The Honey Heist MDP

Prava.co Take-Home Assessment 9/2025

Estimated time: 3–4 hours

A bear raids a beehive over exactly 3 time steps ($t \in \{0, 1, 2\}$). At each step, the bear chooses one action:

- **Scoop**: Attempt to collect honey (immediate reward: 0)
- **Smoke**: Calm the bees (immediate reward: -1)
- **Scoot**: Escape with current honey c (immediate reward: $3c$)

No free terminal salvage. You only get $3c$ if you choose **scoot**; otherwise there is no salvage at the horizon (even at $t = 2$).

Model Specification

States

Each state is (a, c, t) where:

- $a \in \{0, 1, 2\}$: bee agitation (0 = calm, 1 = alert, 2 = critical)
- $c \in \{0, 1, 2\}$: honey scoops collected
- $t \in \{0, 1, 2\}$: time step

Initial state: $(0, 0, 0)$ (calm bees, no honey, time 0)
Terminal conditions:

- Taking **scoot** ends the episode immediately
- Getting stung (from failed **scoop**) ends the episode immediately
- After any action at $t = 2$, the episode ends

Action Constraints

- When $c < 2$: all three actions available
- When $c = 2$: only **smoke** and **scoot** available (bucket full)

Transition Dynamics

Scoop (when $c < 2$):

- With probability $s(a)$: get stung, receive -9 reward, episode ends
- With probability $1 - s(a)$: successfully collect honey
  - Agitation increases: $a' = \min(a + 1, 2)$
  - Honey increases: $c' = c + 1$
  - Time advances: $t' = t + 1$

Smoke:

- With probability r: bees calm down, $a' = \max(a - 1, 0)$
- With probability $1 - r$: no effect on agitation, $a' = a$
- Always: $c' = c$, $t' = t + 1$

Scoot:

- Receive reward $3c$ and episode ends immediately

> **Fixed Parameters:**
>
> - Sting probabilities: $s(0) = 0$, $s(1) = \frac{1}{3}$, $s(2) = \frac{2}{3}$
> - Smoke effectiveness: $r = \frac{2}{3}$ (base model)
> - Discount: none (finite horizon)

Special Case: Actions at $t = 2$

At the final time step, the episode ends immediately after the action, so:

- **Scoop**: Get reward $-9 \cdot s(a)$ (expected sting penalty)
- **Smoke**: Get reward -1 (no future benefit since episode ends)
- **Scoot**: Get reward $3c$

$t = 2$ **dominance.** For any (a, c): $Q_2(\text{scoot}) = 3c \geq 0$, $Q_2(\text{smoke}) = -1$, $Q_2(\text{scoop}) = -9s(a) \leq 0$. Thus **scoot weakly dominates scoop** (strictly unless $a = 0, c = 0$) and **strictly dominates smoke**. With our tie-break: **scoot everywhere at** $t = 2$.

Reachable States

From initial state $(0, 0, 0)$, only certain states are reachable. You need only compute values for:

Time	Reachable States (a, c)
$t = 0$	$(0, 0)$
$t = 1$	$(0, 0)$, $(1, 1)$
$t = 2$	$(0, 0)$, $(0, 1)$, $(1, 1)$, $(2, 2)$

Problems

Format Requirements

- Present value tables $V_t(a, c)$ with rows for agitation a, columns for scoops c
- Present policy tables $\pi_t(a, c)$ showing the optimal action for each state
- Use exact fractions or decimals to 3 places; apply tie-break **after rounding**
- For ties between equally good actions, prefer: **scoot** > **scoop** > **smoke**

1. Backward Induction (40 points)

Compute the optimal value function $V_t(a, c)$ and optimal policy $\pi_t(a, c)$ for all reachable states at each time $t \in \{2, 1, 0\}$.

Value function: $V_t(a, c) = \max\{Q_t(\text{scoop}; a, c), Q_t(\text{smoke}; a, c), Q_t(\text{scoot}; a, c)\}$ over feasible actions.

Start with $t = 2$: Since the episode ends after the action, the Q-values are simply the immediate rewards:

$$Q_2(\text{scoop}; a, c) = -9 \cdot s(a) \quad \text{if } c < 2 \tag{1}$$

$$Q_2(\text{smoke}; a, c) = -1 \tag{2}$$

$$Q_2(\text{scoot}; a, c) = 3c \tag{3}$$

For $t < 2$: Use the Bellman equations:

$$Q_t(\text{scoop}; a, c) = -9 \cdot s(a) + (1 - s(a)) \cdot V_{t+1}(\min(a+1, 2), c+1) \tag{4}$$

$$Q_t(\text{smoke}; a, c) = -1 + r \cdot V_{t+1}(\max(a-1, 0), c) + (1 - r) \cdot V_{t+1}(a, c) \tag{5}$$

$$Q_t(\text{scoot}; a, c) = 3c \tag{6}$$

Present your results as tables. For example:

$V_2(a, c)$

$a \backslash c$	0	1	2
0	?	?	–
1	–	?	–
2	–	–	?

$\pi_2(a, c)$

$a \backslash c$	0	1	2
0	?	?	–
1	–	?	–
2	–	–	?

("–" = unreachable; fill only the listed reachable cells)

2. Monotonicity Analysis (20 points)

Prove that for any fixed c and t, if **scoop** is optimal at agitation level a, then **scoop** is also optimal at all lower agitation levels $a' < a$.

Hint: For fixed c and $t < 2$, show each of $Q_t(\text{scoop}; a, c)$ and $Q_t(\text{smoke}; a, c)$ is **non-increasing** in a (since $s(a)$ is non-decreasing and, by induction, V_{t+1} is non-increasing in a), while $Q_t(\text{scoot}; a, c)$ is independent of a. Therefore the max, $V_t(a, c)$, is non-increasing.

3. Sensitivity Analysis (20 points)

Recompute the optimal policy at $t = 1$ under these modified parameters (analyze each separately):

(a) Higher sting risk: Set $s(1) = \frac{1}{2}$ and $s(2) = 1$ (keeping $s(0) = 0$)

- First recompute V_2 under new parameters
- Then compute V_1 and π_1
- Identify which states change their optimal action and explain why

(b) State-dependent smoke: Make smoke less effective at higher agitation:

- $r(0) = \frac{3}{4}$: probability of calming at $a = 0$
- $r(1) = \frac{2}{3}$: probability of calming at $a = 1$
- $r(2) = \frac{1}{2}$: probability of calming at $a = 2$

Again, recompute V_2 first, then V_1 and π_1.

4. Expected Value (10 points)

Report $V_0(0, 0)$: the expected total reward under optimal play from the initial state. Briefly describe the optimal strategy in words.

5. Extension: Risk vs Reward (10 points, optional)

Consider a risk-averse bear who uses the utility function $U(r) = \sqrt{r + 10}$ for total reward r.

- How does this change the optimal policy at $t = 1$, state $(1, 1)$?
- Explain intuitively why risk aversion might favor certain actions

Deliverables

Submit either:

- A PDF with your solutions, showing key calculations
- A Jupyter notebook or script with your code and output

Include:

1. Value and policy tables for each time step
2. Brief proof for Question 2
3. Sensitivity analysis comparison tables
4. 2–3 sentence interpretation of your results

Questions? Email careers@prava.co