Stanford University

AA228/CS238: Decision Making Under Uncertainty

Fall 2026

Prof. Mykel J. Kochenderfer • Online • email: *mykel@stanford.edu*

PROJECT 3 **Due date: November 20, 2026 (5pm Pacific)**

Project 3 will be submitted on Gradescope. Correct answers will receive full credit, and partial credit may be awarded based on work shown. Hence, we encourage you to show your work.

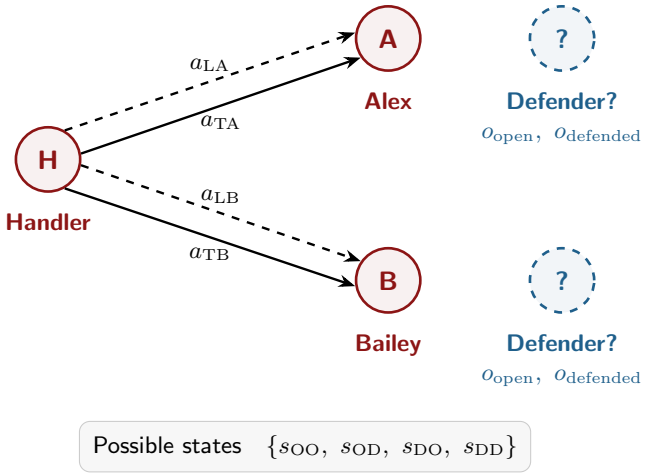
Reading the Defense in Ultimate Frisbee

The Stanford Intelligent Systems Laboratory (SISL) is preparing for the annual department sports competition, the Durand Cup, by running a decision-making drill. The handler (the person with the disc) has two downfield cutters (people trying to receive the disc), Alex and Bailey. Each cutter is either open or defended, but the handler cannot observe this perfectly. We model the situation as a POMDP.

- States: The state indicates whether Alex and Bailey are open or defended.
  - s_{OO}: Alex is open; Bailey is open.
  - s_{OD}: Alex is open; Bailey is defended.
  - s_{DO}: Alex is defended; Bailey is open.
  - s_{DD}: Alex is defended; Bailey is defended.

  Use the ordering $(s_{OO}, s_{OD}, s_{DO}, s_{DD})$ for all belief and alpha vectors.
- Actions: The handler can either look toward a cutter or throw to a cutter.
  - a_{LA}: Look toward Alex.
  - a_{LB}: Look toward Bailey.
  - a_{TA}: Throw to Alex.
  - a_{TB}: Throw to Bailey.
- Observations: An observation comes after an action is taken and after the defense moves.
  - If the handler looks toward a cutter, the handler observes either o_{open} or o_{defended}.
  - If the handler throws to a cutter, the handler observes either o_{complete} or $o_{\text{incomplete}}$.

  The observation probabilities are as follows:
  - If the handler looks toward a cutter who is open, the cutter appears open with probability 0.8.
  - If the handler looks toward a cutter who is defended, the cutter appears open with probability 0.2.
  - If the handler throws to a cutter who is open, the pass is completed with probability 0.8.
  - If the handler throws to a cutter who is defended, the pass is completed with probability 0.4.
- Transitions: After the handler selects an action but before receiving an observation, the defense moves. Each cutter changes status independently according to the following probabilities:
  - A cutter who is open remains open with probability 0.85 and becomes defended with probability 0.15.
  - A cutter who is defended becomes open with probability 0.25 and remains defended with probability 0.75.



Because Alex and Bailey change status independently, the resulting transition matrix is

$$T = \begin{bmatrix} 0.7225 & 0.1275 & 0.1275 & 0.0225 \\ 0.2125 & 0.6375 & 0.0375 & 0.1125 \\ 0.2125 & 0.0375 & 0.6375 & 0.1125 \\ 0.0625 & 0.1875 & 0.1875 & 0.5625 \end{bmatrix}.$$

Rows represent the current state and columns represent the next state, using the ordering

$$(s_{OO}, s_{OD}, s_{DO}, s_{DD}).$$

Thus, entry T_{ij} is the probability of transitioning from state s_i to state s_j. The transition matrix is the same for every action.

- Rewards: The reward is determined by the state when the action is selected.
  - Looking toward either cutter always gives a reward of -1, as it takes time away from making a throw.
  - Throwing to a cutter who is open when the action is selected gives a reward of $+20$.
  - Throwing to a cutter who is defended when the action is selected gives a reward of -20.
- Discount Factor (γ): $\gamma = 0.9$.

Question 1 (Looking toward Alex)**.** Starting from a uniform prior, the handler takes action a_{LA} and observes o_{open}. What is the updated belief? Round all values in your final answer to two decimal places and maintain the following order in your belief vector: $s_{OO}, s_{OD}, s_{DO}, s_{DD}$.
Note: *[a, b, c] is interpreted as a column vector (as discussed in the textbook), while a row vector [a b c] is written without the commas.*

Question 2 (Reading between the cutters)**.** Suppose the handler's current belief about whether the cutters are defended or not is $\mathbf{b} = [0.05, 0.45, 0.45, 0.05]$. The handler takes action a_{LB}, but while taking that action, the defense shifts. Before the handler receives an observation from this action a_{LB}, but after the defense has moved, the new predicted belief is $\bar{\mathbf{b}} = [0.2305, 0.3195, 0.3195, 0.1305]$. The handler then observes that Bailey appears defended; that is, the handler receives the observation o_{defended}.

(a) Compute the probability that Alex is open before and after receiving the observation that Bailey is defended.

(b) Briefly explain why an observation about Bailey changes the handler's belief about Alex.

Question 3 (Throw now or keep looking?)**.** For this question, suppose the current belief is $\mathbf{b} = [0.20, 0.55, 0.05, 0.20]$. Compute the expected immediate reward $R(\mathbf{b}, a) = \sum_s b(s)R(s, a)$ for each of the four actions. Which action would a one-step greedy policy select? Briefly explain your answer.

Question 4 (Calling the next play)**.** For this question, assume that the drill continues after every action and that each action is followed by exactly two possible observations. A look action is followed by either o_{open} or o_{defended}, while a throw action is followed by either o_{complete} or $o_{\text{incomplete}}$.

(a) Draw one possible two-step conditional plan. Label each node with its action and each edge with the corresponding observation.

(b) What is the ratio of the number of three-step plans to the number of two-step plans?

Question 5 (Evaluating the playbook)**.** Write down the alpha vector associated with each of the four one-step conditional plans, using the state ordering $(s_{OO}, s_{OD}, s_{DO}, s_{DD})$. At the belief $\mathbf{b} = [0.40, 0.30, 0.20, 0.10]$, compute the value of each alpha vector. Which one-step conditional plan is optimal, and what is its value?

Question 6 (Playing it safe)**.** Compute the best-action worst-state (BAWS) lower bound for this POMDP and write down its corresponding alpha vector. Recall that $r_{\text{BAWS}} = \frac{1}{1-\gamma} \max_a \min_s R(s, a)$.

Question 7 (Planning with imperfect information)**.** Using the one-step alpha vectors from Question 5, perform one QMDP backup and one fast informed bound (FIB) backup. Compute the resulting component for action a_{LA} and state s_{DD} under each method. Briefly explain why the two values differ.

Question 8 (Look, then throw)**.** The handler considers a two-step plan: look toward Alex, then throw to Alex if Alex appears open and to Bailey otherwise. Compute the alpha vector for this plan and its expected discounted return from a uniform prior.

Question 9 (Is looking worth it?)**.** Starting from the uniform belief $\mathbf{b}_0 = [0.25, 0.25, 0.25, 0.25]$, the handler performs forward search using

$$Q_1(\mathbf{b}, a) = R(\mathbf{b}, a) + \gamma \sum_o P(o \mid \mathbf{b}, a) U_L\big(\text{Update}(\mathbf{b}, a, o)\big),$$

where the leaf heuristic is $U_L(\mathbf{b}) = \max_{a' \in A} R(\mathbf{b}, a')$. Since Alex and Bailey are symmetric, it is sufficient to compare a_{LA} with a_{TA}. For each row below, let

$$p_A = P(A = \text{open} \mid \mathbf{b}_0, a, o), \qquad p_B = P(B = \text{open} \mid \mathbf{b}_0, a, o).$$

The relevant probabilities, rounded to two decimal places, are provided below.

Root action	Observation	$P(o \mid \mathbf{b}_0, a)$	p_A	p_B
a_{LA}	o_{open}	0.53	0.83	0.55
a_{LA}	o_{defended}	0.47	0.23	0.55
a_{TA}	o_{complete}	0.62	0.71	0.55
a_{TA}	$o_{\text{incomplete}}$	0.38	0.29	0.55

Compute the leaf utility associated with each row. Then compute $Q_1(\mathbf{b}_0, a_{LA})$ and $Q_1(\mathbf{b}_0, a_{TA})$ and determine whether the handler should look first or throw immediately.

Question 10 (Updating the playbook)**.** The handler uses point-based value iteration with beliefs that assign nearly all probability to one state. Why might the resulting policy perform poorly when the handler is uncertain about which cutter is open? Suggest one way to improve the selected belief set. Keep your answer to 2–3 sentences.

Question 11 (Knowing when to stop searching)**.** An online planner has computed the following lower and upper bounds on the value of each action at the current belief:

Action	Lower bound	Upper bound
a_{LA}	14	17
a_{LB}	12	17
a_{TA}	16	26
a_{TB}	2	7

Which actions can be pruned using branch and bound? Briefly explain your answer.

Question 12 (One more simulation)**.** The handler uses Monte Carlo tree search with exploration constant $c = 2$. At the current root history h, the visit counts and value estimates are:

Action	$N(h, a)$	$Q(h, a)$
a_{LA}	16	8.0
a_{LB}	2	7.5
a_{TA}	20	8.5
a_{TB}	2	6.0

Using the UCB1 score

$$Q(h, a) + c\sqrt{\frac{\ln N(h)}{N(h, a)}}, \qquad N(h) = \sum_{a} N(h, a),$$

compute each action's score and identify the action selected for the next simulation. Round your scores to two decimal places.