NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER I EXAMINATION 2024–2025

MH4521 – REINFORCEMENT LEARNING

Nov/Dec 2024 Time Allowed: 2 hours

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FIVE (5)** questions and comprises **SEVEN (7)** printed pages.
2. Answer each question beginning on a **FRESH** page of the answer book.
3. This is a **RESTRICTED OPEN BOOK** exam. You are only allowed to bring into the examination hall **ONE DOUBLE-SIDED A4-SIZE REFERENCE SHEET WITH TEXTS HANDWRITTEN OR TYPED ON THE A4 PAPER WITHOUT ANY ATTACHMENTS** (e.g. sticky notes, post-it notes, gluing or stapling of additional papers)
4. Calculators may not be used.

QUESTION 1 **(20 marks)**

Consider a policy π and a bandit environment ν with a finite set $\mathcal{A}$ of actions. The regret R_n is defined as

$$R_n(\pi) = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right],$$

where $\mu^* = \max_{a \in \mathcal{A}} \mu_a$ and where X_t is the reward at round t.

(a) Show that $R_n(\pi) \geq 0$ for all policies π.

(b) Show that, if $R_n(\pi) = 0$ for some policy π, then $\mathbb{P}(\mu_{A_t} = \mu^*) = 1$ for all $t \in \{1, \ldots, n\}$.

(c) For any action $a \in \mathcal{A}$, Let Δ_a be the suboptimality gap and let $T_a(t)$ be the number of times action a was chosen after the end of round t. Prove the regret decomposition lemma, that is, prove that

$$R_n(\pi) = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)].$$

QUESTION 2 **(20 marks)**

Consider the MDP $M = (\mathcal{S}, \mathcal{A}, p, r)$, as characterised in Table 1, for which it holds that $r_\mathrm{f} > r_\mathrm{w} \geq 0$, that $\alpha, \beta \in [0,1]$, and that $\gamma \in (0,1)$. This MDP corresponds to a recycling robot, which can have high (`h`) or low (`l`) battery, which can either fetch (`f`) an item to recycle, wait (`w`) or possibly charge (`c`).

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
`h`	`f`	`h`	α	r_f
`l`	`f`	`l`	β	r_f
`h`	`f`	`l`	$p(\mathtt{l} \mid \mathtt{h}, \mathtt{f})$	r_f
`l`	`f`	`h`	$p(\mathtt{h} \mid \mathtt{l}, \mathtt{f})$	-1
`h`	`w`	`h`	1	r_w
`l`	`w`	`l`	1	r_w
`l`	`c`	`h`	1	0

Table 1: Transitions and deterministic rewards in the MDP M.

(a) What are the probabilities $p(\mathtt{l} \mid \mathtt{h}, \mathtt{f})$ and $p(\mathtt{h} \mid \mathtt{l}, \mathtt{f})$?

(b) Define the action sets $\mathcal{A}(\mathtt{h})$ and $\mathcal{A}(\mathtt{l})$ explicitly. Can we define an MDP $M' = (\mathcal{S}, \mathcal{A}', p', r')$ with $\mathcal{A}'(\mathtt{h}) = \mathcal{A}'(\mathtt{l})$ such that M and M' have the same optimal state-value function v_*? If yes, state sufficient conditions on $\mathcal{A}'$, p', and r'. If no, explain why.

(c) Write the Bellman equation for the optimal state-value function v_* for the MDP M as a function of γ, α, β, r_f, r_w, and v_* itself, simplifying as much as possible.

(d) Consider the case where $r_\mathrm{f} = 1$, $r_\mathrm{w} = 0$, and $\alpha = \beta = \gamma = 1/2$. Apply one pass of the value iteration algorithm starting at $v_0(\mathtt{h}) = 8/5$ and $v_0(\mathtt{l}) = 4/5$, then estimate the policy based on the obtained state-value function. What can you say about this policy?

QUESTION 3 **(20 marks)**

Consider the case where M episodes have been played with a policy π_{b} and, for any $i \in \{1, \ldots, M\}$, denote by T_i the terminal time of the i-th episode and $G_{i,t}$ the return at time $t \in \{0, \ldots, T_i - 1\}$ in the i-th episode. We want to evaluate a policy π that is different from π_{b} in general. Define $\rho_{i,t}$ as the ratio

$$\rho_{i,t} = \frac{\pi(a_{i,t} \mid s_{i,t})}{\pi_{\mathrm{b}}(a_{i,t} \mid s_{i,t})}$$

with $(s_{i,t}, a_{i,t})$ the state-action pair at time t in the i-th episode. Finally, define $\rho_{i,t:T_i-1}$ as the product $\prod_{t'=t}^{T_i-1} \rho_{i,t'}$.

(a) For a given state $s \in \mathcal{S}$, let $\mathcal{T}(s)$ be the set of pairs (i, t) such that the state at time t in episode i is equal to s. Consider the following two estimators of the value $v_\pi(s)$

$$V_1(s) \doteq \frac{\sum_{(i,t)\in\mathcal{T}(s)} \rho_{i,t:T_i-1} G_{i,t}}{|\mathcal{T}(s)|} \qquad \text{and} \qquad V_2(s) \doteq \frac{\sum_{(i,t)\in\mathcal{T}(s)} \rho_{i,t:T_i-1} G_{i,t}}{\sum_{(i,t)\in\mathcal{T}(s)} \rho_{i,t:T_i-1}}.$$

Compare these two estimators in the simple case where $\mathcal{T}(s)$ contains a single element equal to $T_1 - 1$, i.e., $\mathcal{T}(s) = \{(1, T_1 - 1)\}$ by

i) checking whether they are unbiased and,

ii) considering what happens when π and π_{b} differ significantly.

(b) Let $G_1, G_2, \ldots, G_n$ be realisations of the return from the same state $s \in \mathcal{S}$ and define an estimate of the value at s as

$$\hat{V}_n \doteq \frac{1}{C_n} \sum_{k=1}^{n} W_k G_k,$$

for $n \geq 2$, with $C_n = \sum_{k=1}^{n} W_k$.

i) Relate $\hat{V}_n$ to either $V_1(s)$ or $V_2(s)$ and define n, W_k, and G_k based on $\mathcal{T}(s)$, $\rho_{i,t}$, and $G_{i,t}$, with $i \in \{1, \ldots, N\}$ and $t \in \{0, \ldots, T_i - 1\}$.

ii) A new realisation G_{n+1} of the return at s is made available, together with the corresponding weight W_{n+1}. Express $\hat{V}_{n+1}$ and C_{n+1} as a function of $\hat{V}_n$, C_n, G_{n+1} and W_{n+1}.

(c) i) For this question, we consider a given episode $i \in \{1, \ldots, N\}$ and omit i from the notations for the sake of simplicity. Prove that

$$\mathbb{E}_{\pi_{\mathrm{b}}}[\rho_{t:T-1} R_{t+k} \mid S_t = s] = \mathbb{E}_{\pi_{\mathrm{b}}}[\rho_{t:t+k-1} R_{t+k} \mid S_t = s].$$

ii) Propose an expression for $V_1(s)$ of the form

$$V_1(s) = \frac{1}{|\mathcal{T}(s)|} \sum_{(i,t)\in\mathcal{T}(s)} \tilde{G}_{i,t},$$

by explicitly defining $\tilde{G}_{i,t}$, simplifying as much as possible.

QUESTION 4 **(20 marks)**

We consider an episodic task in which the policy is parameterised via the exponential soft-max distribution as

$$\pi(a \mid s, \theta) \propto \exp(h(s, a, \theta)),$$

for some parameter θ and some preference function h.

(a) Give and explain two of the advantages of this approach when compared to approximating the action-value function directly?

(b) Given that the gradient of the objective function $J(\theta)$ can be expressed as

$$\nabla J(\theta) = \sum_s \eta(s) \sum_a \nabla \pi(a \mid s, \theta) q_\pi(s, a),$$

with $\eta(s)$ the expected number of time steps spent in state s in a single episode, prove that

$$\nabla J(\theta) \propto \mathbb{E}_\pi \big[G_t \nabla \log \pi(A_t \mid S_t, \theta) \big].$$

(c) Let $f(x) = 1/(1 + e^{-x})$ be the logistic function and consider the case where only two actions $\{1, 2\}$ are available in all states, with $h(s, 1, \theta) - h(s, 2, \theta) = \theta^\intercal x(s)$ for some feature vector $x(s)$. Express the policy $\pi(\cdot \mid s, \theta)$ via the logistic function.

(d) Now consider the case where the action space is $\mathcal{A}(s) = \mathbb{R}$ for all $s \in \mathcal{S}$, with

$$\pi(a \mid s, \theta) \propto \frac{1}{\sigma(s, \theta)} \exp\bigg(- \frac{1}{2\sigma(s, \theta)^2} (a - \mu(s, \theta))^2 \bigg).$$

i) Is this policy part of the family of policies parameterised via the exponential soft-max distribution? If yes, define the function h corresponding to it. If no, explain what additional assumptions need to be made for it to be the case.

ii) Consider the case where $\sigma(s, \theta)$ does not actually depend on θ and is written as $\sigma(s)$. Let $\mu(s, \theta) = \theta^\intercal x(s)$ with $x(s)$ some given feature vector. How would the vector θ_t in the t-th iteration of the REINFORCE algorithm be updated? Write the recursion explicitly as a function of x, σ, μ, a learning rate α, and the current state-action pair (S_t, A_t).

QUESTION 5 **(20 marks)**

Proximal Policy Optimisation (PPO) is an algorithm that updates the parameters of the policy network π_θ via

$$\theta_{k+1} = \arg\max_{\theta} J_{\theta_k}(\theta)$$

where the objective function is $J_{\theta_k}(\theta) = \mathbb{E}_{\pi_{\theta_k}}\big[L(S, A, \theta_k, \theta)\big]$ and where

$$L(s, a, \theta_k, \theta) = \min\left\{\frac{\pi_\theta(a \mid s)}{\pi_{\theta_k}(a \mid s)} C_{\theta_k}(s, a),\ g(\epsilon, C_{\theta_k}(s, a))\right\},$$

with $C_{\theta_k}(s, a)$ the advantage function and with

$$g(\epsilon, x) = \begin{cases} (1+\epsilon)x & \text{if } x \geq 0 \\ (1-\epsilon)x & \text{if } x < 0, \end{cases}$$

for any $x \in \mathbb{R}$ and for some given $\epsilon \in (0, 1)$.

(a) Study the behaviour of $L(s, a, \theta_k, \theta)$, i.e., when and how much it increases or decreases, considering separately the cases where the advantage function is non-negative and when it is negative.

(b) What are the benefits when using the objective function $J_{\theta_k}(\theta)$ in general, as well as compared to TRPO?

(c) A parameterised value function V_ϕ is required to define the advantage function. If $\{\tau_i\}_{i=1}^N$ is a collection of trajectories, with trajectory τ_i having return $\hat{G}_{i,t}$ at time $t \in \{0, \ldots, T_i - 1\}$, for any $i \in \{1, \ldots, N\}$, then how can the value function V_ϕ be trained?

END OF PAPER