0.1) + 0.1(0.5) + 0.8(0.9) = 0.01 + 0.05 + 0.72 = 0.78.$$

   Thus, the probability of winning is $\boxed{0.78}$.

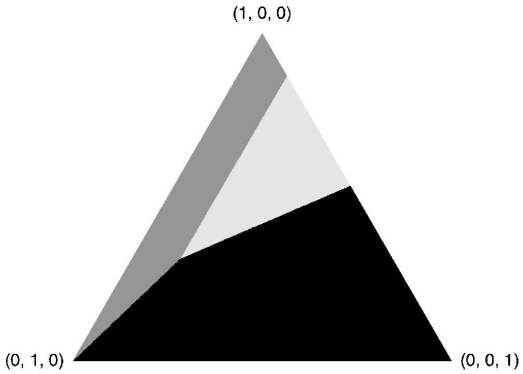
(b) (6 points) The following figure contains the optimal horizon 1 through 5 policies. The diagrams have three colors: *black*, *dark gray*, and *light gray*, with each corner of the triangle indicating absolute belief in p_1 being equal to $0.1, 0.5$ or 0.9.



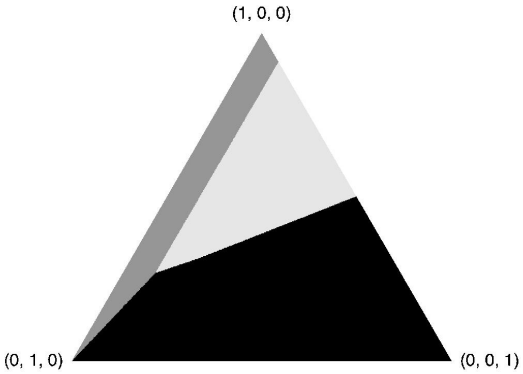
Horizon 1



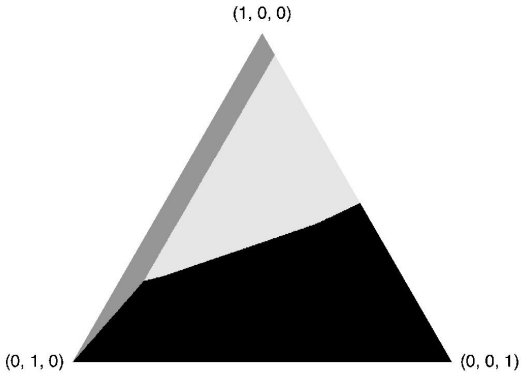
Horizon 2



Horizon 3



Horizon 4



Horizon 5

i. (3 points) We forgot the legend! Which action corresponds to which color?

Dark Gray: ◯ pull arm 1 √ **don't play** ◯ call vinnie

Light Gray: ◯ pull arm 1 ◯ don't play √ **call vinnie**

Black: √ **pull arm 1** ◯ don't play ◯ call vinnie

ii. (4 points) Why is the black region bigger for horizon 2 than horizon 1?

Solution: We gain information from playing.

iii. (4 points) The stripe along the left side of the triangle is getting thinner as we increase the horizon. Will it ever go completely away? Explain why, in terms of the problem.

√ **Never fully go away** ◯ Eventually go away

Solution: No. The stripe corresponds to not playing. The longer the horizon, the larger the potential reward is for playing if you are playing an arm with a success probability equal to 0.9. Thus, even if there is a small chance that you might have the 0.9 arm instead of the 0.5 or 0.1 arm, it becomes more and more worthwhile to call Vinnie for the small chance your belief is incorrect. However, there will always be a level of confidence in the arm not being 0.9 that it is not worth while to play.

Now, you have **two machines**, each with exactly the same set-up as the single one we have studied so far, including the ability to call Vinnie about either machine.

(c) (6 points) Your friend Sydney thinks that an optimal horizon H strategy in this case would be to compute the optimal horizon H Q functions for both machines, and then to pick the action that has the highest Q value among all 6 choices (three actions on each machine). **Provide an example belief in which this is not true and explain**.
Hint: *Consider a scenario where you are fairly certain one machine is good, but highly uncertain about the other.*

Solution: Consider belief $(0, 0, 1)$ on machine 1 (certain $p_1 = 0.9$) and $(1/3, 1/3, 1/3)$ on machine 2 (fully uncertain). Sydney's single-machine Q-function for machine 2 assigns positive value to calling Vinnie or pulling, because it assumes any information learned will be exploited later. But since machine 1 is known to be great, you'll always prefer pulling it over machine 2, so any money spent exploring machine 2 is wasted.

5 Mini-Project Check

5. In miniproject 4, we were trying to rescue a patient and take them to the hospital in a domain with fast-moving fire. We assumed that the state of the fire was completely observable, but it spread in the domain using random transition dynamics.

   For each of the techniques below, **indicate whether we used it** in MP4. If we did, **explain why it was a reasonable** choice for the problem. If we didn't, **explain why it would have been a poor choice**.

   (a) (3 points) Value iteration

   ◯ We did √ **We didn't**

   Solution: The state space of the problem is far too large, exponential in the number of spaces. This makes value iteration infeasible, as it would require a huge amount of time and space.
   The transition model was not hidden or random; it was fixed and known, just a bit complicated.

   (b) (3 points) Most likely observation

   ◯ We did √ **We didn't**

   Solution: Most likely observation would plan using the single most likely fire state. Since we modeled each cell by independent marginal fire probabilities, this could choose a path where every cell has a probability below 0.5 of being on fire, even though the whole path is very risky.

   (c) (3 points) Monte-Carlo tree search

   √ **We did** ◯ We didn't

   Solution: MCTS was a reasonable choice because it can use many rollouts to estimate the risk of different actions under stochastic fire spread before committing to a move but *does not require* building the whole expectimax tree.