Online and Reinforcement Learning (2025–2026)
Home Assignment 2

Hamsa Mohamed (XXXXXXXX)
your@email.com

Deadline: 18 February 2026, 20:59

Submission checklist

- PDF report with detailed answers (no full code dumps).
- Separate `.zip` with well-commented source code and a main file per question (or one main).

Contents

1 Short Questions (8 points)

(1) In a finite discounted MDP, every possible policy induces a Markov Reward Process.
Answer: ☐ True ☑ False
Justification: every policy stationary random policy $\pi \in \Pi^{SR}$ induces a MRP. However a non-stationary policy may not adhere to the Markov property.

(2) Consider a finite discounted MDP, and assume that π is an optimal policy. Then, the action(s) output by π does not depend on history other than the current state (i.e., π is necessarily stationary).
Answer: ☐ True ☑ False
Justification: We two category of policies $\pi \in \Pi^{HR}$ and $\pi \in \Pi^{HD}$ which have history-dependent policies. We may have an optimal policy that depence on history

(3) In a finite discounted MDP, a greedy policy with respect to optimal action-value function $Q^\star$ corresponds to an optimal policy.
Answer: ☑ True ☐ False
Justification: Missing

(4) Policy Iteration (PI) may return a near-optimal policy.
Answer: ☑ True ☐ False
Justification: Theorem 7

2 Introduction of New Products (25 points)

2.1 Problem statement

Briefly restate Exercise 7.6 in your own words (1–3 lines), then proceed.

2.2 Solution

Result

Final answer (clean statement):
Write the final bound / algorithm / conclusion clearly here.

3 Empirical Comparison of FTL and Hedge (25 points)

3.1 Experimental setup

- Dataset / synthetic setting:
- Losses / feedback model:
- Horizon T, number of experts N:

- Learning rate / tuning:
- Random seeds / repetitions:

3.2 Methods

FTL. Describe the update briefly.

Hedge. Describe the update briefly (weights, learning rate, normalization).

3.3 Results

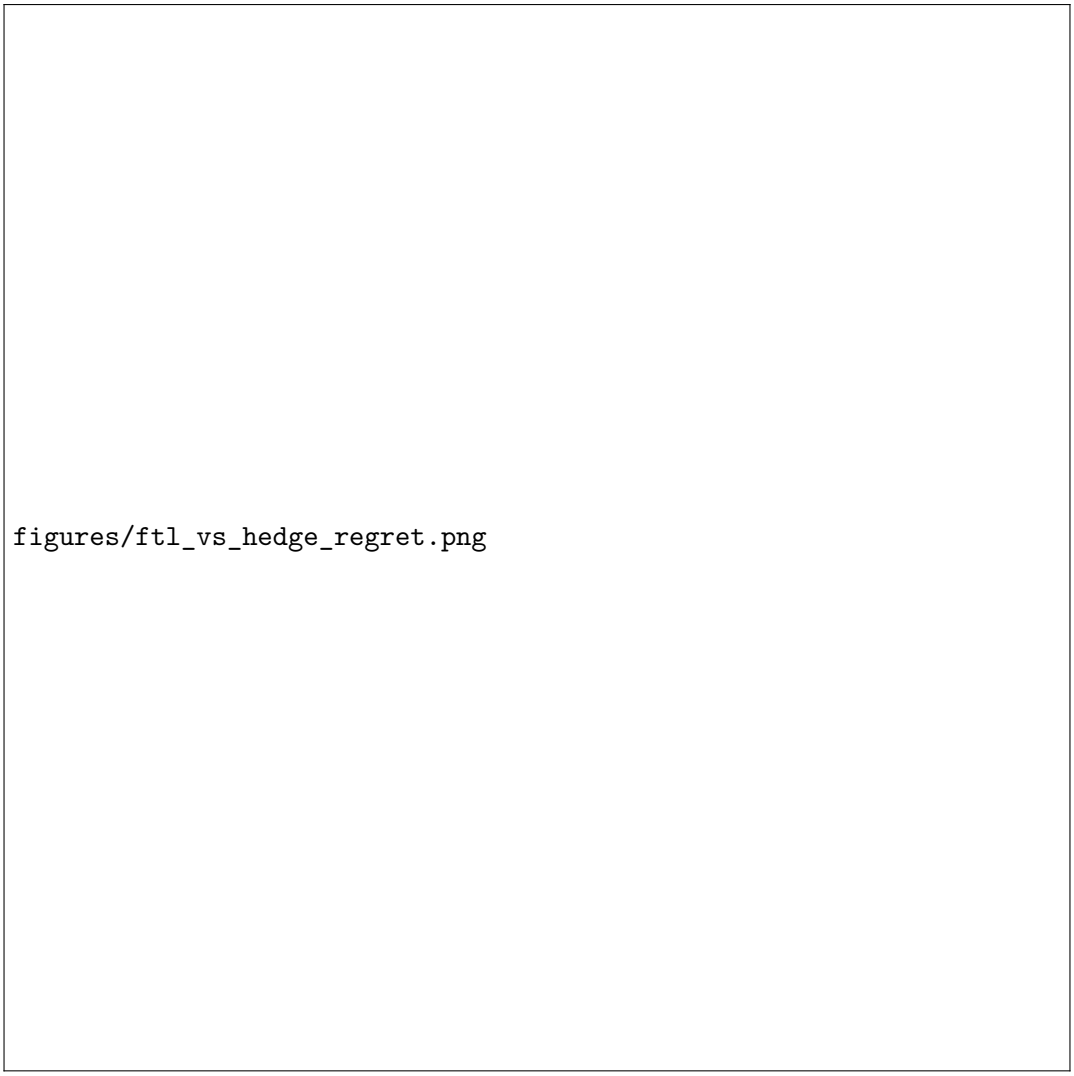

Figure 1: Example: Regret curves for FTL vs Hedge (fill in).

3.4 Discussion

Explain what you observe and why (a few paragraphs).

4 Value Function Bounds (22 points)

4.1 (i) Proof of the inequality

Given. Let $M_1 = (\mathcal{S}, \mathcal{A}, P_1, R_1, \gamma)$ and $M_2 = (\mathcal{S}, \mathcal{A}, P_2, R_2, \gamma)$ be finite discounted MDPs, with rewards in $[0, 1]$ and for all (s, a):

$$|R_1(s,a) - R_2(s,a)| \le \alpha, \qquad \|P_1(\cdot \mid s,a) - P_2(\cdot \mid s,a)\|_1 \le \beta.$$

Goal: show that for any stationary deterministic policy π and any (s, a),

$$|Q_1^\pi(s,a) - Q_2^\pi(s,a)| \le \frac{\alpha}{1-\gamma} + \frac{\gamma\beta}{(1-\gamma)^2}.$$

Proof.

4.2 (ii) Interpretation

Explain in words: how value differences scale with reward mismatch and transition mismatch, and why the factor blows up as $\gamma \to 1$.

5 Solving a Discounted Grid-World (20 points)

5.1 Environment summary

- Grid size: 7×7, accessible states: 20 (after walls).
- Start: upper-left (red). Reward state: lower-right (yellow), absorbing with reward 1 forever.
- Actions: up/left/down/right. Slippery: intended 0.7, stay 0.1, perpendiculars 0.1 each.
- Discount: $\gamma = 0.97$ (then $\gamma = 0.998$ for part (iii)).

5.2 (i) Policy Iteration (PI)

Implementation notes. State what code you used (file name), and any parameters.

Optimal policy and $V^\star$. Include:

- A visualization of the policy (arrows in a matrix).
- A visualization/table of $V^\star$.

5.3 (ii) Value Iteration (VI) with $\varepsilon = 10^{-6}$

Algorithm 1 Value Iteration (VI)

Input: MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, accuracy ε
Initialize $V_0(s) \leftarrow 0$ for all s
repeat
  for each state s **do**
    $V_{\text{new}}(s) \leftarrow \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V(s')\right)$
  end for
  $\Delta \leftarrow \max_s |V_{\text{new}}(s) - V(s)|$
  $V \leftarrow V_{\text{new}}$
until $\Delta < \varepsilon$
Extract policy $\pi(s) \in \arg\max_a \left(R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V(s')\right)$
Output: π, V

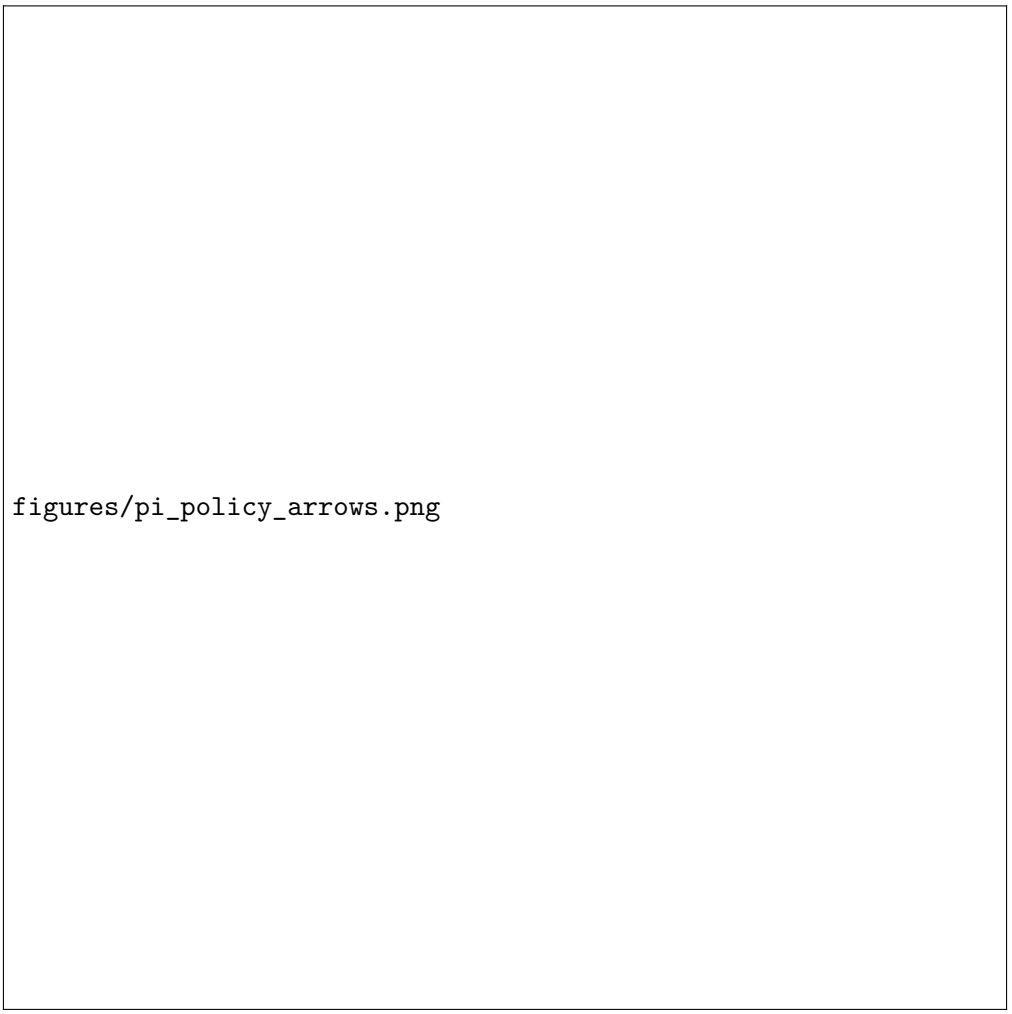

Figure 2: Optimal policy from PI (arrows).

Algorithm.

Results. Show the VI policy and value function (like in part (i)) and confirm it matches PI.

5.4 (iii) VI with $\gamma = 0.998$ and convergence discussion

Report:

- Number of iterations to converge (or runtime) for $\gamma = 0.97$ vs $\gamma = 0.998$.
- A short explanation of why higher γ slows convergence (contraction factor).

Optional: Improved Parametrization of UCB1 (0 points)

If you solved it, place it here. Otherwise omit this section in your final PDF.

A Reproducibility notes (optional)

- How to run: `python main_q3.py` / `python main_q5.py`
- Dependencies: `requirements.txt`
- Random seeds used:

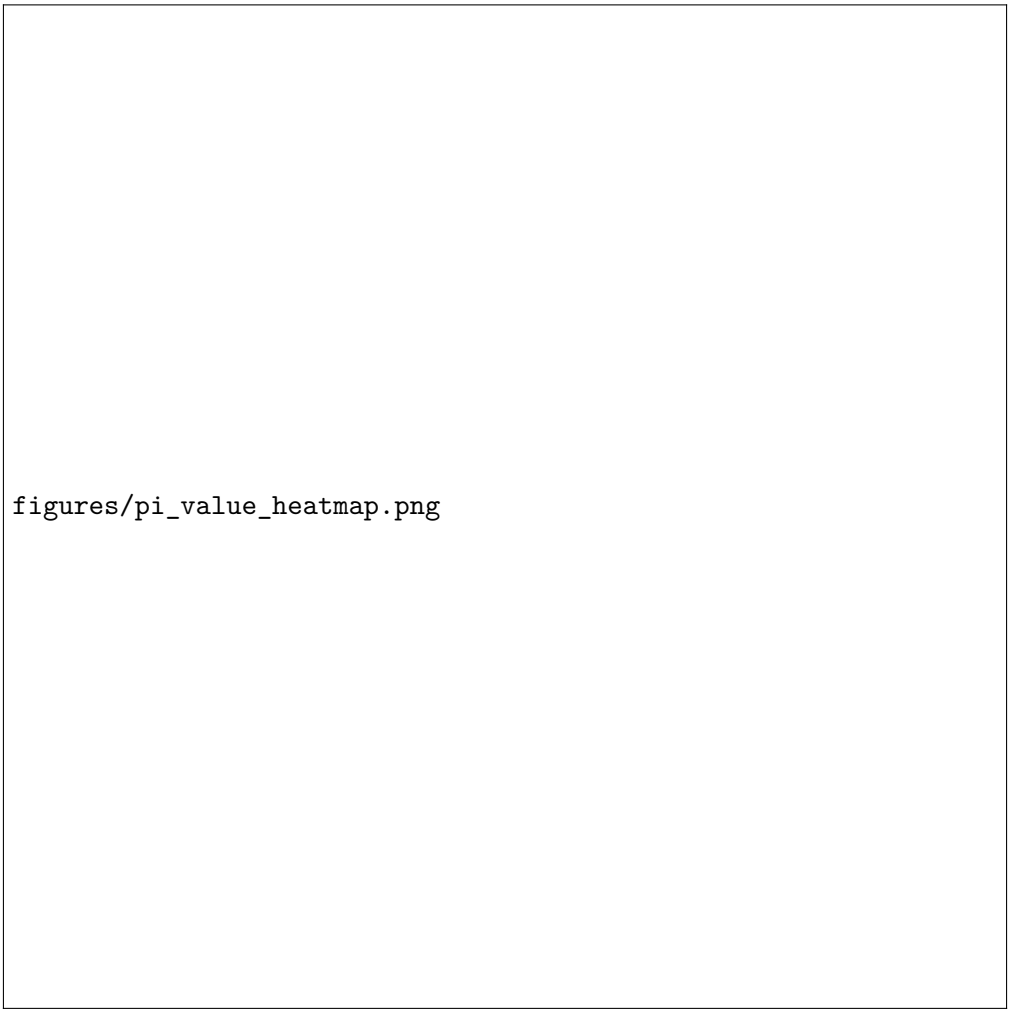

Figure 3: Optimal value function $V^\star$ from PI.