CSE 542: Statistical Reinforcement Learning

Homework 1 Spring 2026

University of Washington Instructor: Kevin Jamieson

Due: 11:59 PM April ~~19~~ 26, 2026 **Points:** 100 (4 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).

Problem 1 (Reward Translation and Scaling (20 pts))**.** Let $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$ be a discounted infinite-horizon MDP with $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. Let V^*, Q^* denote the optimal value and action-value functions, and π^* any optimal policy.

(a) **Reward translation.** Define $M' = (\mathcal{S}, \mathcal{A}, P, r + c, \gamma, \mu)$ where $c \in \mathbb{R}$ is a constant (i.e. $r'(s, a) = r(s, a) + c$ for all (s, a)).

(i) Show that $V^*_{M'}(s) = V^*(s) + \dfrac{c}{1-\gamma}$ for all $s \in \mathcal{S}$.

(ii) Conclude that π^* is optimal in M' if and only if it is optimal in M.

(b) **Reward scaling.** Define $M'' = (\mathcal{S}, \mathcal{A}, P, \alpha r, \gamma, \mu)$ where $\alpha > 0$.

(i) Show that $V^*_{M''}(s) = \alpha\, V^*(s)$ for all $s \in \mathcal{S}$.

(ii) Conclude that π^* is optimal in M'' if and only if it is optimal in M.

(c) **Combining both.** Define $\widetilde{M} = (\mathcal{S}, \mathcal{A}, P, \alpha r + c, \gamma, \mu)$, where $\alpha > 0$ and $c \in \mathbb{R}$. An algorithm returns a policy $\hat{\pi}$ satisfying

$$V^*_{\widetilde{M}}(s) - V^{\hat{\pi}}_{\widetilde{M}}(s) \le \varepsilon \qquad \text{for all } s \in \mathcal{S}.$$

Express this guarantee as a bound on $V^*(s) - V^{\hat{\pi}}(s)$ in terms of ε, c, α, and γ.

Problem 2 (Performance Difference Lemma: Discounted Setting (25 pts))**.** The *Performance Difference Lemma* (PDL) quantifies the gap between two policies via a local advantage comparison. Let $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$ be a discounted infinite-horizon MDP. For a stationary policy π, define the discounted state-action occupancy

$$d^\pi(s, a) \;=\; (1-\gamma) \sum_{t=0}^{\infty} \gamma^t\, \mathbb{P}_\pi(s_t = s,\, a_t = a \mid s_0 \sim \mu),$$

and the *advantage function*

$$A^\pi(s, a) \;=\; Q^\pi(s, a) - V^\pi(s).$$

Also define

$$V^\pi(\mu) := \mathbb{E}_{s_0 \sim \mu}[V^\pi(s_0)]\,.$$

Note that $A^\pi(s, \pi(s)) = 0$ for all s, since $V^\pi(s) = Q^\pi(s, \pi(s))$.

(a) **(15 pts)** Prove that for any two stationary policies π, π' and initial distribution μ,

$$V^{\pi'}(\mu) - V^{\pi}(\mu) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim d^{\pi'}}[A^{\pi}(s,a)]\,.$$

In the special case that π' is deterministic, this can be written as

$$V^{\pi'}(\mu) - V^{\pi}(\mu) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_S^{\pi'}}\big[A^{\pi}(s,\pi'(s))\big]\,,$$

where $d_S^{\pi'}$ is the state marginal of $d^{\pi'}$.

Hint: Write the return difference as a telescoping sum by adding and subtracting $\gamma^t V^{\pi}(s_t)$ inside the expectation under trajectory of π':

$$V^{\pi'}(\mu) - V^{\pi}(\mu) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi'}[r(s_t,a_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)]\,.$$

Recognize the summand as $A^{\pi}(s_t,a_t)$ and write $V^{\pi'}(s_0) - V^{\pi}(s_0) = \sum_{t=0}^{\infty}\gamma^t \mathbb{E}_{\pi'}[A^{\pi}(s_t,a_t) \mid s_0]$. Then rewrite the sum using $d^{\pi'}$.

(b) **(10 pts)** Let $\hat{Q}^*$ be an estimate of Q^* and let $\hat{\pi}(s) = \arg\max_a \hat{Q}^*(s,a)$. Using the PDL from part (a), prove

$$V^*(s) - V^{\hat{\pi}}(s) \;\le\; \frac{2}{1-\gamma}\,\left\|Q^* - \hat{Q}^*\right\|_\infty \qquad \text{for all } s\in\mathcal{S}.$$

Hint: Apply the PDL with $\pi = \pi^*$, $\pi' = \hat{\pi}$, and initial distribution $\mu = \delta_s$. The resulting expression involves $A^*(s,\hat{\pi}(s)) = Q^*(s,\hat{\pi}(s)) - V^*(s)$. To lower-bound $Q^*(s,\hat{\pi}(s))$, use: (i) the definition of $\hat{\pi}$ as greedy w.r.t. $\hat{Q}^*$, so $\hat{Q}^*(s,\hat{\pi}(s)) \ge \hat{Q}^*(s,\pi^*(s))$, and (ii) Q^* and $\hat{Q}^*$ differ pointwise by at most $\left\|Q^* - \hat{Q}^*\right\|_\infty$.

Problem 3 (Performance Difference Lemma: Finite-Horizon Setting (20 pts))**.** Let $M = (\mathcal{S}, \mathcal{A}, \{P_h\}, \{r_h\}, \mu)$ be a finite-horizon MDP with H stages, $h = 0,\dots,H-1$. The value and action-value functions of a non-stationary policy $\pi = (\pi_0,\dots,\pi_{H-1})$ are

$$V_h^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=h}^{H-1} r_t(s_t,a_t) \;\middle|\; s_h = s\right], \quad V_H^{\pi} \equiv 0,$$
$$Q_h^{\pi}(s,a) = r_h(s,a) + \mathbb{E}\big[V_{h+1}^{\pi}(s_{h+1}) \mid s_h = s,\, a_h = a\big]\,,$$

and the advantage function is $A_h^{\pi}(s,a) = Q_h^{\pi}(s,a) - V_h^{\pi}(s)$.

(a) **(10 pts)** Prove that for any two policies π, π' and fixed initial state s_0,

$$V_0^{\pi'}(s_0) - V_0^{\pi}(s_0) \;=\; \sum_{h=0}^{H-1} \mathbb{E}_{\pi'}\Big[A_h^{\pi}(s_h,\, \pi'_h(s_h)) \;\Big|\; s_0\Big]\,.$$

Hint: Analogously to Problem 2(a), write the return difference as:

$$\begin{aligned} V_0^{\pi'}(s_0) - V_0^{\pi}(s_0) &= \mathbb{E}_{\pi'}\left[\sum_{h=0}^{H-1} r_h(s_h,a_h) \;\middle|\; s_0\right] - V_0^{\pi}(s_0) \\ &= \sum_{h=0}^{H-1} \mathbb{E}_{\pi'}\big[r_h(s_h,a_h) + V_{h+1}^{\pi}(s_{h+1}) - V_h^{\pi}(s_h) \mid s_0\big]\,. \end{aligned}$$

where the sum telescopes because $V_H^{\pi} \equiv 0$.

(b) **(10 pts)** Let $\hat{Q}_h^*$ be estimates of Q_h^* for each h, and let $\hat{\pi}_h^*(s) = \arg\max_a \hat{Q}_h^*(s,a)$. Using the PDL from part (a), prove

$$V_0^*(s_0) - V_0^{\hat{\pi}^*}(s_0) \;\le\; 2 \sum_{h=0}^{H-1} \left\| Q_h^* - \hat{Q}_h^* \right\|_\infty .$$

Hint: Apply the PDL with $\pi = \pi^*$ and $\pi' = \hat{\pi}^*$. At each stage h, bound $V_h^*(s_h) - Q_h^*(s_h, \hat{\pi}_h^*(s_h))$ by the same two-step argument as in Problem 2(b).

Problem 4 (Finite-Horizon Error Propagation (35 pts))**.** Consider a finite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, \{P_h\}, \{r_h\}, \mu)$ with stages $h = 0, \dots, H-1$. Suppose we have a generative model and collect N independent transitions from $P_h(\cdot \mid s,a)$ for every (s,a,h). Let $\hat{P}_h$ be the empirical transition kernel. Define $\hat{V}_H^* \equiv 0$ and for $h = H-1, \dots, 0$,

$$\hat{Q}_h^*(s,a) = r_h(s,a) + \hat{P}_h \hat{V}_{h+1}^*(s,a), \qquad \hat{V}_h^*(s) = \max_a \hat{Q}_h^*(s,a),$$

where $\hat{P}_h f(s,a) = \sum_{s'} \hat{P}_h(s' \mid s,a)\, f(s')$. Write $e_h = \left\| Q_h^* - \hat{Q}_h^* \right\|_\infty$ for the Q-estimation error at stage h.

(a) **(8 pts)** Show that the error satisfies the recursion

$$e_h \;\le\; \left\| (P_h - \hat{P}_h) V_{h+1}^* \right\|_\infty + e_{h+1}.$$

Hint: Write $Q_h^* - \hat{Q}_h^* = (P_h - \hat{P}_h)V_{h+1}^* + \hat{P}_h(V_{h+1}^* - \hat{V}_{h+1}^*)$ and bound the second term using $\left\| V_{h+1}^* - \hat{V}_{h+1}^* \right\|_\infty \le e_{h+1}$ (which follows from $|\max_a f(a) - \max_a g(a)| \le \max_a |f(a) - g(a)|$).

(b) **(7 pts)** Unroll the recursion from (a) to obtain

$$e_0 \;\le\; \sum_{h=0}^{H-1} \left\| (P_h - \hat{P}_h) V_{h+1}^* \right\|_\infty .$$

(c) **(10 pts)** Apply Hoeffding's inequality to each term $\left\| (P_h - \hat{P}_h) V_{h+1}^* \right\|_\infty$, using the fact that $V_{h+1}^*(s') \in [0, H-h-1]$. Union-bound over all (s,a,h) triples to show that with probability at least $1-\delta$,

$$e_0 \;\le\; C\, \frac{H^2}{\sqrt{N}},$$

where C is a constant you should make explicit in terms of $|\mathcal{S}|, |\mathcal{A}|, H, \delta$.

(d) **(10 pts)** Now translate the Q-estimation guarantee into a policy performance guarantee.

(i) Invoke Problem 3(b) to conclude

$$V_0^*(s_0) - V_0^{\hat{\pi}^*}(s_0) \;\le\; 2 \sum_{h=0}^{H-1} e_h.$$

(ii) Using part (a), show that for every h,

$$e_h \le \sum_{h'=h}^{H-1} \left\| (P_{h'} - \hat{P}_{h'}) V^*_{h'+1} \right\|_\infty .$$

Deduce that

$$\sum_{h=0}^{H-1} e_h \;\le\; \sum_{h'=0}^{H-1} (h'+1) \left\| (P_{h'} - \hat{P}_{h'}) V^*_{h'+1} \right\|_\infty \;\le\; H \sum_{h'=0}^{H-1} \left\| (P_{h'} - \hat{P}_{h'}) V^*_{h'+1} \right\|_\infty .$$

(iii) Combining (i), (ii), and part (c), state the final $O(\cdot)$ bound on $V_0^*(s_0) - V_0^{\hat{\pi}^*}(s_0)$ in terms of H and N. How does it compare to the $O(H^2/\sqrt{N})$ rate for Q-estimation? What is the source of the extra factor?