EL2805 Reinforcement Learning

Exam – January 14, 2021

Division of Decision and Control Systems
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Exam (tentamen), January 14, 2021, kl 8.00 - 13.00

Aids. Slides of the lectures (**not exercises**), lecture notes.

Observe. Do not treat more than one problem on each page. Each step in your solutions must be motivated. Write a clear answer to each question. Write name and personal number on each page. Please only use one side of each sheet. Mark the total number of pages on the cover.
You will **stop writing at 12:40** to ensure you have enough time to upload your solution on canvas.

The exam consists in 5 problems. The grading policy will be adapted to the fact that the zoom exam is 20 mins shorter than the on-site exam.

Grading.

Grade A: ≥ 43	Grade B: ≥ 38
Grade C: ≥ 33	Grade D: ≥ 28
Grade E: ≥ 23	Grade Fx: ≥ 21

Responsible. Alexandre Proutiere 087906351

Results. Posted no later than January 28, 2021

Good luck!

Problem 1

Answer the following questions and **justify your answers** (unjustified answers will bring no points).

1. Let $(B_t)_{t\geq 1}$ denote a sequence of i.i.d. Bernoulli random variables with mean p (i.e., $\mathbb{P}[B_t = 1] = p$ and $\mathbb{P}[B_t = 0] = 1 - p$). Define the sequence $(Y_t)_{t\geq 1}$ as follows. $Y_0 = 0 = Y_1$, and for all $t \geq 1$, $Y_{t+1} = Y_t - Y_{t-1} + B_t$. Is $(Y_t)_{t\geq 1}$ a Markov chain? Is $(X_t)_{t\geq 1} = (Y_{t+1}, Y_t)_{t\geq 1}$ a Markov chain? [2 pts]

2. Many RL problems are concerned with infinite-horizon discounted MDPs. Such an MDP is characterized by λ, the discount factor, and stationary transition probabilities $p(s'|s, a)$ and reward function $r(s, a)$. Is there an undiscounted stationary MDP with terminal state equivalent to this MDP? Equivalent in the sense that the objective functions (cumulative expected rewards) are the same. If so, provide its transition probabilities and reward function. [2 pts]

3. Consider an infinite horizon discounted MDP. Both the transition probabilities and rewards are known to you. What methods would you use to solve the problem? [1 pt]

4. Give two examples of algorithms you have seen in the course that are based on stochastic approximation. Provide the functions they try to find the root of. [3 pts]

5. Propose a way so that under the SARSA algorithm with ϵ-greedy, the Q-values converge to the true Q-function. [1 pt]

6. Let $(X_t)_{t\geq 1}$ be a stochastic process with values in $[-1, 1]$ with slowly evolving mean $\mu_t = \mathbb{E}[X_t]$. Our goal is to *track* this moving average, i.e., to build, from the observations of the process, an estimate $\hat{\mu}_t$ of its evolving mean that is accurate at all time. We propose: $\hat{\mu}_0 = 0$, and at time t, after observing X_t, we update our estimate as follows

$$\hat{\mu}_{t+1} = \hat{\mu}_t + \alpha_t(X_t - \hat{\mu}_t).$$

   α_t is a *learning rate*. Would you choose a constant learning rate, e.g., $\alpha_t = 0.1$ or a decreasing one, e.g., $\alpha_t = \frac{1}{t+1}$? [1 pt]

Problem 1 - Solution

1. $(Y_t)_{t\geq 1}$ is not a Markov chain because $\mathbb{P}[Y_{t+1} = x|Y_t, Y_{t-1}] \neq \mathbb{P}[Y_{t+1} = x|Y_t]$ (the value of Y_{t-1} impacts that of Y_{t+1}. Then we have: if we denote $X_t = (X_t(1), X_t(2))$,

$$\begin{cases} X_{t+1}(1) = X_t(1) - X_t(2) + B_{t+1} \\ X_{t+1}(2) = X_t(1) \end{cases}$$

   Hence we can write $X_{t+1} = f(X_t) + (B_{t+1}, 0)$ where F is a deterministic function. $(X_t)_{t\geq 1}$ is a Markov chain.

2. Yes. The new state space is $\mathcal{S}' = \mathcal{S} \cup \{\emptyset\}$ where $\emptyset$ represents the terminal state. The transition probabilities p' are defined as follows:

$$p'(s'|s, a) = \lambda p(s'|s, a), \quad \text{if } s, s' \neq \emptyset.$$

$$p'(\emptyset|s, a) = 1 - \lambda, \quad \text{if } s \neq \emptyset.$$

$$p'(\emptyset|\emptyset, a) = 1.$$

   The same reward function as the initial MDP, with $r'(\emptyset, a) = 0$.

3. Policy Iteration and Value Iteration algorithms.

4. SARSA with ϵ-greedy and Q-learning are two algorithms based on Stochastic approximation. For Q-learning the fixed point equation we try to solve is

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} \quad Q(s, a) = r(s, a) + \lambda \sum_{s' \in \mathcal{S}} p(s'|s, a) \max_{a' \in \mathcal{A}} Q(s', a').$$

   For SARSA with ϵ-greedy, the policy converges to a randomized policy that is ϵ-greedy w.r.t. to the limiting Q-values. The fixed point equation we try to solve is

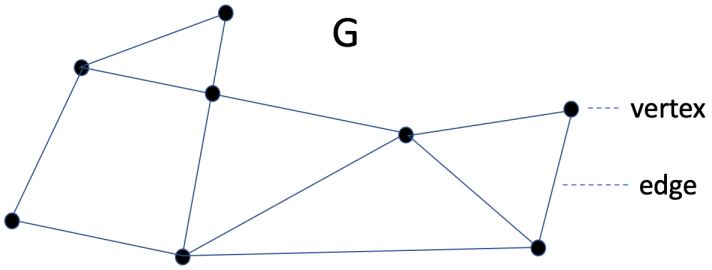
$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} \quad Q(s, a) = r(s, a) + \lambda \sum_{s' \in \mathcal{S}} p(s'|s, a) \left(\frac{\epsilon}{A_{s'}} \sum_{b \in A_{s'}} Q(s', b) + (1 - \epsilon) \max_{b \in A_{s'}} Q(s', b) \right)$$

5. Decrease ϵ slowly (all (state, action) pairs should be visited infinitely often), with limit $\varepsilon \to 0$.

6. Select a constant learning rate. This will result in an exponential weighted decay moving average algorithm. On the other hand choosing $1/(t + 1)$ will resulting in a stochastic approximation algorithm that will converge to some fixed value.

Problem 2 – Modelling using MDPs

1. **Strategic vaccinations.** We model the population as a finite undirected graph $G = (V, E)$ with n vertices/nodes, where V is the set of vertices/nodes and E the set of edges (see below an example of graph). All nodes are initially healthy but not vaccinated. At the beginning of day 0, node s_0 gets infected. At the beginning of each day $t \geq 1$, the node infected in the previous day selects one of its uninfected neighbors uniformly at random, and this node becomes infected if it is not vaccinated. If it is vaccinated, the epidemic stops spreading. At the end of each day, you observe the set of infected and vaccinated nodes and the node that just became infected, and you can select a node in the graph and vaccinate it. The goal is to optimally select nodes to vaccinate so as to minimize the expected number of infected nodes at the end of day T. Model this problem as a MDP. [4 pts]



2. **Clinical trials – A Bayesian formulation of bandits.** You are a doctor. You receive patients sequentially. These patients suffer from the same disease, and you have two treatments A and B. When a patient arrives, you decide to administrate either A or B; the effect of the treatment is immediate, and the patient either recovers or dies. The success probabilities of the two treatments p_A and p_B are unknown, and chosen initially by nature uniformly at random in [0,1]. For example $\mathbb{P}[p_A \in (u, v)] = \int_u^v ds = v - u$. The successes of the treatments are independent across patients and treatments. By sequentially selecting the two treatments, you will gather information about p_A and p_B, and in turn learn their approximate values. Your objective is to cure a maximal expected number of patients among the T first patients. We are going to model this problem as an MDP. But before we do that, we explain how the information gathered allows us to update our knowledge of p_A and p_B.

   (a) Assume that you have selected $N_A = n$ times treatment A. What is the likelihood (i.e., the probabiltiy) that it was successful for $S_A = m \leq n$ patients given that the success rate of A is equal to $u \in [0, 1]$? Show that the likelihood (i.e., the probabiltiy) that it was successful for $S_A = m \leq n$ patients (averaged over all possible values of p_A) is $\int_0^1 \binom{n}{m} u^m (1-u)^{n-m} du$. [1 pt]

   (b) What is the *posterior* distribution of p_A given that you have selected $N_A = n$ times treatment A, and that it was successful for $S_A = m \leq n$ patients? In other words compute $\mathbb{P}[p_A \in (u, v) | m$ patients cured out of $n]$. This distribution is called the beta distribution $\beta(m+1, n-m+1)$ (because its normalizing constant is the beta function). [1 pt]

   (c) Prove that the mean of a random variable with distribution $\beta(m + 1, n - m + 1)$ is $(m + 1)/(n + 2)$. *[Hint: use the facts that for a, b strictly positive integers, $\int_0^1 s^{a-1}(1 - s)^{b-1} ds = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ and $\Gamma(a + 1) = a\Gamma(a)$.* We can interpret this mean as the probability that treatment A is successful for the next patient if we have already selected $N_A = n$ times treatment A, and if it was successful for $S_A = m \leq n$ patients. [1 pt]

   (d) Using the answers to the previous questions, model the clinical problem as an MDP (Remember that importantly, at the beginning the success probabilities of the treatments are unknown). You are allowed to use random rewards in the MDP. [3 pts]

Problem 2 - Solution

1. **States.** Before deciding which node to vaccinate, you observe the sets of infected and vaccinated nodes, you further observe the node that just became infected. Hence the state can be (I, S, v) where I and S are the sets of infected and vaccinated (safe) nodes, and v is the node infected at the beginning of the day. When the epidemic stops, we take $v = \emptyset$.

   Actions space. Let $x = (I, S, v)$ be a state. The set $\mathcal{A}_x$ of actions available in this state is $V \setminus (I \cup S)$.

   Transition probablities. Let $v \neq \emptyset$. Let $x = (I, S, v)$ be a state, and let $w \in \mathcal{A}_x$ denote an action available in this state. The next state can be either $x' = (I \cup \{w'\}, S \cup \{w\}, w')$ for $w' \neq w$ and w' neighbor of v in G, or $x = (I, S \cup \{w\}, \emptyset)$ if w is a neighbor of v in G and the epidemic stops.

   Denote by N_v the number of neighbors of v in G that are not in $I \cup S$. Denote by $\mathcal{N}_v$ this set of neighbors. Then: if w is not a neighbor of v,

$$\forall w' \in \mathcal{N}_v, \quad p\bigg((I \cup \{w'\}, S \cup \{w\}, w')|(I, S, v), w\bigg) = \frac{1}{N_v},$$

   and if w is a neighbor of v in G,

$$\forall w' \in \mathcal{N}_v \setminus \{w\}, \quad p\bigg((I \cup \{w'\}, S \cup \{w\}, w')|(I, S, v), w\bigg) = \frac{1}{N_v},$$

$$\text{and} \quad p\bigg((I, S \cup \{w\}, \emptyset)|(I, S, v), w\bigg) = \frac{1}{N_v}.$$

   Now the transitions after the epidemic stops are trivial:

$$p\bigg((I, S \cup \{w\}, \emptyset)|(I, S, \emptyset), w\bigg) = 1.$$

   Rewards. the reward is collected at the end: $r_t(x, w) = 0$ for all state x, action w and $t < T$; and $r_T((I, S, v), w) = |I|$ (the cardinality of the set I.

2. (a) Because of independence, $\mathbb{P}[m \text{ patients cured out of } n | p_A = u] = \binom{n}{m} u^m (1-u)^{n-m}$. Now, averaging over p_A (uniformly distributed over $[0, 1]$), we get $\mathbb{P}[m \text{ patients cured out of } n] = \int_0^1 \binom{n}{m} s^m (1-s)^{n-m} ds$

2. (b) We have:

$$\begin{aligned}\mathbb{P}[p_A \in (u, v) | m \text{ patients cured out of } n] &= \frac{\mathbb{P}[p_A \in (u, v) \text{ and } (m \text{ patients cured out of } n)]}{\mathbb{P}[m \text{ patients cured out of } n]} \\ &= \frac{\int_u^v \binom{n}{m} s^m (1-s)^{n-m} ds}{\int_0^1 \binom{n}{m} s^m (1-s)^{n-m} ds}.\end{aligned}$$

2. (c) In view of the above density, the mean is equal to:

$$\begin{aligned}\frac{\int_0^1 s \times s^m (1-s)^{n-m} ds}{\int_0^1 s^m (1-s)^{n-m} ds} &= \frac{\int_0^1 s^{m+1} (1-s)^{n-m} ds}{\int_0^1 s^m (1-s)^{n-m} ds} \\ &= \frac{\Gamma(m+2)\Gamma(n-m+1)}{\Gamma(n+3)} \frac{\Gamma(n+2)}{\Gamma(m+1)\Gamma(n-m+1)} \\ &= \frac{\Gamma(m+2)}{\Gamma(m+1)} \frac{\Gamma(n+2)}{\Gamma(n+3)} = \frac{m+1}{n+2}.\end{aligned}$$

2. (d) We model the problem as a finite-time MDP with time-horizon T.

State space. The state will represent the past observations (this is what the decision maker is taking into account to decide the next action). In round $t = 1, \dots, T$, we can represent the past observations by (N_A, S_A, S_B) where N_A is the number of times treatment A has been chosen, S_A is the number of times this treatment has been successful, and S_B is the number of times treatment B has been successful (note that we do not need to record N_B since $N_B = t - N_A$).

We can choose the state space $\mathcal{S} = \{(n_A, s_A, s_B) \in \mathbb{N}^3 : \exists t \in \{1, \dots, T\}, s_A \leq n_A, s_B \leq t - n_A\}$.

Action space. $\mathcal{A} = \{A, B\}$.

Transition probabilities. Let $t \in \{1, \dots, T\}$, let (N_A, S_A, S_B) be a feasible state at time t ($S_A \leq N_A \leq t$ and $S_B \leq t - N_A$):

- If A is selected,

$$p\Big((N_A + 1, S', S_B) | (N_A, S_A, S_B), A\Big) = \begin{cases} \frac{S_A+1}{N_A+2} & \text{if } S' = S_A + 1, \\ 1 - \frac{S_A+1}{N_A+2} & \text{if } S' = S_A. \end{cases}$$

- If B is selected,

$$p\Big((N_A, S_A, S') | (N_A, S_A, S_B), B\Big) = \begin{cases} \frac{S_B+1}{t-N_A+2} & \text{if } S' = S_B + 1, \\ 1 - \frac{S_B+1}{t-N_A+2} & \text{if } S' = S_B. \end{cases}$$

Rewards. Consider random rewards with Bernoulli distributions: in state $x = (N_A, S_A, S_B)$, the reward is Bernoulli with mean $\frac{S_A+1}{N_A+2}$ if A is selected, and $\frac{S_B+1}{t-N_A+2}$ if B is selected.

Problem 3 – Selling under pandemic risk

You have to sell your bike within the next T months. Your task is complicated by the risk of the start of a pandemic: at the beginning of each month, a pandemic starts with probability $c > 0$, and always remains active. At the end of the t-th month, you receive an offer x_t. The offers are statistically independent across months, but their distributions depend on the presence of the pandemic. Without pandemic, $x_t \in \{1, \dots, X\} \subset [0, 1]$ and has a distribution f_0 ($\mathbb{P}[x_t = x] = f_0(x)$), whereas in presence of the pandemic, $x_t \in \{1, \dots, X\}$ and has a distribution f_1. Typically f_1 would be of lower mean than f_0. If you did not manage to sell your bike within T months, the bike is lost. Your objective is to maximize the expected price at which you sell your bike.

1. Model the problem as an MDP. [2 pts]
2. Establish that the optimal policy is of the *double-threshold* kind: this means that you accept the offer only if it is larger than or equal to a threshold. *Double-threshold* means that there are two different threshholds, one (denoted by β_t at the end of the t-th month) if there is a pandemic, and one (denoted by α_t) in absence of pandemic. [2 pts]
3. Derive a recursive expression for the threshold β_t used at the end of t-th month in presence of the pandemic. [1 pt]
4. Derive a recursive expression for the thresholds (α_t, β_t) used at the end of t-th month. [2 pts]
5. Assume that T is very large. Further assume that f_0 (resp. f_1) can be approximated by the uniform distribution over $[0, 1]$ (resp. $[0, 1/2]$), and that $c = 1/2$. By that, we mean that for example, for all function g, $\sum_x f_0(x)g(x) \approx \int_0^1 g(u)du$ and $\sum_x f_1(x)g(x) \approx 2\int_0^{1/2} g(u)du$. What is the optimal policy during the first months? What is the optimal policy the few first months after the start of the pandemic? You can assume here that the mapping giving (α_t, β_t) from $(\alpha_{t+1}, \beta_{t+1})$ identified in the previous question is a contraction, so that (in view of Banach fixed point theorem) for any fixed t, (α_t, β_t) converges when T grows large.[2 pts]
6. Now suppose that if the pandemic starts at all, it disappears at the beginning of each month with probability $d > 0$ and never appears again. Briefly explain how would the optimal policy look like? [1 pt]

Problem 3 - Solution

1. Finite time horizon MDP with horizon T.
   States. The state should include the current offer and the information about the pandemic, active '1' or inactive '0'. State (B, x) where $B = 0$ or 1 (pandemic information), x (offer). We include the state $\emptyset$ to represent the fact we have accepted an offer.

   Actions. Accept (A) or Reject (R).

   Transition probabilities. Starting from a state where the pandemic is active:

$$\begin{cases} p((1,x')|(1,x),R) = f_1(x') & \forall x, x', \\ p(\emptyset|(1,x),A) = 1 & \forall x. \end{cases}$$

   Starting from a state where the pandemic is inactive:

$$\begin{cases} p((0,x')|(0,x),R) = (1-c)f_0(x') & \forall x, x', \\ p((1,x')|(0,x),R) = cf_1(x') & \forall x, x', \\ p(\emptyset|(0,x),A) = 1 & \forall x. \end{cases}$$

   Finally, $p(\emptyset|\emptyset,\cdot) = 1$.

   Rewards. $r_t((\cdot,\cdot),R) = 0$ (no reward if you reject; $r_t((\cdot,x),A) = x$; finally, $r_t(\emptyset,\cdot) = 0$.

2. Let us write Bellman's equation. Let $u_t^\star(z)$ denote the value of being in state z at time t. We have: $u_T^\star((\cdot,x)) = x$ (since we need to accept the last offer) and $u_T^\star(\emptyset) = 0$. For all $1 \le t < T$,

$$\begin{aligned} u_t^\star((0,x)) &= \max\left\{x, c\sum_{x'} f_1(x')u_{t+1}^\star((1,x')) + (1-c)\sum_{x'} f_0(x')u_{t+1}^\star((0,x'))\right\}, \\ u_t^\star((1,x)) &= \max\left\{x, \sum_{x'} f_1(x')u_{t+1}^\star((1,x'))\right\}, \\ u_t^\star(\emptyset) &= 0. \end{aligned}$$

   From this first equation, we deduce that at time t, it is optimal to accept an offer x when the pandemic is inactive if and only if:

$$x \ge \alpha_t := c\sum_{x'} f_1(x')u_{t+1}^\star((1,x')) + (1-c)\sum_{x'} f_0(x')u_{t+1}^\star((0,x')).$$

   Similarly, from the second equation, we deduce that at time t, it is optimal to accept an offer x when the pandemic is active if and only if:

$$x \ge \beta_t := \sum_{x'} f_1(x')u_{t+1}^\star((1,x')).$$

   Thus, the optimal policy if of double-threshold kind.

3. Note first that $\alpha_T = 0 = \beta_T$ since we accept the last offer whatever it is. Then for $1 \le t < T$, let us re-use Bellman's equation to get recursive expressions of the threshold β_t. Indeed note that by definition of the threshold, we have:

$$u_t^\star((1,x)) = \max\{x, \beta_t\}.$$

   We deduce that:

$$\begin{aligned} \beta_t &= \sum_{x'} f_1(x')u_{t+1}^\star((1,x')) \\ &= \sum_{x'} f_1(x')\max\{x', \beta_{t+1}\}. \end{aligned}$$

4. Similarly, by definition of α_t, we have:

$$u_t^\star((0,x)) = \max\{x, \alpha_t\}.$$

Hence:

$$\begin{aligned}\alpha_t &= c\beta_t + (1-c)\sum_{x'} f_0(x') u_{t+1}^\star((0,x')) \\ &= c\beta_t + (1-c)\sum_{x'} f_0(x') \max\{x', \alpha_{t+1}\}.\end{aligned}$$

We have established the following recursion for (α_t, β_t):

$$\alpha_T = 0 = \beta_T,$$
$$\text{for } 1 \le t < T, \begin{cases} \alpha_t = c\sum_{x'} f_1(x')\max\{x', \beta_{t+1}\} + (1-c)\sum_{x'} f_0(x')\max\{x', \alpha_{t+1}\} \\ \beta_t = \sum_{x'} f_1(x')\max\{x', \beta_{t+1}\}. \end{cases}$$

5. When T is very large, the sequence (α_t, β_t) for $t = T, T-1, \ldots$ would converge as t decreases. To prove this, we first write the recursions explicitly:

$$\beta_t = 2\int_0^{1/2} \max\{x, \beta_{t+1}\} dx = \beta_{t+1}^2 + \frac{1}{4}.$$

The function $f(\beta) = \beta^2 + \frac{1}{4}$ is monotonic and a strict contraction (derivate strictly less than 1) on $(0, 1/2)$. The fixed point theorem implies that β_t converges to $\beta \in [0, 1/2]$ at t decreases, where

$$\beta = \beta^2 + \frac{1}{4}.$$

Which gives $\beta = 1/2$. Now consider the thresholds α_t. The recursion writes:

$$\alpha_t = c\beta_t + (1-c)\frac{1}{2}(\alpha_{t+1}^2 + 1).$$

We can see that as for β_t, α_t is going to converge as T increases to α where:

$$\alpha = c\beta + (1-c)\frac{1}{2}(\alpha^2 + 1).$$

We get:

$$\alpha = 2\left(1 - \sqrt{1 - \frac{2\beta+1}{4}}\right) = 2\left(1 - \sqrt{1/2}\right) \approx 0.59.$$

You see that the risk of pandemic significantly decreases this initial threshold (it would have been 1 if $c = 0$).

You can replace the above reasoning by just invoking Banach fixed point theorem. Finally, we have shown that when t is small, we have an optimal static policy with the thresholds (α, β).

6. There will be a third threshold function γ_t for the case after the disappearance of the pandemic. Let $(2, x)$ be the state after the pandemic disappears when the offer is x. Bellman's equations would be:

$$\begin{aligned}
u_t^\star((0,x)) &= \max\left\{x, c\sum_{x'} f_1(x') u_{t+1}^\star((1,x')) + (1-c)\sum_{x'} f_0(x') u_{t+1}^\star((0,x'))\right\}, \\
u_t^\star((1,x)) &= \max\left\{x, d\sum_{x'} f_0(x') u_{t+1}^\star((2,x')) + (1-d)\sum_{x'} f_1(x') u_{t+1}^\star((1,x'))\right\}, \\
u_t^\star((2,x)) &= \max\left\{x, \sum_{x'} f_0(x') u_{t+1}^\star((2,x'))\right\}, \\
u_t^\star(\emptyset) &= 0.
\end{aligned}$$

Problem 4 - Variations around SARSA

A friend of yours is trying to control a small robot using Reinforcement Learning techniques. He is considering a stationary MDP with terminal state with discount factor λ. The MDP has finite state and action spaces, $\mathcal{S} = \{1, 2, \dots, S\}$ and $\mathcal{A}$. He has tried to implement SARSA with an ε-greedy policy. The policy, denoted by μ_t, used to select actions at time t is ε-greedy with respect to the current Q values. Unfortunately, his algorithm, whose pseudo-code is presented below (Alg. 1), does not work. Perhaps you can help him?

Algorithm 1 Faulty SARSA with ε-greedy policy

Input Episodes number $N \in \mathbb{N}$; discount factor $\lambda \in (0, 1]$; learning rate $\alpha_n = 1/\log(n+1)$; $\varepsilon = 0.1$

Initialize matrices $Q(s, a)$ and $n(s, a)$ to 0 for all (s, a)
for episodes $k = 1, 2, \dots, N$ **do**
  $t \leftarrow 1$
  Initialize s_1 and choose a_1 according to a uniform distribution over the actions.
  while Episode k is not finished **do**
    Take action a_t: observe reward r_t and next state s_{t+1}
    Choose a_{t+1} from s_{t+1} using μ_t: a ε-greedy policy with respect to Q
    Compute target value: if the episode is terminal then $y_t = 0$ otherwise

$$y_t = r_t + \max_a Q(s_{t+1}, a)$$

    $n(s_t, a_t) \leftarrow n(s_t, a_t) + 1$
    Update Q function: $Q(s_{t+1}, a_{t+1}) \leftarrow Q(s_t, a_t) - \alpha_{n(s_t,a_t)}(y_t - Q(s_t, a_t))$
    $t \leftarrow t + 1$
  end while
end for

1. Spot all the mistakes in Alg. 1. Motivate why those are mistakes, and correct them. [3 pts]

Your friend came across a variant of SARSA, which is defined through a sequence of policies $(\pi_t)_{t\geq 1}$, and consists in just changing, in Algorithm 1 **after corrections**, the way the target is computed. The target becomes:

$$y_t = r_t + \lambda \sum_a \pi_t(a|s_{t+1}) Q(s_{t+1}, a),$$

where $\pi_t(a|s)$ denotes the probability that a is selected in state s under π_t.

2. What sequence of policies $(\pi_t)_{t\geq 1}$ should you choose so that the corresponding variant of SARSA is on-policy? This variant is called Expected SARSA. [1 pt]

3. Consider an off-policy variant of SARSA corresponding to a stationary policy $\pi = \pi_t$ for all t. Under this algorithm, do the Q values converge? If so, what are the limiting Q values? Justify your answer. [1 pt]

4. Assume that $\mathcal{S} = \{1, 2, 3\}$ and $\mathcal{A} = \{A, B\}$. Consider the stationary policy π defined by $\pi(A|s) = \frac{1}{2s}$ for all $s \in \{1, 2, 3\}$. Your friend implements the SARSA variant with this stationary policy.

   4.(a) Your friend observes a trajectory $(1, A, 1); (3, B, 2); (1, B, 3); (1, B, 0); (2, B, 1); (3, \dots)$ where each triplet represents the state, the selected action and the corresponding reward. Suppose the Q function is initialized at 0 for each (state, action)-pair, and $\lambda = 1, \alpha = 0.5$. What are the Q values, and the greedy policy, after the observed trajectory? [2 pts]

   4.(b) Assume that the Q-values you found are those obtained after convergence of the algorithm. Compute the state-value function of π, i.e., $V^\pi(s)$ for each state. [1 pt]

5. Consider a stationary policy, under which in each round, we apply the decision rule π. Consider the targets y_t used in (i) the SARSA algorithm where the action selections are made under π in each round, and in (ii) the variant of SARSA corresponding to a stationary policy π (and where the actions are still selected using the ε-greedy policy w.r.t. the current Q values).

   5.(a) The *bias* of these targets is defined as $B(s,a) = Q^\pi(s,a) - \mathbb{E}[y_t | s_t = s, a_t = a]$, where Q^π denotes the (state, action) value function of π. Show that the two targets have the same bias. [1 pt]

   5.(b) Are the two algorithms converging to the same Q values? [1 pt]

Problem 4 - Solution

(1)

1. The learning rate α_t does not satisfy $\sum_{t\geq 1}\alpha_t^2 < \infty$ since we have $\log^2(t+1) \leq t+1$ for $t \geq 0$ and $\sum_{t\geq 1}\frac{1}{\log^2(t+1)} \geq \sum_{t\geq 1}\frac{1}{t+1} = \infty$. Alternatively one can use $\alpha_t = 1/t$.
2. When the episode is terminal the target value is $y_t = r_t$, not 0.
3. In the target value the discount factor in front of $\max_a$ is missing
4. In the target value it's not $\max_a Q(s_{t+1}, a)$ but $Q(s_{t+1}, a_{t+1})$.
5. In the Q function update we update $Q(s_t, a_t)$ not $Q(s_{t+1}, a_{t+1})$
6. in the Q update there is a $-$ that should be a $+$.

(2) It's on-policy if the policy π_t used to compute y_t is the same that is used to compute the action a_{t+1} (otherwise it's off policy).

(3) Off-policy algorithms, given that the behavior policy samples all states-action pairs infinitely often, will learn the Q-values of the policy used in the target value computation. To see this, define the value of a policy π as $V^\pi(s) = \mathbb{E}^\pi[\sum_{t\geq 0}\lambda^t r(s_t, a_t)|s_0 = s]$. We can expand the expectation to find out that

$$\begin{aligned} V^\pi(s) &= \sum_a \pi(a|s)\left(r(s,a) + \lambda\mathbb{E}^\pi[\sum_{t\geq 1}\lambda^{t-1}r(s_t,a_t)|s_0 = s, a_0 = a]\right) \\ &= \sum_a \pi(a|s)\underbrace{\left(r(s,a) + \lambda\mathbb{E}^\pi_{s'\sim P(\cdot|s,a)}[V^\pi(s')|s,a]\right)}_{Q^\pi(s,a)} \\ &= \sum_a \pi(a|s)Q^\pi(s,a) \end{aligned}$$

Then the Q values of this policy are defined as $Q^\pi(s,a) = r(s,a) + \lambda\mathbb{E}^\pi_{s'\sim P(\cdot|s,a)}[V^\pi(s')|s,a]$. Therefore Q^π is the fixed point of an operator H:

$$HQ^\pi(s,a) = r(s,a) + \lambda\mathbb{E}^\pi_{s'\sim P(\cdot|s,a)}[V^\pi(s')|s,a]$$

and we can see that the stochastic approximation algorithm will converge to Q^π since

$$\mathbb{E}[y_t|s_t,a_t] = r(s_t,a_t) + \lambda\mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}[\sum_a \pi(a|s_{t+1})Q^t(s_{t+1},a)]$$

notice that the term inside the expectation is just the estimate at time t of the value of π. Therefore in the limit the target value will converge to $Q^\pi(s,a)$ using the stochastic approximation algorithm.
(4.1) $(1,A,1);(3,B,2);(1,B,3);(1,B,0);(2,B,1);(3,\dots)$ with $\lambda = 1, \alpha = 1/2$. The sequence of Q values is as follows

$$Q^{(0)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \end{array}.$$

$Q^{(1)}(1,A) = 0.5$

$$Q^{(1)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \end{array}.$$

$$Q^{(2)}(3,B) = 0.5[2 + \pi(A|1)Q(1,A) + \pi(B|1)Q(1,B) - 0] = 0.5[2 + Q(1,A)/2] = 0.5[2 + 1/4] = 9/8$$

$$Q^{(2)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 9/8 \end{bmatrix} \end{array} .$$

$$Q^{(3)}(1,B) = 0.5[3 + \pi(A|1)Q(1,A) + \pi(B|1)Q(1,B) - 0] = 0.5[3 + Q(1,A)/2] = 0.5[3 + 1/4] = 13/8$$

$$Q^{(3)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1/2 & 13/8 \\ 0 & 0 \\ 0 & 9/8 \end{bmatrix} \end{array} .$$

$$Q^{(4)}(1,B) = 13/8 + 0.5[\pi(A|2)Q(2,A) + \pi(B|2)Q(2,B) - 13/8] = 13/16$$

$$Q^{(4)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1/2 & 13/16 \\ 0 & 0 \\ 0 & 9/8 \end{bmatrix} \end{array} .$$

$$Q^{(5)}(2,B) = 0.5[1 + \pi(A|3)Q(3,A) + \pi(B|3)Q(3,B)] = 0.5[1 + 45/48] = 93/96$$

$$Q^{(5)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1/2 & 13/16 \\ 0 & 93/96 \\ 0 & 9/8 \end{bmatrix} \end{array} .$$

Therefore the greedy policy is $\pi(1) = B, \pi(2) = B, \pi(3) = B$.

(4.2) The values are

$$V(1) = \sum_a \pi(a|1)Q(1,a) = (1/2 + 13/16)/2 = 21/32$$
$$V(2) = \sum_a \pi(a|2)Q(2,a) = \frac{3 \cdot 93}{4 \cdot 96} = \frac{93}{128}$$
$$V(3) = \sum_a \pi(a|3)Q(3,a) = \pi(2|3)9/8 = 45/48$$

(5) 1.Let y_t be the target value with SARSA and y'_t be the target value using the variant of SARSA. The difference in bias $\mathbb{E}[y_t|s_t = s, a_t = a] - \mathbb{E}[y_{t'}|s_t = s, a_t = a]$ is equal to

$$= \mathbb{E}_{s_{t+1} \sim P(\cdot|s,a), a_{t+1} \sim \pi(\cdot|s_{t+1})}[Q(s_{t+1}, a_{t+1})] - \mathbb{E}_{s_{t+1} \sim P(\cdot|s,a)}[\sum_{a'} \pi(a'|s_{t+1})Q(s_{t+1}, a')]$$
$$= \mathbb{E}_{s_{t+1} \sim P(\cdot|s,a)}[\mathbb{E}_{a' \sim \pi(\cdot|s_{t+1})}[Q(s_{t+1}, a')|s_{t+1}]] - \mathbb{E}_{s_{t+1} \sim P(\cdot|s,a)}[\mathbb{E}_{a' \sim \pi(\cdot|s_{t+1})}[Q(s_{t+1}, a')|s_{t+1}]]$$
$$= 0$$

2. Depends on the policy π. The variant of SARSA will converge to the true Q-values of π, since the behavior policy is ε-greedy. SARSA on the other hand is on policy. Therefore the only way that the two algorithms converge to the same Q-values is that π is exploring enough (i.e., $\pi(s,a) > 0$ for all state-action pairs).

Problem 5 – Towards natural policy gradient

Consider a RL problem whose underlying MDP has finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Suppose you parametrize your policy π_θ using an exponential family of probability distributions of the form:

$$\pi_\theta(s,a) = \exp\left(\theta^\top T(s,a) - A(\theta,s)\right). \tag{1}$$

Here $\pi_\theta(s,a)$ denotes the probability of selecting action a in state s under π_θ. The parameter θ has dimension d: $\theta \in \mathbb{R}^d$. $T : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a function called the sufficient statistic and the log-partition function $A : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ is a smooth real-valued function of θ.

1. Derive A the log-partition function so that π_θ is a probability distribution over actions. [1 pt]
2. Compute the score function $\nabla_\theta \log \pi_\theta(s,a)$ as a function of T, A and their gradients. [1 pt]
3. Show that $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[[\nabla_\theta \log \pi_\theta(s,a)][\nabla_\theta \log \pi_\theta(s,a)]^\top] = -\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\nabla^2_\theta \log \pi_\theta(s,a)]]$. Recall that the Hessian $\nabla^2_\theta f(\theta)$ of a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ at θ is the $d\times d$ matrix whose i-th line, j-th column entry is $\frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\theta)$. You may first establish the following identity:

$$\nabla^2_\theta(\log f)(\theta) = \frac{\nabla^2_\theta f(\theta)}{f(\theta)} - \frac{\nabla_\theta f(\theta)\nabla_\theta f(\theta)^\top}{f(\theta)^2}.$$

[2 pts]

4. Show that the Fisher information $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[[\nabla_\theta \log \pi_\theta(s,a)][\nabla_\theta \log \pi_\theta(s,a)]^\top]$ is given by $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[[\nabla_\theta \log \pi_\theta(s,a)][\nabla_\theta \log \pi_\theta(s,a)]^\top] = \nabla^2_\theta A(\theta,s)$. [1 pt]

Consider a discounted MDP with stationary rewards and transition probabilities. You decide that Actor-Critic algorithms are the way to proceed. Hence, for π_θ as in (1), you introduce a so-called compatible function approximator. That is, we assume that the critic has the following parametrization (in ω): $Q^{\pi_\theta}_\omega(s,a) = \omega^\top[\nabla_\theta \log \pi_\theta(s,a)]$. To track the (state, action) value function of π_θ, the critic aims at minimizing over ω the Mean-Squared-Error (MSE):

$$\varepsilon(\omega) = \sum_s d(s) \sum_a \pi_\theta(s,a)[Q^{\pi_\theta}_\omega(s,a) - Q^{\pi_\theta}(s,a)]^2,$$

for some function $d(s)$ (that will be specified later).

5. With π_θ defined as in (1), show that if $\omega^\star$ minimizes $\varepsilon(\omega)$ then:

$$\omega^\star = \left(\sum_s d(s)[\nabla^2_\theta A(\theta,s)]\right)^{-1} \left[\sum_s d(s) \sum_a \pi_\theta(s,a)(\nabla_\theta \log \pi_\theta(s,a))Q^{\pi_\theta}(s,a)\right].$$

You can assume that indeed, the matrix $\sum_s d(s)[\nabla^2_\theta A(\theta,s)]$ is invertible. *Hint: Write out the first order necessary condition to get a minimizer of ε.* [3 pts]

6. What choice of function $d(s)$ does ensure that

$$\omega^\star = \left(\sum_s d(s)[\nabla^2_\theta A(\theta,s)]\right)^{-1} \nabla_\theta J(\theta)$$

where $J(\theta)$ is the expected cumulative discounted reward under policy π_θ? [2 pts]

As a final remark, in this problem, you have shown that a reasonable choice of parameters for the critic corresponds to the so-called natural policy gradient.

Problem 5 - Solution

1) We must have

$$1 = \sum_a \pi_\theta(s,a) = \sum_a \exp\left(\theta^\top T(s,a) - A(\theta,s)\right)$$

equivalently, by taking the log of both sides and re-arranging

$$A(\theta,s) = \log \sum_a \exp(\theta^\top T(s,a)).$$

2) We have that

$$\nabla_\theta \log \exp\left(\theta^\top T(s,a) - A(\theta,s)\right) = \nabla\left(\theta^\top T(s,a) - A(\theta,s)\right) = T(s,a) - \nabla_\theta A(\theta,s)$$

3) Observe that for any sufficiently smooth f, by the quotient rule, it holds that

$$\begin{aligned}\nabla_\theta^2 \log f(\theta) &= \frac{\nabla^2 f(\theta)}{f(\theta)} - \frac{\nabla_\theta f(\theta)[\nabla_\theta f(\theta)]^\top}{[f(\theta)]^2} \\ &= \frac{\nabla^2 f(\theta)}{f(\theta)} - \nabla_\theta \log f(\theta)[\nabla_\theta \log f(\theta)]^\top.\end{aligned}$$

The result follows by applying the above indentity to $\log \pi_\theta(s,a)$ and noting that

$$E_{a\sim\pi_\theta(\cdot|s)}\left[\frac{\nabla_\theta^2 \pi_\theta(s,a)}{\pi_\theta(s,a)}|s\right] = \sum_a \frac{\nabla^2 \pi_\theta(s,a)}{\pi_\theta(s,a)}\pi_\theta(s,a) = \nabla_\theta^2 \sum_a \pi_\theta(s,a) = \nabla_\theta^2 1 = 0.$$

4) By 2. we have that

$$\nabla_\theta \log \pi_\theta(s,a) = T(s,a) - \nabla_\theta A(\theta,s)$$

and thus by 3., after differtiating once more and taking into account the sign change

$$E_{a\sim\pi_\theta(\cdot|s)}[[\nabla_\theta \log \pi_\theta(s,a)][\nabla_\theta \log \pi_\theta(s,a)]^\top|s] = -\nabla_\theta[T(s,a) - \nabla_\theta A(\theta,s)] = \nabla_\theta^2 A(\theta,s).$$

5) The first order condition for a minimum is

$$\begin{aligned}0 &= \nabla_\omega \sum_s d(s) \sum_a \pi(s,a)[Q_\omega^{\pi_\theta}(s,a) - Q^{\pi_\theta}(s,a)]^2 \\ &= \sum_s d(s) \sum_a \pi_\theta(s,a)\nabla_\omega[\omega\nabla_\theta \log \pi_\theta(s,a) - Q^{\pi_\theta}(s,a)]^2 \\ &= 2\sum_s d(s) \sum_a \pi_\theta(s,a)\left[[\nabla_\theta \log \pi_\theta(s,a)][\nabla_\theta \log \pi_\theta(s,a)]^\top\omega - \nabla_\theta \log \pi_\theta(s,a)Q^{\pi_\theta}(s,a)\right]\end{aligned}$$

which we re-arrange as

$$\begin{aligned}\omega &= \left(\sum_s d(s) \sum_a \pi(s,a)\left[[\nabla_\theta \log \pi_\theta(s,a)][\nabla_\theta \log \pi_\theta(s,a)^\top]\right)^{-1}\left[\sum_s d(s) \sum_a \pi(s,a)\nabla_\theta \log \pi_\theta(s,a)Q^{\pi_\theta}(s,a)\right] \\ &= \left(\sum_s d(s)[\nabla_\theta^2 A(\theta,s)]\right)^{-1}\left[\sum_s d(s) \sum_a \pi(s,a)\nabla_\theta \log \pi_\theta(s,a)Q^{\pi_\theta}(s,a)\right]\end{aligned}$$

6) $d(s) = \sum_{t=1}^\infty \lambda^{t-1} P(s_t = s|s_1,\pi_\theta)$ and the result follows immediately from 5. by the policy-gradient theorem.