CSE 542: Statistical Reinforcement Learning

Homework 2 Spring 2026

University of Washington Instructor: Kevin Jamieson

Due: 11:59 PM May 10, 2026 **Points:** 100 (3 problems)

Instructions. Write up solutions clearly and concisely in LaTeX. You may collaborate with other students or LLMs, but each student must write their own solution independently and follow the collaboration policy defined on the course website (i.e., citations required). For Problem 3, submit your code as an appendix (see website for a convenient latex package).

Problem 1 (Gap-Dependent Regret for Contextual UCB (30 pts))**.**
Setting. At each round $t = 1, 2, \dots, T$:

(i) Nature reveals a context $x_t \in \mathcal{X}$, drawn i.i.d. from some distribution ν.

(ii) The learner selects an action $a_t \in \mathcal{A}$.

(iii) The learner observes a reward $r_t \in [0, 1]$ with conditional mean $\mu(x_t, a_t)$.

The action and context sets are finite: $|\mathcal{X}| = X$ and $|\mathcal{A}| = A$. Define

$$\mu^\star(x) := \max_{a\in\mathcal{A}} \mu(x, a), \qquad \pi^\star(x) := \arg\max_{a\in\mathcal{A}} \mu(x, a), \qquad \Delta(x, a) := \mu^\star(x) - \mu(x, a) \;\geq\; 0.$$

The pseudo-regret over T rounds is

$$R(T) \;:=\; \sum_{t=1}^{T} \Delta(x_t, a_t).$$

The UCB algorithm. Initialize $n_1(x, a) = 0$ for all (x, a). At round t, given the past $(x_1, a_1, r_1), \dots, (x_{t-1}, a_{t-1}, r_{t-1})$, define

$$n_t(x, a) := \sum_{s=1}^{t-1} \mathbf{1}\{(x_s, a_s) = (x, a)\}, \quad \widehat{\mu}_t(x, a) := \frac{1}{n_t(x, a)} \sum_{s=1}^{t-1} r_s \, \mathbf{1}\{(x_s, a_s) = (x, a)\},$$

where $\widehat{\mu}_t(x, a)$ is set arbitrarily when $n_t(x, a) = 0$. Fix $\delta \in (0, 1)$ and set

$$\iota := \log\Big(\tfrac{2XAT}{\delta}\Big), \qquad b_t(x, a) := \begin{cases} +\infty, & n_t(x, a) = 0, \\ \sqrt{\dfrac{\iota}{2\, n_t(x, a)}}, & n_t(x, a) \geq 1. \end{cases}$$

At round t, the algorithm plays $a_t \in \arg\max_{a\in\mathcal{A}} \big[\widehat{\mu}_t(x_t, a) + b_t(x_t, a)\big]$, breaking ties arbitrarily.

Remark: This problem is a stage-1 reduction of UCB-VI. The bonus b_t plays the role of the per-stage bonus b_h^k, and $\Delta(x, a)$ is the contextual analogue of $\Delta_h(s, a)$. The proof structure mirrors the gap-dependent UCB-VI argument from lecture, minus the downstream value-error term.

(a) **Optimism event** (6 pts)**.** Define

$$\mathcal{E} := \bigcap_{t \le T+1} \bigcap_{(x,a):\, n_t(x,a) \ge 1} \Big\{ |\widehat{\mu}_t(x,a) - \mu(x,a)| \le b_t(x,a) \Big\}.$$

Show that $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$.

Hint: For fixed (x,a) and $n \ge 1$, the first n rewards observed at (x,a) are i.i.d. in $[0,1]$ with mean $\mu(x,a)$, so Hoeffding gives a deviation bound for that prefix of length n. Union-bound over $n \in \{1,\dots,T\}$ and over $(x,a) \in \mathcal{X} \times \mathcal{A}$.

(b) **Each sub-optimal pull is paid for by the bonus** (10 pts)**.** On the event $\mathcal{E}$, show that for every round t with $a_t \neq \pi^\star(x_t)$,

$$\Delta(x_t, a_t) \;\le\; 2\, b_t(x_t, a_t).$$

Hint: Chain together (i) optimism upper bound on $\mu^\star(x_t)$, (ii) the greedy choice of UCB at x_t, and (iii) optimism lower bound on $\mu(x_t, a_t)$.

(c) **Bound on visits to sub-optimal arms** (6 pts)**.** On the event $\mathcal{E}$, show that for every (x,a) with $\Delta(x,a) > 0$,

$$n_{T+1}(x,a) \;\le\; 1 + \frac{2\,\iota}{\Delta(x,a)^2}.$$

Hint: If $a_t = a$ at some round t with $n_t(x,a) \ge 1$, part (b) implies $\Delta(x,a) \le 2b_t(x,a) = \sqrt{2\iota / n_t(x,a)}$; rearrange. The single "free" pull accounts for the case $n_t(x,a) = 0$.

(d) **Gap-dependent regret bound** (6 pts)**.** Conclude that on the event $\mathcal{E}$,

$$R(T) \;\le \sum_{(x,a):\, \Delta(x,a) > 0} \Big(\Delta(x,a) \;+\; \frac{2\,\iota}{\Delta(x,a)} \Big).$$

(e) **Minimax form** (2 pts)**.** Show that there is a universal constant $c > 0$ such that on the event $\mathcal{E}$,

$$R(T) \;\le\; c\sqrt{XAT\,\iota}.$$

Hint: Split arms by gap into "small" ($\Delta(x,a) \le \theta$) and "large" ($\Delta(x,a) > \theta$); bound the contribution of small-gap arms by $T\theta$ and the contribution of large-gap arms by $\sum_{(x,a):\Delta>\theta} 2\iota/\Delta(x,a) \le 2XA\iota/\theta$. Optimize θ.

Problem 2 (Why Exploration Matters: ε-Greedy versus UCB-VI (25 pts))**.**
The deep-exploration MDP. Fix $H \ge 3$. The MDP M_H has state space

$$\mathcal{S} = \{s_0,\; s_L,\; s_R^{(1)}, s_R^{(2)}, \dots, s_R^{(H-1)},\; s_\perp\},$$

action space $\mathcal{A} = \{a_1, a_2\}$ (so $A = 2$), and deterministic transitions:

- From s_0: $a_1 \to s_L$, and $a_2 \to s_R^{(1)}$.
- From s_L: any action remains at s_L.
- From $s_R^{(h)}$ for $h = 1, \dots, H-2$: there is a fixed correct action $a_h^\star \in \mathcal{A}$; under $a_h^\star$ the next state is $s_R^{(h+1)}$, otherwise it is $s_\perp$.

- From $s_R^{(H-1)}$ and from $s_\perp$: any action transitions to $s_\perp$.

The rewards are

$$r_h(s_L, \cdot) = \tfrac{1}{2} \text{ for all } h, \quad r_{H-1}(s_R^{(H-1)}, a_{H-1}^\star) = H-1, \quad r_h(s,a) = 0 \text{ otherwise.}$$

The initial state is s_0 deterministically. Each episode consists of H stages, $h = 0, 1, \ldots, H-1$, with s_h the state at stage h.

ε-Greedy. Recall the ε-greedy algorithm: at each step h in episode k, with probability $1-\varepsilon$ the agent plays $\arg\max_a \widehat{Q}_h^k(s_h, a)$ (ties broken uniformly at random), and with probability ε the agent plays uniformly over $\mathcal{A}$. Here $\widehat{Q}_h^k$ is the certainty-equivalence (plug-in MLE) estimate of $Q_h^\star$ formed from all transitions and rewards observed in episodes $1, \ldots, k-1$, with $\widehat{Q}_h^1 \equiv 0$. Despite its lack of theoretical guarantees, ε-greedy is the default exploration scheme used in much of practical deep RL (e.g., DQN and its descendants).

The point of this problem is to formalize, in a small example, why ε-greedy is a poor idea for exploration and how UCB-VI's optimism circumvents the issue.

(a) **Value computation** (5 pts)**.**

(i) Compute $V_0^\star(s_0)$ on M_H.

(ii) Let π_L denote the policy "always play a_1." Compute $V_0^{\pi_L}(s_0)$.

(iii) Conclude that the per-episode regret of *any* policy that plays a_1 at s_0 is at least $(H-1)/2$.

(b) **ε-greedy has linear regret** (13 pts)**.** Fix any constant $\varepsilon \in (0,1)$ (independent of H and K). Argue that the expected regret of ε-greedy over K episodes is linear in K:

$$\mathbb{E}[R(K)] \;\geq\; c_\varepsilon\,(H-1)\,K \qquad \text{whenever} \qquad K \;\leq\; \frac{2^{H-2}}{\varepsilon},$$

for some constant $c_\varepsilon \in (0,1)$ depending only on ε. A *high-level* argument is sufficient—you do not need to formalize every step. Your write-up should clearly state and justify each of the following key facts. (You may assume the convention $\widehat{Q}_h^k(s,a) = 0$ for any (s,a,h) that has never been visited in episodes $1, \ldots, k-1$.)

(1) After at most a constant number of episodes, with high probability the empirical estimates satisfy $\widehat{Q}_0^k(s_0, a_1) \geq (H-1)/2$ and $\widehat{Q}_0^k(s_0, a_2) = 0$, so the greedy action at s_0 is a_1. From then on, ε-greedy plays a_2 at s_0 with probability exactly $\varepsilon/2$.

(2) Conditional on the agent reaching a right-branch state $s_R^{(h)}$ that has not yet been fully traversed in any prior episode, $\widehat{Q}_h^k(s_R^{(h)}, \cdot) \equiv 0$, so greedy ties are broken uniformly. Conclude that the agent plays the correct next action $a_h^\star$ with probability exactly $1/2$.

(3) Combine (1) and (2) to argue that the per-episode probability of fully traversing the right branch is at most $\varepsilon \cdot 2^{-H}$.

(4) Use a union bound over K episodes to argue that, for $K \leq 2^{H-2}/\varepsilon$, with probability at least $3/4$ the right branch is *never* fully traversed during the K episodes.

(5) On the event "right branch never traversed," bound the per-episode return of ε-greedy by $(H-1)/2$, so the per-episode regret is at least $(H-1)/2$. Combine with (4) to conclude expected regret $\Omega(KH)$.

(c) **UCB-VI escapes the trap** (7 pts)**.** Recall the basic UCB-VI regret bound from lecture: with probability at least $1 - 2\delta$,

$$\sum_{k=1}^{K} \big(V_0^\star(s_1^k) - V_0^{\pi^k}(s_1^k)\big) \;\le\; H^2 S\sqrt{8AK\log(2KHSA/\delta)}.$$

(i) Plugging in $S = H+2$ and $A = 2$ for M_H, give an explicit polynomial-in-H upper bound on UCB-VI's regret over K episodes (ignoring log factors).

(ii) Determine an explicit threshold $K_0(H)$ such that for $K \ge K_0(H)$, UCB-VI's average per-episode regret is at most $(H-1)/4$. Compare this $K_0(H)$ with the K-regime $K \le 2^{H-2}/\varepsilon$ from part (b).

(iii) In one paragraph, explain qualitatively *why* UCB-VI succeeds. Specifically, for an unvisited triple $(s_R^{(h)}, a, h)$, what is the value of the bonus $b_h^k(s_R^{(h)}, a)$, and what does the resulting $\widehat{V}_h^k(s_R^{(h)})$ look like? Why does this make $\widehat{V}_0^k(s_0)$ for the a_2 branch larger than the a_1 branch, even before any positive reward is observed in the right chain?

Problem 3 (MLE, Pessimism, and Optimism: An Empirical Comparison (45 pts))**.**
In this problem you will implement and compare three offline-RL estimators on a small finite-horizon MDP. Throughout, the rewards $r_h(s,a)$ are deterministic and known; only the transitions P_h must be estimated from data. Use any programming language you like; submit your code as an appendix or alongside your write-up. All experiments must be averaged over at least 50 random seeds.

The test MDP. Use the following H-stage chain MDP. The state space is $\mathcal{S} = \{s_0, s_1, \ldots, s_N\}$ (so $S = N+1$), the action space is $\mathcal{A} = \{a_R, a_L\}$, and transitions are:

- From s_i with $i < N$: $a_R \to s_{i+1}$ deterministically; $a_L \to s_{\max(i-1,0)}$ deterministically.
- From s_N: any action stays at s_N.

Rewards: $r_h(s_N, \cdot) = 1$ for all h; $r_h(s_0, a_L) = 0.05$ for all h; $r_h(s,a) = 0$ otherwise. The initial state is s_0 deterministically. Use $H = 10$ and $N = 5$ unless otherwise specified.

(a) **Backward dynamic programming with a signed bonus** (10 pts)**.** Implement a single function with the signature

```
policy_from_model(P, r, b, H)
```

that takes

- `P`: array of shape (H, S, A, S) where $P[h, s, a, s']$ is the (true or empirical) probability $P_h(s' \mid s, a)$;
- `r`: array of shape (H, S, A);
- `b`: array of shape (H, S, A) representing a *signed* bonus added to the per-step Bellman update;
- `H`: the horizon;

and returns $(\widehat{Q}, \widehat{V}, \widehat{\pi})$ defined by the backward recursion

$$\widehat{V}_H \equiv 0, \qquad \widehat{Q}_h(s,a) = r_h(s,a) + b_h(s,a) + \sum_{s'} P_h(s' \mid s,a)\, \widehat{V}_{h+1}(s'), \qquad \widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a),$$

with $\widehat{\pi}_h(s) \in \arg\max_a \widehat{Q}_h(s,a)$. By choosing P and b appropriately, this single function computes:

- the optimal policy $\pi^\star$ (using $P = P^{\text{true}}$, $b \equiv 0$);
- the MLE plug-in policy $\widehat{\pi}$ (using $P = \widehat{P}$, $b \equiv 0$);
- the pessimistic policy $\tilde{\pi}$ (using $P = \widehat{P}$, $b = -|b|$);
- the optimistic policy $\bar{\pi}$ (using $P = \widehat{P}$, $b = +|b|$).

(b) **True policy evaluation** (5 pts)**.** Implement a function

`evaluate_policy(P_true, r_true, pi, H)`

that returns $V_0^\pi(s_0)$ under the *true* MDP via the policy-Bellman recursion

$$V_h^\pi(s) = r_h(s, \pi_h(s)) + \sum_{s'} P_h^{\text{true}}(s' \mid s, \pi_h(s))\, V_{h+1}^\pi(s'), \qquad V_H^\pi \equiv 0.$$

You will use `evaluate_policy` on the policies returned by `policy_from_model` to assess their true performance.

(c) **Three logging policies** (10 pts)**.** Implement and use the following three data-collection schemes. Each produces a dataset of N_{ep} episodes (i.e. $N_{\text{ep}} \cdot H$ transitions). For each, you will form the empirical kernel $\widehat{P}_h(s' \mid s,a) = n_h(s,a,s')/n_h(s,a)$, taking $\widehat{P}_h(\cdot \mid s,a)$ to be uniform on $\mathcal{S}$ when $n_h(s,a) = 0$ (this choice will not matter on the event used in part (d), since unvisited triples receive bonus H and are clipped).

(i) *Uniform.* At every step, $\pi_h^{\text{unif}}(\cdot \mid s) = \text{Unif}(\mathcal{A})$.

(ii) *Reasonable but biased.* The deterministic policy that takes a_R for $h \le \lfloor H/2 \rfloor$ and a_L for $h > \lfloor H/2 \rfloor$. This policy has good coverage in the early stages and zero coverage of a_R in the late stages.

(iii) *UCB-VI online.* Run UCB-VI for N_{ep} episodes against the true MDP, with per-episode bonus

$$b_h^k(s,a) := H\sqrt{\frac{\log(2N_{\text{ep}}HSA/\delta)}{2\,\max(n_h^k(s,a),\,1)}}, \quad \delta = 0.05,$$

and clipped at H. Collect all $N_{\text{ep}} \cdot H$ transitions into the dataset.

For each dataset, define the offline bonus

$$b_h^{\text{off}}(s,a) := \min\left\{ H,\ \ H\sqrt{\frac{\log(SAH/\delta)}{2\,\max(n_h(s,a),\,1)}} \right\}, \quad \delta = 0.05,$$

where unvisited triples $n_h(s,a) = 0$ receive bonus H.

(d) **Sweep and compare** (15 pts)**.** For each logging policy, sweep $N_{\text{ep}} \in \{50,\ 200,\ 1000,\ 5000\}$. For each pair (logger, N_{ep}):

(i) Compute the MLE plug-in policy $\widehat{\pi}$, the pessimistic policy $\tilde{\pi}$ (using $b = -b^{\text{off}}$), and the optimistic plug-in policy $\bar{\pi}$ (using $b = +b^{\text{off}}$) from `policy_from_model`.

(ii) Compute the sub-optimality $V_0^\star(s_0) - V_0^{\widehat{\pi}}(s_0)$ for each of $\widehat{\pi}, \tilde{\pi}, \bar{\pi}$ via `evaluate_policy`.

(iii) Average over ≥ 50 random seeds.

Plot all three sub-optimalities versus N_{ep} on log-log axes, with one panel per logging policy (three panels total). On each panel include all three estimators with clear legend.

(e) **Discussion** (5 pts)**.** Answer the following in a few sentences each.

(i) Why does pessimism beat MLE under the "reasonable but biased" logger but not under the uniform logger? Make your answer concrete by identifying which specific (s, a, h) triples are responsible for the gap.

(ii) Why is the optimistic plug-in policy $\bar{\pi}$ *worse* than MLE for offline deployment, even though optimism is essential for online exploration in UCB-VI? Frame your answer in terms of which value functions are concentration-controlled (the optimal $V^\star$ versus the learned $V^{\widehat{\pi}}$) and what the three estimators each guarantee in the offline setting.

(iii) Does the MLE policy under UCB-VI's collected data perform comparably to pessimism? What does this say about the role of pessimism vs. the quality of the data-collection policy?